GREG BOGNAR AND IWAO HIROSE

THE ETHICS OF HEALTH CARE RATIONING

An Introduction

Second Edition

ROUTLEDGE

The Ethics of Health Care Rationing

The rationing of health care is universal and inevitable, taking place in both poor and affluent countries, in publicly funded and private health care systems. Someone must budget for as well as dispense health care while aging populations severely stretch the availability of resources.

The Ethics of Health Care Rationing is a clear, timely, and much-needed introduction to this important topic. Substantially revised and updated, this second edition includes new chapters on disability discrimination and age discrimination, and on the price of drugs and medical therapies. Beginning with a helpful overview of why rationing is an ethical problem, the authors examine the following key topics:

- What sort of distributive principles should we rely on when thinking about health care rationing?
- What is the relation between ethics and cost-effectiveness in health care?
- How should we think about controversies surrounding discrimination over disability and age?
- How should we approach controversies surrounding rationing and the price of pharmaceutical drugs and medical therapies?
- Should patients be held responsible for their health?
- Why does the debate on responsibility for health lead to issues about socioeconomic status and social inequality?

Throughout the book, examples from the United States, the United Kingdom, and other countries are used to illustrate the ethical issues at stake. Additional features such as chapter summaries, annotated further reading, and discussion questions have also been updated, making this an ideal starting point for students new to the subject, not only in philosophy but also in closely related fields such as politics, health economics, public health, medicine, nursing, and social work.

Greg Bognar is Senior Lecturer in Practical Philosophy at Stockholm University and Senior Researcher at the Stockholm Centre for Healthcare Ethics (CHE), Sweden. He is currently working on an edited volume with Axel Gosseries on the ethics of age limits and age discrimination.

Iwao Hirose is Professor and Canada Research Chair in Value Theory and the Philosophy of Public Policy at McGill University, Canada. He is the author of *Egalitarianism* (Routledge, 2015) and *Moral Aggregation* (2015), and a coeditor of *The Oxford Handbook of Value Theory* (2015) and *Weighing and Reasoning* (2015).

Praise for the first edition

"Most contemporary publications related to health care rationing are written for specialized, often academic, audiences, but this work is an introduction to the topic for general readers. It is accessible to those with no prior knowledge of philosophy, bioethics, or health policy. Suggested readings are available at the end of each chapter for both new and advanced readers to explore chapter topics in more depth. *Summing Up: Recommended.* All health sciences students, researchers/faculty, professionals/practitioners, and general readers."

– M. L. Charleroy, CHOICE

"Against the background of ineluctable scarcity in health care resources, this important, accessible and provocative book introduces readers to pressing issues concerning how, morally speaking, we ought to determine who gets what. If we want an informed public debate on health care rationing, I don't know of a better place to start."

– Samuel Kerstein, University of Maryland, USA

"A great introduction to the field, combining philosophical sophistication with economic literacy to deliver profound insights into the resource allocation dilemmas facing health care decision makers. A valuable resource for students and health professionals alike."

– Richard Cookson, University of York, UK

"Bognar and Hirose illuminate and make accessible the most pressing and entrenched controversies in health care rationing. This book spans political philosophy, health economics and bioethics, grounding arguments in vividly described cases. It delivers complex ideas in a relaxed style perfectly suited to drawing us all in to a long overdue common inquiry."

– Monique Jonas, University of Auckland, New Zealand

"*The Ethics of Health Care Rationing is* . . . a 'must read' for students of bioethics and other interested parties, as it sheds a light on issues that we often tend to avoid or ignore in our field. Moreover, it is an agreeable read and the mix of real-life and fictional examples works particularly well."

– Kristien Hens, Ethical Perspectives

The Ethics of Health Care Rationing

An Introduction

Second Edition

Greg Bognar and Iwao Hirose

Routledge
Taylor & Francis Group
LONDON AND NEW YORK

Cover image: © Daniel Chetroni/Getty Images

Second edition published 2022
by Routledge
4 Park Square, Milton Park, Abingdon, Oxon OX14 4RN

and by Routledge
605 Third Avenue, New York, NY 10158

Routledge is an imprint of the Taylor & Francis Group, an informa business

First edition published by Routledge, 2014

British Library Cataloguing-in-Publication Data
A catalogue record for this book is available from the British Library

Library of Congress Cataloging-in-Publication Data
Names: Bognar, Greg, author. | Hirose, Iwao, author.
Title: The ethics of health care rationing : an introduction / Greg Bognar and Iwao Hirose.
Description: Second edition. | Milton Park, Abingdon, Oxon ; New York, NY : Routledge, 2022. | Includes bibliographical references and index.
Identifiers: LCCN 2021042607 (print) | LCCN 2021042608 (ebook) | ISBN 9780367505219 (hardback) | ISBN 9780367505127 (paperback) | ISBN 9781003050216 (ebook)
Subjects: LCSH: Health care rationing—Moral and ethical aspects.
Classification: LCC RA410.5 .B65 2022 (print) | LCC RA410.5 (ebook) | DDC 174.2—dc23/eng/20211021
LC record available at https://lccn.loc.gov/2021042607
LC ebook record available at https://lccn.loc.gov/2021042608

ISBN: 978-0-367-50521-9 (hbk)
ISBN: 978-0-367-50512-7 (pbk)
ISBN: 978-1-003-05021-6 (ebk)

DOI: 10.4324/9781003050216

Typeset in Garamond
by Apex CoVantage, LLC

Contents

Acknowledgments vi
Preface to the second edition viii

Introduction 1

1 Ethics and health care 7

2 The value of health 28

3 Ethics and cost-effectiveness 51

4 Disability discrimination 76

5 The question of age 94

6 The aggregation of health benefits 112

7 Responsibility for health 133

8 The price of drugs 155

Conclusion 174

Glossary 180
Bibliography 184
Index 190

Acknowledgments

This book grew out of our discussions and collaborative research over many years. We first learned about the issues in this book while we were Post-Doctoral Research Fellows at the Harvard University Program in Ethics and Health. We thank Dan W. Brock, Norman Daniels, and Daniel Wikler at the Program for their guidance and encouragement over the years. In addition, we thank Daniel Wikler for his permission to use the vaccination program example that we describe in Chapter 1.

Many people sent us written comments on earlier versions of the first edition. We learned a great deal from these comments, even if we could not always do justice to them in the final version. We also had to leave out several suggestions for including additional topics simply because of space constraints. With these caveats, we thank Yukiko Asada, Rob Baltussen, Kristine Bærøe, Dan Hausman, Samuel J. Kerstein, Ingrid Miljeteig, Erik Nord, Ole Frithjof Norheim, and three anonymous reviewers for Routledge. Many more people helped us with their questions in conversation and at public lectures. We thank audiences at Charles Sturt University, the Eindhoven University of Technology, the Fondation Brocher, Georgia State University, the Program in Ethics & Health at Harvard University, La Trobe University, Monash University, the Blavatnik School of Government at the University of Oxford, the Universidad Nacional Autonoma de Mexico, University of Bergen, University of Connecticut at Storrs, University of Leeds, University of Melbourne, the Conference on Distributive Justice in Health at the Centre for Humanistic Studies of the University of Minho, and the Ethics Center at the University of Zurich.

We would also like to thank our students who served as guinea pigs for this material. Their feedback has improved the book in many ways. In the fall semester of 2012, Hirose taught the material in this book to undergraduates at McGill University, and Bognar taught it at La Trobe University. Bognar also taught some of the material in the book at the "Justice: Theory and Applications" summer university program of the Central European University in 2013.

We want to express our special gratitude to Fondation Brocher. The Brocher Visiting Fellowship enabled Hirose to complete Chapters 5 and 6 in

spring 2012, and the 2012 Brocher Summer Academy brought Bognar to the Villa Brocher for three days of discussion and joint revision of the first edition. We also want to thank the Swedish Collegium for Advanced Study for hosting Hirose's visit to Sweden in the spring of 2021, which enabled us to work together and complete the second edition, and the University Center for Human Values at Princeton University where Bognar completed the revision of the first few chapters in the spring of 2020.

Bognar acknowledges financial support from the Swedish Research Council and thanks the journal *Bioethics* for its permission to use material from his article, "Fair Innings" (2015). Hirose gratefully acknowledges financial support from the Canadian Institutes of Health Research, the Social Sciences and Humanities Research Council, and the Canada Research Chairs Program.

Despite some relative disagreements about other philosophical issues, we have no disagreement about what is presented in this book. Bognar is primarily responsible for Chapters 2 through 5, and Hirose is for Chapters 6 through 8. Bognar and Hirose are jointly responsible for the Introduction, Chapter 1, and the Conclusion.

Last, but not least, we thank our editors at Routledge, Tony Bruce and Adam Johnson, for their support of this project.

Preface to the second edition

For the new edition, we revised the whole text, fixing typos (of which there were unpleasantly many), updating empirical data where necessary, and making many small stylistic changes along the way to make the text more lucid. In addition, we made two major changes. First, we split Chapter 4 in the first edition into two chapters: one on disability (Chapter 4) and one on age (Chapter 5). In the first edition, Chapter 4 covered both disability discrimination and the problem of age. In this edition, both issues get a chapter of their own. The new chapters have a large amount of new material. They reflect how our views have evolved in the last few years. Second, we added a chapter on a topic that was not addressed in the first edition. Chapter 8, on the price of drugs, is completely new and contains original material.

The second edition was prepared during the COVID-19 pandemic. We recognized early on that it would be impossible to do justice to the complexity of the many ethical issues that the pandemic has raised. The best we could do was to point to the connections between these issues and the topics in the first edition by incorporating examples that will now be familiar to readers. But the ethical problems of COVID-19 deserve their own book.

The first edition of this book was dedicated to Dan Brock. Sadly, he passed away before the second edition was completed. This edition is dedicated to his memory and to his contributions to establishing population-level bioethics.

Introduction

Rationing health care, we suspect, sounds like a horrible idea. For some, the word *rationing* conjures up images of wartime hardships – long lines waiting at distribution points for basic necessities such as bread, sugar, cooking oil, or gasoline. For others, health care rationing sounds like the government intruding on people's private lives with its bureaucrats lording over life and death, deciding whether Grandma can get her medicines or the life-saving treatment she needs. In some countries, the idea of rationing raises fears about privatizing cherished universal health care systems, destroying social solidarity and reducing people to commodities.

Rationing, in its broadest sense, is the controlled allocation of some scarce resource or good. It implies that limits are placed on the good's availability. People who need or want the rationed good are restricted to getting it in a certain quantity or size or at a certain time. They are not free to use or consume it in the way they want.

In health care, rationing can apply to treatments, services, pharmaceuticals, medical procedures, and so on. When health care resources are rationed, patients may be restricted to certain treatments. They might be placed on waiting lists. There might be limits on how often they are eligible for diagnostic procedures or screening tests. And, in the worst cases, patients may be denied beneficial or even life-saving treatments and interventions. No doubt, many people feel that rationing health care is not just a nuisance: it can seriously affect quality of life, and it might even, literally, be a matter of life and death.

When health care is rationed, then somewhere, someone makes a decision about the limits of what is provided or how it is provided. For instance, someone decides that hospitals cannot perform a particular kind of surgery. Someone decides that a particular type of medicine is not subsidized. Someone organizes patients into a waiting list. As we will say, someone makes a *priority setting* decision, choosing which beneficial treatments or interventions are more important than others, which have the best value, and which are not important at all. All of these decisions interfere with our freedom to decide, together with our doctors, what sort of intervention or treatment or medicine or medical technology we need or want. All of them interfere with

DOI: 10.4324/9781003050216-1

our freedom as patients and health care consumers. They are choices that are imposed on us. The priorities of those who made the decisions may often conflict with our own.

So, if rationing health care is a horrible idea, the *ethics* of rationing health care is even worse. It sounds like an oxymoron. If rationing health care is horrible, how can it be ethical?

Our aim in this book is to show you that health care rationing not only *can* be ethical, but also it *must* be. Our case is very simple. We shall argue that the rationing of health care resources is inevitable. It takes place in all health care systems – public or private, rich or poor. It is not only inevitable, it is actually ubiquitous as well. So you might consider it a necessary evil. But then it is crucial that it is done as ethically as possible to reduce its evil effect. Hence, you should care about the ethics of rationing health care. It is not an oxymoron.

Actually, we will also make a stronger argument. We will argue that rationing in health care is not only inevitable and widespread, but it is also *desirable*. Health care rationing, or setting priorities between alternative resource uses, is far from being a necessary evil – it is a good thing. We all benefit when health care resources are allocated in a morally defensible way. This is another reason why you should care about the ethics of rationing health care.

To many people, these claims may sound incredible. They associate health care rationing with poor countries. It is not something, they believe, that takes place – or should take place – in affluent countries. But the truth is that the health care systems of the most developed countries do ration health care one way or another. As a matter of fact, the careful rationing of health care is one of the factors that make a health care system work well. The best health care systems in the world do it.

Other people associate health care rationing with governments. It is, they believe, something that takes place only in single-payer, government-run health care systems. Some people who have this belief probably have private health insurance. So they believe they are not affected by rationing.

The truth is that privately run health care is rationed just as much as publicly run health care. The rationing is done by the companies providing health insurance. They might offer a choice between different plans, but they all involve limits and controls on what they offer. Rationing is not confined to governments only.

When we were planning this book, many people advised us against using the word *rationing*. They were worried about its negative, and often political, connotations. Some philosophers have recently stopped using the "R-word" altogether. We believe this is a mistake. It's a perfectly accurate word for the subject. It should not be yielded to those who attempt to use it to raise public fears for their own political gain. It should be defended. Health care is too important to allow the muddying of the waters by a fear to call things by their names.

Our central claim is that the rationing of health care is an ethical problem. Setting priorities in health care must be based on sound moral principles. This book provides an introduction to this complex topic. While there are excellent books on health care rationing in philosophy, health economics, and health policy, they tend to be written with a specialist audience in mind. We are unaware of any other entry-level book. In fact, our topic has, until recently, received little attention in ethics.

The area of philosophy that is closest to our concerns is bioethics. Traditionally, bioethics has focused on ethical issues that arise in the doctor–patient relationship and in medical research. It has addressed topics such as the permissibility of abortion or physician-assisted suicide, embryonic stem cell research, respect for patient autonomy, or the protection of research subjects. More recently, as health care has become an important focus of public debate all around the world, some bioethicists have started to address questions that arise at the population level – for instance, questions about increasing international and domestic inequalities in health, the health-related causes and consequences of poverty, the aging of societies, and the allocation of health care resources. This relatively new area of philosophy has become known as *population-level bioethics*. We have learned a lot from people working in this area, and we point the reader to their works at the end of each chapter in the *Further readings* section.

The problem of health care rationing is complex. For one thing, most examples of rationing in health care are rather mundane, uncontroversial, and even boring. They concern setting levels of subsidies for pharmaceuticals, levels of co-payments for health care services, reimbursement policies for medical devices, and similar decisions within complex administrative institutions. They do not make for striking examples. The examples that are usually employed in discussions of rationing are not only more fascinating, but also more unusual, and hence less representative. They concern expensive cancer drugs that provide a few months of remission at enormous costs, patients on waiting lists for scarce transplantable organs, or priority lists for vaccinations during an influenza pandemic. We ourselves will use such examples in this book. But it is important to keep in mind that most examples of health care rationing are much more pedestrian.

In addition, health care institutions differ from country to country. To keep our discussion concise, we ignore many of these differences. Our aim is to highlight the general ethical questions and moral principles that apply equally to different settings and health care systems. We focus on the general issues that any attempt of priority setting must face. For instance, we do not have anything to say about whether a health care system should be run publicly or rely on private health insurance, or how taxpayers and patients should share the costs of health care. But the questions we do raise are relevant to various institutional arrangements. Our hope is to furnish readers with a clear understanding of at least the basics of the complex problems surrounding the ethics of health care rationing.

Still, there are many questions and ideas that we have to address in this book. Before we embark on the journey, it is worth having a road map in our hands.

We will begin in the next chapter by defending the two claims that we have already made. We show why health care rationing is widespread and explain why it is also inevitable. The explanation has to do with the unavoidability of scarcity. We present some general ways in which scarcity arises in health care. In later chapters, we will give more specific real-life examples of resource scarcity and the way rationing can address it. Before that, we also give a very brief introduction to moral argument and explain some central ethical concepts and ideas. This will provide the necessary background for later discussions.

If the rationing of health care resources is inevitable, then we must be able to compare different resource allocations as better or worse, acceptable or unacceptable, and so on. Since the goods and resources that are allocated in health care are diverse, we need common criteria for their evaluation. Chapter 2 addresses this issue. Naturally, you might think that a common criterion should be health: one way of allocating resources is better than another if it results in better health for people. But health is not a quantity that can be measured, like weight or height. Its measurement consists in considering its value through its impact on quality of life. We explain how researchers try to measure the value of health by examining the quality of life judgments that people make about the badness of health states. These judgments can help compare alternative allocations. But measuring the value of health is riddled with problems and puzzles. We present some of these problems and puzzles in connection with two of the most widespread measures of the value of health: quality-adjusted life years (QALYs) and disability-adjusted life years (DALYs).

For readers new to our topic, Chapter 2 is probably going to be the most tedious in this book. We apologize for that. Nevertheless, it introduces concepts and ideas without which the material in subsequent chapters would be much more difficult to understand.

Chapter 3 is about cost-effectiveness analysis (CEA). This is the policy maker's main tool for evaluating the costs and benefits of different interventions and health care services. But the use of cost-effectiveness analysis for setting priorities among different uses of health care resources is controversial – not only among academics and policymakers, but also among the general public. In this chapter, we explain how cost-effectiveness analysis works, address the main ethical problems of its use, and correct some misunderstandings that often appear in discussions. We also present several examples.

The following two chapters address two problems for cost-effectiveness analysis. One of these problems is discrimination against people with disabilities and chronic health conditions. We take up this topic in Chapter 4. Some people believe that if health care resource allocation is based on a principle that directs you to maximize health benefits, you will often give priority to people without disabilities, and the health care needs of people with disabilities will be neglected. We show that this worry is based on misunderstandings. But the objection does raise some important ethical issues, which would be ignored if we focused only on costs and benefits.

The second problem is discrimination by age, the subject of Chapter 5. Some people believe that age should be a relevant consideration in the allocation of health care resources. In particular, the health needs of younger people should have higher priority than the health needs of the elderly. We try to provide a coherent formulation of this view, but we ultimately leave the question of age discrimination open. People have different moral intuitions about particular cases, and controversies about the role of age and disability in resource allocation have arisen in many practical applications. With aging societies and the ever-growing prevalence of chronic illness, these controversies are going to become more and more acute.

One of these controversies has already arrived. When we started working on this edition, most of the world was under lockdown because of the COVID-19 pandemic. What happened in 2020 and 2021 brought health care rationing to the center of attention like never before. The allocation of ventilators, access to testing, and vaccinations were regularly discussed in the news. Rationing became a common household word. The theoretical discussions in books like this one suddenly became urgent practical matters. Thus, in Chapter 5, we use the example of COVID-19 to give an overview of the ethics of health care rationing in public health emergencies, with a special focus on the issue of age. As we explain, the most important difference between "normal" and emergency medicine with respect to rationing is that in the latter you must choose between particular patients. Unsurprisingly, these tragic choices give rise to vexing moral issues. And because of the unequal distribution of the risk of death between age groups, the example of COVID-19 allows us to test our moral intuitions about the role of age.

Chapter 6 broadens the discussion by connecting the problem of health care rationing to more general debates in ethical theory. In this chapter, you will also encounter striking, imagined and real-life examples of deciding who should live and who should die. Our aim is to show how some of the moral principles used in health care resource allocation lead to familiar, but deeply controversial, problems in ethical theory. These problems concern the aggregation of benefits across different people, the moral justification of taking into account the number of those who benefit, and the use of lotteries in life-and-death cases.

Chapter 7 broadens the discussion even further. It begins by focusing on a controversial issue in public health: should individual responsibility for health and healthy lifestyles be taken into account in the provision of health services and treatment decisions? Some influential theories in political philosophy hold that inequalities are not a matter of justice if they are the result of choices for which individuals can be held responsible. It is not unjust if some people end up with disadvantages through their own choice or fault.

Society, as a matter of justice, is not required to come to their aid to reduce their disadvantages.

The theme of individual responsibility is becoming more and more prevalent in public debates. The application of some theories of distributive justice to health care seems to suggest that individual responsibility should have a central role in health care rationing. But very few authors have brought together the philosophical and the practical arguments on this topic. We will connect the two debates. We will also emphasize that the question of responsibility quickly leads to broader issues about the relation between health and behavior, class, race and socioeconomic status. We give a very brief account of the growing literature on the social determinants of health, which examines these issues.

This edition of the book contains an additional chapter on a topic that was absent from the first edition. In the last few years, the relentless growth of the costs of health care and, in particular, the prices of pharmaceutical drugs and medical therapies, have taken center stage in discussions of health care. Many innovative drugs and therapies are so expensive that health care systems struggle to provide them. At the same time, we still lack medicines for some common and disabling conditions because pharmaceutical companies lack the incentives to develop them. Many people think that something has gone wrong with the way we develop, price, and distribute medicines and therapies, especially in light of global disparities in health. Chapter 8 takes up these issues.

In the Conclusion, we return to the claim we made a few pages back: that the rationing of health care is not only inevitable and widespread, but it is both morally defensible and desirable as well. It is a good thing from which we all benefit. Of course, this is the case only if rationing choices are based on sound ethical principles and made transparently and accountably. We conclude by defending this idea.

The arguments and ideas in the following chapters are sometimes complex and might require some patience on the part of the reader. We have attempted to present them as clearly as possible. We do not assume any prior knowledge of philosophy, health economics, medicine, or health policy. At times, we use numerical examples. They never rely on anything beyond the most basic math skills. For those who want to explore the topics in greater depth, each chapter ends with a list of further readings and discussion questions.

1 Ethics and health care

1.1 The vaccination programs

Imagine that your team of public health experts has a contract with the government of a remote, tiny island state to vaccinate children against a fatal disease. The disease threatens only children, and each child has an equal chance of contracting it. The vaccination has no side effects and provides total immunity against the disease.

Altogether, there are 1,000 children on the island. Eight hundred of them live on the coastal plains, and 200 live in remote mountains. It costs only $1 to vaccinate a child who lives near the coast, but $4 to vaccinate a child who lives in the mountains. It costs four times as much to vaccinate the children in the mountains because it is difficult to reach them.

The problem is that you are only given $800 for this work (this is a very poor country). Your team cannot vaccinate all the children. Because of logistical reasons, you have to choose between two ways of organizing your vaccination campaign. The two programs are:

A. vaccinating every child living on the coastal plains, but none of the children living in the mountains;
B. vaccinating half of the children who live on the coastal plains, and half of those who live in the mountains.

Which program would you choose?

If you choose Program A, 800 children will be vaccinated. They will be protected against the disease. If you choose Program B, half of the children on the plains and half of the children in the mountains will be selected randomly. In the end, 500 children will be vaccinated – 400 on the plains and 100 in the mountains.

We often present this example to our students. We ask them to make a choice between these hypothetical programs. We get very consistent results. A majority of the students in any class chooses Program A, but there is always a fairly large minority that chooses Program B. Students disagree about the

DOI: 10.4324/9781003050216-2

right choice. We have never met a class where there was anything approaching consensus in favor of either program.

Next, we ask a follow-up question from those who are in favor of Program A.

Here is the question. Suppose that just as you are about to leave the island with your team, you get a call from the Ministry of Health. They are happy to tell you that the government has given you another $800 for a second round of vaccinations. Even better, they also have vaccinations available against a second disease. This disease is just like the first: it only affects children, it is invariably fatal, all children have the same chance of contracting it, and anyone's chance of contracting it is equal to the chance of contracting the first disease.

At this point, you have a meeting with your team to discuss your options. For logistical reasons, you must choose between the following two programs:

C. vaccinating all the children who live in the mountains against the first disease;
D. vaccinating all the children who live on the coastal plains against the second disease.

If you choose Program C, you will vaccinate all the 1,000 children living on the island against the first disease. If you choose Program D, you will vaccinate 800 children against both diseases. In the first case, you will provide 1,000 vaccinations to 1,000 children; in the second case, you will provide 1,600 vaccinations to 800 children.

Would you choose Program C or Program D?

In our experience, an overwhelming majority of the students who favored Program A chooses Program C. We have met very few students who favor A *and* D.

Those who favor Program A in the first question usually give the following explanation for their choice. The vaccination confers a great benefit – immunity against a fatal disease. It is very important to provide this benefit to as many children as possible. Of course, Program A leaves out the children who live in the mountains. But for each child that you could vaccinate in the mountains, you can vaccinate four children on the plains. Choosing Program A is justified by the benefits that would be bestowed on a greater number of children.

Those who favor Program B have a different explanation. They argue that it is wrong to exclude the children living in the mountains. It is not their fault that they live in a remote place. There is something unfair about discriminating against some of the children merely because they are growing up in less accessible places. If not all of the children can be vaccinated, you should at least give an equal chance to all of those who live on the coastal plains and all of those who live in the mountains. To these students, this seems to be a requirement of fairness.

Remarkably, those who choose A *and* C tend to give a similar explanation for choosing Program C in the second question. For these students, maximizing the benefits of vaccination is the most important consideration in the first question. But the consideration of fairness appears in the second question and becomes more important than benefit maximization – even if very few students might be able to explain what precisely they mean by fairness.

At this point, those who favored Program B – vaccinating half of the children on the plains and half of the children in the mountains – also get a second question. Here is their question. Just as you are preparing to leave the island, you get a call from the Ministry of Health. You are given another $800 for vaccinations. You have to decide between two programs:

E. vaccinating the remaining half of the children who live on the coastal plains and the remaining half of those who live in the mountains;
F. vaccinating all the children who live on the coastal plains against the second disease.

If you choose Program E, you will end up vaccinating all the 1,000 children living on the island against the first disease. You will end up giving out 1,000 vaccinations to 1,000 children. If you choose Program F, you will end up vaccinating one half of the children on the plains against both diseases, the other half of the children on the plains against the second disease only, and one half of the children in the mountains only against the first disease. Altogether, you end up providing 1,300 vaccinations (800 + 400 + 100) to 900 children.

Which program would you choose?

In our experience, a large majority of those who favored Program B in the first question favors Program F in the second question – even if there are typically some holdouts favoring B and E. Program F usually gets a comfortable majority. When asked to explain their choices, students often say that although they continue to believe that it is important to avoid the unfairness of choosing Program A in the first question, they acknowledge that the greater benefits of Program F can tilt the balance in the second question. After all, if you implement programs B and F, you will vaccinate nine-tenths of the children against at least one disease and a significant minority against two.

When we present these questions, we emphasize that we are not looking for "right" or "wrong" answers. Rather, what matters is what we can learn from the answers about our moral beliefs. And the lesson is clear: most people who consider this example believe that it is important to choose the course of action that will bring about the greatest benefits – but they also believe that it is important to allocate resources in a fair way. Of course, we will need to say a lot more about the requirement of fairness. But one thing we can already say: fairness and benefit maximization can, and often do, conflict. It is important to find the right balance between them. This book is about how we can do that.

1.2 The ubiquity of rationing health care

The story of the vaccination programs is a thought experiment. By asking you to make moral judgments in hypothetical situations, it is designed to shed light on the ethical principles that are relevant to the distribution of benefits in conditions when resources are scarce. Philosophers often use thought experiments to help analyze difficult questions. They often involve an element of science fiction: you are asked to imagine that you are a brain in a vat or you are teleported to another planet or you are deceived by an evil demon. But the vaccination story is different. It is not entirely fictional. It is modeled on a real-life ethical dilemma.

In 2003, the World Health Organization (WHO) and the Joint United Nations Programme on HIV and AIDS (UNAIDS) launched the "3 by 5" program. The aim of the program was to provide antiretroviral therapy to three million people with HIV/AIDS living in developing countries before the end of 2005. Even if successful, the program would have reached only a fraction of those who could have benefited from the therapy. In the end, the target was met only in 2007. At the end of 2011, around 6.65 million eligible patients in developing countries received antiretroviral therapy, up from 400,000 in 2003. But still less than half of eligible patients had access to therapy. (The targets of the recent 90–90–90 program are that by 2020, 90 per cent of all people living with HIV should know their HIV status, 90 per cent of all people with diagnosed HIV infection should receive sustained antiretroviral therapy, and 90 per cent of all people receiving antiretroviral therapy should have viral suppression. Although there has been a lot of progress, the world has missed these targets too.)

One controversial aspect of rolling out the initial 3 by 5 program was whether delivery should focus on urban or rural populations. In developing countries, there is a shortage of HIV clinics and health facilities. Concentrating delivery in urban areas ensured that more patients could be reached, but it made the program inaccessible to rural populations. Patients living far from cities could not reach the facilities because of long distances, bad roads, and their inability to pay for transport. Some experts argued that the program should focus on those areas where the infrastructure is already in place in order to reach as many people as quickly as possible. Others argued that rural populations should not be neglected, even if fewer patients can be served as a consequence.

Thus, policymakers faced the same dilemma as our students in the classroom. The choices they made, however, had real consequences. For some people, they were a matter of life and death. But national guidelines for implementing the program often treated such dilemmas as merely technical questions: matters that require the expertise of medical doctors, economists, and policymakers. The ethical nature of the dilemmas was rarely acknowledged, and the choices were made without consulting the citizens of these countries.

It is understandable that hard ethical choices are sometimes treated as technical questions. The policymakers who were responsible for broadening access to antiretroviral therapy had to set priorities among competing resource uses. They had to engage in the rationing of health care. But the idea of rationing health care makes people uncomfortable. It entails that there are patients who could benefit from care but have to do without it. Many people get upset when they hear or read stories in which someone is denied potentially beneficial (maybe even life-saving) medical care. In many countries, the very idea of rationing health care is taboo. Politicians who talk about it risk their prospects for reelection. So it is not surprising that rationing choices are often hidden behind technical or medical language.

Still, it is not right. It is the responsibility of policymakers to reflect on the values they take into account when they make choices about the use of social resources – both in health care and beyond. It is our right as citizens to demand that social choices that can potentially have a great effect on our lives are made in a transparent and accountable manner. It is also our responsibility to think through the ethical issues faced by our society. We should have the chance to contribute to their discussion and resolution. To do that, we need a basic understanding of medical and economic matters; but what we need most is ethical argument. Medical doctors and economists can help us understand technical matters, and philosophers can help us with the ethical argument.

So, the first point we want to make is that the rationing of health care is an ethical issue. We all have a stake in getting it right. Next, we want to argue that health care rationing is ubiquitous. It affects all of us.

Some readers might think that the rationing of health care has little to do with their society. Where they live, there is a well-functioning health care system. They might think that rationing is something that takes place mainly in resource-poor environments or the least developed countries. True, the 3 by 5 program targeted middle- and low-income countries. But it would be a mistake to conclude from this one example that only these countries should be concerned with rationing. In fact, rationing is universal. It takes place in poor as well as affluent countries, in publicly funded health care systems as well as in private health insurance.

Other readers might associate the rationing of health care with government – in particular, with faceless bureaucrats in drab offices making life-and-death choices. In the vaccination program, you probably assumed that you were working for the government or perhaps an NGO (nongovernmental organization). But you would have faced the very same choices if you were a private contractor with expertise in public health campaigns. The need to set priorities in health care is not limited to government-run health care systems. Private actors, including insurance providers, need to do it just as much.

Neither is it the case that rationing is an exception in affluent countries, rather than the rule. Most people in affluent countries could probably mention organ transplantation as an example of health care rationing. Because there are many more patients than available organs, patients everywhere are placed on waiting lists. Tragically, some of them die before a suitable donor is found. Waiting lists are a form of rationing. It is not difficult to see how they raise ethical issues. Should priority be given to the patients who have waited the longest or to those who need an organ most urgently or to those whose survival prospects are the best? Clearly, these are partly ethical, rather than merely medical questions.

And surely, all of our readers can mention the COVID-19 pandemic that put the issue of health care rationing in the limelight. As we are writing this, health care systems around the world are still under enormous pressure. We already know that tragic choices had to be made. Moreover, the vaccines that have been developed against the disease could not immediately be produced in sufficient quantities. It was necessary to set priorities among different patient groups. Who should be vaccinated first? Should it be the young or the old? Is it fair to give priority to essential workers or those who have dependents? Clearly, these are partly ethical, rather than merely technical questions.

These examples are familiar. But they are also the most unusual. They concern extreme cases of scarcity and public health emergencies. In such cases, rationing might be unavoidable. But what about our claim that health care rationing is ubiquitous?

In a way, what is rationed in these examples are people. Patients are matched to resources. The examples present choices about who gets medical treatment or who gets it before others. They make good topics for debate, but they are far from being ordinary. Most rationing choices are not like this. They do not concern setting priorities among patients. They concern setting priorities among treatments, services, pharmaceuticals, medical procedures, and so on. They concern *what* to provide in the health care system and how to provide it, and not to whom to provide it.

Health care rationing is the controlled allocation of scarce health care resources. Occasionally, it takes the form of selecting particular patients or patient groups. But usually it takes the form of setting priorities among interventions. By "intervention," we mean any use of resources in the health care system that aims to address health problems or the risks of health problems. By "resource use," we mean any mobilization of human, physical, financial, or other sorts of assets to achieve these aims.

Thus, when the government decides which pharmaceuticals to subsidize from the health care budget, it engages in rationing. When it decides in which city to build a hospital or clinic, it is an example of rationing. When it introduces a cancer-screening program, it is rationing health care. All of these decisions require resources that could be spent elsewhere. Implicitly or indirectly, all such decisions determine who will benefit. Patients of subsidized medicines have to spend less than others. Residents of the city in which the hospital is built have better access to specialist services than others.

Private health insurance is no different. When an insurance provider decides which treatments to include in its plans, it engages in rationing.

When it determines the co-payments, its choice is an example of rationing. When it refuses to provide coverage for people with preexisting conditions, it is, obviously, rationing health care by excluding these people.

Most of us experience the consequences of rationing at some point in our life. When a doctor prescribes a medicine for you that is not subsidized by the health care system or your insurance provider, you might be facing the consequences of a rationing decision that was made by others. When you are told that you need a procedure but it will take many months before you can get it, it might be because of choices made about the use of health care resources. You have to wait because resources are scarce, and hence their allocation is controlled. Perhaps the procedures are scheduled on a "first-come, first-served" basis. This in itself is a form of rationing. The procedures could be scheduled on some other basis.

Most of us are unaware that rationing decisions take place all the time because of the enormous complexity of modern health care systems. Rationing is almost never a matter of simple choices between this or that intervention. Rather, it is a matter of trying to achieve different (and often conflicting) objectives, of making trade-offs between different resource uses, and of trying to create as much benefit as possible from limited resources. When we are faced with the consequences of rationing decisions, the consequences are often indirect and sometimes unintended. Indeed, it is not easy to find simple real-life examples of health care rationing. But that is because real-life examples are complex, not because they are rare.

It would be a mistake to think that rationing decisions are intended to make your life harder or to deny you benefits to which you should be entitled. On the contrary, the rationing of health care ought to serve the purpose of benefiting everyone. But not everyone can be benefited all the time. Your medicine might not be subsidized because it provides little benefits to patients, and it is better to spend the money on medicines that provide more significant benefits. But you are more likely to take notice when you have to pay the full costs, and less likely to take notice when you benefit from not having to pay the full costs. You will not be thinking on these occasions about the benefits of health care rationing. If health care resources are allocated fairly and efficiently, everyone benefits. But people take the benefits for granted.

This is why the ethics of health care rationing is so important. If health care resources are allocated unfairly and inefficiently, many people will fail to receive benefits that they should, and could, get. This is morally wrong. But even when the allocation of health care resources is fair and efficient, there must be limits on what can be provided. Some patients will be disadvantaged by these limits. Thus, the limits must be morally justified. Otherwise, they impose unacceptable burdens on those patients.

Our discussion so far has left one question unaddressed. Why is rationing in health care inevitable? The answer is *scarcity*. Health care resources are scarce. This is why we must set priorities. But this answer just leads to another question. Why are health care resources scarce? Why can't we simply spend more on health care so that there is no need for rationing resources?

1.3 The inevitability of rationing health care

There are many reasons for the scarcity of health care resources. Some of these are technological: in the last few decades, medicine has made enormous advances. It is now possible to cure many previously fatal diseases and to manage long-term chronic conditions. We can now do more than ever before to restore and maintain health. But being able to do more also means spending more. The expansion of health services accounts for most of the increase in health care spending that has taken place over the last 50 years or so. This will continue. Our increasing understanding of genetics, for example, promises to lead to new and usually more costly therapies. As our armory to fight diseases expands, the pressure on health care budgets is going to increase further.

Other reasons are demographic. Life expectancies are increasing almost everywhere. Meanwhile, in many countries, fewer children are born. Aging societies spend more on health. Usually, people need health care the most in their very first years and then in their last few years. In many countries, aging accounts for a substantial share of the increase in health care spending. Since the populations of more and more countries are beginning to age, we can expect the growth of spending to continue.

There are also economic reasons. It is difficult to design a health care system that works efficiently. As a patient, you usually do not know enough about your condition to decide on the best treatment. That decision needs expert knowledge. Thus, you are not like a consumer looking for something to eat for lunch, who can use her experience and easily available information to make an informed choice. You rely on your doctor to tell you what you need. Moreover, patients often do not directly bear the costs (or all of the costs) of health care services. Since they are less sensitive to costs, they tend to demand more. When it comes to your health, it is better to be sure. An additional diagnostic procedure may just bring you peace of mind, even if, from a medical perspective, it is unlikely to be useful.

At the same time, doctors are often in a difficult situation. They are obligated to give you the best diagnosis and treatment. But they are also expected to act as gatekeepers – making sure that you use only the services that you really need. There can be a tension between their obligations to you as a patient and their role in ensuring that medical resources are used well. It is difficult to find the appropriate balance between these obligations. If doctors, hospitals, or other actors in the health care system are not sensitive to costs, they are more likely to contribute to the misuse and waste of resources. The problem, to put it in the economist's terms, is that incentives are often distorted in the health care system. Controlling costs is difficult.

Overuse of resources can obviously lead to scarcity. But sometimes this can happen in striking ways. In recent years, researchers have raised the alarm that our indiscriminate use of antibiotics might lead to the emergence of resistant strains of bacteria. For instance, there are worries that extensively drug-resistant tuberculosis might lead to an epidemic in the future. If cheap, easily available antibiotics do not work anymore, health care systems have to rely on more costly alternatives. This is harmful for everyone – each dollar that has to be spent on more expensive antibiotic treatments could have been used elsewhere in the health care system. Everyone would benefit if the use of antibiotics was more tightly controlled.

The way health care resources are distributed can itself contribute to scarcity. According to the latest available data (2019), the United States spends 17.7 per cent of its GDP on health care, roughly double the average of the OECD countries, a rich-country club. Yet, almost 30 million people – around 9 per cent of the population – have no health insurance. (The number of uninsured decreased by 20 million in the ten years after the Affordable Care Act was enacted in 2010.) Worse, average life expectancy in the United States is lower than in many other countries, including some that are much poorer. Americans do not get better health for the extra dollars they spend on health care. This suggests that a lot of their spending is less efficient than it should be.

Our example about the vaccination programs illustrates yet another way that scarcity can arise. Interventions and health care services are seldom sufficiently divisible to ensure equal access. If you are worried about the distribution of income, you can, in principle, redistribute it any way you like (since money can be divided up as finely as you want). But you cannot redistribute health care resources the same way. You cannot build a hospital in every village. Decisions about the location of health care infrastructure and the organization of health care delivery inevitably create inequalities of access, which can itself be a source of scarcity. In the vaccination example and the 3 by 5 program, the costs of reaching some populations increased scarcity.

Hence, scarcity and access are closely related. We can agree that everyone should have access to basic health care services; no one should be excluded from the health care system. But equal access cannot mean access to everything by everyone. Limits must be set, and they inevitably create restrictions on access to particular interventions and services.

Unequal access is problematic for a further reason. People are generally more tolerant of income inequalities than inequalities in health and access to health care. They believe that inequalities in income and wealth, at least within certain limits, might be beneficial for society: they create incentives or reflect differential effort. But very few people believe that similar considerations apply to health. Health inequalities have no beneficial social effects, and they rarely, if ever, reflect "effort" (an issue to which we will return in Chapter 7). Thus, many people would consider inequalities in health and access to health care much more troubling than other forms of inequality. Even those who do not consider income inequality unfair might be worried about inequalities in health and in the delivery of health care.

For the reasons listed, scarcity is inevitable in health care. Since it is inevitable, rationing is indispensable: societies must try to allocate the available health care resources efficiently and equitably. This is the only way to avoid inefficiency, waste, and unfairness.

Some people want to resist this conclusion. Scarcity in health care, they argue, should not be managed. Instead, it should be eliminated. Since health is important, we should spend more on it. This objection basically says that there is always more money. You just have to find it.

There is an element of truth in this objection. Surely, sometimes the right response to scarcity is to get rid of it. There are still many people in low-income countries who do not get antiretroviral therapy. More should be spent to ensure that they do. Affluent countries could, and arguably should, do much more to help achieve this.

Even so, the objection underestimates the gravity of the problem. Suppose you become a powerful but benevolent dictator. Since you want to use your power to help people, you decide to eliminate scarcity in health care. You decide to spend enough to keep up with technological developments and scientific breakthroughs. You spend enough to meet every medical need in a rapidly aging society. You manage to eliminate economic inefficiencies and distorted incentives from the health care system. You introduce policies to provide equal access to everyone. Have you overcome the need for rationing?

For several reasons, you have not. First, even if the most apparent forms of scarcity are eliminated, others remain. Even if every potentially beneficial intervention is available, you still have to decide which to offer first – you cannot offer everything, everywhere, all the time. You still have to decide whether to organize a cancer-screening program or a maternal health campaign first. Time is a scarce resource. There is only so far you can go to eliminate scarcity by spending more money.

Second, you will soon realize that in health care, resource use and its costs can easily spiral out of hand. Suppose you introduce more successful cancer treatments. Because better treatments are available, more people are screened. Since more people are screened, more cases are found and treated. No doubt, it is better that fewer people die prematurely because of cancer. But screening and treatment have increased your costs exponentially, creating scarcity elsewhere in the health care system. So you have to increase spending further, which may in turn lead to further scarcity. Paradoxically, better health services can increase scarcity.

Third, eliminating scarcity itself requires rationing choices, since the only way you can get rid of scarcity is by setting priorities. You can only avoid inefficiencies if you spend resources the most efficient way. You can only achieve better health outcomes for the whole population if you take into account the benefits and the costs of interventions. You can only reduce health inequalities if you set the right priorities among the needs of different groups within the population. These are all rationing decisions.

At the end of the day, your "war on scarcity" is likely to leave you with a depleted budget. At this point, you are faced with the question: was it worth it? Health care competes with other social goods. When resources are spent on health, there is less for education, infrastructure projects, and national defense. Sometimes, resources spent on health would do more good elsewhere. Priorities must be set both within health care and between health and other social objectives.

So, no matter where you turn, you face the need for rationing. Even for a benevolent dictator, this must be very annoying.

Some people might object that since health care saves lives, expenditures on health (or at least on life-extending interventions) should have absolute priority. But there are other ways to save lives. Highway safety regulations are a more effective way to do it. Moreover, few people would agree, on reflection, that saving lives should always have priority. Later in the book, we will describe some examples of drugs that can extend the lives of people with terminal cancer by a few weeks or months – at enormous expense. All the money spent on these treatments has to come from somewhere. There is nothing morally objectionable in asking whether these treatments are worth their costs. There might be a point at which saving lives is just not worthwhile any more.

Other people reject the need for rationing for another reason. Health, they argue, is fundamentally important to well-being. Because it is so important, you are entitled to it: you have a right to health. If you have a right to something, then you should be provided with it, and you should be provided with it even when the cost–benefit calculation is unfavorable.

But the idea of a right to health is ambiguous at best. At some point in your life, you will inevitably fall ill. You will die someday. Are your rights violated then? Who violates them? It is better to treat the right to health as a right to *health care*. But this is problematic too. Do you have a right to all sorts of health care, no matter how little the benefits? What about the costs? (Do you have to bear the costs yourself? If you have a right to something, should you bear its costs? What if you cannot afford it?) The right to health care had better not be interpreted as the right to any amount or form of health care. At the most, it should be interpreted as a right to *basic* health care: as the right to fundamentally important forms of health care.

Interpreted this way, this proposal just takes us back to the original issue. You have to decide which interventions and services are basic or fundamentally important. Surely, those that have the greatest benefits or prevent the greatest loss in health should belong to this group. Interventions and services that bring little benefit should not. But making this distinction requires you to settle questions of priority. It does not liberate you from the need to face the question of rationing – in fact, it requires it. Treating some forms of health care as a matter of rights does not avoid the problem. It just conceals it.

Scarcity is always present in health care systems. Therefore, rationing is inevitable. If you try to eliminate or minimize scarcity, you have to set priorities. That also requires rationing. You cannot escape it. Since rationing is inevitable, it is all the more important to get it right. Since it is a moral issue, "getting it right" requires thinking carefully through the ethical questions that it raises. This is what the following chapters will help you do.

Before we embark on this project, we need to introduce some general ideas about ethics. How can we settle moral questions? How should ethical argument proceed? What are the main ethical concepts and theories that are relevant to the topic of this book? We will now look at these questions before returning to health in Chapter 2.

1.4 Moral argument

In everyday life, we all make moral judgments. We say it is wrong to tell a lie; it is wrong to bully classmates; it is right to donate to charities working on alleviating poverty and suffering. Ethics is the branch of philosophy that studies right and wrong. It is concerned with developing and defending principles and theories that can be used to determine which acts or policies are right and which are wrong. It helps us decide which moral judgments to accept and which to reject.

Many moral judgments are uncontroversial. Few people would deny that lying or bullying is wrong and that giving to charity is right. Other moral judgments are the source of deep disagreement. Some people believe, for instance, that nontherapeutic abortion is morally wrong; others believe it is morally permissible (i.e. not morally wrong). Part of what makes their disagreement so divisive is that moral principles apply to everyone – even those who do not accept them. That is, moral judgments are interpersonally valid. If lying is wrong, then it is wrong no matter who does it. It is wrong for you to tell a lie even if you think there is nothing wrong with lying. By the same token, if abortion is in fact morally permissible (or wrong), then it is morally permissible (or wrong) regardless of what your own view happens to be. It is possible to be mistaken in one's moral judgments.

This basic point sometimes surprises those who are new to philosophy. They think that ethics is subjective. By "subjective," they usually mean that ethics is a matter of preference or opinion – so that what is right for one person may be wrong for another person in the same circumstances, depending on what they think or how they feel. For example, those who take ethics to be subjective would claim that the moral judgment about abortion depends on personal opinion. Some people believe abortion is morally wrong, so for them it's wrong; others believe it is permissible, so for them it's permissible. The moral judgment about abortion comes down to personal choice.

This train of thought is incorrect. In many countries, women have a legal right to abortion in the first trimester. Legally speaking, they have the right to decide whether they carry the fetus to term or terminate their pregnancy. Within the legal limits, abortion is a matter of personal choice. From this fact, however, it does not follow that the moral judgment about abortion is a matter of personal choice. Deciding what falls under the category of personal choice and what falls under the category of moral judgment is itself an ethical issue. You have to make moral judgments to determine the borderline around personal choice. That is, you must justify the moral judgment that the decision about abortion should be a matter of personal choice. You need a moral argument. You cannot take it for granted that abortion within the first trimester is a matter of personal choice.

Ethics is easily confused with something else. Sometimes ethics is confused with the law. But moral and legal judgments are different. For example, some people might think that morally wrong acts are just those that the law prohibits. Killing an innocent person is illegal, and this is why murder is morally wrong. But insofar as an action is legally permitted, it is morally permissible.

This train of thought is also incorrect. The law does not subsume ethics. Here is an example. Canada is one of the most liberal countries when it comes to nontherapeutic abortion (i.e. selective abortion). In most countries where it is legal, abortion beyond the first trimester is not legally permitted unless there is a serious risk to the health of the pregnant woman. In Canada, however, there is no legal restriction on abortion. It is permissible for a woman to request the termination of her pregnancy at any stage. There are only two practical restrictions. First, the termination of the pregnancy must be requested by the woman herself. Second, the abortion must be performed by a registered physician. Furthermore, in most provinces, the public health care system covers the full or partial cost of abortion. Thus, almost all cases of abortion are legally permissible.

Does this mean that every case of abortion is morally permissible? Clearly, it does not. Suppose a woman requests abortion simply because she prefers a boy to a girl, and a prenatal screening test predicts a baby girl. The reason for this particular request for an abortion is pure prejudice. Is abortion for sexist reasons morally permissible? There are no morally relevant differences between men and women. Abortion on the basis of sex is a form of discrimination against women. If that is correct, it follows that abortion for purely sexist reasons is morally wrong. Thus, there are some cases of abortion that might be morally wrong. Sometimes, legally permitted acts are not morally permissible. It follows that moral judgments are not equivalent to legal judgments.

Sometimes, people wonder: do the same moral principles apply to all people? Are moral principles universal? These questions concern the issue of *moral relativism*. Simply put, moral relativism holds that there are no moral principles or judgments that are valid or true in all societies and cultural and historical settings. What is right in one society may not be right in other societies. Some philosophers support relativism. Others reject it and support universal ethics. In this book, we will not take sides. But we do want to point out a difficulty for relativism. If you believe that the truth of moral judgments is relative to particular societies or cultures, you still need to explain what makes them true. For, presumably, you do not want to say that moral questions can be answered simply by taking a poll. You cannot seriously believe that the view that the majority holds is always automatically right merely because the people who hold that view outnumber the people who disagree. It is obvious that this cannot be correct.

In any case, moral relativism does not imply that "anything goes." Moral relativism is not the view that moral judgments are subjective or arbitrary. Relativists and those who hold that moral principles are universal both believe that moral judgments can be true or false and that there are good and bad moral arguments.

This point leads us to a more general question: what is distinctive about moral arguments? How can moral claims be defended?

In many respects, moral arguments are similar to other sorts of arguments. They require a valid inference from premises to the conclusion; they are good arguments only if the premises are true and mutually consistent. One thing that is distinct about moral arguments is that they have normative premises: claims about what is good or bad, permissible or prohibited, what ought or ought not to be done. Often, these premises are based on *moral intuitions* – strong convictions about the rightness (or wrongness) of some kinds of actions that just "seem right" (or wrong).

A lot of ethics is concerned with discovering, clarifying, evaluating, and systematizing moral intuitions in order to use them in moral arguments. Hence, moral intuitions need not be arbitrary or unjustified at all. (In this respect, talk of intuitions can be very misleading.) In later chapters, we will engage in the discovery, clarification, evaluation, and systematization of such intuitions. We will engage in moral argument. It is helpful to highlight some of the features and methods that are common in constructing moral arguments.

First, moral argument often uses thought experiments. We have already seen one example of this: the example of the vaccination programs was a thought experiment. It was designed to discover your intuitions in order to identify relevant moral considerations. Thought experiments usually ask you to assume that "other things are equal." This expression is used to simplify an example and enable you to focus on one particular feature. Especially in difficult choices, people are sometimes tempted to "solve" the moral question by introducing some additional assumption into the thought experiment. But changing the example is against the rules of moral argument. It defeats the purpose of the thought experiment.

Here is another example. When we discuss the relevance of age in the allocation of health care resources, we will consider examples where age is the only difference and all other features are equal. One example might be this:

> Imagine that you are faced with a choice between saving the life of John and saving the life of George. John is 20 years old, and George is 70 years old. Whoever is saved would live for another ten years in full health. Everything else is equal. What is the right thing to do?

As we will discuss in Chapter 5, if your moral intuition is that John should be saved, then it might reflect the idea that age is a morally relevant consideration in deciding whom to save. If your intuition is that we should be indifferent between (but not toward) saving John's life and George's life, then it might reflect the idea that age is a morally irrelevant consideration. This toy example is set up in such a way that the only difference between the two people is their age. This enables you to focus on the moral relevance or irrelevance of age. This is why you are asked to assume that all the other features are equal. If you assume that George is a Nobel Prize winner and John is a criminal, the example becomes more complicated. To avoid complication, you must keep in mind that other things are equal.

Second, we use examples and thought experiments not only to identify moral intuitions but also to test whether they are consistent or not. Many people have conflicting moral intuitions, and the conflict can be discovered by looking at different examples. You have a certain intuition in one example. You consider the idea behind the intuition. Now you are asked to think of another example and consider what that idea implies in this second example. If you have conflicting moral intuitions in the two examples, a genuine moral problem arises. You have to find a way to reconcile your intuitions.

Here is a famous illustration. Many people have conflicting intuitions about abortion and infanticide. One is that abortion is morally permissible. The other is that infanticide is morally wrong. Why are these intuitions in conflict? A typical explanation for the permissibility of abortion is that the fetus is not a person. Those who have this view believe that destroying a fetus is different from killing a person. Killing a person is wrong, but destroying a fetus is not wrong. What is the difference between a fetus and a person? One answer is that a fetus does not have self-consciousness whereas a person does. That is, self-consciousness is the criterion for drawing the moral difference between a fetus and a person. But if this idea is correct, how is it possible to justify the intuition that infanticide is morally wrong? A five-day-old newborn does not have self-consciousness. The only difference between a fetus and a newborn is that the fetus is in the womb, and the newborn is outside of it. Some infants are born prematurely, sometimes 15 weeks before the due date, or late, sometimes four weeks after their due date. Thus, the date of birth is an arbitrary cut-off point. Such an arbitrary cut-off point does not suffice to make any moral difference. So there does not seem to be any morally relevant difference between a fetus and a newborn baby. Therefore, the moral judgment about abortion must be the same as the judgment about infanticide: if abortion is morally permissible, then it must be the case that infanticide is also morally permissible.

There are two options now. The first is to give up one of the intuitions. That is, you have to give up either the intuition that abortion is morally permissible or the intuition that infanticide is morally wrong. By giving up either intuition, you can keep the consistency of your moral judgments.

The second option is to give up the idea that self-consciousness is the criterion of personhood. If you take this option, you must come up with a different view about what it is to be a person. This is a serious philosophical task. We will not go into it any further. The point is that once you identify your moral intuition in one case, you must ask how far your intuition can go in terms of consistency by testing it in other cases.

You might ask: why does consistency matter? Why should you want to make your moral beliefs consistent? The answer is that consistency is indispensable in any theory in any academic discipline. Who would believe in an inconsistent theory? A set of inconsistent intuitions is too arbitrary and never constitutes an ethical theory. There is no reason to accept arbitrary moral judgments. This is why one of the main tasks of ethics is to systematize intuitions by justifying them in a coherent moral framework. This is what moral principles and ethical theories attempt to do.

The third feature of moral argument that we need to highlight is the issue of the "burden of proof." Let us stick to the example of abortion and infanticide. Those who think that both abortion and infanticide are morally wrong have consistent judgments on these issues. They argue that infanticide is morally wrong (as many people agree) and that there is no morally relevant difference between a fetus and a newborn baby. It follows, according to their argument, that abortion is morally wrong. There is no inconsistency problem for those who take both abortion and infanticide to be wrong. They do not need to prove anything. The problem is for those who support abortion and reject infanticide. They must establish a morally relevant difference between fetus and newborn if they are not willing to give up one of their intuitions. The burden of proof is on them. Obviously, this does not mean that they have lost the argument. It just means that the ball is on their side of the court. They have to make the next move.

Let us return to the vaccination example with which we began this chapter. It seems that very few people hold that the only relevant consideration is to maximize the benefits of the vaccination programs, regardless of the way the benefits are distributed. Similarly, few people agree that the only relevant consideration is to give an equal chance of getting the benefits to all the children (which we interpreted as a consideration of fairness). Most of us have the intuition that both benefit maximization and fairness are morally relevant. So the burden of proof is on us to fit these considerations into a coherent moral framework. As a first step, we can try to see how far taking only one of these considerations into account can take us. We will follow this methodology in the next couple of chapters, focusing on benefit maximization. But we will also broaden our analysis as we go along, in part by identifying our intuitions about fairness.

But since the two considerations – benefit maximization and fairness – are commonly thought to fall into different categories of normative concepts, we should briefly introduce these two broad categories. One of them is the category of *deontic* concepts. They include "right" and "wrong," "fair" and "unfair." The other category is that of *axiological* concepts. They include "good" and "bad," "benefit" and "harm."

There are two broadly defined approaches in ethics, corresponding to the way deontic and axiological concepts are related. One approach is *consequentialism*. On consequentialist theories, deontic concepts depend on axiological concepts. For instance, consequentialists might say that the rightness or wrongness of an act depends solely on the goodness or badness of its consequences. So for consequentialists, right and wrong are solely a matter of consequences.

The second approach is *deontology*. According to deontological theories, deontic concepts are independent of axiological concepts. That is, the rightness or wrongness of an act can be determined independently of good and bad – in particular, the goodness or badness of the act's consequences. For consequentialists, rightness is a matter of the value of outcomes; for deontological approaches, rightness and wrongness are a matter of something else. For instance, they may be a matter of individual rights, or they may be a matter of fairness. However, one complication is that it does not follow that the goodness or badness of consequences does not matter for deontology. Most contemporary deontologists accept that the goodness of the consequences of an act is one factor that can affect the ethical status of that act. They just insist that it is not the *only* factor.

Here is an example to make this clearer. Think of the moral judgment about torture. Many people think that torture is wrong. Why? Typically, there are two types of explanation. The first is that the act of torturing itself is wrong and that this moral judgment has nothing to do with how much good it would produce. According to this explanation, the act of torturing is wrong under any circumstances, regardless of the goodness of the consequences that torture can bring about. For example, torturing a terrorist who set a bomb to kill innocent people is wrong, even if it is the only way to obtain information on where the bomb is hidden, so that you can defuse the bomb and save the lives of a thousand innocent people. This type of explanation is based on deontology.

The second type of explanation for the wrongness of torture is that the badness of torture outweighs the goodness of its consequences. According to this explanation, torture is wrong in most cases, but it can be permissible in some cases when the good effects of torture outweigh its bad effects. For instance, if you can defuse the bomb and save the lives of a thousand innocent people, the goodness of saving the lives of a thousand innocent people can outweigh the badness of the suffering of the terrorist. This type of explanation is based on consequentialism.

The difference between the two approaches can also be thought of in the following way. Consequentialists believe that ethics is about promoting the good. The right act is that which has the best consequences. Deontologists believe that ethics is primarily about complying with duties that may have nothing to do with the goodness of consequences. For them, ethics is concerned with *constraints* on the promotion of the good. Rights and fairness are constraints: an act that would have the best consequences may nevertheless be wrong if it violates a right or if it leads to unfairness.

In this regard, the approach taken in this book is not purely consequentialist. Even though we will argue that the allocation of health care resources should aim at the best consequences – to achieve the best health outcomes – we will also take into account various constraints on the maximization of health benefits. But once again, things are more complicated: perhaps at a higher level of abstraction, deontic concepts, such as fairness, can be given a consequentialist justification. We will not address this issue. Ultimately, we remain neutral between the two approaches.

One argument that will emerge from the rest of this book is that fairness has a central role in the ethics of health care rationing. The maximization of health benefits is certainly important; however, it must be constrained and qualified by the concern for fairness. The notion of fairness crops up in every stage of this book. But philosophers disagree over what fairness demands in different contexts. Therefore, we do not attempt to give an account of fairness. Rather, we will examine it in different contexts of health care rationing and then attempt to draw some lessons in the Conclusion.

Chapter summary

Health care rationing is the controlled allocation of health care resources. It is ubiquitous in every health care system, even if rationing choices are not always readily apparent. Because of the scarcity of health care resources, rationing is also inevitable. Health care rationing is an ethical issue, and it needs to be governed by ethical principles. Two relevant, basic moral ideas are the maximization of the benefits from the use of health care resources and the fairness of the distribution of those benefits.

Discussion questions

1 "The health care system and health insurance markets are regulated by the government. In many countries, democratically elected politicians run the government. The decision concerning health care rationing should therefore be made by politicians. We do not need any ethical principle for rationing health care." Do you agree with this argument? Why or why not?

2 Imagine that you are in a position of making decisions about the use of health care resources. You have $100,000 in hand. You can use it either to fund an expensive treatment of a rare form of disease for Jessica – a young child whose plight has been presented on national television – or you can use the money to fund a public health program to reduce children's risk of exposure to lead paint. Statistically, the program is expected to save the lives of two children, who are not yet identified. Some people

value helping an identifiable victim more than a statistical victim, hence preferring to use the funds to help Jessica. Is such a bias toward identifiable victims ethically justifiable?

3 Many people are upset by the idea of rationing health care. At the same time, funding bodies regularly set priorities for the allocation of funds for medical research. This is a form of rationing, too. Medical research can benefit patients, but when some research projects get low priority, some patients can be disadvantaged. Yet, no one objects to the controlled allocation of resources for medical research. Is rationing more acceptable in the setting of research priorities than in health care? Why or why not? What is the moral difference between them?

4 People disagree about many moral questions, and there is a great diversity of moral beliefs among different cultures. Do you think that the facts of disagreements and diversity provide an argument for moral relativism? Why or why not?

5 When Seattle's Swedish Hospital started offering kidney dialysis to a limited number of outpatients in 1962, a committee consisting of laypeople was set up to make decisions concerning who should receive the treatment from a pool of needy patients. This committee became known as the *God Committee*. The following is a part of the committee members' discussion, famously reported in Alexander (1962, 110). In your view, which considerations in this discussion are reasonable, and which are unreasonable, for determining who should receive the treatment?

LAWYER: The doctors have told us they will soon have two more vacancies at the Kidney Center, and they have submitted a list of five candidates for us to choose from.

HOUSEWIFE: Are they all equally sick?

Dr. MURRAY (John A. Murray, M.D., Medical Director of the Kidney Center): Patients Number One and Number Five can last only a couple more weeks. The others probably can go a bit longer. But for purposes of your selection, all five cases should be considered of equal urgency, because none of them can hold out until another treatment facility becomes available.

LAWYER: Are there any preliminary ideas?

BANKER: Just to get the ball rolling, why don't we start with Number One – the housewife from Walla Walla.

SURGEON: This patient could not commute for the treatment from Walla Walla, so she would have to find a way to move her family to Seattle.

BANKER: Exactly my point. It says here that her husband has no funds to make such a move.

LAWYER: Then you are proposing we eliminate this candidate on the grounds that she could not possibly accept treatment if it were offered?

MINISTER: How can we compare a family situation of two children, such as this woman in Walla Walla, with a family of six children such as patient Number Four – the aircraft worker?

STATE OFFICIAL: But are we sure the aircraft worker can be rehabilitated? I note he is already too ill to work, whereas Number Two and Number Five, the chemist and the accountant, are both still able to keep going.

LABOR LEADER: I know from experience that the aircraft company where this man works will do everything possible to rehabilitate a handicapped employee . . .

HOUSEWIFE: If we are still looking for the men with the highest potential of service to society, I think we must consider that the chemist and the accountant have the finest educational backgrounds of all five candidates.

SURGEON: How do the rest of you feel about Number Three – the small businessman with three children? I am impressed that his doctor took special pains to mention this man is active in church work. This is an indication to me of character and moral strength.

HOUSEWIFE: Which certainly would help him conform to the demands of the treatment . . .

LAWYER: It would also help him to endure a lingering death . . .

STATE OFFICIAL: But that would seem to be placing a penalty on the very people who perhaps have the most provident . . .

MINISTER: And both these families have three children too.

LABOR LEADER: For the children's sake, we've got to reckon with the surviving parents' opportunity to remarry, and a woman with three children has a better chance to find a new husband than a very young widow with six children.

SURGEON: How can we possibly be sure of that? . . .

6. In countries where health insurance is provided by private insurance companies (as well as in countries where people can purchase private health insurance in addition to public health insurance), health insurance markets are highly regulated. Some of the regulations concern the kind of rationing that private insurers are permitted to undertake (e.g. with regard to exclusions based on preexisting health conditions). If you were to advise the government, what forms of rationing would you recommend the regulations should and should not permit? What forms of rationing should or should not health insurance companies be allowed to use?

Further readings

The vaccination program example is discussed in the context of HIV/AIDS and in much more detail by Johansson and Norheim (2011). The example originally comes from Daniel Wikler, who generously agreed to lend it to us. Sreenivasan (2012) offers an argument for health care rationing from the perspective of justice; see also the other contributions to the volume of which it is part (Rhodes *et al.* 2012). The problem of abortion and infanticide is introduced by Tooley (1972). If you have never studied ethics before, a good introduction is given by Driver (2007). Timmons (2013) provides not only a slightly more advanced, but also more detailed, introduction.

2 The value of health

2.1 Well-being and health

The rationing of health care resources includes the controlled allocation of things such as subsidies for medicines, operating costs for hospitals, places on waiting lists, organs for transplantation, or funds for public health programs and medical research. It also involves deciding which interventions and services should be covered by health insurance packages. In emergencies, it might be necessary to ration beds in intensive care units, vaccines in areas affected by epidemics, or emergency medical personnel to different locations. These allocation choices must be efficient and fair: they must lead to the best consequences while taking into account relevant moral constraints. But the things that are allocated are very different. How can we decide which ones of the many possible allocations are fair and do the most good?

The answer to this question may at first seem straightforward. The objective of health care is to restore and maintain health and to prevent and alleviate suffering due to ill-health. Obviously, you cannot literally redistribute health itself. Unlike income, health cannot be taken from one person and given to another. You cannot restore a patient's health by taking some health from someone else and giving it to her. Still, you might be able to use health as a *metric* to compare different resource allocations. You can try to measure the degree to which different allocations contribute to restoring and maintaining health. Thus, you might be able to say that heart surgeries do more to restore and maintain health than hip replacements. They are more important, so they should have higher priority.

But this answer faces a difficulty. Health is not some sort of natural quantity that can be measured on a common scale – as opposed to distance or blood pressure. Compare a person who has poor eyesight to another who has poor hearing. Neither of them can function as well as others. Both of them fall short of what is typical of healthy human beings. But which of them falls short *more*? Which of them is unhealthier? Good health is made up of many kinds of physical, biological, mental, and psychological functions, which do not have a common metric. You cannot simply look at the "amount" of health that people have, because health does not come in one sort of quantity.

DOI: 10.4324/9781003050216-3

To be sure, some comparisons are easy enough to make: a person who has asthma or a broken arm is less healthy than a person in perfect health. But other comparisons seem intractably difficult. Are you less healthy if you have asthma or if you have migraines? Are you less healthy if you have a broken arm or a broken leg? Are you less healthy if you are deaf or if you are blind?

Consequently, you cannot compare alternative uses of health care resources by measuring the extent to which they contribute to restoring and maintaining health or to preventing ill-health. It is impossible to directly compare different aspects of health. There is no metric of health that helps you determine whether asthma medication restores functioning to a greater degree than back pain medication. They restore different functions. If you have to choose between providing asthma medication or back pain medication, you cannot make your choice by determining which makes people healthier. Both asthma and back pain medication can improve health, but it is not possible to directly determine which leads to a greater improvement.

Some readers may find this argument too hasty. After all, surely asthma is worse than back pain. Breaking a leg leaves you worse off than breaking an arm. Being deaf is less of a disadvantage than being blind.

These comparisons may or may not be true. But the thing to note is that they are not comparisons of health. They are not claims about which conditions represent more or less health. Rather, they are comparisons of *value*. What they say is that it is *worse* to be with some of these conditions than to be with others. So they provide no objection to our argument. Instead, when people make such claims, what they mean is that a condition is worse when it makes life more difficult, when it leads to less well-being, when it creates disadvantage. The distinction is important. It is one thing to try to measure health; it is another to measure the *value* of health.

Fortunately, a measure of health is not needed for the purposes of resource allocation. For, ultimately, we do not much care about health itself. What we do care about is its value for us: the way it affects our well-being or quality of life. (In this book, we use these terms interchangeably.) Disease and injury lead to a loss of quality of life: they cause pain, worsen functioning, or shorten lives. The point of medical interventions and health care services is to make life better by alleviating pain, restoring functioning and prolonging life. Consequently, when we allocate health care resources, we should be interested in their impact on quality of life. In other words, what matters is the impact of health on well-being.

Of course, the next question is what well-being is. Unfortunately, there is no generally accepted theory in philosophy. There are many rival theories. Fortunately, however, we can remain neutral between them. Here's why. No one would deny that health makes a major contribution to our well-being. As philosophers say, it has *instrumental value* for us. On any plausible theory, health will be important because of its instrumental value. But on some theories, health will additionally have *intrinsic value*: good health in itself is one of those things that make life good. On these views, health is part of well-being.

Whichever kind of view you take, health will be important. So, we can remain noncommittal in this book about theories of well-being, since our interest is in health. More precisely, our interest is in *health-related quality of life*: that fraction of overall well-being that is determined by health. To discuss it, we do not have to take a view on whether health has intrinsic or instrumental value. For simplicity, we shall simply say that health is a *component* of well-being, leaving it open whether it just contributes to it or is itself a part of it.

Let us take stock. We have argued for the following so far. The allocation of health care resources is an ethical problem. Because scarcity is inevitable, resources should be distributed in a way that does the most good. (Another aim, to be addressed as we go along, is that they should be distributed fairly.) As it is often put, the utilization of health care resources should provide "the best value for money." But how do you find out which interventions and services provide better value for money than others? A straightforward answer is that you can do this by measuring health. But health is not a quantity; it involves many functions which are impossible to compare. So you have to proceed indirectly, by measuring the value of health – its impact on quality of life. Other things being equal, an intervention provides more value if it has a greater positive impact on quality of life. The more it increases health-related quality of life, the greater its value.

We should note that not everyone agrees with the way we think about well-being and health. For example, the World Health Organization (WHO) defines health as "a state of complete physical, mental, and social well-being, and not merely the absence of disease or infirmity." Our view about well-being and health is incompatible with this definition of health. But we should not accept this definition. For one thing, it is implausibly expansive: it *identifies* health with well-being. But, obviously, there are other things beside health that contribute to physical, mental, and social well-being. Happiness and an adequate material standard of living are two plausible candidates. The WHO's definition would turn them into matters of health. We should be more modest: health is valuable because it is a component of well-being, not because it exhausts it. When we measure the value of health, we do not measure all of well-being.

Yet, there is a serious difficulty for our view that the ethics of health care rationing should focus on health-related quality of life. The view assumes that you can put a value on the impact of health on overall well-being – that you can measure health-related quality of life independently of other components of well-being. But this requires separating the contribution of health to well-being from the contribution of other components. It requires taking a measure of a person's overall well-being and telling how much of it is due to her happiness, standard of living, health, and so on.

We are sure it can immediately be seen that it is very unlikely that well-being can be measured this way. Even if you could independently measure health-related quality of life, happiness-related quality of life, standard of living, and so on, it is incredible that well-being is simply the sum of them, or that they can be put together in some other simple way to make up overall well-being. This is because different components of well-being interact: their impacts are inseparable from one another. Asthma is worse for someone who enjoys working outdoors. A finger injury is worse for a concert pianist than an opera singer. Back pain is worse for someone who takes care of small children than someone who works in an office. The value of health is not separable from the value of other components of well-being.

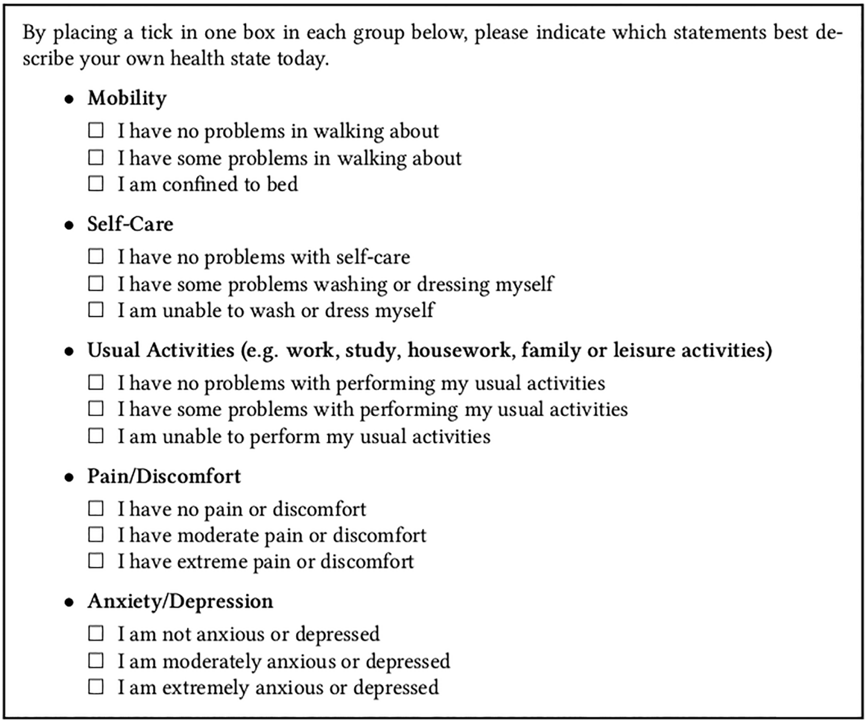
This problem is widely recognized. No one disputes that the impact of health cannot be separated from the impact of other components of well-being. But there is much less agreement on how serious the problem is. Some philosophers have argued that since we cannot evaluate health as a component of well-being, we should simply give up and try to measure overall well-being instead. We could then choose between alternative resource allocations by determining which makes the lives of people go best, all things considered.

But it seems a bit extravagant to try to allocate health care resources by taking into account all the different ways in which health can interact with other components of well-being. Consider just the information needs of such a proposal. The effectiveness of every intervention and medical procedure would depend not only on how they improve functioning, whether they remove all symptoms, or how many years they add to a patient's life, but also on how important the improvement of a particular function is for the particular patient, how it would affect that patient's happiness, standard of living, or any other component of her well-being. You would have to collect all this information for every single patient! Needless to say, this would be prohibitively costly.

This is not to say that such information is never relevant. A physician in a hospital or in general practice can and should take into account the impact of a condition and its treatment on particular patients, with their different values and circumstances. But in the allocation of health care resources, our focus is on populations. We need to abstract away from the differences of individual patients and consider the badness of a condition in general terms. We should acknowledge that any measure of health-related quality of life is an approximation. While the inseparability problem cannot be avoided, a measure of health-related quality of life can be interpreted as expressing the typical or average impact of health on well-being. After all, it sounds plausible that on first approximation a broken wrist is just as bad for you as for me. Of course, on second look, it may be worse for you if you are a concert pianist. But even though every patient is different, in large-scale resource allocation choices we are forced to abstract away from such individual differences.

2.2 Health-related quality of life

How can we put a value on the impact of health on well-being? How can we measure health-related quality of life? Consider particular diseases and injuries. Diabetes, asthma, depression, or HIV have very different impacts



By placing a tick in one box in each group below, please indicate which statements best describe your own health state today.

- **Mobility**

  □ I have no problems in walking about
  □ I have some problems in walking about
  □ I am confined to bed

- **Self-Care**

  □ I have no problems with self-care
  □ I have some problems washing or dressing myself
  □ I am unable to wash or dress myself

- **Usual Activities (e.g. work, study, housework, family or leisure activities)**

  □ I have no problems with performing my usual activities
  □ I have some problems with performing my usual activities
  □ I am unable to perform my usual activities

- **Pain/Discomfort**

  □ I have no pain or discomfort
  □ I have moderate pain or discomfort
  □ I have extreme pain or discomfort

- **Anxiety/Depression**

  □ I am not anxious or depressed
  □ I am moderately anxious or depressed
  □ I am extremely anxious or depressed

Figure 2.1 The EQ-5D (3L) questionnaire.

on a patient's life. They affect differently the way a patient is able to function biologically, psychologically, or socially. How can we express their impact in a single, summary value?

Broadly speaking, there are two approaches. In this section and the next, we present the most commonly used approach. In Section 2.4, we present an alternative.

The first approach focuses on *health states* rather than particular diseases and injuries. A health state is a description of different levels of functioning that patients can achieve in the presence of particular health conditions. It is a constellation of different functional limitations. Rather than measuring the badness of particular conditions, this approach evaluates health states directly. To see how this works, consider the EQ-5D, a widely used questionnaire for describing health states. It is reproduced here as Figure 2.1.

As apparent at first glance, this is a very simple questionnaire. Patients are asked to describe how well they function within five "dimensions" or aspects of health. Their answers to the questions define their health state. For instance, a patient who has no problems with walking, self-care, and performing daily activities, but has moderate pain and anxiety, will be in a different health state than a patient who has some problems walking, washing and dressing, and performing other daily activities, but no pain or anxiety.

The EQ-5D is intended to be simple, easy to fill out, and quick. It includes only five dimensions of health. It does not allow fine discrimination between different levels of functioning within these dimensions – in this version, there are only three descriptions to choose from. Even so, notice the large number of health states that can be described: three levels in five dimensions defines 3^5 = 243 different health states! The five-level version of the EQ-5D, which allows two additional levels of functioning for the five dimensions, describes 3,125 different health states. Even more detailed instruments are able to differentiate between tens of thousands of health states.

There are countless similar questionnaires. Some of them are general; others are targeted to the circumstances of particular patient groups. Some focus on the health outcomes of particular treatments and interventions. Some of them are short and simple, like the EQ-5D; others are longer and much more comprehensive.

It is important to note that the health states defined by the EQ-5D and similar questionnaires are simply descriptions. When a patient ticks the first box for the first three questions, and the second for the last two, her health state is very different from the health state of a patient who ticks the second box for the first three questions, and the first box for the last two. But based on this information, you cannot tell whose health-related quality of life is lower. The health states still need to be evaluated.

Thus, the respondents are given a second task. They are presented with a vertical scale that looks very much like a thermometer. It has a hundred grades, numbered in increments of 5 between 0 and 100, where 100 is defined as "the best health you can imagine," and 0 is defined as "the worst health you can imagine." The respondents are asked to put a mark on the scale that indicates their current health. Thus, in the first step, the researchers learn the health state a patient or respondent is currently in. In the second step, they learn how she evaluates that health state.

Suppose there are three respondents. The first one has no problem in any of the dimensions in the questionnaire. She rates her health as 100. The second respondent has some problems with performing daily activities and some moderate pain or discomfort, but she has no problems with mobility and self-care, and she is not anxious or depressed. This respondent rates her own health at 76. The third respondent ticked the middle box for all the questions: she has some problems with walking, self-care, and other daily activities, and she is also in moderate pain and moderately depressed. She gives the value of 52 to her own health.

We now have evaluations of three health states. For the sake of simplicity, we can put these values on a scale between 0 and 1, where full health has the value of 1 and the worst imaginable health state, not better than death, has the value of 0. Thus, the health-related quality of life of the first respondent is 1; the health-related quality of life of the second respondent is 0.76; and the health-related quality of life of the third respondent is 0.52. (Of course, in practice, values like 0.52 and 0.76 are averages, since they are determined by the responses of many people. Thus, researchers do not have to repeat the second task each time. The transformations from descriptions to valuations are already available from previous studies.)

The method for establishing quality of life values for health states that we have just presented is called the *rating scale* method. It is simple to administer for researchers and easy to understand for respondents. But it has a serious limitation.

In our example, the value 0.76 is associated with the health state characterized by some problems with performing daily activities, some moderate pain, but no problems in any of the other dimensions; the value 0.52 is associated with the health state characterized by some problems with walking, self-care, daily activities, and moderate pain and depression. Based on these values, the first health state is less bad than the second. Patients in this health state have a higher health-related quality of life than patients in the second. But can we say anything more than this?

Imagine that these states are the health outcomes for particular patients with and without treatment. Patient A is currently in the first health state: her health-related quality of life is 0.76. You can, however, provide a treatment to her that would restore her to full health – to the health-related quality of life of 1. The treatment would alleviate her pain and restore her ability to carry out daily activities. Patient B's health-related quality of life is currently 0.52. You can also treat her, but you cannot restore her to full health. All you can do for her is restore her mobility and ability to care for herself, as well as curing her depression. But she will be left with some moderate pain and problems with carrying out usual daily activities. The health outcome of her treatment would be the first health state with health-related quality of life at 0.76. Now the question is: which treatment would result in a greater improvement?

A simple answer is that the improvements are the same. Patient A would improve from 0.76 to 1. Patient B would improve from 0.52 to 0.76. The increases look the same – 0.24 in both cases.

But the simple answer is wrong. On the basis of the rating scale method, you *cannot* claim that the increases represent the same improvement in health-related quality of life. You cannot say this because the method establishes a ranking only. The *differences* between the values of this ranking do not have any meaning. They cannot be interpreted as measures of improvement. And that is a problem, since in health care what we are usually interested in is improvement.

The rating scale method provides very little information – it provides only a ranking of health states. To decide whether A and B would achieve the same degree of health-related quality of life improvement, a more precise scale is needed. You need a scale on which the *intervals* between different values can be compared.

To be fair, some researchers argue that the rating scale method does provide you with a scale on which such comparisons are possible. But this is extremely controversial. It is mysterious how having people indicating values on a thermometer-like scale could take you from a ranking to a measure with which the differences between values can be compared. The respondents only provide their rankings. How can you be certain that those rankings carry the necessary information for interval comparisons?

Plainly, you need a test for this. Researchers have developed methods to elicit the sort of valuation from respondents that make the construction of more precise scales possible. These can be used to test the values elicited by the rating scale method. But at this point, the whole problem seems to go away. For notice that once you use the other methods to test the rating scale method, you can simply go ahead and use these methods for the evaluation of health states directly. There is no need for rating scales.

What makes it possible to construct more precise measures on the other elicitation methods is that they use comparisons from the start. Respondents do not directly evaluate health states. Instead, they are asked to make trade-offs between living with different health outcomes. Their evaluation of any particular health state is indirect.

One of the best-known methods is called the *standard gamble*. In this method, respondents are given a health state description. For example, they are told that in one health state, patients have some moderate pain and discomfort, as well as some problems with performing usual daily activities. Then they are asked to make a choice. On the one hand, they can choose to live in this health state for a certain amount of time (e.g. ten years followed by instant death). On the other hand, they can choose to receive a treatment that will either restore them to full health with some probability p for the same amount of time, or lead to instant death with probability $(1 - p)$ (where $0 \leq p \leq 1$). Respondents have to determine the value for p at which they are indifferent between the two options. At this point, they would be just about as willing to take the gamble as to live in the health state.

In other words, p is varied until the respondents are indifferent. Suppose at this point p is 0.76. Respondents would be willing to risk death in order to be cured, as long as they have at least as great of a chance of survival as this. To put it a bit imprecisely, their responses reveal the *relative* value they place on the health state compared to full health and death. (More precisely, it reveals the relative value they place on the *differences* between the health state and full health, and the health state and death.)

We have now established the value of the health state characterized by some moderate pain and some problems with performing usual daily activities. It is 0.76. We can substitute any health state description in the question to determine its value. For instance, suppose that respondents are indifferent between living in a health state characterized by some problems with walking, self-care, and daily activities as well as moderate pain and depression, and a “treatment gamble” in which they have a 52 per cent chance of being restored to full health and a 48 per cent chance of instant death. The value of this health state is 0.52.

The values of health-related quality of life for these two health states are 0.76 and 0.52. Evidently, the second health state is worse than the first – people would be willing to take a greater risk to avoid it. But unlike before, these values provide us with a more precise scale for measuring health-related quality of life. Since they are based on trade-offs, these are relative values. They allow for comparisons of *changes* in health-related quality of life.

Return to our patients A and B. A's health-related quality of life is 0.76; B's is 0.52. You can return A to full health or you can improve B's condition from 0.52 to 0.76. The question was which treatment results in a greater health-related quality of life improvement. Armed with the scale provided by the standard gamble method, you can now say that the change from 0.52 to 0.76 and the change from 0.76 to 1 represent equal improvements – that is, they represent changes of the same magnitude in health-related quality of life. To put it a bit more technically, the standard gamble yields an interval scale that carries much more information than mere rankings. In particular, the ratio of differences – intervals on the scale – can be compared.

If the standard gamble looks a bit complicated, that's because it is. It can be time consuming to explain and administer, and its critics complain that it is not easy for research participants to understand. A similar but simpler method is the *time trade-off* method.

This time, the respondent is not faced with risky choices. Rather, she has to determine how much time (typically, in years of life) she would be willing to give up to avoid living in a health state that is worse than full health. Let us take our stock example again: on the one hand, a person can live for T years with some problems with performing daily activities and some moderate pain; on the other hand, she can live for X years in full health. Plainly, $X < T$, since it is better to spend a given amount of time in full health than to spend the same amount of time with less than full health. Consequently, the value of the health state is determined by X/T. For instance, if respondents consider 7.6 years in full health just as good as ten years with some problems with performing daily activities and some moderate pain, then the value of this health state is 0.76.

The time trade-off method yields the same sort of scale as the standard gamble. It also allows comparing the magnitudes of different improvements in health-related quality of life. Because it does not involve probabilities, it might be a bit easier to understand for respondents. But it is still more complicated than the rating scale method. Nonetheless, the standard gamble and the time trade-off have an advantage. When you are seriously ill, you might have to make trade-offs between longevity and health or take "gambles" between risky treatments. In other words, many patients in real life are faced with just these sorts of choices. From this perspective, the standard gamble and the time trade-off look more realistic than directly estimating the value of health states on some thermometer-like visual scale.

Earlier, we said that the rating scale method is unlikely to provide an interval scale. But if it yields fairly similar valuations, then perhaps you could argue that it can be used as a convenient shortcut to avoid the more complicated elicitation methods. As it happens, this is a much more controversial and complicated issue than it seems, with notoriously persistent disagreements between different researchers. Here are some results from one study that focused on functional limitations on the ability to walk unaided. One of the health states that were examined was described as "needing a walking stick when walking." Using the standard gamble, the value of health-related quality of life in this health state is 0.85; using the time trade-off method, it is 0.78; and using the rating scale method, it is 0.65. In general, it seems that the values are lowest on the rating scale and considerably higher on the time trade-off method and the standard gamble, with the latter yielding the highest values. Just what to conclude from these results is not entirely clear. But perhaps they suggest that the rating scale method is indeed problematic: when they are not forced to consider the need to make sacrifices and trade-offs, people tend to overestimate the badness of health states.

Nevertheless, both the standard gamble and the time trade-off method face their own problems. Admittedly, these problems are a bit technical. We will not go into them in detail. But the basic ideas are not difficult to understand.

Consider the time trade-off method first. In this method, respondents have to determine how many years of life they would be willing to give up to avoid living in a health state that is worse than full health. In the example we have worked with, respondents consider 7.6 years in full health just as good as 10 years in a health state marked by some problems with performing daily activities and some moderate pain. Therefore, the value of this health state is 0.76.

But this conclusion relies on a crucial assumption. The method assumes that when people consider their future health, their evaluations are not distorted by how far they look ahead in the future. To use the technical term, they do not *discount* their future health.

If people discount their future health, then they value good health in the near future more than good health in the distant future. They put a greater value on avoiding a bad health state next year than ten years from now. (Whether this is rational is a separate question that we will not take up here.) The problem is that if people do discount future health, then when they consider 7.6 years in full health just as good as ten years in a worse health state, their valuation may be distorted, because they put a smaller value on health in the further future. They would, for example, place the value of 0.5 only on the health state in itself. But since it appears less bad to them in the further future, the time trade-off that they are willing to make *now* makes the health state look less bad. So you get a distorted result.

This would not be a problem if you knew whether your respondents discount future health. If you knew the rate at which they discount, you could take it into account. But in practice, time trade-off studies have to assume that people do not discount future health at all. Perhaps the assumption is correct. But you cannot find this out from time trade-off questions. Hence you need the assumption.

The problem facing the standard gamble is similar. In this method, people make risky choices between treatment options. The outcomes of these choices are different health states. The method assumes that when respondents consider the options, their choices are determined only by the severity of those health states, rather than the risk itself. More technically, it is assumed that people's risk-attitudes toward different health states are constant. For instance, they do not become more sensitive to risk when they consider worse health states. The results will be distorted if respondents have different levels of willingness to take risks. Just as someone who has $10 and someone who has $1,000 might differ in their willingness to take gambles with their money, people's willingness to take risks with their health might depend on their own health state.

Again, you will not be able to find out from the responses to standard gamble questions whether respondents have the same risk-attitude in different choices. That is why you need to make an assumption.

What should we conclude from the discussion of these problems? On the one hand, it would be easy to become pessimistic about the prospects of measuring the value of health. The rating scale method suffers from a credibility problem: it is hard to believe that people can evaluate the badness of different health states at the required level of precision merely by placing them on a thermometer-like scale. The standard gamble has to assume that people have a constant risk-attitude toward health. The time trade-off method has to assume that people do not discount future health. These assumptions can be questioned.

On the other hand, we should realize that we cannot do without measuring health-related quality of life. It would be impossible to allocate health care resources efficiently and fairly without it. It is true that none of the methods for measuring the value of health is free of problems. Each of them requires simplifying assumptions. But health state valuations are merely approximations. It is inevitable that some imprecision will creep into them.

One thing to ask is how severe these imprecisions are. There are two sorts of answer to this question. First, we can test our methods. Researchers have looked at the consistency between the results given by the same respondents for repeated questions, as well as the consistency of the results from different groups of respondents. Overall, they have found that the responses are fairly consistent – as statisticians put it, they are reliable. This is good news, since it gives us some confidence that our measurements are close to the truth.

The second answer is to remind ourselves that we should not expect a perfect measure. Social science is seldom as exact as the natural sciences. Measuring health-related quality of life is especially difficult, but its difficulties are still less severe than the difficulties faced by other measures of well-being. Consider, for instance, measuring the gross domestic product (GDP) of a country. Think of the simplifications, approximations, and assumptions that go into calculating a country's GDP! The assumptions that must be made when measuring health-related quality of life seem mild indeed compared to them. Yet, GDP is regularly used to represent a country's "economic health," and policies are regularly evaluated by their expected impact on GDP.

For another example, consider the Human Development Index, a measure of the well-being of the population in a country used by the United Nations Development Programme and other organizations. It takes into account life expectancy at birth, mean years of schooling and expected years of schooling, and gross national income per capita. It was designed to overcome the limitations of GDP as a measure of well-being. Yet it is obviously a very crude measure. But that does not mean it is useless. Rather, you should keep in mind its limitations when you use it. The same applies to measures of health-related quality of life. They must play a role in health care resource allocation. But their limitations must be kept in mind.

2.3 Quality-adjusted measures

Suppose you are interested in the health of populations. Perhaps you want to compare the health of people living in two different countries. Or you want to compare the health of people in two socioeconomic groups within one country. You notice that one bad thing that disease and injury do to people is shortening their lives by killing them. Since longevity is valuable, you might decide to make these comparisons by looking at how long the people in these groups live.

But it is unclear what "how long people live" means. A population is made up of people of different ages. So you have to look at some average. But even that is insufficient. People in your population are still alive, so you have to look at how long they can *expect* to live.

It is customary to take life expectancy as a simple measure of health. But since people in your population are at different ages, they obviously have different life expectancies, even on average. Ten-year-old children have a different average life expectancy than 50-year-old adults. So what you could focus on is *life expectancy at birth*.

This is a familiar and widely used measure (it is one component of the Human Development Index, mentioned above). Suppose you discover that the population of one country has a higher average life expectancy at birth than the population of another country. You might also believe it is unfair that there are such differences in people's life prospects. You might argue that it is a matter of justice to help the second country to increase average life expectancy. Similarly, you might discover that people in a more advantaged socioeconomic group within one country can expect to live longer than people belonging to a less advantaged socioeconomic group. You might then argue it is a matter of justice to help the more disadvantaged have better life prospects.

Your argument is based, in both cases, on some moral principle and a simple measure of well-being – in this case, a measure of health-related quality of life, one component of well-being. You could not apply your moral principle if you did not have some measure like that. In fact, in the absence of a measure, you would not even be able to identify the states of affairs that you can consider unfair. You *must* use a measure, despite the methodological difficulties discussed in the previous sections.

But life expectancy at birth is a very crude approximation of well-being or even only of health-related quality of life. It provides too little information. It tells you about the "quantity" of life that people can expect to have, but it does not tell you anything about its quality. It does not tell you how healthy people are during their life. A better measure would take into account the quality of life as well as its quantity.

Fortunately, the health state evaluations that we have been discussing can be used for just this purpose. In the previous section, we looked at how you can assign values to different health states on a scale between 0 and 1. We interpreted 1 as full health. Now we can extend that interpretation: let 1 stand for *spending one year in full health*. Let values that are smaller than 1 stand for *spending one year in a health state that is worse than full health*. (Taking one year as the unit is yet another simplification.) This way, when you look at the health states that people can expect to be in throughout their years of life, you can assign values to them. Any year spent in full health has the value of 1; any year spent in less than full health has a value that is *adjusted* by the health-related quality of life for the health state during that year.

This is the measure of *health-adjusted life expectancy* (HALE). Suppose a person, at birth, can expect to live for 75 years. For most of her life, she can be expected to be completely healthy. In the last ten years, however, she can expect to suffer from chronic conditions. For five years, she can expect to have some problems with performing daily activities and to live with some moderate pain. Her health-related quality of life will be only 0.76 during this period. And for the last five years of her life, she can also expect to have some problems with walking and self-care, as well as being moderately depressed. Her health-related quality of life in these five years will be 0.52.

Her health-adjusted life expectancy is easy to calculate. She expects to spend 65 years in full health. Each of these years has a value of 1. Then she can expect to spend five years at the health-related quality of life level 0.76, followed by five years at 0.52. So her health-adjusted life expectancy is

$$65 \times 1 + 5 \times 0.76 + 5 \times 0.52 = 71.4.$$

Therefore, even though this person's life expectancy at birth is 75 years, her health-adjusted life expectancy at birth is only 71.4 years.

Obviously, there are going to be differences between people in any population. But at least in principle, the data for health-adjusted life expectancies can be collected. They can provide a more precise way to compare the populations of different countries and socioeconomic groups.

Moreover, the idea behind health-adjusted life expectancy can be further extended. It can be applied to any health outcome, including outcomes associated with particular interventions and treatments for different conditions. This general measure is called *quality-adjusted life year* (QALY).

A QALY is a combination of health-related quality of life and years of life: 1 QALY can represent one year of life in full health; or it can represent two years at health-related quality of life level 0.5; or it can represent four years at 0.25. For example, suppose that one treatment for cancer patients provides five years of remission at the health-related quality of life level of 0.4, while another treatment provides three years of remission at 0.7. The outcome of the first treatment is 2 QALYs; the outcome of the second treatment is 2.1 QALYs. The second treatment, taking into account both health-related quality of life and quantity of life, is more valuable. It results in more QALYs.

Note that quality-adjusted measures assume that the value of a health state is proportional to its duration. This does not seem an unreasonable assumption to us. But perhaps it does not always hold. For the time being, we can treat it as another simplifying assumption. If it turns out to be unrealistic, it can be modified.

QALYs enable us to compare all sorts of resource uses in health care. They can represent the value of the health outcomes of different treatments and interventions. They can represent the value of public health programs. They can be used to evaluate the health of particular patients, or patient groups, or even whole populations. In the next chapter, we will take a more detailed look at how they can help us make choices in health care priority setting.

2.4 The burden of disease

At the beginning of Section 2.2, we said there were two approaches to health-related quality of life measurement. The first focuses on the impact of ill-health on the different ways a person functions, defining health states in terms of shortfalls in functioning. This approach forms the foundation of QALYs.

The other approach focuses on diseases, injuries, and risk factors. It begins from a distinction introduced by the World Health Organization – the distinction between impairment, disability, and handicap. An *impairment* is the loss or abnormality in physiological, psychological, or anatomical functioning that is the direct consequence of disease or injury. It can be described in biomedical terms. A *disability* is a loss or restriction of ability, as a result of the impairment, to carry out an activity that is considered normal for human beings. And a *handicap* is the disadvantage that results from the impairment or disability that limits or prevents the individual to fulfill her role in her economic, social, and cultural environment. (The WHO no longer uses this distinction. Nevertheless, we think it remains useful for introducing the ideas below.)

When we evaluate the badness of different conditions, we can consider them as impairments, disabilities, or handicaps. But if we evaluate them merely as impairments, we are unlikely to capture their impact on well-being. And if we evaluate them as handicaps, there is going to be too much variation depending on the economic, social, and cultural circumstances of different people and populations. In order to find a middle ground between taking into account too little and too much information, we should evaluate them as disabilities.

This is the approach taken by the Global Burden of Disease project – an international attempt to measure the harm from mortality and morbidity from disease and injury in the populations of different countries and regions of the world. The harms, or the "burdens," of hundreds of conditions are measured on a common scale, and they are aggregated into a summary value for a given population. The conditions range from mild hearing loss through alcohol-use disorder and HIV/AIDS to acute schizophrenia.

The measure developed by the Global Burden of Disease project is called *disability-adjusted life year* (DALY). The basic idea is similar to the QALY, but the details are different. Since the primary interest of the developers of the DALY was in the harm associated with different conditions – rather than the benefit associated with different interventions – DALYs represent the gap between actual health and some ideal level of health. The gap can be caused both by losing years of life because of disease or injury and by having to live with a disability. DALYs are a combination of *years of life lost* due to disability (YLL) and *years of life lived with a disability* (YLD).

Let us explain. One of the harms of diseases and injuries is premature death – shortening people's lives. If a person is killed by a disease at age 50, one way to represent the harm is to take the difference between the number of years that she has lived and the number of years that she could have lived. But for this, you obviously need to be able to estimate how many more years this person could have lived. One way to do this would be to take the average life expectancy at 50 in the population to which this person belonged. But this leads to a problem: do we really want to say that the death of a 50-year-old person who belongs to a population where average life expectancy is 60 is *less bad* than the death of a person at the same age who belongs to a population where average life expectancy is 80?

The developers of the DALY offered another solution. They argued that the harm of premature mortality should be estimated not on the basis of how many more years a person could have lived in the particular population to which she belonged, but on the basis of what human beings could achieve under reasonably ideal conditions. This way, the harm of dying at 50 is the same no matter where a person lives, since it depends on how long people could ideally live. Of course, you cannot say for certain how many years that would be, but you can look at the population with the greatest life expectancy in the world. You can treat the life expectancy in the population that achieves the greatest life expectancy at birth as the ideal life expectancy for all human beings. Under present conditions, living less than that is the harm caused by premature mortality.

The country with the greatest life expectancy is Japan. In the early 1990s, at the time the Global Burden of Disease studies began, life expectancy at birth in Japan was 82.5 years for females and 80 years for males. Thus, the ideal age to which premature death was compared was set to 82.5 years for females and 80 years for males. The difference between men and women was attributed to the different survival potential of the sexes, which is thought to be at least in part biologically determined. Thus, if a person dies at 50, the burden of premature mortality was the same regardless of whether this person lived in one of the most advanced nations or one of the least developed countries. The only factor that made a difference was the person's sex.

You might want to stop us at this point. You might wonder why there should be different ideal life expectancies for men and women. You might argue that even if men are naturally disposed to have shorter lives – or, perhaps just as importantly, they are more willing to take risks with their own health – this should not make a difference to the burden of premature death. It should not be less bad if a man dies at 50 than if a woman does.

You would be right to make this argument. Mostly for this reason, the ideal life expectancies were changed in more recent updates of the Global Burden of Disease studies. For instance, the ideal life expectancy at birth is 87.9 years both for men and women in the 2017 update. There is no longer a difference in the burden of premature death between the sexes. The change also reflects the gains in life expectancy in the last couple of decades as well as the gradual narrowing of the life expectancy gap between men and women.

After this small detour, let us continue with the calculation of years of life lost. On the current methodology, if a person dies at 50, the years of life lost are 38.7 years. This represents the burden of premature mortality associated with this person's disease or injury. (Of course, 50 plus 38.7 is more than 87.9. But there is no error in the math here: the years of life lost are greater than the 37.9 you might expect, because average life expectancy at birth and average life expectancy at 50 are different. If you survive to 50, you can expect to live longer than you could expect to live at birth. The average increases because some people in your birth cohort have already died. If you survive to 80, your ideal life expectancy is another 12 years.)

The other component of DALYs is years of life lived with disability, used to represent non-fatal health outcomes. If a person has diabetes or if she is blind, her health-related quality of life falls short of perfect health. Each year she spends having the condition is adjusted for her health-related quality of life, just as in the case of QALYs. The *disability weights* that are used in DALYs represent the burden of the disability associated with particular diseases and injuries. (Recall that disabilities are losses of ability to carry out normal human activities.)

Recent updates of the Global Burden of Disease studies include almost 400 different diseases and injuries. These are the *causes* that lead to particular pathological conditions. These pathological conditions are called *sequelae*, and there are well over a 1,000 of them. They include such diverse conditions as anemia due to malaria, heart failure due to ischemic heart disease, measles, major depressive disorders, and so on. In many cases, the treated and untreated forms of a disease are also distinguished. AIDS with antiretroviral treatment is treated as a separate condition from AIDS without antiretroviral treatment. In other cases, the phases of a condition are also distinguished: a cancer can be controlled, metastatic, or in the terminal phase.

Not all of these conditions need to be assigned a separate disability weight, however, since multiple conditions can lead to the same health outcome. For instance, anemia can have a genetic cause or it can be caused by vitamin deficiency, iron deficiency, by certain chronic or infectious diseases, and so on. Anemia due to malaria is a separate condition from anemia due to maternal hemorrhaging or anemia due to peptic ulcer disease. But all of these conditions can be given a common disability weight insofar as they lead to very similar health outcomes. Thus, for instance, the 2010 update of the Global Burden of Disease study associates over 1,000 conditions with 220 different health outcomes. The health outcomes represent the disabilities that result from the pathological conditions – which can be caused by roughly 300 different diseases and injuries.

This might sound more complicated than it is. In a nutshell, the idea is to identify the diseases and injuries that are the ultimate causes of many pathological conditions and then associate all these conditions with health states that represent similar levels of disability. In the final step, these disabilities are assigned weights that represent their burden.

Thus, the disability weights express lost health-related quality of life. They are measured on a scale between 0 and 1. Compared to QALYs, the scale, however, is inverted: full health is represented by 0, death is represented by 1, and disabilities are represented by weights between 0 and 1. The smaller the weight, the smaller the burden of the disability. This is because DALYs represent harm: the greater the disability weight, the greater the harm.

For an example, consider anemia again. It can have several different causes, and it is associated with several conditions. But in terms of health-related quality of life, what matters is the resultant disability. Evidently, however, anemia can cause different levels of disability. So anemia comes with mild, moderate, and severe forms. The disability weight of mild anemia is 0.004; the weight of moderate anemia is 0.052; and the weight of severe anemia is 0.149.

Here are some other examples of disability weights: AIDS without antiretroviral treatment has a disability weight of 0.582; severe dementia has 0.449; uncontrolled asthma 0.133. Severe motor- and cognitive impairment with blindness due to malaria has a disability weight of 0.625.

DALYs are the sum of years of life lost due to disability and the years of life lived with disability. The burden of the years of life lived with disability is determined by the disability weights. For example, suppose that a person at 40 is struck by a disease whose disability weight is 0.5 and which kills him at age 50. The burden of this condition is 38.7 years of life lost, as well as ten years of life spent with a disability whose weight is 0.5. Altogether, these are about 44 DALYs (0.5 × 10 and the years of life lost due to death at 50). This is the burden of this person's disease. Since this is a harm, the smaller this number is, the smaller is the burden.

Therefore, 1 DALY can represent one year of lost life, or two years with a disease whose disability weight is 0.5, four years with a disease whose disability weight is 0.25, and so on. If you add up the burden of disease for each person within a population, you get a summary measure of the overall burden for that population. This can then be compared with similar measures for other countries or regions of the world.

You can also examine the burden of disease globally, according to different causes – tracing back the DALYs associated with different conditions to the diseases and injuries that are responsible for them. According to 2010 data, the top causes of the global burden of disease are ischemic heart disease, followed by lower respiratory infections, stroke, diarrhea, and HIV/AIDS. The list continues with lower back pain and malaria. Notice that not all of these are important causes of mortality. Lower back pain, for instance, has a great health burden, but it is negligible as a cause of death. Lung cancer is a major cause of premature mortality, but because of the high average age of death and the low number of years lived with disability associated with it, it is not a major cause of the overall burden of disease. Ischemic heart disease, lower respiratory infections, and stroke, however, are important causes both of premature mortality and the overall burden of disease.

The data can also show changes over time. In 2004, the top five leading causes of the burden of disease worldwide were lower respiratory infections (94.5 million DALYs), diarrheal diseases (72.8 million DALYs), unipolar depressive disorders (65.5 million DALYs), ischemic heart disease (62.6 million DALYs), and HIV/AIDS (58.5 million DALYs). In recent years, an increasing proportion of the global disease burden is attributable to chronic disease compared to infectious disease, and this trend is expected to continue. In 2017, the leading causes were ischemic heart disease, lower respiratory infections, chronic obstructive pulmonary disease (COPD), diarrheal diseases, neonatal conditions, and lower back pain.

Furthermore, there continue to be enormous variations between different countries and regions of the world. In high-income countries, the causes of the burden of disease in 2004 were unipolar depressive disorders (10 million DALYs), ischemic heart disease (7.7 million DALYs), and cerebrovascular disease (4.8 million DALYs). At the same time, in low-income countries, the leading causes were lower respiratory infections (76.9 million DALYs), diarrheal diseases (59.2 million DALYs), and HIV/AIDS (42.9 million DALYs). Note the difference between the magnitude of these numbers in the two groups of countries. Finally, there are enormous variations in the distribution of the burden of disease between age groups. According to the

2004 study, 36 per cent of the total disease burden in the world falls on children under 15 years – almost all of them living in low- and middle-income countries.

Originally, when the Global Burden of Disease project began in the early 1990s, the disability weights were determined on the basis of studies with various trade-off questions, using groups of health care professionals from different countries as respondents. The developers of the DALY argued that health care professionals are the best placed to determine disability weights, since they are the most familiar with a wide range of health conditions. This procedure has received a lot of criticism.

One of the objections was that it is unlikely that the weights determined by health care professionals have any intercultural validity. Disabilities, it was argued, come with different burdens in different social and cultural settings. It is unlikely, therefore, that it is possible to assign the same weight to the same disability in different settings. The measurement of DALYs, therefore, lacks intercultural validity.

In response, disability weights were re-estimated for the 2010 update of the Global Burden of Disease studies. Instead of health care professionals, general population samples were used. The researchers undertook two kinds of surveys. In the first one, household surveys were carried out in five different countries. They interviewed almost 13,000 individuals in Bangladesh, Indonesia, Peru, Tanzania, and the United States. The second survey was web-based. Anyone could take part in it. Overall, almost 30,000 respondents helped evaluate the burden of particular disabilities on the basis of pairwise comparisons of health states. In the years since, many more surveys have been carried out.

The most important finding of the surveys was a high degree of agreement in the responses. Respondents from very different social, economic, and cultural backgrounds gave very similar evaluations. The high degree of agreement serves as evidence that DALYs can be applied in different settings. People from different backgrounds largely agree on the badness of different health conditions.

Nevertheless, the new surveys did not address another objection. According to some critics, the badness of different conditions should be evaluated only by those who are the most familiar with those conditions: the patients.

2.5 Whom to ask?

One unresolved issue in health-related quality of life measurement concerns the role of the respondents whose evaluations are used to determine the quality-adjustment factors in QALYs and other health-related quality of life measures. Normally, quality of life researchers use a random sample of respondents. In the United Kingdom, for instance, responses to the EQ-5D questionnaire from samples of the general population have been used to evaluate new interventions and health care services.

The disadvantage of this approach is that it is unlikely that everyone can evaluate different health states equally well. Some people are more familiar with a given health state than others. They have experience of it or know someone who has experienced it. This was one of the reasons the developers of DALYs originally used responses from health care professionals to assign disability weights. Health care professionals are more familiar with diseases and injuries than members of the general public. They know better what it is like to live with them.

But the best knowledge might be the patients' knowledge. After all, they have first-hand experience of what it is like to live with a condition. In fact, many health-related quality of life questionnaires are designed for patients, especially those that are concerned with the treatment outcomes associated with particular conditions. But general measures (like QALYs and DALYs) seldom use patient evaluations.

One reason for this is practical. If health states were evaluated only by those who have experience of them, different groups of respondents would have to be used for each of them. This would be prohibitively expensive. But there is also a deeper problem here. People who have a chronic illness or permanent functional limitation often *adapt* to their condition: they cope with it by changing their aims, adjusting their plans, and learning new ways to live with their limitations. Adaptation means that often, though not always, they judge their own health state as *less bad* than others. (Some forms of mental illness are exceptions. It is impossible to adapt to unipolar major depression.) In general, health professionals and family members of people living with a particular condition rate that condition worse than the people who themselves have the condition, and members of the general public rate it even worse than health professionals and family members.

Adaptation is another manifestation of the inseparability problem. When people living with a condition judge their health less bad than others, their evaluation is influenced by other factors. Their lower health-related quality of life does not necessarily lead to proportionally lower overall well-being, because they are able to compensate for their health limitations within other components of well-being. But people who do not have experience of the condition disregard the possibility of successful adaptation.

The discrepancy between the evaluations of patients and the general public leads to a paradox. On the one hand, if the values of patients who have successfully adapted to their condition are used, then the quality-adjustment factors will be higher – the health state turns out to be *less bad*. That means that its prevention, treatment, and rehabilitation will be considered less urgent. It will have lower priority in the allocation of health care resources.

On the other hand, if the values of the general public are used in determining the health-related quality of life associated with particular conditions, their prevention and treatment will be more urgent, because of the lower quality-adjustment factor that results from the responses of people less familiar with those conditions. But it would be a bit peculiar to say that less-informed respondents have a more accurate view on the urgency of the prevention and treatment of disability and chronic disease.

Some people believe it is self-evident that the values of patients should be used. They know more about their conditions. They know it is possible to adapt to them and to lead happy and successful lives. But you have to be careful with this argument: adaptation is not always desirable or admirable. You can adapt to limitations by finding other worthwhile goals and activities. But you can also adapt to limitations by giving up your goals and activities and learning to be content with less. Adaptation is not always a healthy way to cope with adverse circumstances.

One solution to this problem may be to use a *deliberative process* in health state evaluation. Some health economists have suggested that respondents should be given a chance to discuss, reflect on, and even revise their evaluations in the light of further information and discussion with patients. This could lead to more agreement on quality adjustment factors and allow respondents to consider whether adaptation is desirable in particular cases.

Others have argued that health care resource allocation concerns the choices that particular societies make about the use of scarce common resources. These choices should reflect the values of the general population, rather than particular patient groups. Some also add that we should worry less about funding treatments that should not be funded and more about *not* funding treatments that should be. Therefore, when in doubt, you should use the lowest quality-adjustment factors. You should, that is, err on the side of caution.

As this very brief survey shows, there is no generally accepted solution to this problem.

Chapter summary

In order to allocate health care resources fairly and efficiently, we need to be able to measure the value of health: its contribution to quality of life. Although the impact of health on quality of life is difficult to separate from the impact of other components of quality of life, researchers have developed a number of methods for measuring health-related quality of life. These are typically based on surveys for describing and evaluating different health states. On the basis of these evaluations, it is possible to construct general measures of health-related quality of life, including quality-adjusted life years (QALYs) and disability-adjusted life years (DALYs). This chapter has provided a survey of these procedures. We also presented some of their underlying assumptions as well as some of their problems.

Discussion questions

1. Consider the EQ-5D questionnaire in Figure 2.1. In your view, is it capable of adequately describing and distinguishing health states? Why or why not? What questions would you add or remove?

2. We argued in this chapter that the standard gamble and the time trade-off may be more adequate methods for eliciting health state valuations than the rating scale, because patients are sometimes required to make treatment gambles or make trade-offs between health states. Evaluate this argument. Is it relevant to the assessment of different methods?
3. When health-adjusted life expectancies are calculated, it is assumed that each health state has the same value at different ages (i.e. the quality-adjustment factors are the same). Do you agree with this assumption?
4. In calculating the burden of disease in a population, you can use either actual life expectancies or some ideal life expectancy for representing the harm of premature mortality. What are the advantages and disadvantages of these alternatives? Which one should be used?
5. Suppose you are trying to determine the burden of a particular health condition. For determining the quality adjustment factor of this condition, you can survey either a sample of the general population or a sample of patients who have had the condition. Which sample should you use? Why? What would be the considerations for and against your view?
6. If you study this book together with others, form three groups. First, select a health state from the EQ-5D questionnaire. Working independently, have the members of the first group evaluate the health state using the rating scale method, members of the second group using the standard gamble, and members of the third group using the time trade-off. (If there are more than one members in a group, take the average of their evaluations.) Discuss the results together. How close are the values given by the different methods? What do you think are the reasons for the differences?

   Second, select another health state from the questionnaire. Working now only in two groups, have members of one group evaluate the health state using the standard gamble and members of the other group using the time trade-off. Together, consider the *differences* between the values of the two health states using these methods. On which method is the improvement from the worse health state to the better one greater? What do the methods imply about the urgency of treating people in these health states?

References and further readings

The World Health Organization's definition of health is quoted from the Preamble to the Constitution of the World Health Organization, 1946, available at http://apps.who.int/gb/bd/PDF/bd47/EN/constitution-en.pdf.

An excellent introduction into health evaluation is the collection by Murray *et al.* (2002). Broome (2002) and Brock (2002) provide two discussions of the inseparability problem in that volume. See also Bognar (2008a) on this topic. A good discussion of philosophical issues in health-related quality of life measurement is given by Brock (1993). A lot has been written on health state evaluation methods and the construction of QALYs; see, for instance, Froberg and Kane (1989a, 1989b, 1989c, 1989d) or Weinstein *et al.* (2009). For philosophical assessments and arguments, see Broome (1999), Nord *et al.* (2009) – which reports the study on the comparison of values of health states obtained by the rating scale, standard gamble, and time trade-off methods – and especially the comprehensive book by Daniel Hausman (2015).

For DALYs and the Global Burden of Disease Project, see Murray (1996) – this is now, however, outdated since the publication of the 2010 update, which introduced a number of major methodological and philosophical revisions. (For this and more recent updates, see the dedicated website of the journal *The Lancet*, available at https://www.thelancet.com/gbd. For current global burden of disease data and for data visualizations, see also the website of the Institute for Health Metrics and Evaluation (IHME), at https://www.healthdata.org/.) A good introduction to the current state of the project is given by Vos (2020). The volume in which this paper appears (Eyal *et al.* 2020) contains discussions of the philosophical aspects of measuring the burden of disease.

For the problem of adaptation, see Menzel *et al.* (2002) and Wolff *et al.* (2012).