Methods for the Economic Evaluation of Health Care Programmes

Fourth Edition

MICHAEL F. DRUMMOND
MARK J. SCULPHER
KARL CLAXTON
GREG L. STODDART
GEORGE W. TORRANCE

OXFORD

Methods for the Economic Evaluation of Health Care Programmes

Methods for the Economic Evaluation of Health Care Programmes

FOURTH EDITION

Michael F. Drummond
Professor, Centre for Health Economics,
University of York, UK

Mark J. Sculpher
Professor, Centre for Health Economics,
University of York, UK

Karl Claxton
Professor, Department of Economics and Related Studies and
Centre for Health Economics, University of York, UK

Greg L. Stoddart
Professor Emeritus, McMaster University, Hamilton, Canada

George W. Torrance
Professor Emeritus, McMaster University, Hamilton, Canada

OXFORD
UNIVERSITY PRESS

OXFORD
UNIVERSITY PRESS

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

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

First edition published in 1987
Second edition published in 1997
Third edition published in 2005
Fourth edition published in 2015

Impression: 1

Published in the United States of America by Oxford University Press
198 Madison Avenue, New York, NY 10016, United States of America

British Library Cataloguing in Publication Data

Data available

Library of Congress Control Number: 2015938217

ISBN 978–0–19–966587–7 (hbk.)
ISBN 978–0–19–966588–4 (pbk.)

Printed and bound by
CPI Group (UK) Ltd, Croydon, CR0 4YY

In memory of our friend and colleague, Bernie O'Brien

Preface to the fourth edition

The first question in anyone's mind reading the fourth edition of any book will be 'What has changed from previous editions?' The most obvious change, in keeping with the tradition we have maintained since the outset, is the addition of a new coauthor, Karl Claxton. Like all those before him, Karl has questioned aspects of the work and provoked changes which otherwise may not have been made.

The other change, of course, is that the field itself has moved on in the 10 years since the last edition. The new edition reflects these changes. Chapters 5 and 6, on measuring and valuing effects, reflect the growth in the literature on the measurement of health gain and other benefits of health care. In addition, we include two new chapters (10 and 11) discussing the methods of evidence synthesis and characterizing uncertainty, given their growing importance in economic evaluation.

However, considering the 28-year period since the original publication of the book, the most fundamental change relates to the role of economic evaluation in health care decision-making. Back in 1987 our emphasis was on explaining the methods used in economic evaluations so that readers could critically appraise them and potentially embark on their own studies. As the role of economic evaluation in decision-making expanded we added a chapter on 'Presentation and use of economic evaluation results', discussing the use of cost-effectiveness thresholds and the transferability of data from one setting to another. However, in discussing the content of the fourth edition, we realised that this was no longer sufficient because, owing to the international growth in the use of economic evaluation, it has become apparent that the use of particular methods is best discussed in the context of the decision problem being faced.

Therefore, in this edition we have added two new chapters (2 and 4) which emphasize that, in health care decision-making, it is important to be clear on what we are trying to maximize (for example, health or welfare), the constraints that we face, and the importance of opportunity cost. This enables us to give additional insights on the role of the various analytic approaches, given the decision-making context. In essence, the choice of methods and the use of study results are now integrated throughout the book, rather than being discussed in separate chapters.

We hope that readers feel that the new edition represents an improvement on previous editions and that it leads to further advances in both methods and decision-making processes in the future.

Michael F. Drummond
Karl Claxton
Greg L. Stoddart
York, UK and Hamilton, Canada

Mark J. Sculpher
George W. Torrance

Acknowledgements

We would like to thank a number of people who have helped us in important ways in developing the 4th edition of this book. The following individuals provided comments and suggestions on draft chapters which greatly improved the book: Bernard van den Berg, John Brazier, Tony Culyer, Catherine Claudius Cole, Richard Grieve, Jonathan Karnon, Andrew Lloyd, Andrea Manca, Gavin Roberts, Marta Soares and Beth Woods. Needless to say, none of the above are responsible for the final views expressed. Rita Faria and Sebastian Hinde helped develop the critical appraisal exercises for Chapter 3. Gill Forder, Frances Sharp, and Gillian Robinson provided a range of assistance in formatting and checking materials. We are grateful to all.

Contents

List of abbreviations

AAA	abdominal aortic aneurysm
AD	Alzheimer's disease
AQoL	Assessment of Quality of Life (measure)
ARM	age-related maculopathy
BSC	best supportive care
CABG	coronary artery bypass grafting
CCA	cost consequences analysis; *or* complete case analysis
CCyR	complete cytogenetic response
CDR	clinical dementia rating
CEA	cost-effectiveness analysis
CEAC	cost-effectiveness acceptability curve
CEAF	cost-effectiveness acceptability frontier
CER	comparative effectiveness research
CHEERS	Consolidated Health Economic Evaluation Reporting Standards
CI	confidence interval
CLT	central limit theorem
CMA	cost-minimization analysis
CRF	case report forms
CRT	cardiac resynchronization therapy
CT	computed tomography (scan)
CUA	cost–utility analysis
CV	compensating variation
DALY	disability-adjusted life-year
DCE	discrete choice experiment
DES	discrete event simulation; *or* drug-eluting stent
DIRUM	Database of Instruments for Resource Use Measurement
DRG	diagnosis-related group
DVT	deep vein thrombosis
EC	expected costs
EDSS	expanded disability status scale
ENBS	expected net benefit of sample
EORTC	European Organisation for Research and Treatment of Cancer
EVPI	expected value of perfect information
EVSI	expected value of sample information
GBD	global burden of disease
GDP	gross domestic product
GLM	general linear models
GORD	gastro-oesophageal reflux disease
HAQ	health assessment questionnaire
HEED	health economic evaluations database
HMO	health maintenance organization
HRQoL	health-related quality of life
HTA	Health Technology Assessment
HTAi	Health Technology Assessment International
HUI	health utilities index
HYE	healthy-year equivalents
ICD	International Classification of Diseases
ICER	incremental cost-effectiveness ratio
ISM	individual sampling models
ISPOR	International Society for Pharmacoeconomics and Outcomes Research
IVF	*in vitro* fertilization
MACE	major adverse cardiac events
MAR	missing at random
MCAR	missing completely at random
MCDA	multi-criteria decision analysis
MEPS	Medical Expenditure Panel Survey

MI	multiple imputation
MI	myocardial infarction
MMR	major molecular response
MNAR	missing not at random
NB	net benefit
NHB	net health benefit
NHMS	National Health Measurement Study
NHS	National Health Service (UK)
NICE	National Institute for Health and Care Excellence (UK)
NIHR	National Institute for Health Research (UK)
NMB	net monetary benefit
NPV	net present value
OLS	ordinary least squares
PBAC	Pharmaceutical Benefits Advisory Committee
PBC	programme budget categories
PBS	Pharmaceutical Benefits Schedule
PRO	patient-reported outcome
PSA	probabilistic sensitivity analysis
PTO	person trade-off
QALY	quality-adjusted life-year
QoL	quality of life
RCT	randomized controlled trial
SAVE	saved young life equivalents
SD	standard deviation
SD	strong dominance
SMDM	Society of Medical Decision Making
SUIT	stress urinary incontinence treatment
TAU	treatment as usual
TTO	time trade-off
UKPDS	UK prospective diabetes study
VAS	visual analogue scale
VPF	value of a prevented fatality
WHO	World Health Organization
WTA	willingness to accept
WTP	willingness to pay
YHL	years of healthy life

Chapter 1

Introduction to economic evaluation

1.1 Some basics

Those who plan, provide, receive, or pay for health services face an incessant barrage of questions such as the following.

- Should clinicians check the blood pressure of each adult who walks into their offices?
- Should planners launch a scoliosis screening programme in secondary schools?
- Should individuals be encouraged to request annual check-ups?
- Should local health departments free scarce nursing personnel from well-baby clinics so that they can carry out home visits on lapsed hypertensives?
- Should hospital administrators purchase each and every piece of new diagnostic equipment?
- Should a new, expensive drug be listed on the formulary?

These are examples of general, recurring questions about who should do what to whom, with what health care resources, and with what relation to other health services.

The answers to these questions are most strongly influenced by our estimates of the relative merit or value of the alternative courses of action they pose. This book focuses on the evaluation of alternative policies, services, or interventions which are intended to improve health. Since the effects of choosing one course of action over another will not only have effects on health, but also on health care resources as well as other effects outside health care, informing health care decisions requires consideration of costs and benefits. For this reason this type of evaluation is most commonly referred to as *economic evaluation*. The purpose of economic evaluation, however, is to inform decisions, so the key inputs to any economic evaluation are evidence about the effects of alternative courses of action. Much of this evidence will draw on the results of clinical evaluations (e.g. randomized clinical trials). The evidence from clinical studies needs to be sought in a systematic way, interpreted appropriately (including an assessment of its relevance and potential for bias) and then, when appropriate, synthesized to provide estimates of key parameters (see Chapter 10). Therefore, economic and clinical evaluations are not alternative approaches to achieve the same end but complements. Economic evaluation provides a framework to make best use of clinical evidence through an organized consideration of the effects of all the available alternatives on health, health care costs, and other effects that are regarded as valuable.

For these reasons an understanding of the core principles of clinical epidemiology and the criteria for assessing the relevance and potential for bias in clinical evidence of the effect of an intervention is very important and has been described elsewhere (Guyatt et al. 2008; Stevens et al. 2001). These guides and other introductory texts in clinical epidemiology provide suitable background, so we do not review them here. However, later chapters of the book draw on, and develop, these core principles.

1.2 Why is economic evaluation important?

To put it simply, resources—people, time, facilities, equipment, and knowledge—are scarce. Choices must and will be made concerning their deployment, and methods such as 'what we did last time', 'gut feelings', and even 'educated guesses' are rarely better than organized consideration of the factors involved in a decision to commit resources to one use instead of another. This is true for at least four reasons.

1. *Without systematic analysis, it is difficult to identify clearly the relevant alternatives.* For example, in deciding to introduce a new programme (rehabilitation in a special centre for chronic lung disease), all too often little or no effort is made to describe existing activities (episodic care by family physicians in their offices) as an alternative 'programme' to which the new proposal must be compared. Furthermore, if the objective is, indeed, to reduce morbidity due to chronic lung disease then preventive programmes (e.g. cessation of cigarette smoking) may represent a more efficient avenue and should be added to the set of programmes being considered in the evaluation. Of course, in practice the range of alternative programmes compared may be restricted to those that are the responsibility of a particular decision-maker (e.g. a given decision-maker may be responsible for cancer treatment, but not for cancer prevention). Also, if a new programme is compared to 'existing care', it is important to consider whether existing care is itself cost-effective. This may not be the case, for example, if there is an alternative, lower-cost, programme that is just as effective. Although it may not possible to consider all conceivable alternatives in a given study, an important contribution of economic evaluation is to minimize the chances of an important alternative being excluded from consideration, or a new programme being compared to a baseline which is not cost-effective.
2. *The perspective (or viewpoint), assumed in an analysis is important.* A programme that looks unattractive from one perspective may look significantly better when other perspectives are considered. Analytic perspectives may include any or all of the following: the individual patient, the specific institution, the target group for specific services, the Ministry of Health budget, the government's overall budget position (Ministry of Health plus other ministries), and the wider economy or the aggregation of all perspectives (sometimes called the 'societal' perspective).
3. *Without some attempt at quantification, informal assessment of orders of magnitude can be misleading.* For example, when the American Cancer Society endorsed a protocol of six sequential stool tests for cancer of the large bowel, most analysts would have predicted that the extra cost per case detected would increase markedly with each test. But would they have guessed that it would reach $47 million for

the sixth test, as Neuhauser and Lewicki (1975) demonstrated? Admittedly, while this is an extreme example, it illustrates that without measurement and comparison of outputs and inputs we have little upon which to base any judgement about value for money. In fact, the real cost of any programme is not the number of dollars appearing on the programme budget, but rather the value of the benefits achievable in some other programme that has been forgone by committing the resources in question to the first programme. It is this 'opportunity cost' that economic evaluation seeks to estimate and to compare with programme benefits.

4. *Systematic approaches increase the explicitness and accountability in decision-making*. Economic evaluation offers an organized consideration of the range of possible alternative courses of action and the evidence of their likely effects. It also requires that the scientific judgements needed to interpret evidence are made explicitly so they can be scrutinized and the impact of alternative, but plausible, views examined. Possibly more importantly, it can provide a clear distinction between the questions of fact and the unavoidable questions of social value. Indeed, the main contribution of economic evaluation may not be in changing the decisions that are ultimately made but *how* they are made. By making the scientific and social value judgements explicit it offers the opportunity for proper accountability for the social choices made on behalf of others. (These issues are discussed further in Chapter 2.)

1.3 The features of economic evaluation

Economic evaluation seeks to inform the range of very different but unavoidable decisions in health care. Whatever the context or specific decision, a common question is posed: are we satisfied that the additional health care resources (required to make the procedure, service, or programme available to those who could benefit from it) should be spent in this way rather than some other ways? The other ways these resources could be used might include providing health care for other patients with different conditions, reducing the tax burden of collectively funded health care, or reducing the costs of social or private insurance premiums.

Economic evaluation, regardless of the activities (including health services) to which it is applied, has two features. First, it deals with both the inputs and outputs, which can be described as the *costs* and *consequences*, of alternative courses of action. Few of us would be prepared to pay a specific price for a package whose contents were unknown. Conversely, few of us would accept a package, even if its contents were known and desired, until we knew the specific price being asked. In both cases, it is the linkage of costs (what must be given up) and consequences (the overall benefits expected to be received) that allows us to reach our decision.

Second, economic evaluation concerns itself with choices. Resources are limited, and our consequent inability to produce all desired outputs (including efficacious therapies), necessitates that choices must, and will, be made in all areas of human activity. These choices are made on the basis of many criteria, sometimes explicit but often implicit, especially when decisions are made on our own behalf using our own resources.

Economic evaluation seeks to identify and to make explicit the criteria (social values) that are applied when decisions are made on others' behalf; when the consequences

accrue to some, but some or all of the costs will be borne by others. It can also provide useful information to patients and their clinicians when making choices about their own health care, since they are not necessarily best placed to identify and to synthesize all relevant evidence and undertake the computation required fully to assess all the effects of the alternative courses of action available, and especially so at the point of care.

These two characteristics of economic evaluation lead us to define economic evaluation as *the comparative analysis of alternative courses of action in terms of both their costs and consequences*. Therefore, the basic tasks of any economic evaluation are to identify, measure, value, and compare the costs and consequences of the alternatives being considered. These tasks characterize all economic evaluations, including those concerned with health services (see Box 1.1).

Box 1.1 Economic evaluation always involves a comparative analysis of alternative courses of action

Figure 1.1 illustrates that an economic evaluation is usually formulated in terms of a choice between competing alternatives. Here we consider a choice between two alternatives, A and B. The comparator to Programme A, the programme of interest, does not have to be an active treatment. It could be doing nothing. Even when two active treatments are being compared, it may still be important to consider the baseline of doing nothing, or a low-cost option. This is because the comparator (Programme B) may itself be inefficient. (As mentioned earlier, it is important that the evaluation considers all relevant alternatives.)

The precise nature of the costs and consequences to be considered, and how they might be measured and valued, will be discussed in later chapters of the book. However, the general rule when assessing programmes A and B is that the *difference* in costs is compared with the *difference* in consequences, in an incremental analysis.

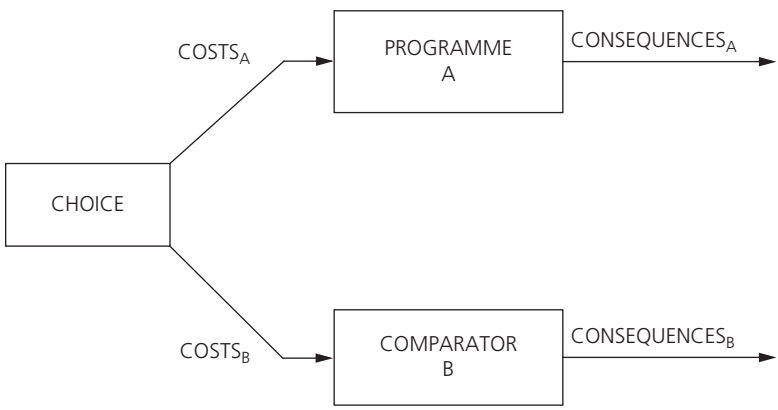


Fig. 1.1 Economic evaluation always involves a comparative analysis of alternative courses of action.

However, not all of the studies measuring costs constitute economic evaluations. The large literature on *cost of illness*, or *burden of illness*, falls into this category. These studies describe the cost of disease to society, but are not full economic evaluations because alternatives are not compared (Drummond 1992). Some studies do compare alternatives but just consider costs. An example of such a study is that by Lowson et al. (1981) on the comparative costs of three methods of providing long-term oxygen therapy in the home: oxygen cylinders, liquid oxygen, and the oxygen concentrator (a machine that extracts oxygen from air). Such studies are called *cost analyses*. The authors argued that a cost analysis was sufficient as the relative effectiveness of the three methods was not a contentious issue. However, a full economic evaluation would explicitly consider the relative consequences of the alternatives and compare them with the relative costs.

1.4 Do all economic evaluations use the same techniques?

The identification of various types of costs and their subsequent measurement in monetary units is similar across most economic evaluations; however, the nature of the consequences stemming from the alternatives being examined may differ considerably. Let us consider three examples to illustrate how the nature of consequences affects their measurement, valuation, and comparison to costs.

1.4.1 Example 1: cost-effectiveness analysis

Suppose that our interest is the prolongation of life after renal failure and that we are comparing the costs and consequences of hospital dialysis with kidney transplantation. In this case the outcome of interest—life-years gained—is common to both programmes; however, the programmes may have differential success in achieving this outcome, as well as differential costs. Consequently we would not automatically lean towards the least-cost programme unless, of course, it also resulted in a greater prolongation of life. In comparing these alternatives we would normally calculate this prolongation and estimate incremental cost per unit of effect (that is, the extra cost per life-year gained of the more effective and more costly option). Such analyses, in which costs are related to a single, common effect that may differ in magnitude between the alternative programmes, are usually referred to as *cost-effectiveness analyses (CEAs)*. Note that the results of such comparisons may be stated either in terms of incremental cost per unit of effect, as in this example, or in terms of effects per unit of cost (life-years gained per dollar spent).

It is sometimes argued that if the two or more alternatives under consideration achieve the given outcome to the same extent, a *cost-minimization analysis* (CMA) can be performed. However, it is not appropriate to view CMA as a form of full economic evaluation (see Box 1.2).

There are many examples of CEA in the early literature on economic evaluation. Ludbrook (1981) provided an estimate of the cost-effectiveness of treatment options for chronic renal failure. In addition, a number of studies compare the cost-effectiveness of actions that do not produce health effects directly, but that achieve other clinical objectives that can be clearly linked to improvements in patient outcome. For example, Hull et al. (1981) compared diagnostic strategies for deep vein thrombosis in terms of the

Box 1.2 The death of cost-minimization analysis?

Economic evaluations are sometimes referred to in the literature as *cost-minimization analyses (CMAs)*. Typically this is used to describe the situation where the consequences of two or more treatments or programmes are broadly equivalent, so the difference between them reduces to a comparison of costs.

It can be seen from Figure 1.2 that there are nine possible outcomes when one therapy is being compared with another. In two of the cases (boxes 4 and 6) it might be argued that the choice between the treatment and control depends on cost because the effectiveness of the two therapies is the same.

However, Briggs and O'Brien (2001) point out that, because of the uncertainty around the estimates of costs and effects, the results of a given study rarely fit neatly into one of the nine squares shown in the diagram. Also, because of this uncertainty, CMA is not a unique study design that can be determined in advance.

The only possible application of CMA is in situations where a prior view has been taken, based on previous research or professional opinion, that the two options are equivalent in terms of effectiveness. However, here one might question the basis on which this view has been formed. It is likely only to be justifiable in situations where the two therapies embody a near-identical technology (e.g. drugs of the same pharmacological class).

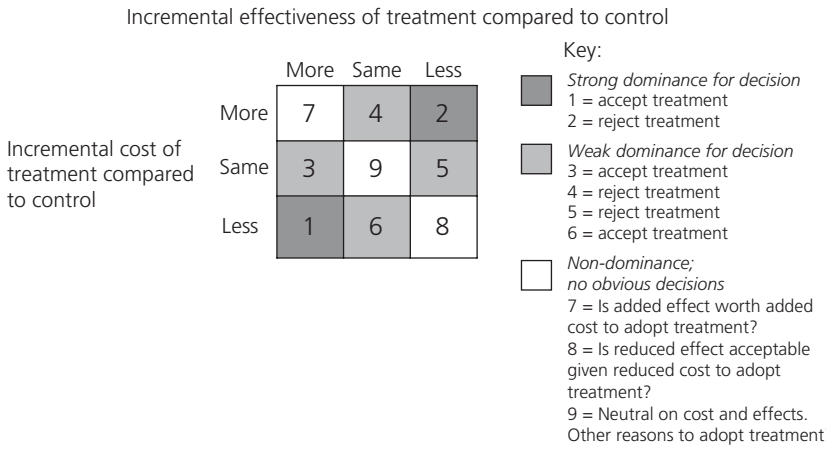


Fig. 1.2 The death of cost-minimization analysis?

cost per case detected. Similarly, Logan et al. (1981) compared work-site and regular (physician office) care for hypertensive patients in terms of the cost per mmHg drop in diastolic blood pressure obtained. Sculpher and Buxton (1993) compared treatments for asthma in terms of the cost per episode-free day.

The more recent literature contains a lower proportion of CEAs, probably because of influential sets of methods guidelines, such as those produced by the Washington

Panel (Gold et al. 1996), or the official requirements for the conduct of economic evaluations in some jurisdictions, such as the United Kingdom (NICE 2013). Many of these guidelines recommend the use of cost–utility analysis, with quality-adjusted life-years (QALYs) as the measure of benefit (see Section 1.4.2).

Of the CEAs that are published, many are conducted alongside a single clinical study and use the chosen clinical endpoint as the measure of benefit in the economic study. Examples of this approach are the study by Haines et al. (2013) on the cost-effectiveness of patient education for the prevention of falls in hospital (which used 'number of falls prevented' and 'reduction in the number of patients who fell' as the denominator of the cost-effectiveness ratio) and the study by Price et al. (2013) on the cost-effectiveness of alternative asthma treatments (which used 'number of patients who experienced severe exacerbations' and 'number of patients with risk domain asthma control' as the measures of benefit).

Other frequent examples of cost-effectiveness studies are those of prevention or diagnostic interventions. These tend to focus on the specific impact of the intervention as opposed to the broader health of the patient. Examples of this approach are the study by Rabalais et al. (2012) of the CEA of positive emission tomography (PET)-CT for patients who had oropharyngeal cancer of the neck (which used 'patients free from disease in the neck after one year' as the benefit measure) and the study by Pukallus et al. (2013) on the cost-effectiveness of a telephone-delivered education programme to prevent early childhood caries (which used 'reduced number of caries' as the benefit measure). Another feature of many of these studies is that they do not necessarily calculate cost-effectiveness ratios, rather they present differences in cost (between the alternative programmes) alongside the other outcomes.

Finally, some CEAs are conducted in jurisdictions where QALYs are not recommended as the measure of benefit in economic studies. An example is the study by Dorenkamp et al. (2013) on the cost-effectiveness of paclitaxel-coated balloon angioplasty in patients with drug-eluting stent restenosis, conducted from the perspective of the German Statutory Insurance. This used 'life-years gained' as the denominator in the incremental cost-effectiveness ratio.

Cost-effectiveness analysis is of most use in situations where a decision-maker, operating with a given budget, is considering a limited range of options within a given field. For example, a person with the responsibility for managing a hypertension treatment programme may consider blood pressure reduction to be a relevant outcome; a person managing a cancer screening programme may be interested in cases detected. However, even in these situations, these outcomes may be insufficient. For example, the benefits from detecting a cancer will depend on the type of cancer and the stage of its development. Similarly, the benefits from reducing blood pressure by a given amount will depend on the patient's pretreatment level.

However, the biggest limitation of these analyses is that, because of the specific measures of effect used in evaluating a given treatment or programme, it is difficult to assess the *opportunity cost* (i.e. benefits forgone) in other programmes covered by the same budget. In order to make an informed decision, the decision-maker needs to compare the benefits gained from introducing the new intervention with those lost from any existing programmes that will be displaced. This requires the use of a generic measure of benefit that is relevant to all the interventions for which the decision-maker is responsible.

1.4.2 Example 2: cost–utility analysis

Another term you might encounter in the economic evaluation literature is *cost–utility analysis (CUA)* (NICE 2013). These studies are essentially a variant of cost-effectiveness and are often referred to as such. The only difference is that they use, for the consequences, a generic measure of health gain. As we will argue later, this offers the potential to compare programmes in different areas of health care, such as treatments for heart disease and cancer, and to assess the opportunity cost (on the budget) of adopting programmes. In this literature the term 'utility' is used in a general sense to refer to the preferences individuals or society may have for any particular set of health outcomes (e.g. for a given health state, or a profile of states through time). Later, in Chapter 5, we shall be more specific about terminology, because utility has specific connotations in economics. The various methods to elicit health state preferences to construct measures of health-related quality of life might be better thought of as measures of outcome that attempt to capture effects on different aspects of health.

The notion that the value of an outcome, effect, or level of health status is different from the outcome, effect, or level of health status itself can be illustrated by the following example. Suppose that twins, identical in all respects except occupation (one being a signpainter and the other a translator), each broke their right arm. While they would be equally disabled (or conversely, equally healthy), if we asked them to rank 'having a broken arm' on a scale of 0 (dead) to 1 (perfect health) their rankings might differ considerably because of the significance each one attaches to arm movement, in this case due to occupation. Consequently, we would expect that their assessments of the value of treatment (i.e. the degree to which treatment of the fractures improved the quality of their lives) would also differ.

The estimation of preferences for health states is viewed as a particularly useful technique because it allows for health-related *quality-of-life* adjustments to a given set of treatment outcomes, while simultaneously providing a generic outcome measure for comparison of costs and outcomes in different programmes. The generic outcome, usually expressed as QALYs, is arrived at in each case by adjusting the length of time affected through the health outcome by the preference weight (on a scale of 0 to 1) of the resulting level of health status (see Box 1.3). Other generic outcome measures, such as the *healthy years equivalent (HYE)* (Mehrez and Gafni 1989), the *disability-adjusted life-year (DALY)* (Tan-Torres Edejer et al. 2003), and the *saved-young-life equivalent* (Nord 1995), have been proposed as alternatives to the QALY. These are discussed further in Chapter 5.

The results of CUAs are typically expressed in terms of the cost per healthy year gained, or cost per QALY gained, by undertaking one programme instead of another. Examples of CUAs include the study by Boyle et al. (1983) on neonatal intensive care for very-low-birth-weight infants, that by Oldridge et al. (1993) on a formal post-myocardial infarction rehabilitation programme, and that by Torrance et al. (2001) on the incorporation of a viscosupplementation product into the treatment of knee osteoarthritis.

Cost–utility analyses now represent the most widely published form of economic evaluation. Recent examples include the study by Stranges et al. (2013) of two alternative drug regimens for treating *Clostridium difficile* infection in the United States, the study by

Box 1.3 QALYs gained from an intervention

In the conventional approach to QALYs the quality-adjustment weight for each health state is multiplied by the time in the state (which may be discounted, as discussed in Chapter 4) and then summed to calculate the number of QALYs. The advantage of the QALY as a measure of health output is that it can simultaneously capture gains from reduced morbidity (quality gains) and reduced mortality (quantity gains), and integrate these into a single measure. A simple example is displayed in Figure 1.3, in which outcomes are assumed to occur with certainty. Without the health intervention an individual's health-related quality of life would deteriorate according to the lower curve and the individual would die at time Death 1. With the health intervention the individual would deteriorate more slowly, live longer, and die at time Death 2. The area between the two curves is the number of QALYs gained by the intervention. For instruction purposes the area can be divided into two parts, A and B, as shown. Then part A is the amount of QALY gained due to quality improvements (i.e. the quality gain during time that the person would have otherwise been alive anyhow), and part B is the amount of QALY gained due to quantity improvements (i.e. the amount of life extension, but adjusted by the quality of that life extension).

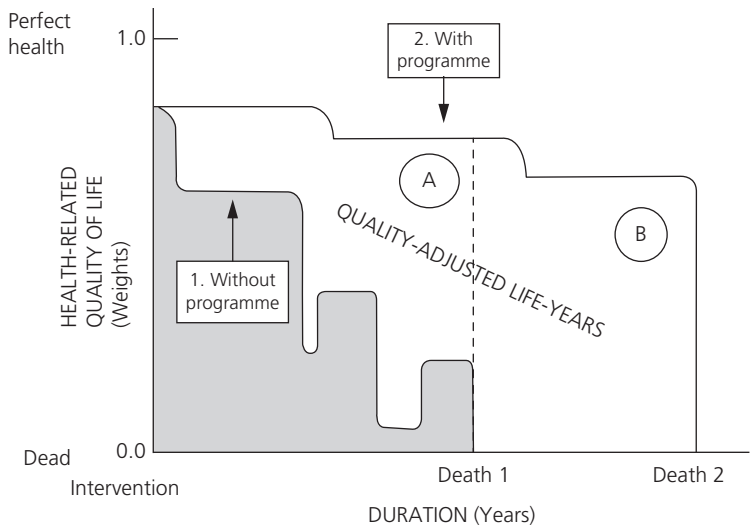


Fig. 1.3 QALYs gained from an intervention.

Reproduced from Gold, M.R. et al. (ed.), *Cost-effectiveness in health and medicine*, Figure 4.2, p. 92, Oxford University Press, New York, USA, Copyright © 1996, with permission of Oxford University Press, USA. Source: data from Torrance, G.W., Designing and conducting cost–utility analyses, pp. 1105–11, in B. Spilker (ed.), *Quality of life and pharmacoeconomics in clinical trials*, 2nd edition, Lippincott-Raven, Philadelphia, USA, Copyright © 1996.

Pennington et al. (2013) comparing three types of prosthesis for total hip replacement in adults with osteoarthritis, and the study by McConnachie et al. (2014) on the long-term impact on costs and QALYs of statin treatment in men aged 45–64 years with hypercholesterolaemia. A good source of published cost–utility studies is the CEA Registry, maintained by the New England Medical Center (<https://research.tufts-nemc.org/cear4>). In addition, economic evaluations of all types are summarized on the Health Economic Evaluations Database (HEED), published by Wiley (<http://onlinelibrary.wiley.com/book/10.1002/9780470510933>). This can also be accessed via the Cochrane Library.

1.4.3 Example 3: cost–benefit analysis

Both CEAs and CUAs are techniques that relate to constrained maximization; that is, where a decision-maker is considering how best to allocate an existing budget. In this situation a decision to expand one programme, to increase the number of cancers detected or to increase the QALYs gained, has an opportunity cost in terms of benefits forgone in other programmes covered by the budget. However, is there a form of economic evaluation that can address whether it is worthwhile expanding the budget?

One approach would be to broaden the concept of value and to express the consequences of an intervention in monetary terms in order to facilitate comparison to programme costs. This, of course, requires us to translate effects such as disability days avoided, life-years gained, medical complications avoided, or QALYs gained, into a monetary value that can be interpreted alongside costs. This type of analysis is called *cost–benefit analysis* (CBA) and has a long track record in areas of economic analysis outside health such as transport and environment. The results of such analyses might be stated either in the form of a ratio of costs to benefits, or as a simple sum (possibly negative) representing the net benefit (loss) of one programme over another.

With CBA, monetary valuation of the different effects of interventions is undertaken using prices that are revealed in markets. Where functioning markets do not exist, individuals can express their hypothetical willingness to pay for (or accept compensation to avoid) different outcomes. The literature contains a number of studies that assess individuals' *willingness to pay* for health benefits. For example, Johanneson and Jönsson (1991) give estimates for willingness-to-pay for antihypertensive therapy, Neumann and Johanneson (1994) give them for *in vitro* fertilization, and O'Brien et al. (1995) give them for a new antidepressant. A comprehensive CBA of health care interventions would use this approach to value the health benefits. Although there are many examples of studies using willingness-to-pay methods, very few CBAs incorporating these estimates have so far been published. See O'Byrne et al. (1996) for an example of such a study in the field of asthma and a pilot study by Haefeli et al. (2008), exploring the use of willingness-to-pay estimates in a CBA of spinal surgery.

The measurement characteristics of the various forms of economic evaluation are summarized in Table 1.1. However, it is important to note that the more fundamental differences between the various techniques relate not to their measurement characteristics, but to the value judgements implied in following each approach and their appropriateness for addressing particular resource allocation problems. This is explored in more depth in Chapter 2. Each approach to economic evaluation embodies a series of normative judgements and it is important to appreciate these when conducting a study.

Table 1.1 Measurement of costs and consequences in economic evaluation

Type of study	Measurement / valuation of costs in both alternatives	Identification of consequences	Measurement/ valuation of consequences
Cost analysis	Monetary units	None	None
Cost-effectiveness analysis	Monetary units	Single effect of interest, common to both alternatives, but achieved to different degrees	Natural units (e.g. life-years gained, disability days saved, points of blood pressure reduction, etc.)
Cost–utility analysis	Monetary units	Single or multiple effects, not necessarily common to both alternatives	Healthy years (typically measured as quality-adjusted life-years)
Cost–benefit analysis	Monetary units	Single or multiple effects, not necessarily common to both alternatives	Monetary units

1.5 Use of economic evaluation in health care decision-making

Over the past 20 years, two factors have led to an increased prominence of economic evaluation within health care decision-making. First, increasing pressures on health care budgets have led to a shift in focus from merely assessing clinical effectiveness, to one on assessing both clinical effectiveness *and* cost-effectiveness. Secondly, decision-making processes have emerged in several jurisdictions that enable the results of economic evaluations to be used as an integral part of funding, reimbursement, or coverage, decisions.

Although economic evaluation can be applied to all health technologies, including drugs, devices, procedures, and systems of organization of health care, in the main the formal requirement for assessment of cost-effectiveness has been applied to pharmaceuticals. In 1991 the Commonwealth of Australia announced that, from January 1993, economic analyses would be required in submissions to the Pharmaceutical Benefits Advisory Committee, the body that advises the minister on the listing of drugs on the national formulary of publicly subsidized drugs, the Pharmaceutical Benefits Schedule (PBS). A new set of submission guidelines, including economic analyses, was produced (Department of Health, Commonwealth of Australia 1992) and submissions were invited initially on a voluntary basis.

Since that time this policy has become fairly widespread, with approximately half the countries in the European Union, plus Canada and New Zealand, requesting economic analyses of pharmaceuticals, and sometimes other health technologies, to varying degrees. In the last 5 years several payers in the United States and countries in Latin America and Asia have also expressed an interest in receiving economic data. However,

in Africa the majority of economic evaluations are still conducted alongside projects commissioned by international agencies.

Some jurisdictions have requested economic evaluations for technologies other than drugs, including the United Kingdom, where the National Institute for Health and Care Excellence (NICE) assesses the clinical and cost-effectiveness of a wide range of technologies, including public health interventions, before issuing guidance for their use in the National Health Service.

When using economic evaluation, it is important to be clear on the decision-making context. This may be dependent on the broader financing and organization of the health care system in the jurisdiction concerned. For example, in single-payer systems such as a national health service or national health insurance, the reimbursement decision is being made on behalf of the population covered. Therefore, it makes sense to have a centralized process in which the available evidence is considered and a decision made according to a given decision rule. Therefore, it is no surprise that economic evaluations have been much more prominent in jurisdictions with single-payer systems, most notably Australasia, Canada, the Scandinavian countries, and the United Kingdom.

In contrast, in multi-payer systems, such as those operating in the United States and many middle-income countries in Asia and Latin America, the role of health technology assessment (HTA) is more diverse. Also, it is more common in multi-payer systems to have substantial patient copayments, which strengthen the role and legitimacy of the patient as a decision-maker. Thus, decisions on the adoption and use of health technologies tend to be made in a more decentralized fashion, reflecting individual preferences, as opposed to applying a population-wide decision rule. For example, in the United States individual health plans have used economic evaluation to construct a value-based formulary for drugs, where the level of patient copayment is linked to the cost-effectiveness of the product (Sullivan et al. 2009). There is also a growing interest in value-based insurance design (Chernew et al. 2007). Nevertheless, the overall use of HTA and economic evaluation in reimbursement decisions tends to be lower in multi-payer systems.

Another influence on the opportunity to use economic evaluation relates to how the reimbursement decisions are made for different categories of health technology. One of the reasons why economic evaluation has been widely applied to pharmaceuticals is that there is usually a clear decision-making process for including drugs on national or local formularies (e.g. the PBS in Australia). It was thus relatively easy to incorporate an economic component into a decision that was previously made solely on clinical grounds. On the other hand, the reimbursement for medical devices or (surgical) procedures is usually through a broader process of financing hospitals or compensating clinical professionals. While in principle it is possible to incorporate cost-effectiveness principles into these payment processes (e.g. in setting diagnosis-related group (DRG) tariffs or physician fees), it is inherently more complex (Sorenson et al. 2015). Therefore, in jurisdictions where there is an interest in conducting HTAs of non-pharmacological technologies, attempts are made to introduce economic considerations via other routes, such as the development of clinical practice guidelines (e.g. <http://guidance.nice.org.uk/CG/Published>).

1.6 How to use this book

There is a growing literature on economic evaluation in health care. Studies have been conducted by economists, medical researchers, clinicians, and multidisciplinary teams containing one or more of these parties. Several textbooks in health economics contain a discussion of economic evaluation in health care. This book is intended as a supplement to such texts, and not a replacement for them. It aims to take readers past the stage of general appreciation of the methods involved, and towards preparing them for some *hands-on* experience in undertaking an evaluation, perhaps as part of a multidisciplinary team including economists, epidemiologists, and clinicians. We do not claim to provide a comprehensive methods 'cookbook', nor that after reading this book the uninitiated could work without support. Rather, we seek to provide a well-equipped 'tool kit' which, based on our own experience of undertaking economic evaluations, we believe will result in the reader being better prepared to meet most situations.

Chapter 2 introduces the reader to the type of decisions faced in health care and explains why they are often social choices made on behalf of others. Empirical questions of fact are distinguished from the unavoidable but disputed questions of value that are posed, such as which effects should count and how should they be measured and valued. Three broad approaches to the ways in which these questions of value have been addressed are examined by considering whether the objective of health care ought to be health itself, a broader view of welfare based on individual preferences or values revealed in other ways. The purpose is to motivate an understanding of normative principles by demonstrating how they help navigate important debates about methods of economic evaluation, indicating to the reader what to look out for in later chapters.

Chapter 3 discusses the critical assessment of economic evaluations. There is a checklist of ten questions, designed for those wanting to conduct or to critically appraise an economic evaluation. In this way, the chapter serves as a brief introduction to all the methods issues that are discussed in more detail in later chapters. The checklist is then applied in the critical appraisal of two published studies, chosen to illustrate elements of good and bad evaluation practice. One study uses a decision-analytic model and the other bases the economic evaluation on a single clinical dataset, such as a randomized controlled trial, or observational database. The major methodological flaws encountered in published studies are illustrated, such as the failure to consider all relevant alternatives, the inadequate synthesis of the available evidence on effectiveness, the problems in extrapolating beyond the evidence observed in the clinical study, the issues relating to transferring evidence from trials to practice or from one geographical setting to another, and the inadequate characterization of uncertainty.

Chapter 4 explains how summary measures of cost-effectiveness can be used to inform decisions in health care. The first part of the chapter explains the decision rules available when considering a choice between two alternative interventions, introducing the concepts of incremental cost-effectiveness, net health and net money benefit. How opportunity costs might differ in different contexts and the implications for an appropriate cost-effectiveness threshold are explored before examining the relationship between cost-effectiveness and cost–benefit analysis. Later parts of the chapter generalize the decision rules to multiple alternatives, covering the concepts of dominance and

extended dominance. Finally, the chapter uses the framework to examine how to deal with future health care costs, how to discount costs and health benefits and consider whether economic evaluation should be restricted to the perspective of the health care system.

Chapter 5 discusses the ways in which the health effects of health care programmes can be measured and valued, focusing on measures of health state preference. It shows how these can be used to construct measures of health gain, such as QALYs or DALYs. The chapter also discusses the history of utility theories and the notions of utility, value, and preference. It further discusses whether health state preference measures, as usually constructed, can be viewed as 'health utilities'. Finally, the chapter discusses the main approaches for estimating health state preference values and the main generic instruments that have been developed using multi-attribute utility theory.

Chapter 6 discusses other methods for measuring and valuing health effects, often with the objective of expressing these in monetary terms, linking back to the discussion of normative principles, and the use of the cost-effectiveness thresholds discussed in Chapter 2. The main question posed in the chapter is 'what is the social value of the effects of an intervention?' The various approaches for valuing the outputs of health care interventions are discussed, including revealed preferences and stated preferences. A number of stated preference approaches are discussed in detail, including contingent valuation (willingness to pay) and discrete choice experiments.

The chapter also discusses whether the valuation of the benefits from health care should go beyond the valuation of health gain and the implications this raises, including a discussion of the implied social welfare function.

Chapter 7 discusses the measurement and valuation of costs, making a distinction between health care costs and costs outside the health care system. The controversies over the inclusion, in economic evaluations, of productivity costs and costs in added years of life are discussed, as are key concepts such as marginal costs. In discussing costs borne by the patient and family, the possibility of treating some of these impacts as consumption effects is raised, making links to the discussion of measuring and valuing health effects in previous chapters.

Chapter 8 discusses the use of clinical studies (such as randomized trials) as vehicles for economic evaluation. Using individual patient data offers advantages over summary data from secondary sources. These include more rigorous quantification of uncertainty and heterogeneity to inform resource allocation decisions. This chapter also includes an update on statistical developments relating to the analysis of individual patient data for economic evaluation. It also introduces some more fundamental material which has been developed in recent years. Specifically, it discusses recent developments in econometric methods designed to estimate relative effectiveness (and hence cost-effectiveness) from non-randomized studies where confounding is a major concern. It also discusses the value of observational data in economic evaluation and illustrates this by using examples from economic evaluation studies using these methods.

Chapter 9 discusses economic evaluation using decision analytic models, where data from a number of different sources are brought together. The chapter covers recent developments in modelling. In particular, there is a detailed consideration of the design of decision models, the appropriate level of complexity and how to establish the

boundaries for a model. In addition, there is a description of cohort and individual patient sampling models, with more examples of their use. There is also a discussion of dynamic transmission models and their importance in the economic evaluation of interventions for infectious diseases, the use of modelling in the evaluation of diagnostic technologies, and the critical review of decision models.

A fundamental tenet of evidence-based decision making is the need to use all relevant evidence in an analysis undertaken to inform decisions and policy. Hence an understanding of the principles of systematic review and evidence synthesis is essential to practitioners in the field. Chapter 10 covers these principles, including the different uses of systematic reviews in economic evaluation (of economic evaluations and of the different parameters in an economic model), standard 'pairwise' meta-analysis using random and fixed effects, meta-regression and its use in economic evaluation, network meta-analysis, and the use of aggregate versus individual patient data in meta-analysis. The chapter provides extensive references to the wider literature on systematic review and synthesis. The main focus of the chapter is on how and why these methods are used in economic evaluation.

Chapter 11 discusses the characterization, reporting and interpretation of uncertainty in economic evaluation, introducing the reader to more recent methods of analysis which can be used to address a range of important policy questions. The following questions are addressed: why does uncertainty matter; how can parameter uncertainty be represented; how can other sources of uncertainty be represented; is more evidence needed; what type of evidence is needed; what type of research design would be most useful? The chapter concludes with a discussion and illustration of how answers to these questions can be used to inform whether approval or coverage should be withheld until further research is available or whether research should be conducted while a technology is approved for widespread use. In doing so, it is possible to illustrate the impact of irrecoverable costs on approval or coverage decisions and to explore the relationship between uncertainty and price.

Finally, Chapter 12 discusses how the reader can take matters further, having mastered the contents of the book.

References

Boyle, M.H., Torrance, G.W., Sinclair, J.C., and Horwood, S.P. (1983). Economic evaluation of neonatal intensive care of very-low-birth-weight infants. *New England Journal of Medicine*, **308**, 1330–7.

Briggs, A.H. and O'Brien, B.J. (2001). The death of cost-minimisation analysis? *Health Economics*, **10**, 179–84.

Chernew, M.E., Rosen, A.B., and Fendrick, A.M. (2007). Value-based insurance design. *Health Affairs*, **26**(2), w195–w203.

Department of Health, Commonwealth of Australia (1992). *Guidelines for the pharmaceutical industry on preparation of submissions to the pharmaceutical benefits advisory committee, including submissions involving economic analyses*. Canberra: Australian Government Publishing Service.

Dorenkamp, M., Boldt, J., Leber, A.W., et al. (2013). Cost-effectiveness of paclitaxel-coated balloon angioplasty in patients with drug-eluting stent restenosis. *Clinical Cardiology*, **36**, 407–13.

Drummond, M.F. (1992). Cost of illness studies: a major headache? *PharmacoEconomics*, **2**, 1–4.

Gold, M.R., Siegel, J.E., Russell, L.B., and Weinstein, M.C. (ed.) (1996). *Cost-effectiveness in health and medicine*. New York: Oxford University Press.

Guyatt, G., Rennie, D., Meade, M.O., and Cook, D.J. (ed.) (2008). *Users' guides to the medical literature*, 2nd edition. Chicago: American Medical Association.

Haefeli, M., Elfering, A., McIntosh, E., Gray, A., Sukthankar, A., and Boos, N. (2008). A cost-benefit analysis using contingent valuation techniques: a feasibility study in spinal surgery. *Value in Health*, **11**, 575–88.

Haines, T.P., Hill, A.M., Hill, K.D., et al. (2013). Cost effectiveness of patient education for the prevention of falls in hospital: economic evaluation from a randomized controlled trial. *BMC Medicine*, **11**, 135.

Hull, R., Hirsh, J., Sackett, D.L., and Stoddart, G.L. (1981). Cost-effectiveness of clinical diagnosis, venography and non-invasive testing in patients with symptomatic deep-vein thrombosis. *New England Journal of Medicine*, **304**, 1561–7.

Johannesson, M. and Jönsson, B. (1991). Economic evaluation in health care: is there a role for cost–benefit analysis? *Health Policy*, **17**, 1–23.

Logan, A.G., Milne, B.J., Achber, C., Campbell, W.P., and Haynes, R.B. (1981). Cost-effectiveness of a work-site hypertension treatment programme. *Hypertension*, **3**, 211–18.

Lowson, K.V., Drummond, M.F., and Bishop, J.M. (1981). Costing new services: long-term domiciliary oxygen therapy. *Lancet*, **ii**, 1146–9.

Ludbrook, A. (1981). A cost-effectiveness analysis of the treatment of chronic renal failure. *Applied Economics*, **13**, 337–50.

McConnachie, A., Walker, A., Robertson, M., et al. (2014). Long-term impact on healthcare resource utilization of statin treatment, and its cost effectiveness in the primary prevention of cardiovascular disease: a record linkage study. *European Heart Journal*, **35**, 290–8.

Mehrez, A. and Gafni, A. (1989). Quality-adjusted life-years, utility theory and health years equivalents. *Medical Decision Making*, **9**, 142–9.

NICE [National Institute for Health and Care Excellence] (2013). *A guide to the methods of technology appraisal*. Available at <http://www.nice.org.uk/aboutnice/howwework/devnicetech/guidetothemethodsoftechnologyappraisal.jsp> (Accessed 5 December 2013).

Neuhauser, D. and Lewicki, A.M. (1975). What do we gain from the sixth stool guaiac? *New England Journal of Medicine*, **293**, 226–8.

Neumann, P. and Johannesson, M. (1994). The willingness to pay for *in vitro* fertilization: a pilot study using contingent valuation. *Medical Care*, **32**, 686–99.

Nord, E. (1995). The person-trade-off approach to valuing health care programs. *Medical Decision Making*, **15**, 201–8.

O'Brien, B.J., Novosel, S., Torrance, G., and Streiner, D. (1995). Assessing the economic value of a new antidepressant: a willingness-to-pay approach. *PharmacoEconomics*, **8**, 34–5.

O'Byrne, P., Cuddy, L., Taylor, D.W., Birch, S., Morris, J., and Syrotuik, J. (1996). Efficacy and cost–benefit of inhaled corticosteroids in patients considered to have mild asthma in primary care practice. *Canadian Respirology Journal*, **3**, 169–75.

Oldridge, N., Furlong, W., Feeny, D., et al. (1993). Economic evaluation of cardiac rehabilitation soon after acute myocardial infarction. *American Journal of Cardiology*, **72**, 154–61.

Pennington, M., Grieve, R., Sekhon, J.S., Gregg, P., Black, N., and van der Meulen, J.H. (2013). Cemented, cementless, and hybrid prostheses for total hip replacement: cost effectiveness analysis. *BMJ*, **346**, f1026.

Price, D., Small, I., Haughney, J., et al. (2013). Clinical and cost effectiveness of switching asthma patients from fluticasone-salmeterol to extra-fine particle beclometasone-formoterol: a retrospective matched observational study of real-world patients. *Primary Care Respiratory Journal*, **22**, 439–48.

Pukallus, M., Plonka, K., Kularatna, S., et al. (2013). Cost-effectiveness of a telephone-delivered education programme to prevent early childhood caries in a disadvantaged area: a cohort study. *BMJ Open*, **3**, e002579.

Rabalais, A., Walvekar, R.R., Johnson, J.T., and Smith, K.J. (2012). A cost-effectiveness analysis of positron emission tomography-computed tomography surveillance versus up-front neck dissection for management of the neck for N2 disease after chemoradiotherapy. *Laryngoscope*, **122**, 311–14.

Sculpher, M.J. and Buxton, M.J. (1993). The episode-free day as a composite measure of effectiveness. *PharmacoEconomics*, **4**, 345–52.

Sorenson, C., Drummond, M.F., Torbica, A., Callea, G. and Mateus, C. (2015). The role of hospital payments in the adoption of new medical technologies: an international survey of current practice. *Health Economics Politics and Law*, **10**, 133–59.

Stevens, A., Abrams, K., Brazier, R., Fitzpatrick, R. and Lilford, R. (ed.) (2001). *The advanced handbook of methods in evidence-based healthcare*. London: Sage.

Stranges, P.M., Hutton, D.W., and Collins, C.D. (2013). Cost-effectiveness analysis evaluating fidaxomicin versus oral vancomycin for the treatment of *Clostridium difficile* infection in the United States. *Value in Health*, **16**, 297–304.

Sullivan, S.D., Watkins, J., Sweet, B., and Ramsey, S.D. (2009). Health technology assessment in healthcare decisions in the United States. *Value in Health*, **12**(Suppl. 2), S39–44.

Tan-Torres Edejer, T., Baltussen, R., Adam, T., et al. (2003). *WHO guide to cost-effectiveness analysis*. Geneva: World Health Organization.

Torrance, G.W., Raynauld, J.P., Walker, V., et al. (2001). A prospective, randomized, pragmatic, health outcomes trial evaluating the incorporation of hylan G-F 20 into the treatment paradigm for patients with knee osteoarthritis (part 2 of 2): economic results. *Osteoarthritis and Cartilage*, **10**, 518–27.

Chapter 2

Making decisions in health care

2.1 Some basics

As outlined in Chapter 1, those who plan, provide, receive, or pay for health services face an incessant barrage of recurring questions about who should do what, to whom, and with what health care resources. These questions are not academic ones but represent unavoidable decisions which will be made whether they are based on evidence, analysis and explicit social values within an accountable process; or if they are implicit decisions based on 'what was done before', 'gut feelings', 'educated guesses', or even 'what would cause the least difficulties'.

Informing decisions by providing answers to these questions requires the range of possible alternative courses of action to be identified, as well as the evidence of the likely effects of each to be identified and interpreted (see Sections 2.3.1 and 2.3.2). It then becomes possible to estimate the expected effects of each course of action and consider whether the additional benefits offered (compared to the other courses of action available) are sufficient to justify any additional costs. This assessment depends critically on the value of what is given up by others as a consequence (opportunity costs—see Sections 2.3.3 and 2.3.4).

Deciding which effects should count, and how they should be measured and valued depends on disputed questions of social value. For example, is the purpose of health care to improve health itself or should other effects on patients and wider society also count and, if so, how should they be valued relative to health? Similarly, considering whether the costs of a course of action are justified by the benefits offered requires an assessment of what is likely to be given up as a consequence of the additional costs and how these things should be valued relative to the additional health benefits offered. Therefore, informing decisions in health care requires a careful distinction between questions of fact (what are the effects likely to be including what will be given up, see Section 2.3), and the unavoidable but quite naturally disputed questions of social value (which effects should count and how should they be measured and valued). We outline three broad approaches to the ways in which these important questions of value have been addressed in Section 2.4.

2.2 Informing health care choices

In most contexts decisions about health care are made on behalf of others. For example, should a new oncology drug be funded? Should a screening programme be extended to other lower-risk groups in the population? Should access to smoking cessation programmes be made more widely and easily available? These decisions will affect the type

of health care available for the potential beneficiaries and the nature of the health outcomes they can expect from the care they receive. However, the costs of providing the service or intervention to these patients will fall on others. Where there are constraints on the growth in health care expenditure, these additional costs are resources that are no longer available to offer effective care that would benefit other patients with very different health care needs.

Therefore, although the type of specific questions posed in health care appear at first sight very practical and in many respects routine (as they are), they also pose the most profound questions about how social choices ought to be made. Indeed, it is health care decisions that pose difficult and disputed questions of social choice most starkly, since the real question is often who is to live a little longer and who is to die a little sooner than they otherwise might. For these reasons, decisions about health care should be, and often are, subject to the greatest possible scrutiny by a range of stakeholders including patient groups representing those who can benefit from the intervention, and their clinicians concerned that there is access to the range of interventions that can improve the health of patients they are responsible for.

It also includes the manufacturers of pharmaceuticals and devices concerned about the prices they can achieve for their products and the returns they can expect from future research and development. Relevant stakeholders also include, however, those with responsibility for other patients with legitimate claims on health care resources beyond the immediate beneficiaries (e.g. hospital administrators, administrators of health care plans, reimbursement authorities, and ultimately national ministries of health). Wider society also has a stake in these decisions in two respects. Firstly, that decisions are made in a way that is 'fair' to all those eligible to make a claim on the resources of the health care system, since all are potential future patients. Secondly, that those ultimately responsible for health care provision—for example, by directly raising and allocating tax revenue or by setting the rules for how a private or social insurance health care system functions—are representing society's values in what health care can and cannot be provided and to whom.

In most circumstances outside health care, it is individuals (consumers) who assess the potential value of the benefits likely to be offered by a product and then decide whether or not to purchase it using resources available to them. In doing this, they take account of the benefits offered by the other things that they could have purchased instead. In other words, most choices are made by individuals who receive the benefits and incur the costs. Why is this generally not the case in health care?

One difference is that the consumers (patients) are not in a position to know what type of health care is needed. Neither do they know what the benefits are likely to be, especially when they are at the point of needing health care. Patients require specialist expertise (trained and certified clinicians) to diagnose, to advise, and to help select alternative courses of action, by clinicians acting as an agent for the patient. The marked asymmetry of information between the patient and clinician, and the limited opportunities for the patient to find out if they have been well or poorly advised, means that this 'agency relationship' is not necessarily perfect, especially when the interests of the patient and the agent conflict (Johnson 2014; Pita Barros and Olivella 2014). Even when there is no conflict of interests, the agents (doctors) may not necessarily be best

placed to identify and to synthesize all relevant evidence and to undertake the computation required to fully assess all the effects of the alternative courses of action available.

Even if such assessments were possible, however, and clinicians acted as perfect agents for their patients, there are other difficulties. As well as facing uncertainty about the benefits of the alternative interventions at the point of care, the individual also faces uncertainty about when and what type of health care might be needed, so they incur a risk of the potentially catastrophic costs associated with it. For this reason, some form of insurance is present in almost all health care systems; whether this is provided by private insurance, social insurance, or public health care provision funded through general taxation. The difficulty is that, once 'insured' in these ways, at the point of selecting between alternative courses of action much or all of the cost will fall on others (Nyman 2014; Rice 2014).

For these reasons economic evaluation often informs decisions taken on behalf of others by those with responsibility for other patients as well as the immediate beneficiaries. For example, reimbursing a new oncology drug, including it in a benefits package, or approving its use in a public health care system will have a direct impact on access to care for other patients when there is some restriction on health care expenditure.

Alternatively, even when there are no restrictions it will increase the costs of social or private insurance; for example, increasing premiums, or copayments and deductibles. Such an increase in the costs of health care will also reduce access and health outcomes for others; for example, individuals may be unable to afford higher copayments, employers may be unwilling to offer health insurance at all or may select a more restricted benefits package. Whether or not these indirect health effects of increased health care costs are of concern depends on whether one believes that the value of the health effects is fully reflected in these individual choices about whether to pay the higher costs. That is, whether one believes that the health that is lost is forgone because it is less valuable than the increase in cost, but that the health that is gained by incurring these additional costs is gained because it is more valuable (see Section 2.4.2).

It should be emphasized that economic evaluation is also useful to clinicians when advising patients about appropriate treatment at the point of care, and to individuals when considering the choice between alternative insurance plans which offer different packages of access at different costs. It is important to note, however, that the costs that are relevant to these alternative types of decision-maker will differ. Also, which costs fall where and on whom will depend on the nature of the health care system.

Given what is at stake in decisions about health care, it seems wholly inappropriate to abdicate responsibility for these difficult choices and be content with implicit decisions based on opaque scientific and social value judgements such as 'gut feelings' or 'educated guesses', or more arbitrary pressures such as 'what would cause the least difficulties'. Economic evaluation offers an organized consideration of the range of possible alternative courses of action and the evidence of the likely effects of each. This is more likely to lead to better decisions that improve overall social value. It also requires that the scientific judgements needed to interpret evidence are made explicitly so they can be scrutinized and the impact of alternative but plausible views examined. Possibly more importantly, it can provide a clear distinction between these questions of fact and the unavoidable questions of value. Indeed, the main contribution of economic

evaluation may not be in changing the decisions that are ultimately made but how they are made. By making the scientific and social value judgements explicit, it offers the opportunity for proper accountability for choices made on behalf of others.

2.3 **Requirements for economic evaluation**

Regardless of the activities (including health services) to which it is applied, economic evaluation requires a comparison of two or more alternative courses of action, while considering both the inputs (costs) and outputs (consequences) associated with each. These two essential features of any economic evaluation can be used to distinguish and to classify the other types of study that are commonly encountered in the health literature. In Figure 2.1, the answers to two questions are examined: (1) is there comparison of two or more alternatives; and (2) are both costs (inputs) and consequences (outputs) of the alternatives examined. This defines a six-cell matrix.

In cells 1A, 1B, and 2 there is no comparison of alternatives. Such studies do not offer an evaluation, but a description of a single service or intervention. Studies falling in cell 1A only describe the health consequences so offer an *outcome description*, while studies falling in cell 1B offer a *cost description.* The large literature on *cost of illness*, or *burden of illness*, falls into these categories. These types of study describe the resource cost or health consequences of disease to society, but cannot inform the choice between alternative courses of action. Some studies, falling in cell 2, describe both the outcomes and costs of a single service or programme. This type of *cost–outcome description* could also be described as an audit of a service or an intervention.

Cells 3A and 3B identify situations in which two or more alternatives are compared, but in which the costs and consequences of each alternative are not examined simultaneously. In cell 3A, only the consequences of the alternatives are compared. This includes the large and important clinical evaluation literature which provides estimates of the efficacy and effectiveness of the alternative interventions (e.g. randomized clinical trials). In cell 3B, only the costs of the alternatives are examined so represent *cost analyses*.

All the studies described in cells 1–3 cannot in isolation adequately inform a choice between alternative courses of action. Although they are not themselves sufficient to inform decisions, they often provide the key evidence required for

Are both costs (inputs) and consequences (outputs) of the alternatives examined?

Is there a comparison of two or more alternatives?		**No**		**Yes**
	No	Examines only consequences	Examines only costs	
		1A Partial Evaluation Outcome description	**1B Partial Evaluation** Cost description	**2 Partial Evaluation** Cost-outcome description
	Yes	**3A Partial Evaluation** Efficacy or effectiveness evaluation	**3B Partial Evaluation** Cost analysis	**Full economic evaluation** What should count? How should it be measured? How should it be valued?

Fig. 2.1 Distinguishing characteristics of health care evaluations.

decision-making—especially evidence from comparative clinical studies in cell 3A. It is only the economic evaluation in cell 4, however, that provides a comparison of the inputs and outputs of the alternatives. How the type of evaluation might be undertaken will depend on questions of value (which effects should count, how they should be measured and valued), and the questions of fact that follow (which methods of analysis might be most useful in different circumstances and how their results can be interpreted). Addressing these questions is the subject of subsequent chapters. However, we briefly introduce some of the key issues that are important irrespective of the methods used or the type of values that are applied (see Section 2.4).

2.3.1 Which alternatives should be compared?

Informing a particular decision requires identifying the possible alternative courses of action that could be taken to improve the health of patients who find themselves in a particular situation—for example, patients with a particular diagnosis, at a specific stage of disease, and after treatment with other interventions. In other words the alternatives that need to be compared are 'mutually exclusive' in the sense that a patient in that situation can only receive one of them—the decision is 'either/or'. In many situations an intervention can be offered in combination with others, in which case there may be a question about where and in what sequence the intervention should be offered with others currently available, or how diagnosis and treatment should be combined. In these situations the relevant comparison is between alternative strategies (e.g. combinations, sequences of treatment or alternative combinations of diagnostic criteria and treatment).

Different decisions about which alternative (or strategy) to offer can be made for different types of patients who may have the same condition or indication but where the effects of interventions are likely to differ. This might include, for example, patients who have not responded to other treatments, have additional medical conditions, or are believed to be at higher risk due to other characteristics such as past history. Often there will be number of such subgroups of patients, where the effects of the alternative courses of action are likely to differ and for which different decisions about the use of an intervention can be made. The relevant alternatives to compare are the mutually exclusive alternatives within each subgroup, not comparisons between the subgroups. This is because a health system could decide to offer an intervention to all or only to some subgroups (these are not mutually exclusive 'either/or' decisions). These distinctions are discussed a greater length in Chapter 4 (Sections 4.2 and 4.4).

The range of potentially relevant alternatives could be very large and often extend beyond those that are compared in any single clinical study. This is one of the reasons why decision modelling and methods of evidence synthesis are increasingly used to make these comparisons (see Chapters 9 and 10). One important alternative will be 'existing care', but it is important to consider whether existing care is itself is the 'best' that could be done in the absence of the intervention being considered; for example, it may not be if there is a lower cost alternative that is just as effective. Without considering the other alternatives that are available, which may not necessarily be part of current practice, there is a danger that the system could wrongly conclude that a new intervention

is worthwhile only because it has been compared to an alternative that is more costly and/or less effective than others that are available. In other words, any alternative can look 'good' when only compared to something that is sufficiently 'bad' (see Section 4.4).

In principle, relevant alternatives include all those that have some possibility that they might be worthwhile (some conceivable alternatives can be safely ruled out on these grounds). Relevant alternatives might also be restricted by the responsibilities of a particular decision-maker. For example, a given decision-maker may be responsible for cancer treatment but not cancer prevention. The decision being considered may also be restricted to how to treat, for example, a particular form of cancer at a particular stage of disease, rather than how to diagnose and treat at earlier as well as later stages. Even so it can be a challenge to consider all potentially relevant alternatives. An important contribution of economic evaluation is to minimize the chances of an important alternative being excluded from consideration, and not to restrict comparison to what is currently done or what has been compared in other studies.

2.3.2 What evidence is currently available?

The critical inputs to any economic evaluation are evidence about the effects of alternative course of action. Much of this evidence is drawn from the results of clinical studies, especially randomized clinical trials. However, other types of study also provide important evidence about the risk of important clinical events and the type of resource use associated with them (see Chapter 10). Just as it would be inappropriate to only consider a single alternative course of action when others are available, it would also be wrong to select a single clinical study to estimate the likely effects when other relevant evidence is available. There are circumstances when a single clinical study is the key or only piece of relevant evidence, so economic evaluation can be conducted alongside or within a single study (see Chapter 8). More commonly, however, there are a number of relevant studies, several alternatives to compare (not all of which are compared in single studies), and a selection of different types of evidence required from different types of studies to estimate the longer-term effects on health and costs (see Chapters 9 and 10).

Therefore, a systematic approach to searching for published evidence is needed so that the evidence used is not selected in a potentially biased way. Methods of systematic review are well developed and guides to these methods are available elsewhere (Guyatt and Rennie 2002). Once identified, the results from relevant studies must be extracted, interpreted and then, where appropriate, combined or synthesized to provide estimates of the key parameters required to estimate expected effects (see Chapter 10). Such methods of meta-analysis are well developed and introductory guides to their use and interpretation of methods are available (Borenstein et al. 2009). Therefore, economic evaluation provides a means to bring together and to make best use of the published results of clinical studies and other relevant evidence, through systematic review and meta-analysis, so it can more directly inform the decisions that will be made.

2.3.3 What perspective should be adopted?

Which costs and consequences should count, and how they should be measured and valued, depends to a large extent on which of the many different types of decision-maker

in health care is intended to be informed by economic evaluation. There are a number of different types of potential decision-makers including: individual patients and their clinicians; those with a wider responsibility for other patients beyond the immediate beneficiaries (e.g. hospital administrators, administrators of health care plans, reimbursement authorities, and ultimately national ministries of health); and those ultimately responsible for health care provision (through directly allocating resources or by setting the rules for how the health care system functions) and for other socially valuable activities (e.g. education, defence, reductions in tax, etc.).

Clearly the costs that are relevant to these different types of decision-makers will differ and which costs fall where, and on whom, will depend on the nature of the health care system. Alternatives that might appear attractive from one viewpoint or perspective may appear unattractive from others. Therefore, an important question is which of these very different perspectives will be most appropriate? One pragmatic answer to this question is: the perspective of those who commissioned, or who are intended to be informed by, the analysis. In which case, the focus should be decisions that are within their remit and the costs and consequences that are relevant to them. However, there remains a bigger question of which perspective ought to inform the type of social choices described in Section 2.2. This is discussed at greater length in Section 4.5.3, but will depend to a large extent on questions of value: what is the primary purpose of health care (see Section 2.3); and should the constraints on the resources available for health be respected as a revealed expression of value (see Section 2.4)?

2.3.4 What will be given up as a consequence of additional costs?

The methods of analysis described in subsequent chapters make it possible to estimate the expected costs and consequences of each of the alternative courses of action available. However, to decide whether the additional benefits offered (compared to the other courses of action available) are sufficient to justify any additional costs depends critically on the value of what is given up by others as a consequence—the opportunity costs. To conclude that an alternative which imposes additional costs represents a 'cost-effective' use of resources requires some comparison with the opportunity costs. Without this the results of economic evaluation cannot be put to use and inform decisions—it remains only a description of costs and consequences.

Opportunity costs will depend on what is likely to be given up and the value placed on it. What is likely to be given up depends to some extent on the nature of the health care system. For example, where there is a budget for the public provision of health care or where there are other constraints on the growth in health care expenditure, the opportunity costs will fall, at least in part, on health outcomes. This is because the additional costs are resources that are no longer available to offer effective care that would benefit other patients with very different health care needs. In these circumstances an estimate of the health expected to be given up as a consequence of the additional cost is required (i.e. an estimate of the cost-effectiveness 'threshold'; see Section 4.3.1). If all costs are health care costs, then considering whether an alternative is cost-effective is equivalent to asking whether the additional health benefits offered are greater than the health expected to be lost as a consequence of the additional cost.

Alternatively, if there are no restrictions on health care expenditure then the opportunity costs will fall on other consumption opportunities outside the health care system. In these circumstances, deciding whether the health benefits are worth the loss of other consumption opportunities requires some estimate of how much consumption should be given up to improve health (i.e. a consumption value of health or willingness to pay for health improvement, see Section 4.3.2 and Chapter 6). In most health care systems some of the opportunity costs will fall on health, even when there are no administrative budget constraints, and some will fall on consumption even when there is a fixed budget for public provision. This question of how opportunity costs might be assessed, what a cost-effectiveness 'threshold' ought to represent, and how it might be estimated is discussed at greater length in Section 4.3.

2.3.5 How uncertain is the decision and is more evidence needed?

Economic evaluation provides a framework for the organized consideration, based on existing evidence, of the likely costs and consequences of alternative courses of action. However, the expected costs and consequences will be uncertain, partly because of the uncertainty in the estimates of inputs or parameters of the type of analysis commonly used to estimate them (see Chapters 8–11). In the face of this uncertainty a decision-maker must nonetheless come to a view about which alternative course of action is expected to be worthwhile. An assessment of the implications of uncertainty surrounding this decision is an essential part of any decision-making process for a number of reasons. First, to assess the potential value of acquiring additional evidence that could better inform this decision in the future. Secondly, to identify the type of evidence that might be needed and how further research might be designed. Thirdly, to consider whether a decision to approve or reimburse an intervention or to invest in a new service should be delayed until the additional evidence required becomes available. How uncertainty can be characterized and used to inform these aspects of decisions is dealt with in Chapter 11.

2.3.6 Does this type of analysis lead to better decisions?

The purpose of economic evaluation is not to predict the future costs and consequences of choosing a particular course of action but to inform a decision *at a particular point in time*. The question is whether a 'better' decision will be made using economic evaluation at the time the decision must be made, not necessarily how well the analysis predicts future costs and consequences. In principle this could be tested by randomly allocating decision-makers to those using and those not using the results of an economic evaluation, and following them up to see whether the outcomes (costs and consequences) are 'better' for the group where decisions were informed by such analysis. Therefore, the predictive accuracy of the analysis is not the real test—an analysis that ultimately proved to be highly accurate may not have changed decisions and improved outcomes, but one that proved less accurate may have led to changes in decisions with substantially better outcomes overall.

Economic evaluation should combine the relevant evidence available at the time the decision is made with the current understanding of disease processes and the health

care system. Since evidence and understanding accumulates over time, all quantitative analysis will ultimately be 'wrong' with hindsight. The appropriate question is 'was it useful at the time, and did it lead to better decisions?'

'Science is at no moment quite right, but it is seldom quite wrong, and has, as a rule, a better chance of being right than the theories of the unscientific. It is, therefore, rational to accept it hypothetically' (Russell 1959, p.13). If 'organized consideration of the range of possible alternative courses of action incorporating the evidence of the likely cost and consequences of each', is substituted for 'science' and implicit decision-making for 'the theories of the unscientific' then, this is precisely what is being claimed.

Of course, no quantitative analysis, no matter how sophisticated or assiduously conducted, can capture all aspects of value or reflect all reasonably held scientific judgements; not least because they are quite naturally disputed. The question is whether an economic evaluation can directly inform the assessments required when decision-makers are faced with unavoidable choices in health care and does it offer a useful starting point for deliberation about the relative value of alternative courses of action that is accountable to reason, existing evidence, and widely held values. Economic evaluation enables clear distinctions to be made between these questions of fact and value. By making these explicit it offers the opportunity for 'better' decisions to be made in the sense that they can be subject to proper accountability. It is for this reason that the real value of economic evaluation is not simply changing which decisions are ultimately made but how these choices are made on others' behalf (Culyer 2012a).

2.4 What is the purpose of health care interventions?

How social choices about health care should be made on the behalf of others requires decisions about which effects should count and how they should be measured and valued. These disputed questions are really about what the purpose of health care is believed to be: is it health itself or should other effects on patients and wider society also count and, if so, how should they be valued relative to health? We outline three broad approaches to the ways in which these important questions of social value have been addressed.

2.4.1 Improving health?

Improving health is most natural answer to the question of what is the primary purpose of health care. Certainly, improving population health is often the stated objective of policies, health care institutions, and clinicians. It is also the social objective that underpins much of the economic evaluation that is undertaken. Although a relatively simple and narrow social objective compared to the broader notions of welfare discussed in Sections 2.4.2 and 2.4.3, it does, nonetheless, pose some difficult questions of what aspects of health are important and how they should be measured and weighted.

2.4.1.1 How can we measure health?

There are a vast number of different measures of the health effect of interventions used in the clinical evaluation literature. A simple taxonomy of health outcome measures is illustrated in Figure 2.2. Many of the measures reported in clinical studies are not really

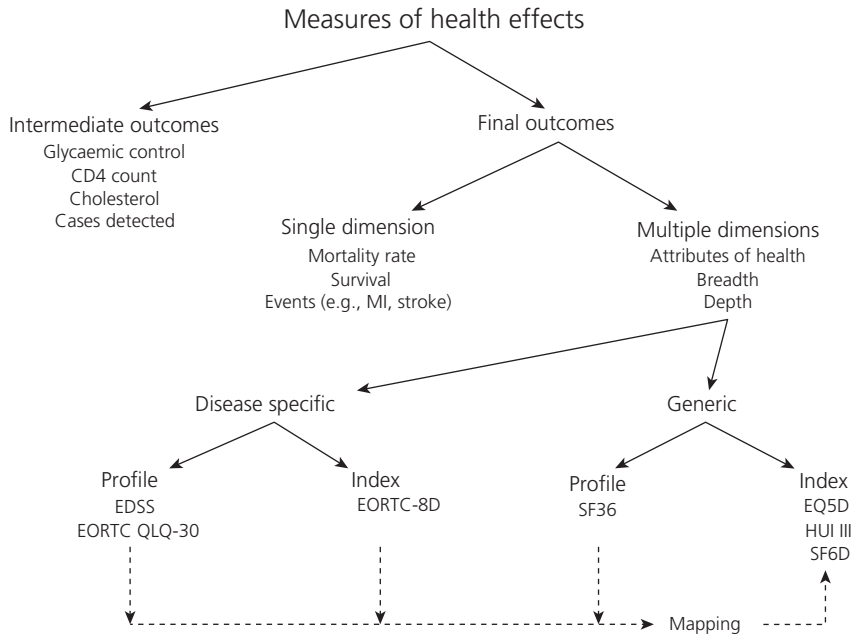


Fig. 2.2 A taxonomy of measures of health outcome.

measures of health outcome itself, but of intermediate outcomes or surrogates that can be linked to changes in health outcome (e.g. improved glycaemic control in diabetes). Although such measures of intermediate outcome might be adequate to establish whether one intervention is more effective than another, such measures cannot, by themselves, indicate the magnitude of any improvement in health offered by an intervention. To do that, changes in the intermediate outcome or surrogate must be linked to changes in measures of health outcome itself (see Chapters 5 and 9).

Some measures of health outcome are restricted to a single aspect or dimension of health: for example, effects on mortality and survival. However, although length of life is clearly an important aspect of health, the quality in which it is lived is also important. There are very many measures which describe the different dimensions or attributes of health-related quality of life (HRQoL) and the different levels achieved within each. Many of these descriptions are specific to particular disease areas and attempt to describe those aspects of health that are most important for patients with a particular condition. Others attempt to offer a generic description where any state of health can be represented as a level of performance on each of the attributes. So even the apparently simple task of describing health poses difficult questions of which dimensions or attributes are important.

2.4.1.2 How should aspects of health be weighted?

Different aspects of HRQoL can be described and changes in the different dimensions recorded using those descriptions. In addition, these different aspects of HRQoL need

to be weighted in order to provide a measure that can be used to identify whether or not health has improved or deteriorated and to what extent. Some multidimensional measures of HRQoL do not attempt to so and simply provide a profile or description of performance across the different attributes. The problem is that, when interpreting these profiles, coming to a view about whether health has improved requires some weighting, whether this is done explicitly or implicitly. Many of these types of profiles, whether specific to a particular disease (e.g. Expanded Disability Status Scale (EDSS) in multiple sclerosis) or generic (e.g. SF36), do provide a score based on performance on each attribute, so it may be tempting to use the overall score as a measure of the magnitude and direction of changes in health. The problem is that such scores are likely to be unrelated to the importance to these different aspects of HRQoL. Therefore, they can misrepresent the magnitude of any change (i.e. the measure will not have *cardinal* properties; Culyer 2014). Indeed, they might also misrepresent the direction of change: for example, if a less important attribute improves but another very important one deteriorates, the overall score might rise when health has in fact declined (i.e. the measure may not even have *ordinal* properties; Culyer 2014).

Therefore, how should weights be assigned to the different aspects of health to represent their relative importance? One way to do so is to ask people to provide weights that reflect how they would rate each of the possible health states relative to full health. Here health states would represent combinations of attributes at particular levels. In this way each possible health state can weighted relative to full health (with a score of 1). There are many ways to elicit such health state valuations (see Chapter 5), although choice-based methods reflect the real trade-offs in many health care decisions. An example of a choice-based method is where people are asked how much time in full health they would be willing to give up in order to avoid time in a particular health state. By assigning weights in this way not only are the different aspects of HRQoL weighted, but it also enables effects on length of life to be combined with the quality in which it will be lived (see Box 1.3).

As well as different ways to elicit health state valuations, there are different ways to construct measures of HRQoL by combining the resulting set of weights (or a 'tariff') with the descriptions of health states (see Chapter 5). At each stage different assumptions are made about people's preferences, some of which could be relaxed with more evidence (e.g. by valuing profiles, or sequences of health states rather than single states). The question is whether the measure is an adequate measure of health for the purpose of informing health care decisions by providing a useful starting point for the deliberation required, not whether it captures all possible aspects and fully reflects individual's preferences (such a measure is sometimes described as 'utility'; see Chapter 5).

Another important question is who should be asked to provide these weights. Should it be a representative sample of the general population or should patients with the condition be asked to weight the different attributes? Again this is, in part, a question of social value: whose preferences should be used judge the benefits offered by an intervention? On the one hand current patients are probably the best proxy for how people are likely to feel should they find themselves in that health state. The general population might have little experience of the health state they are valuing and consequently find it difficult to anticipate how they would adapt to it: as a consequence their valuation will

tend to be lower. On the other hand current patients may have experience of the health states they are valuing and the adaptations that are possible, but they might have little or no recent experience of full health and might struggle to imagine what that would be like (e.g. due to chronic illness or generally poor health). As a consequence they will tend to provide higher values. Whether it is more appropriate to use values which do not account for likely adaptation and experience or values which may embed the effect of poor health experience and poor expectations is not self-evident (Dolan 1999).

2.4.1.3 Why do we need a generic measure?

Many interventions will have effects on different aspects of health outcome associated with the specific disease or condition, including unintended consequences (adverse events), which can be measured in a variety of different ways (clinical- and disease-specific measures). Some interventions in one disease area (e.g. diabetes) will have effects on outcomes in other areas (e.g. cardiovascular, wound management, ophthalmology), each with specific measures of outcome. Therefore, to identify the additional health benefits offered by the alternative courses of action available, a measure of health is required that can summarize an often complex prospect of effects.

Any summary must implicitly or explicitly weight the different aspects of outcome. If the consistency and accountability of decisions are important, some explicit weighting of different aspects of health is necessary, which might reflect the preferences of future potential patients. Of course any measure is necessarily a simplification since some complete and universal description of all aspects of health is unattainable. To be feasible within time and resource constraints, some assumptions about individual preferences and social values will be inevitable.

But why should a generic rather than a disease-specific measure be used? There are a number of reasons why a generic measure of health outcome has advantages over one that is specifically designed to measure health in a particular disease. Some comparable generic measure is required when informing a particular decision where some or all the alternative courses of action may have effects on health outcomes in other types of diseases, or where there are side effects of treatment which have impacts on different aspects of health. However, even if effects on health are limited to a particular disease or even a single dimension of outcome, there remain two important advantages to a generic and comparable measure.

First, consistency with other decisions relevant to different groups of patients with other conditions, made at different times, is an important aspect of accountability. This requires a measure of health that is comparable across the range of health care decisions.

Secondly, however, there is a more fundamental reason. Any additional cost of the alternatives available will mean that something of value must be given up elsewhere by others. Where the additional costs are resources that will no longer be available to offer effective care that could improve the health outcomes of other patients with very different health care needs, some comparison of different aspects of health across different disease areas is unavoidable. Therefore, informing a particular decision about the alternatives available to patients who find themselves in a particular situation necessarily involves a comparison of the additional health benefits offered with other aspects of health outcome that could have been gained for other patients with different

conditions. A generic and comparable measure enables a comparison of the health expected to be gained with the health expected to be lost elsewhere.

2.4.2 Improving welfare?

Although improving health might be the most natural answer to the question of what is the primary purpose of health care, health improvement is not the only thing that is socially valuable. If this were not true then society would devote all its resources to health-enhancing activities. It does not do so because consumption opportunities for individuals as well as other objectives of public policy and public expenditure (e.g. education or criminal justice) are also valuable. Insofar as decisions about health care have an impact on other valuable activities, then they ought to be taken into account. However, to do so requires having a much broader view of what counts and how they should be measured and valued relative to each other. It requires a definition of social welfare, often described as specifying a social welfare function (Culyer 2014).

2.4.2.1 How can we define welfare?

The task of specifying all the things that count and how they should be valued appears at first sight hopelessly ambitious. This is because specifying any welfare function, even one that represents one's own personal views, would be a considerable task. It also appears futile because, even if it were possible, why should one view be imposed on others? People are likely to have different views about what should count and, even if they agreed on what should count, they are likely to differ on how these things should be valued relative to each other.

Traditionally economics overcomes this problem by focusing only on the preferences of the individual. It founds a definition of welfare on the notion that it is only the individual that can decide whether their welfare has improved or not, and that they make choices based on their preferences to improve their own welfare (they maximize their own utility). Therefore, we can infer whether welfare has improved or not from the preferences that individuals have, which are revealed by the choices that they make. For example, if an individual chooses x rather than y we can say that they prefer x to y so their welfare must be greater with x rather than y.

This is sufficient if the alternatives being considered only improve the welfare of some (effects are preferred by some) and do not reduce the welfare of others (they are indifferent between x and y). That is, if there are benefits for some but at no cost to others—described in economics as a Pareto improvement (Culyer 2014). Unfortunately, this is not particularly useful because most decisions, including those in health care, involve a choice between alternatives where the additional benefits offered will accrue to some but the additional costs will mean that sources of value must be given up by others.

The way economics traditionally deals with this problem is to ask whether those who would gain from choosing a particular course of action could, in principle at least, compensate those who would lose but still remain better off; that is, they would still prefer the course of action even if they paid the compensation. If they can compensate the losers then the course of action can be regarded as a potential Pareto improvement. If the compensation is paid then some will be better off (the beneficiaries still prefer it, so their welfare is higher), but no one will be worse off because those who lose will be

compensated and, by definition, they will be indifferent so their welfare is unchanged (i.e. it would be a Pareto rather than a potential Pareto improvement) (Tsuchiya and Williams 2001).

2.4.2.2 How can changes in welfare be measured?

This definition of welfare, founded entirely on individual preferences, combined with the principle of a 'compensation test', is the foundation of how economics has traditionally informed questions of whether an alternative course of action is 'efficient', by which is meant that it offers an improvement in this view of social welfare.

So how can the compensation that could be offered by the beneficiaries and required by the losers be measured and compared? Individuals reveal their preferences by the choices they make in markets for those things that might be gained (outputs) or given up (inputs). Individuals assess the potential value of the benefits offered by a product and then decide whether or not to purchase it at the market price. They will only choose to purchase it if, at the current market price, it improves their welfare. Therefore, market prices do not simply indicate what things 'cost' but they also represent the social value of the inputs and outputs of an alternative. This is because the price represents the compensation required to give up the product (if it is an input) or the amount they would be willing to offer as compensation to others (if it is an output). The common metric for such measure of compensation is money because it represents the other consumption opportunities available at market prices (Hurley 2000; Tsuchiya and Williams 2001).

In many respects this view suggests there is no role of economic evaluation at all. So long as consumers are fully informed and undistorted markets exist for all the inputs and outputs of the alternative courses of action being considered, then, 'the competitive market acts as a giant (but decentralized) cost–benefit calculator. No second guessing by economists is required' (Pauly 1995, p.103).

However, even when markets exist, observed market prices will need to be 'adjusted' for any distortions that might be present in the relevant markets. For example, such 'adjustment' may be needed where markets are not competitive due to monopoly, where they are distorted by taxation, or where there are effects that are not reflected in market prices, such as environmental damage (Boadway and Bruce 1984). There are a number of reasons why we do not observe competitive markets for health and health care (see Section 2.2). So the health effects of the alternatives considered in health care decisions need to be valued based on a 'price' *as if* there was a competitive and undistorted market. Such 'shadow prices' can be derived in different ways. One is by observing situations where people make choices where health is valued implicitly as an attribute in other markets (e.g. the trade-off between risk and wages in the labour market). More commonly experiments can be undertaken which offer people hypothetical choices to establish how much they are willing to pay for health or the collection of benefits offered by an alternative, or which establish what trade-offs individuals would be willing to make between health and a range of other attributes, one of which can be valued in money terms (see Chapter 6).

The details of such valuations are dealt with in later chapters but the principles of evaluation if this view of welfare is adopted are clear: identify those who gain and lose;

value inputs (losses) and outputs (gains) either at their market prices, where they are believed to be undistorted, or estimate shadow prices that would reflect the outcome of complete and undistorted market to value them (Mishan 1971; Sugden and Williams 1979). This places all effects in the same metric of equivalent consumption opportunities so we can ask whether the consumption value of the benefits exceeds the consumption value of the costs. If they do, then we can conclude that the gainers could in principle compensate the losers, and that the alternative being considered would be an efficient use of resources and would improve social welfare.

This 'welfarist' view implies that:

> Health care programmes should be judged in the same way as any other proposed change: i.e., the only question is do they represent a potential Pareto improvement (as measured by individual utility) not do they improve health outcomes as measured in either physical units or health state utility. It is possible that a programme may increase the health of some but reduce the health of others. If those that gain health outcome can compensate those that lose health (measured by individual willingness to pay) then the programme may be a potential Pareto improvement even if the health outcomes overall are lower. (Pauly 1995).

This is the basis of what is traditionally called 'welfarist economics' (Brouwer et al. 2008; Hurley 2000).

2.4.3 Is there more to welfare than individual preferences?

This view of how welfare can be defined and measured enables strong prescriptions for social choice. These are based only on observed market prices, subject to appropriate adjustments for the effect of distortions, and revealed or hypothetical prices where suitable markets do not exist. It enables claims to be made about what would improve social welfare (what is efficient) and, therefore, what ought to be done. In short, it offers a prescription for social choice without actually having to specify a particular social welfare function. However, the strength of this prescription comes at some cost.

One might object to the prescription offered by a particular analysis on the grounds that the many potential distortions in relevant markets have not been fully accounted for (whether a market is distorted and to what extent is often disputed) or that the hypothetical values are unreliable. Since income determines to a large extent what level of compensation would be required by gainers or can be offered by losers, one might also object to preferences weighted according to the prevailing distribution of income if it is believed to be 'unjust' in some way. For example, even if the prevailing distribution of income is considered to have arisen from individuals maximizing their welfare by making informed, rational, and free choices in competitive markets, those individuals' initial endowments may be not be judged as 'fair'. In these circumstances valuations can be adjusted, assuming that some more preferred distribution of income had been achieved (Sugden and Williams 1979). However, it would require a (non-welfarist) means to identify what the preferred distribution ought to be. Then adjustment can be applied to the valuation of all inputs and outputs, not just health.

The more fundamental objections to the welfarist approach are to question whether individuals are always the best and only judge of their welfare; and to doubt the view that the reasons for individual preferences are irrelevant and that it is only individuals'

welfare, as revealed by their preferences, that defines the social good and is relevant to social choices. There are a number of different types of argument that have been made. First, that some goods and services should be provided or subsidized because they are meritorious (Musgrave 1959). Secondly, that basic goods and services (health included) are so important to other aspects of life that equity in their distribution is important to society (Tobin 1970). Thirdly, that basic or primary goods are necessary to enjoy others and fully participate in society (Rawls 1971); or, alternatively, that a notion of capabilities, that rests on what the provision of goods and services enable people to do and be, is more important than preference (Sen 1979, 2002).

What all these views have in common is the idea that 'mere' preference is too narrow to judge social welfare, and that there are other things beyond individual preferences that ought to play a role in social choices. For this reason the notion of including other characteristics (of people and commodities), in addition to preferences, has been termed 'extra-welfarist' (Culyer 1989, 2012b).

Objections to welfarism also seem to be reflected in the lack of appetite among policy-makers and decision-makers for fully welfarist analysis in the economic evaluation of health care. There are very few examples of fully welfarist analyses in the published literature, and even fewer examples of their use in policy decisions. Indeed, the principle of allocating resources on the basis of preferences weighted according to the prevailing distribution of income seems to have been explicitly rejected by government in many spheres of public policy and, in particular, in health. The prime reason appears to relate to concerns about the existing distribution of income and health.

2.4.3.1 Distribution of income and health

There are, however, examples of economic evaluation of health care that use monetary valuations of the outputs, based on estimates of how much patients are willing to pay for the estimated health effects or the prospect offered by alternative courses of action. However, these studies commonly present estimates adjusted for average income rather than individual values (see Chapter 6). This is often justified by citing equity and distributional concerns. Adjusting valuations for average income would be consistent with a view that the socially 'optimal' distribution is believed to be the average income (generally it is not). Such a view would also require all valuations of inputs and outputs to be adjusted (whether based on market or shadow prices), not just health.

If the distribution of income is judged to be acceptable, but it is the distribution of health that is of concern, then valuing health effects at average income is unlikely to account for health equity concerns. If there are concerns about the distribution of health effects it becomes necessary to trade off alternatives which have net effects on health and health equity against the consumption value of their additional costs. These concerns pose questions such as: how much consumption opportunities or health should society give up to improve health equity (Deaton 2002)? The types of assessments required to make these trade-offs cannot be adequately addressed by using valuations adjusted for average income or any other preferred distribution. The problem is that equity concerns are an additional social objective and need to be evaluated as such, rather than as a restriction on how inputs and outputs are valued (Asiara et al. 2014; Cookson et al. 2014).

2.4.3.2 Distinction between known and unknown lives

Founding the valuation of the health effects on individual preferences requires asking how much consumption individuals are, on average, willing to give up to reduce risk or how much additional consumption is required to compensate for incurring additional risk. This poses real difficulties when it comes to social choices because any finite value of the effect of an alternative on mortality risk rests entirely on the distinction between known and unknown lives. If the life 'saved' or 'lost' is known then no compensation is possible (there is an unbounded value), so finite values are only possible if the life is unknown (the individual faces a change in risk rather than certain death) (Mishan 1985).

It seems to be a common and possibly innate emotional response to favour those who are known and identifiable or with whom we can most closely identify. The question is whether this common emotional response is a sound basis for the type of social decisions required in health care. The distinction between known and unknown lives is dubious because it is often only a question of perspective (someone may be unknown to some but known to others) or ignorance (the currently unknown could, in principle, become known with sufficient effort); and, in any case, the unknown will become known over time (it can only be known lives that are ultimately lost or saved) (Broome 1978). This question becomes especially important when the beneficiaries of an intervention are often identifiable but those that will bear the opportunity costs (give up health as a consequence of the additional costs) are not, but are no less 'real' (Claxton and Culyer 2006).

If social choices are to be consistent and coherent they should use all the information that is available at the time. One piece of information available to decision-makers is that any (ex-ante) finite compensation acceptable for loss of an unknown life is known not to approximate the (ex-post) unbounded compensation when the life is actually lost. Therefore, social decisions based on (ex-ante) individual preferences will be inconsistent and incoherent (Broome 1978). This is not to suggest that life and health cannot or should not be valued in monetary terms; such trade-offs are and will be made. However, such valuation might better rest on other sources of valuation rather than individual preferences measured by the compensation required. For example, where health effects are measured using HRQoL and when making the unavoidable trade-offs between health gained and lost elsewhere, all lives are regarded and treated as if they are known.

2.4.3.3 Informing social decision-making

It is for these reasons that the extra-welfarist rather than the strictly welfarist approach underpins much of the economic evaluation in health care (Coast et al. 2008). The extra-welfarist approach identifies some of the reasons why effects on health might quite reasonably be singled out to be measured and reported separately from other inputs and outputs, without necessarily ignoring or disregarding the impact on other consumption opportunities that might still be valued based on individuals' preferences expressed in relevant markets (see Section 4.5.3). Of course taking this approach poses the question of what other characteristics should be included and what weight should they be given. It also poses the difficult question of what the social welfare function

should look like and which of the very many alternative specifications ought to be adopted and imposed on individuals whose preferences differ.

There are two closely related responses to this problem. The first is to say that it is not the preferences of the individual that matter but the preferences, values, and criteria of the decision-maker responsible for making these choices that should count as they have been given responsibility to make decisions on others' behalf (Sugden and Williams 1979). They may well wish to reflect individual preference in their valuation of inputs and outputs; for example, to use market or shadow prices of the inputs, but use measures of HRQoL based on patient or population values for health states for the output. In these circumstances the role of economic evaluation is to understand the values of the social decision-makers, make sure these are reflected in the analysis so that the results are relevant to them, and ensure they are consistently applied so they, and others, can see and critically reflect on the implications of holding particular values. Under this view, the purpose of economic evaluation is not to impose a particular view of what social welfare ought to be, but to make explicit the implications of the criteria that have been and are being used to make these types of decision (Williams 1993). However, which alternative appears worthwhile will depend on the particular decision-maker and the values they hold, so it poses the question of why their values carry any particular weight compared to others when making social choices.

A useful response is to consider how society tries to solve this difficult problem of how social welfare ought to be defined when there are disputed, conflicting, and contradictory claims that are forever changing. In general, societies try to establish legitimate processes to balance these conflicting and contradictory claims. For example, social democratic processes can be viewed as establishing a socially legitimate higher authority. This is not confined to the executive, but also includes all the checks and balances associated with the range of formal institutions (e.g. elected legislature, independent judiciary) as well as the less formal institutions of civil society (e.g. a free press, public debate, religious freedom).

Decision-making bodies and institutions in health can be seen as the agent of this socially legitimate higher authority (the principal), where the latter is unable to express an explicit, complete, and coherent social welfare function. The agent acts as a delegated authority, but one that cannot be asked to improve social welfare, since it cannot be specified by the principal. Rather, the principal allocates resources and devolves powers to the agent, giving it a responsibility to pursue explicit and specific objectives that are regarded as socially valuable (such as improving health). Since the achievement of objectives needs to be monitored by the principal to hold the agent to account, they must be measurable, so necessarily narrowly defined. The implication of this process (e.g. the current trade-off between health and the resources made available for health care) reveals a partial but legitimate expression of some unknown underlying social welfare function (Paulden and Claxton 2012). This does not require that the agents or the principle are 'perfect' in some way (they may be quite dysfunctional); only that, by revealing the values and exposing the implications, economic evaluation can contribute to a process of accountability and change.

In these circumstances economic evaluation cannot be used to make claims about social welfare or the optimality or otherwise of the resources allocated to health care. Its

role is more modest, claiming to inform social decisions in health rather than prescribing social choice. It is this role that economic evaluation has tended to play in health policy and underpins much of the economic evaluations that have been conducted. The ambition maybe more modest, but it exposes the policy implications of the social values implicit in existing policies and the resources allocated by those who claim some legitimacy to make such decisions. In this sense it contributes to holding higher authorities to account and can contribute to changing priorities, policy, and resourcing through social democratic processes where conflicts with widely held social values are exposed.

2.5 Concluding remarks

Economic evaluation seeks to inform the range of very practical and unavoidable decisions in health care, which will be made whether or not they are based on evidence, analysis and explicit social values. These decisions will affect the type of health care available for the potential beneficiaries and the type of health outcomes they can expect. However, the costs of providing the service or intervention for these patients will mean that resources are not available to be used in other ways, such as providing health care for other patients with different conditions, reducing the tax burden of collectively funded health care, or reducing the costs of social or private insurance premiums. Decisions require consideration of whether what is likely to be given up by others as a consequence of additional costs are justified by the benefits offered to the immediate beneficiaries.

Therefore, although the type of specific questions posed in health care appear at first sight very practical and routine (as they are), they also pose the most profound questions about how choices ought to be made on others' behalf; this is, what effects should count and how should they be measured and valued. Indeed, it is decisions in health care that pose these difficult and disputed questions most starkly, so they should be, and often are, subject to the greatest possible public scrutiny.

2.5.1 Distinguishing questions of fact and value

Given what is at stake in social choices about health care, implicit decisions based on opaque scientific and social value judgements seem inappropriate. Economic evaluation offers an organized consideration of the range of possible alternative courses of action combined with the evidence of the likely effects of each. By doing so it makes explicit the scientific judgements required to estimate the costs and consequences of each. These questions of fact can then be scrutinized, and the impact of alternative but plausible views examined. It also means that they can be carefully distinguished from questions of value: what effects should count and how they should be measured and valued. This ensures some consistency in the way they are applied in other decisions, offers the opportunity for critical reflection on the implications of holding particular values, and provides an opportunity for proper accountability for decision made on others' behalf.

This is not to say that all economic evaluations necessarily offer such clear distinctions or that the process by which economic evaluation is used to inform decisions is sufficiently transparent to allow proper accountability. For example, it can sometimes

be difficult to: properly scrutinize and critically appraise how the questions of fact have been addressed, especially if analysis is poorly reported (see Chapter 3); come to a view about how uncertain decisions are likely to be, if sources of uncertainty have not been characterized; or how sensitive the results might be to other assumptions and judgements if there has been insufficient exploration of other plausible views (see Chapters 9 and 11). Similarly, the questions of value that underpin the analysis are not always obvious and will sometimes need to be teased out. Certainly, whether an analysis is described as a cost-effectiveness analysis, cost–utility analysis, or cost–benefit analysis is not necessarily a good guide to how the questions of value discussed in Section 2.4 have been addressed.

2.5.2 Implications for economic evaluation

There are profoundly different but reasonably held views about the role economic analysis ought to play in social choice in general and particularly in health. The different approaches to questions of social value defines, to a large extent, the role that economic evaluation ought to play—either a prescription of what ought to be done or informing legitimate decision-making processes. These disputed questions also underpin much of the debate about what are appropriate methods of analysis that are discussed in more detail in later chapters. For example, what the appropriate perspective for economic evaluation ought to be (see Chapter 4), including whether future unrelated costs should be included in the analysis (see Chapter 7); at what rates should future costs and benefits be discounted (see Chapters 4 and 7); whether an assessment of what is likely to be given up as a consequence of additional costs (a cost-effectiveness threshold) reflect the health effects of changes in health care expenditure or some social consumption value of health (see Chapter 4); should we try to measure the health effects of alternative courses of action (see Chapter 5) or directly value the prospect of the benefits they offer by how much patients might be willing pay for it (see Chapter 6); and, if we are to measure health outcomes, then whose preferences should count and how concerned should we be about the assumptions that may be required about individual preferences when constructing measures of HRQoL (see Chapter 5)? These issues and debates are raised and discussed in subsequent chapters. Although they appear at first sight technical matters of how the details of analysis should be undertaken, the answers rest quite firmly on the questions of value discussed in this chapter.

Acknowledgement

Text extracts from Pauly M.V., 'Valuing health benefits in monetary terms', in Sloan F.A. (ed.), *Valuing health care: costs, benefits and effectiveness of pharmaceuticals and other medical technologies*, Cambridge University Press, Cambridge, UK, Copyright © Cambridge University Press 1995, reproduced with permission.

References

Asiara, M., Cookson, R., and Griffin, S. (2014). Incorporating health inequality impacts into cost-effectiveness analysis, in A.J. Culyer (ed.), *Encyclopedia of health economics*, pp. 22–6. Amsterdam: Elsevier.

Boadway, R.W. and Bruce, N. (1984). *Welfare economics*. Oxford: Blackwell.

Borenstein, M., Hedges, L.V., Higgins, J.P.T., et al. (2009). *Introduction to meta-analysis*. Chichester: Wiley.

Broome, J. (1978). Trying to value a life. *Journal of Public Economics*, **9**, 91–100.

Brouwer, W.B.F., Culyer, A.J., van Exel, N.J.A., et al. (2008). Welfarism vs extra-welfarism. *Journal of Health Economics*, **27**, 325–38.

Claxton, K. and Culyer, A.J. (2006). Wickedness or folly? The ethics of NICE's decisions. *Journal of Medical Ethics*, **32**, 373–7.

Coast, J., Smith, R., and Lorgelly, P. (2008). Welfarism, extra-welfarism and capability: the spread of ideas in health economics. *Social Science and Medicine*, **67**, 1190–8.

Cookson, R., Griffin, S., and Nord, E. (2014). Incorporation of concerns for fairness in economic evaluation of health programs: overview, in A.J. Culyer (ed.), *Encyclopedia of health economics*, pp. 27–34. Amsterdam: Elsevier.

Culyer, A.J. (1989). The normative economics of health care finance and provision. *Oxford Review of Economic Policy*, **5**, 34–58.

Culyer, A.J. (2012a). Deliberative processes and evidence-informed decision-making in health care—do they work and how might we know?, in R. Cookson and K. Claxton (ed.), *The humble economist. Tony Culyer on health, health care and social decision making*. London: Office of Health Economics / York: University of York.

Culyer, A.J. (2012b). Commodities, characteristics of commodities, characteristics of people, utilities and the quality of life, in R. Cookson and K. Claxton (ed.), *The humble economist. Tony Culyer on health, health care and social decision making*. London: Office of Health Economics / York: University of York.

Culyer, A.J. (2014). *The dictionary of health economics*, 3rd edition. Cheltenham: Edward Elgar.

Deaton, A. (2002). Policy implications of the gradient of health and wealth. *Health Affairs*, **21**, 13–30.

Dolan, P. (1999). Whose preferences count? *Medical Decision Making*, **19**, 482–86.

Guyatt, G. and Rennie, D. (2002). *Users' guides to the medical literature: a manual for evidence-based clinical practice*. Chicago: American Medical Association.

Hurley, J. (2000). An overview of the normative economics of the health sector, in A.J. Culyer and J.P. Newhouse (ed.), *Handbook of health economics*. Amsterdam: Elsevier.

Johnson, E.M. (2014). Physician-induced demand, in A.J. Culyer (ed.), *Encyclopedia of health economics*, pp. 77–82. Amsterdam: Elsevier.

Mishan, E. (1971). *Cost-benefit analysis*. London: Allen & Unwin.

Mishan, E.J. (1985). Consistency in the valuation of life: a wild goose chase?, in E.F. Paul (ed.), *Ethics and economics.* Oxford: Basil Blackwell.

Musgrave, R.A. (1959). *The economics of public finance*. New York: McGraw Hill.

Nyman, J.A. (2014). Demand for and welfare implications of health insurance, in A.J. Culyer (ed.), *Encyclopedia of health economics*, pp. 159–66. Amsterdam: Elsevier.

Paulden, M. and Claxton, K. (2012). Budget allocation and the revealed social rate of time preference for health. *Health Economics*, **21**, 612–18.

Pauly, M.V. (1995). Valuing health benefits in monetary terms, in F.A. Sloan (ed.), *Valuing health care. Costs, benefits and effectiveness of pharmaceuticals and other medical technologies*. Cambridge: Cambridge University Press.

Pita Barros, P. and Olivella, P. (2014). Markets in health care, in A.J. Culyer (ed.), *Encyclopedia of health economics*, pp. 201–20. Amsterdam: Elsevier.

Rawls, J. (1971). *A theory of justice*. Cambridge, MA: Harvard University Press.

Rice, T. (2014). Moral hazard, in A.J. Culyer (ed.), *Encyclopedia of health economics*, pp. 334–40. Amsterdam: Elsevier.

Russell, B. (1959). *My philosophical development*. London: Routledge.

Sen, A. (1979). Personal utilities and public judgments. Or what's wrong with welfare economics. *The Economic Journal*, **89**, 537–58.

Sen, A. (2002). Why health equity? *Health Economics*, **11**, 659–66.

Sugden, R. and Williams, A.H. (1979). *The principles of practical cost-benefit analysis*. Oxford: Oxford University Press.

Tobin, J. (1970). On limiting the domain of inequality. *Journal of Law and Economics*, **13**, 263–77.

Tsuchiya, A. and Williams, A. (2001). Welfare economics and economic evaluation, in M. Drummond and A. McGuire (ed.), *Economic evaluation in health care: merging theory with practice*. Oxford: Oxford University Press.

Williams, A. (1993). Priorities and research strategy in health economics for the 1990s. *Health Economics*, **2**, 295–302.

Chapter 3

Critical assessment of economic evaluation

3.1 Some basics

Those who receive or read an economic evaluation are often faced with the difficult task of assessing study results. The question that readers of evaluations are most likely to ask themselves is 'Are these results useful to me in my decision-making context?' The answer to this question is determined by the answers to the following specific questions:

1 Are the methods employed in the study appropriate and are the results valid?

2 If the results are valid, would they apply to my decision-making context?

This chapter concentrates on question 1, and is designed to assist users of economic evaluation in assessing the validity of the results they encounter.

When assessing the validity of evidence, we normally proceed by examining closely the methods employed to produce the evidence. Often it is helpful to separate the various elements of methods so that each can be scrutinized more closely. In this chapter we identify the key elements of any economic evaluation and discuss methods characteristics that users may expect to find in well-executed studies. A brief summary of relevant questions to ask about an economic evaluation is provided in Box 3.1, and this critical appraisal checklist is then applied to two published articles.

Of course, it is unrealistic to expect every study to satisfy all of the points; however, the systematic application of these points will allow readers to identify and assess the strengths and weaknesses of individual studies.

3.2 Elements of a sound economic evaluation

3.2.1 Was a well-defined question posed in answerable form?

Such a question will clearly identify the alternatives being compared and the perspective(s) from which the comparison is to be made. Questions such as, 'Is a chronic home care programme worth it?' and 'Will a community hypertension screening programme do any good?' solicit the issues of to whom and compared to what. Similarly, questions such as, 'How much does it cost to run our intensive care unit?', and 'What are the costs and outcomes of adolescent counselling by social workers?' are not questions for economic evaluation because they fail to specify the alternatives for comparison. This is not to say that the questions do not provide important accounting or management information; they may do so, but the answers to them do not by themselves inform resource allocation decisions.

Box 3.1 A checklist for assessing economic evaluations

1 Was a well-defined question posed in answerable form?

- 1.1 Did the study examine both costs and effects of the service(s) or programme(s) over an appropriate time horizon?
- 1.2 Did the study involve a comparison of alternatives?
- 1.3 Was a perspective for the analysis stated and was the study placed in any particular decision-making context?
- 1.4 Were the patient population and any relevant subgroups adequately defined?

2 Was a comprehensive description of the competing alternatives given? (i.e. can you tell who did what to whom, where, and how often?)

- 2.1 Were any relevant alternatives omitted?
- 2.2 Was (should) a 'do nothing' alternative (be) considered?
- 2.3 Were relevant alternatives identified for the patient subgroups?

3 Was the effectiveness of the programmes or services established?

- 3.1 Was this done through a randomized controlled clinical trial? If so, did the trial protocol reflect what would happen in regular practice?
- 3.2 Were effectiveness data collected and summarized through a systematic overview of clinical studies? If so, were the search strategy and rules for inclusion or exclusion outlined?
- 3.3 Were observational data or assumptions used to establish effectiveness? If so, were any potential biases recognized?

4 Were all the important and relevant costs and consequences for each alternative identified?

- 4.1 Was the range wide enough for the research question at hand?
- 4.2 Did it cover all relevant perspectives? (Possible perspectives include those of patients and third-party payers; other perspectivess may also be relevant depending on the particular analysis.)
- 4.3 Were capital costs, as well as operating costs, included?

5 Were costs and consequences measured accurately in appropriate physical units prior to valuation (e.g. hours of nursing time, number of physician visits, lost work-days, gained life-years)?

- 5.1 Were the sources of resource utilization described and justified?
- 5.2 Were any of the identified items omitted from measurement? If so, does this mean that they carried no weight in the subsequent analysis?
- 5.3 Were there any special circumstances (e.g. joint use of resources) that made measurement difficult? Were these circumstances handled appropriately?

Box 3.1 A checklist for assessing economic evaluations (continued)

6 Were costs and consequences valued credibly?

6.1 Were the sources of all values clearly identified? (Possible sources include market values, patient or client preferences and views, policy-makers' views, and health professionals' judgements.)

6.2 Were market values employed for changes involving resources gained or depleted?

6.3 Where market values were absent (e.g. volunteer labour), or market values did not reflect actual values (e.g. clinic space donated at a reduced rate), were adjustments made to approximate market values?

6.4 Was the valuation of consequences appropriate for the question posed (i.e. has the appropriate type or types of analysis—cost-effectiveness, cost–benefit—been selected)?

7 Were costs and consequences adjusted for differential timing?

7.1 Were costs and consequences that occur in the future 'discounted' to their present values?

7.2 Was any justification given for the discount rate(s) used?

8 Was an incremental analysis of costs and consequences of alternatives performed?

8.1 Were the additional (incremental) costs generated by one alternative over another compared to the additional effects, benefits, or utilities generated?

9 Was uncertainty in the estimates of costs and consequences adequately characterized?

9.1 If patient-level data on costs or consequences were available, were appropriate statistical analyses performed?

9.2 If a sensitivity analysis was employed, was justification provided for the form(s) of sensitivity analysis employed and the ranges or distributions of values (for key study parameters)?

9.3 Were the conclusions of the study sensitive to the uncertainty in the results, as quantified by the statistical and/or sensitivity analysis?

9.4 Was heterogeneity in the patient population recognized, for example by presenting study results for relevant subgroups?

10 Did the presentation and discussion of study results include all issues of concern to users?

10.1 Were the conclusions of the analysis based on some overall index or ratio of costs to consequences (e.g. cost-effectiveness ratio)? If so, was the index interpreted intelligently or in a mechanistic fashion?

Box 3.1 A checklist for assessing economic evaluations (continued)

10.2 Were the results compared with those of others who have investigated the same question? If so, were allowances made for potential differences in study methodology?

10.3 Did the study discuss the generalizability of the results to other settings and patient/client groups?

10.4 Did the study allude to, or take account of, other important factors in the choice or decision under consideration (e.g. distribution of costs and consequences, or relevant ethical issues)?

10.5 Did the study discuss issues of implementation, such as the feasibility of adopting the 'preferred' programme given existing financial or other constraints, and whether any freed resources could be redeployed to other worthwhile programmes?

10.6 Were the implications of uncertainty for decision-making, including the need for future research, explored?

A well-specified question, for example, might look as follows:

> From the perspective of (a) both the Ministry of Health and the Ministry of Community and Social Services budgets, and (b) patients incurring out-of-pocket costs, is a chronic home care programme preferable to the existing programme of institutionalized, extended care in designated wards of general hospitals?

Note that the perspective for an analysis may be that of a specific provider or providing institution, the patient or groups of patients, a third-party payer (public or private), or a broad perspective (i.e. all costs and consequences to whomsoever they accrue). Often the perspective may be specified by a decision-maker when requesting a study.

For many treatments and programmes, the capacity to benefit will differ for patients with differing characteristics. This could, for example, be because different types of patients have different baseline risks, or because the treatment effect itself systematically varies between different types of patient. Therefore, apart from considering the main treatment alternatives, it is important to consider relevant patient subgroups and to present data on costs and consequences for each. For example, Mark et al. (1995) presented data on the incremental cost-effectiveness of tissue plasminogen activator (t-PA) compared with streptokinase (SK) for patients of different age groups and type of myocardial infarction.

Obviously it is difficult for the analyst to consider every single cost and consequence of a health care programme to all members of society. Indeed, the 'ripple effects' of some programmes may be far reaching and consideration of some items may have to be excluded for practical reasons. However, it is important to recognize that in considering the use of the community's scarce resources, the perspective of the providing institution may not include effects that are considered to be important and a broader perspective should also be considered, including the costs and consequences falling on other public agencies, patients, and their families. For example, it may be that a programme is

preferable from a broader perspective, but not from the perspective of the providing institution. In such a case the Ministry of Health may wish to consider giving an incentive to the providing institution to ensure that the programme goes ahead.

The existence of different perspectives was highlighted by Byford et al. (2003) in their study of treatments for recurrent deliberate self-harm. Costs falling on the following sectors were considered: hospital services, social services, voluntary sector services, community accommodation, and the criminal justice system. Costs resulting from lost productivity, due to time off work, were also estimated. The relative costs of the two treatments depended on the perspective adopted. From a health care perspective, the relative annual costs per patient of the two programmes were fairly similar: £2395 for a new intervention, manual-assisted cognitive behaviour therapy, and £2502 for treatment as usual. However, when a broader perspective was adopted, including all costs, the annual cost difference per patient was £838 higher for treatment as usual. (The challenges of adopting a broader perspective are considered further in Chapter 4.)

3.2.2 Was a comprehensive description of the competing alternatives given?

A full description of the competing alternatives is essential for three further reasons:

- Readers must be able to judge the applicability of the programmes to their own settings.
- Readers should be able to assess for themselves whether any costs or consequences may have been omitted in the analysis.
- Readers may wish to replicate the programme or procedures being described.

Therefore, readers should be provided with information allowing an identification of costs (who does what to whom, where, and how often?) and consequences (what are the results?).

Other relevant alternatives, within a given treatment or programme, could include intensity (i.e. dosage) and length of treatment. For example, Sculpher et al. (2000) compared low and high doses of an angiotensin-converting enzyme inhibitor in patients with chronic heart failure. Finally, the alternatives could consist of various sequences of care or clinical care pathways. In such cases, the main issue may not be whether or not a particular treatment should be used, but to determine its appropriate position in the treatment sequence or strategy. Sometimes the main issue may be to determine the circumstances under which a given treatment should be discontinued, perhaps because of lack of response. (This issue is discussed in more detail in Chapter 9 on modelling, since many of the decision-analytic models used in economic evaluation involve comparing alternative treatment strategies, as opposed to individual treatments or technologies.)

An important contribution of economic analysis is to minimize the risk of important relevant alternatives being excluded from consideration. Therefore the analyst should consider (1) whether a 'do nothing' alternative is a feasible option and (2) whether the formulation of the decision problem is being unduly constrained by the existing effectiveness evidence base or by the professional or managerial responsibilities of the person commissioning the study. In principle, all relevant alternatives should be considered, as comparing with an inappropriate alternative can be misleading, especially

if the new therapy is compared with one which itself is not cost-effective. If it is only possible to consider one alternative, this should be the treatment strategy that best represents current practice in the jurisdiction concerned.

3.2.3 Was the effectiveness of the programmes or services established?

An important component of an economic evaluation is the assessment of the effectiveness of the alternative interventions. Therefore, some indication of the validation of effectiveness should be given. Thus, economic evaluations using decision-analytic modelling typically take estimates of treatment effect from previously conducted randomized clinical trials and in such cases the source(s) of the effectiveness data should be documented. It is also possible that the economic evaluation may have been conducted simultaneously with the evaluation of efficacy or effectiveness. In such cases it will be necessary either to describe the methods of the clinical study concerned, or to provide a reference. Note that economic evaluations, by themselves, are incapable of establishing effectiveness. Precedent or simultaneous evidence of effectiveness is required. Those wishing to know more about the methods of establishing whether a therapy does more good than harm should consult Guyatt et al. (2008).

Evidence on effectiveness may come from a single study, especially when the economic evaluation is carried out alongside a clinical study. Alternatively, it can come from a systematic overview of several clinical studies. In the former case it is important to consider whether the estimate of treatment effect from that particular trial is representative of the whole body of evidence for the treatments concerned (see Chapter 8). In the latter case it is important that the reasons for inclusion or exclusion of studies from the overview are given, so that the reader can assess whether or not a biased subset of the available clinical evidence has been used. For more details of the methodology of systematic reviews see Centre for Reviews and Dissemination (2009) and for guidance on how they should be reported see Moher et al. (2009). (The synthesis of data for use in economic evaluation is discussed in Chapter 10.)

3.2.4 Were all the important and relevant costs and consequences for each alternative identified?

Even though it may not be possible or necessary to measure and value all of the costs and consequences of the alternatives under comparison, a full identification of the important and relevant ones should be provided. The combination of information contained in the statement of perspective and the programme description should allow judgement of what specific costs and consequences or outcomes it is appropriate to include in the analysis. For example, if the perspective being adopted was that of the health care system, costs falling on patients or their carers would be excluded. Also, costs or consequences that are identical for the options under consideration could be excluded, since they would not enter into the calculation of the *difference* in costs or consequences; for example, the costs of establishing a diagnosis of the condition concerned. An overview of the categories of costs and consequences that are potentially relevant to an economic evaluation of health services and programmes is given in Figure 3.1. Four categories of

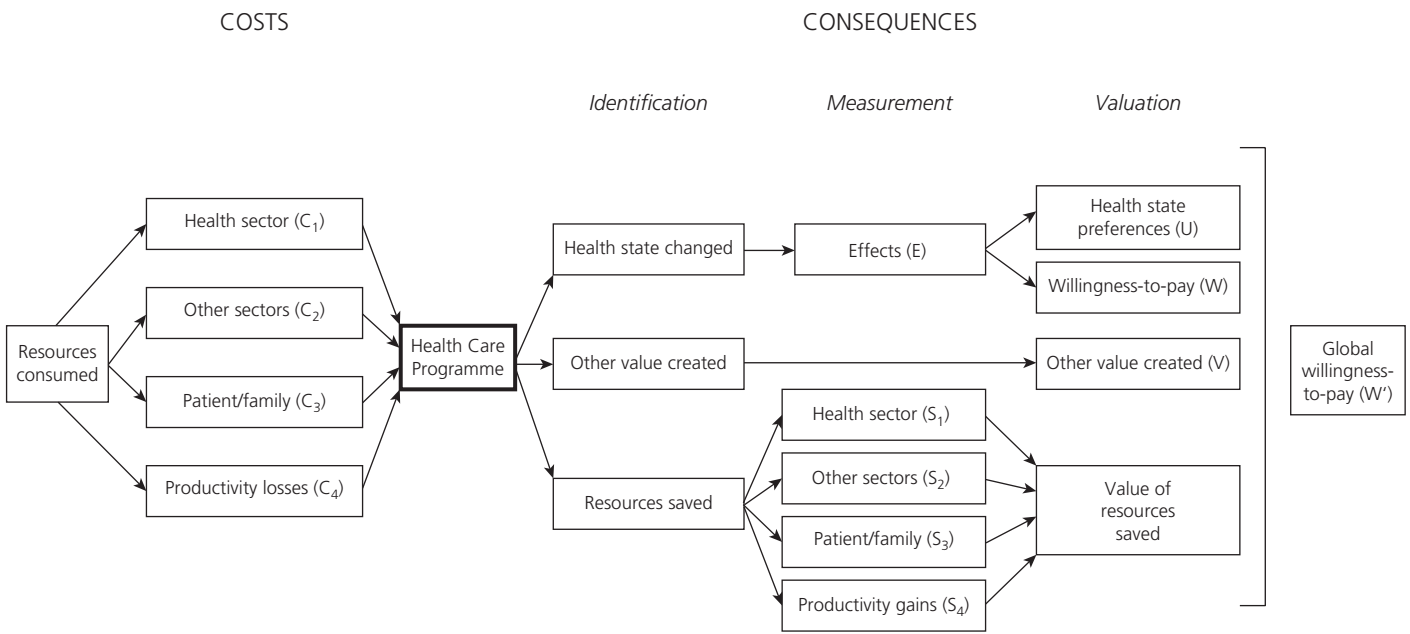


Fig. 3.1 Components of economic evaluation in health care.

cost are identified. The health care resources consumed consist of the costs of organizing and operating the programme, including dealing with the adverse events caused by the programme. The identification of these costs often amounts to listing the *ingredients* of the programme—both variable costs (such as the time of health professionals or supplies) and fixed or overhead costs (such as light, heat, rent, or capital costs). The ways of measuring and valuing these items are discussed in more detail in Chapter 7.

The patient and family resources consumed include any out-of-pocket expenses incurred by patients or family members as well as the value of any resources that they contribute to the treatment process. When patients or family members lose time from work while seeking treatment or participating in a health care programme there could be associated productivity losses. (The measurement and valuation of these is also discussed in Chapter 7.)

A fourth category, resources consumed in other sectors, also warrants mention. Some programmes, such as those for the care of the elderly, consume resources from other public agencies or the voluntary sector. Occasionally it may also be the case that the operation of a health service or programme changes the resource use in the broader economy. Examples of situations where these factors may be important are the following:

- An occupational health and safety programme (perhaps legislated by government) that changes the production process in an automobile manufacturing plant, thereby using up more resources, perhaps in a more labour-intensive way. These costs are passed on in the increased price of cars and are borne by the purchasers of cars, who are likely not the workers for whom the programme was initiated.
- A road speed limit policy for trucks, which reduces morbidity and mortality due to accidents, but increases the price of, for example, fruit which now takes longer to arrive (i.e. a higher wage bill for the truck driver).

In principle, these factors should be considered in an economic evaluation adopting a broad perspective, especially in the evaluation of public health interventions (Weatherly et al. 2009). However, for many health care programmes, differences between the options being compared may be small and in practice few economic analyses take them into account, especially as most of these analyses are undertaken from the perspective of the health care system, or the payer.

Three categories of consequences of health services or programmes are also shown in Figure 3.1. The changes in health state relate to changes in the physical, social, or emotional functioning of individuals. In principle, such changes can be measured objectively, and refer only to an individual's ability to function and not to the significance, preference, or value attached to this ability by the individual, or by others. However, as indicated in Figure 3.1 and discussed in Chapters 5 and 6, values can be attached either by health state preference scores or by estimating individuals' willingness to pay.

In addition, other value may be created by the programme (e.g. increased convenience to patients) and resources may be freed. For example, a vaccination programme may free resources if fewer individuals contract the disease and thus require treatment.

As mentioned in Chapter 2, many of the health effects can be captured in a generic measure of health outcome. But given the wide range of costs and consequences indicated above, it may be unrealistic to expect all relevant items to be measured and

valued in the analysis, due to their small size or influence relative to the effort required to measure or value them accurately; however, it is helpful to users for them to be identified. It is particularly important that the outcomes of interest be identified clearly enough for a reader to judge the appropriateness of the type (or types) of economic evaluation chosen; that is, it should be apparent:

- whether a single outcome is of primary interest as opposed to a set of outcomes
- whether the outcomes are common to both alternatives under comparison
- to what degree each programme is successful in achieving each outcome of interest.

Similarly, it is important to know whether the consequences of primary interest are the therapeutic effects themselves, the change in the health-related quality of life of patients and their families, or the overall value created. Primarily this is determined by the audience(s) for the study and their objectives.

3.2.5 Were costs and consequences measured accurately in appropriate physical units prior to valuation?

While identification, measurement, and valuation often occur simultaneously in analyses, it is a good practice for users of evaluation results to view each as a separate phase of analysis. Once the important and relevant costs and consequences have been identified, they must be measured in appropriate physical and natural units. For example, measurement of the operating costs of a particular screening programme may yield a partial list of 'ingredients' such as 500 physical examinations performed by physicians, 10 weeks of salaried nursing time, 10 weeks of a 100-square-metre clinic, 20 hours of medical research librarian time from an adjoining hospital, and so on. Similarly, costs borne by patients may be measured, for instance, by the amount of medication purchased, the number of times travel was required for treatment, or the time lost from work while being treated.

Notice that situations in which resources are jointly used by one or more programmes present a particular challenge to accurate measurement. How much resource use should be allocated to each programme? And on what basis? A common example of this is found in every hospital, where numerous clinical services and programmes share common overhead services provided centrally (e.g. electric power, cleaning, and administration). In general, there is no non-arbitrary solution to this measurement problem; however, users of results should satisfy themselves that *reasonable* criteria (square metres of floor space, number of employees, number of cases, and so on) have been used to distribute the common costs. Users should definitely ascertain that such shared costs have in fact been allocated to participating services or programmes, as this is a common omission in evaluations! Clinical service directors often argue that small changes in the size of their programmes (up or down) do not affect the consumption of central services. Sometimes it is even argued that overhead costs are unaffected by the service itself. However, though this argument may be intuitively appealing from the viewpoint of a particular programme or service director, the extension of this method to each service in the hospital would imply that the totality of services could be operated without light, heat, power, and secretaries! (The allocation of overhead costs is discussed further in Chapter 7.)

With respect to the measurement of consequences, if the identification of outcomes of interest has been clearly performed, then selection of appropriate units of measurement for programme effects should be relatively straightforward. For example, effects might relate to mortality and be measured in life-years gained or deaths averted; they might relate to morbidity and be measured in reductions in disability days or improvements on some index of health status measuring physical, social, or emotional functioning; they may be even more specific, depending on the alternatives under consideration. Thus, percentage increase in weight-bearing ability may be an appropriate natural measurement unit for an evaluation of a physiotherapy programme, while the number of correctly diagnosed cases may be appropriate for a comparison of venography with leg scanning in the diagnosis of deep vein thrombosis.

In some cases the measurement of consequences will be based directly on the clinical evidence used in the economic evaluation. For example, the evaluation may use data from clinical trials estimating the number of cases detected or life-years gained. However, it is much more common for trials to estimate the number of patients surviving at the end of a follow-up period (e.g. 1 year), or the progression of disease over a fixed period of time. The choice of time horizon is an important methodological consideration in economic evaluations. It should be long enough to capture the major health and economic consequences, both intended effects and unintended side effects. Thus, for the majority of economic evaluations the relevant time horizon is the patient's lifetime.

The measurement of consequences for the economic evaluation may therefore require extrapolation of effectiveness over time. This is particularly true where the alternative interventions have mortality effects, where it is necessary to measure the consequences in terms of life-years or quality-adjusted life-years (QALYs) gained. Extrapolation of effects beyond the end of the trial requires additional data (from long-term observational studies) and may also require several assumptions, such as the likely progression of disease in patients who discontinue therapy. There may be no unambiguously right way to make such extrapolations, but at least the methods used should be transparent and justified, with the uncertainty in the estimates characterized (see Chapter 8). Extrapolation is a central feature of economic evaluations using a decision-analytic modelling approach and is discussed in more detail in Chapter 9.

Changes in quality of life are usually measured using some type of scale. The scales can either be disease (or condition)-specific or generic, covering a broad range of dimensions of health-related quality of life. Most scales have several dimensions along which measurements of changes can be made. In addition, some of the scales have an associated algorithm that can be used to generate an overall score, or measurement of the change in health-related quality of life. The use of quality of life scales is discussed in Chapter 5.

3.2.6 Were costs and consequences valued credibly?

The sources and methods of valuation of costs and consequences should be clearly stated in an economic evaluation. Costs are normally valued in units of local currency, based on prevailing *prices* of, for example, personnel, commodities, and services, and can often be taken directly from programme budgets. All current and future programme costs are normally valued in constant dollars of some base year (usually the present), in order to remove the effects of inflation from the analysis.

It should be remembered that the objective in valuing costs is to obtain an estimate of the worth of resources depleted by the programme. This may necessitate adjustments to some apparent programme costs (e.g. the case of subsidized services or volunteer labour received by one programme instead of another). In addition, valuation of the cost of a day of institutional care for a specific condition is particularly troublesome in that the use of an average cost per day (the widely quoted *per diem*), calculated on the basis of the institution's entire annual caseload, is almost certainly an overestimate or underestimate of the actual cost for any specific condition, sometimes by quite a large amount.

In principle and (with great effort) in practice, it is possible to identify, measure, and value each depleted resource (e.g. drugs, nursing time, light, food, and so on) in treating a specific patient or group of patients. While this may yield a relatively accurate cost estimate, the detailed monitoring and data collection are usually prohibitively expensive. The other broad alternative costing strategy is to start with the institution's total costs for a particular period and then to improve upon the method of simply dividing by the total patient-days to produce an average cost per day. Quite sophisticated methods of cost allocation to individual hospital departments or wards have been developed. An intermediate method involves acceptance of the components of the general *per diem* relating to *hotel* costs (as these are relatively invariant across patients) combined with more precise calculation of the medical treatment costs associated with the specific patients in question. Of course, the effort devoted to accurate *per diem* estimates depends upon their overall importance in the study; however, unthinking use of *per diems* or average costs should be guarded against. (This is discussed further in Chapter 7.)

Usually, the costing methods employed in a given study are influenced by the local availability of financial data. For example, many countries now have available casemix-related costs for episodes of care in hospital; for example, costs by diagnosis-related groups (DRGs). These can be reasonable approximations for the costs of treating different categories of patient.

The choice of study perspective can also affect the costing method. For example, if a payer perspective is adopted, the most relevant cost estimates are the amounts actually paid from the payer's budget. On the other hand, if the aim is to adopt a broad 'societal' perspective, when following a 'welfarist' approach to economic evaluation, not only will a broad range of costs be included, but the estimates based on market values should be adjusted to reflect any market distortions. This is almost never done in practice (see Chapter 2).

In the estimation of health state preference values, we are basically attempting to ascertain how much better the quality of life is in one health situation or 'state' compared with another (e.g. dialysis at home with help from a spouse or friend versus dialysis in hospital). Several techniques are available for making the comparison; the important thing to note is that each will produce an adjustment factor with which to increase or decrease the value of time spent in health situations or 'states', resulting from the alternative in question relative to some baseline. The results of these analyses are usually expressed in *healthy years* or *QALYs* gained, as a result of the programmes being evaluated.

Many economic evaluations incorporate one of the generic preference-based health measures, such as the EuroQoL EQ-5D (EuroQoL Group 1990), the SF-6D (Brazier et al. 2002), or the Health Utilities Index (Feeny et al. 1995). These measures employ a

questionnaire administered to patients in the study, to classify them into one of a predetermined set of health states. The health state preference values, or utilities, are then available from a scoring formula (or tariff) that accompanies the measure. Typically the source of values is the general public.

There are many unresolved issues in the measurement of preferences in health, which readers of economic evaluations should note. Users of such analyses will probably want to know, at minimum, *whose* preferences were used to construct the adjustment factor—the patient's, the provider's, the general public's, or the decision-maker's? If patients' preferences have not been employed, we may want to assure ourselves further that the persons whose preferences did count clearly understood the characteristics of the health state, either through personal experience or through a description of the state presented to them. These issues are taken up in Chapter 5.

There are now several approaches for valuing the various attributes of health care programmes, either relative to one another or in money terms. Many of the same issues arise when estimating willingness to pay, either for a change in health state or for the overall impact of the programme in question. These issues are taken up in Chapter 6.

One of the important consequences of health care programmes is the creation of healthy time. The valuation of this item poses difficulties. Indeed, its categorization, under changes in health state or patient and family resources freed, is uncertain. The reasons for this are as follows:

- The value of healthy time can manifest itself in a number of ways. First, living in a better health state has a value to the individual in its own right (e.g. less pain, better health-related quality of life). Secondly, healthy time can be used in leisure. Thirdly, healthy time can be used for work, which generates income for the individual and productive output for society.
- In some economic evaluations, where the measurement of the effects (E) is purely in clinical terms such as 'disability days avoided', this does not capture the value of healthy time. Therefore, if it is to be included it would have to be estimated separately, as an element of the patient and family resources freed (see Figure 3.1). In an economic evaluation where we are attempting to *value* the consequences of the programme, we might expect that the value of living in a better health state is captured in the health state preference value (U), or the willingness to pay (W). The value of healthy time in leisure is probably also captured in U or W, as it is closely linked with improved health-related quality of life.
- Whether or not the value of using healthy time for work is also included in U or W probably depends on how the scenario (used for valuation) is written. It may be possible, for example, to ask individuals to imagine that their income would not be affected by their health state, as this is covered by unemployment insurance. In such a case, an analyst undertaking an evaluation from a broad perspective may wish to include a separate estimate of productivity gains to society if a person's health state is improved. If so, it would be important to ensure that the person did *not* include this in their own valuation, so as to avoid double counting.

Of course, healthy time can also be consumed by a programme if it requires the individual to spend time seeking or undergoing treatment, perhaps in hospital. Therefore,

in an economic evaluation undertaken from a broad perspective, the value of healthy time lost would have to be estimated separately.

In an economic evaluation attempting to value the benefits of the programme, the way forward would again depend on how the health state scenario is described. If the description contained elements of the process of undergoing care (e.g. 'you will be admitted to hospital for 7 days for treatment'), this may be reflected in the value of *U* or *W*. Otherwise, the value of healthy time lost in therapy may have to be estimated separately.

Therefore, the assembly of components (building blocks) in the analysis is partly dependent on how they are measured and valued. In conducting an economic evaluation from a broad perspective, it is important to avoid both zero counting and double counting. Of course, the estimation of global willingness to pay (W') avoids the problem of categorizing the value of healthy time, either as a component of the change in health state or as a component of patient and family resources freed. Rather, the challenge of this approach is to ensure that individuals appreciate *all* the elements of value created and resources consumed by a health care programme, so that this is reflected in their valuation (W').

Finally, as mentioned in Chapter 2, it should be remembered that the relevance of the various elements of value will depend on the approach being adopted by the decision-maker. Those decision-makers adopting an extra-welfarist approach will prefer to value only the improvements in health.

3.2.7 Were costs and consequences adjusted for differential timing?

Because comparison of programmes or services must be made at one point in time (usually the present), the timing of programme costs and consequences that do not occur entirely in the present must be taken into account. Different programmes may have different time profiles of costs and consequences. For example, the primary benefits of an influenza immunization programme are immediate while those of hypertension screening occur well into the future. The time profile of costs and consequences may also differ within a single programme; for example, although the benefits of a hypertension screening programme will occur mostly in the future, the costs are incurred in the present.

Therefore, future cost and benefit streams are reduced or 'discounted' to reflect the fact that the amounts spent or saved in the future should not weigh as heavily in programme decisions as those spent or saved today. This is primarily due to the existence of *time preference*. That is, individually and as a society we prefer to have money or resources now, as opposed to later, because we can benefit from them in the interim. This is evidenced by the existence of interest rates (as well as the popular wisdom about 'a bird in the hand'). This means that health care resources committed today could be invested with a positive rate of return, generating yet more resources to secure more health in the future. Moreover, because time preference is not exclusively a financial concept, discounting of consequences should also be considered in economic evaluations. The concept of discounting and the determination of the discount rate are discussed in Chapter 4. The mechanics of discounting are discussed in Chapter 7.

3.2.8 Was an incremental analysis of costs and consequences of alternatives performed?

For meaningful comparison, it is necessary to examine the additional costs that one service or programme imposes over another, compared with the additional effects, benefits, or utilities it delivers. This *incremental* approach to analysis of costs and consequences can be illustrated by reference to adjuvant chemotherapy for early breast cancer (Campbell et al. 2011). The risk of recurrence following surgery is reduced by chemotherapy, but increasingly effective regimens are associated with higher costs and toxicity profiles. To investigate this, the analysts undertook a cost-effectiveness study of four treatment strategies: (1) no chemotherapy; (2) chemotherapy using cyclophosphamide, methotrexate, and fluorouracil (CMF) (a first-generation regimen); (3) chemotherapy using epirubicin-CMF (E-CMF) or fluorouracil, epirubicin, and cyclophosphamide (FEC60) (second-generation regimens); and (4) chemotherapy with FEC60 followed by doxetaxel (FEC-D) (a third-generation regimen).

Table 3.1 shows the costs and consequences (in terms of QALYs) of the four treatment strategies. Although one might be tempted to compare the simple ratios of costs to outcomes for the four alternatives, the correct comparison is the one of *incremental* costs over *incremental* outcomes, because this tells us the extra amount we are paying to gain an extra QALY by moving to a more effective chemotherapy regimen. For example, using the figures given in Table 3.1, to switch the treatment regimen from no chemotherapy to E-CMF chemotherapy would result in an extra cost of £1042 (15246 – 14204) per patient and 1.73 (12.66 – 10.93) extra QALYs, an incremental cost-effectiveness ratio (ICER) of £602 per QALY (£1042/1.73). Alternatively, if the current regimen were E-CMF, to switch from this to FEC-D would have an ICER of £14005 (i.e. (18327 – 15246)/(12.88 – 12.66)).

Table 3.1 Economic evaluation of alternative treatment strategies for early breast cancer: expected discounted lifetime costs (2009 GBP) and QALYs for each of the four treatment strategies modelled based on the reference case cohort[a]

	No chemotherapy	CMF chemotherapy	E-CMF/FEC60 chemotherapy	FEC-D chemotherapy
Costs				
Chemotherapy[b]	—	£3113	£3691	£7111
Routine follow-up[c]	£1087	£1220	£1245	£1264
Recurrence	£13 116	£10 743	£10 310	£9 952
Total costs	£14 204	£15 076	£15 246	£18 327
Total QALYs	10.93	12.35	12.66	12.88

[a]Patient is age 40, with one positive node and a grade 2 ER-negative tumour 3 cm in diameter.

[b]Includes chemotherapy drug and toxicity costs.

[c]Costs incurred in disease-free health state.

Sometimes the difference between the incremental and average cost-effectiveness ratios can be quite dramatic. Earlier (in Chapter 1) we pointed out that, in the case of screening for cancer of the colon, there was a big difference between the average cost (per case detected) of a protocol of six sequential tests and the incremental cost of performing a sixth test, having already done five (Neuhauser and Lewicki 1975).

The same principle can be illustrated graphically on a four-quadrant diagram known as the *cost-effectiveness plane* (Black 1990) (see Box 3.2).

Box 3.2 The cost-effectiveness plane

In Figure 3.2, the horizontal axis represents the difference in effect between the intervention of interest (A) and the relevant alternative (O), and the vertical axis represents the difference in cost. The alternative (O) could be the status quo or a competing programme.

If point A is in quadrants II or IV the choice between the programmes is clear. In quadrant II the intervention of interest is both more effective and less costly than the alternative. That is, it *dominates* the alternative. In quadrant IV the opposite is true. In quadrants I and III the choice depends on the maximum cost-effectiveness ratio one is willing to accept. The slope of the line OA gives the cost-effectiveness ratio.

From Black, W.C., The cost-effectiveness plane: a graphic representation of cost-effectiveness, *Medical Decision Making*, Volume 10, Number 3, pp. 212–15, Copyright © 1990 by Society for Medical Decision Making. Reprinted by permission of SAGE Publications.

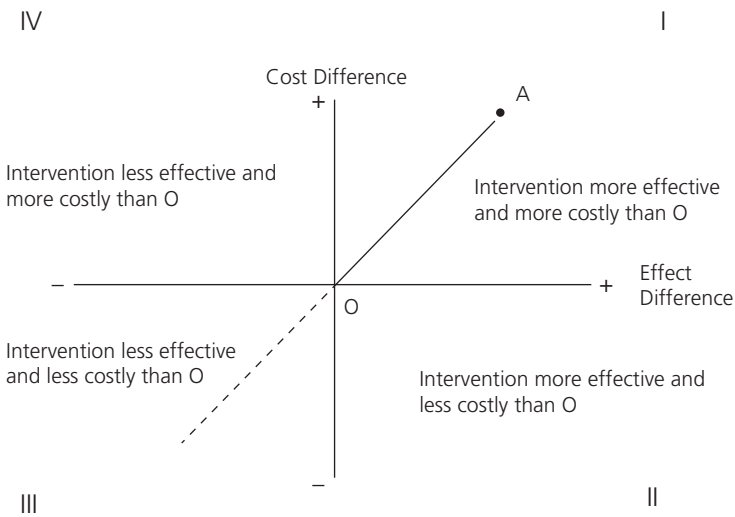


Fig. 3.2 The cost-effectiveness plane.

Adapted from Black, W.C., The cost-effectiveness plane: a graphic representation of cost-effectiveness, *Medical Decision Making*, Volume 10, Number 3, pp. 212–15, Copyright © 1990 by Society for Medical Decision Making. Reprinted by permission of SAGE publications.

In practice the impact of most interventions falls in quadrant I. That is, they add to cost but increase effectiveness, certainly when compared with no intervention. Let us therefore plot the data given in Table 3.1 for no chemotherapy (assuming that this is our current practice) and two of the active treatment regimens, E-CMF and FEC-D (see Figure 3.3). Here we only show quadrant I of the cost-effectiveness plane and have placed our current practice at the origin on the plane. The slopes of the lines from the origin give the cost-effectiveness ratios for the two active treatment regimens as compared with no chemotherapy, which are £602 and £2114 per QALY gained for E-CMF and FEC-D respectively. However, if the decision is whether to give the patient E-CMF, or the more expensive FEC-D, the relevant ratio to consider is the ICER between the two (£14 005 per QALY), the slope of the line joining the two points, since this tells us the cost of the extra QALYs we would 'buy' by opting for the more expensive therapy.

The interpretation of ICERs is discussed further in Chapter 4. Determining whether a given programme is 'cost-effective' in a given jurisdiction normally relies on reference to a local standard, or 'threshold' of the maximum acceptable level of cost-effectiveness. For example, in the United Kingdom, the National Institute for Health and Care Excellence (NICE) currently applies a threshold range of £20 000–30 000 per QALY when issuing guidance on the use of health technologies in the UK National Health Service (Rawlins and Culyer 2004). This is intended to reflect the opportunity cost of the resources, in terms of the health gains produced by the treatments or programmes that would be displaced if the new treatment was adopted.

In jurisdictions where there is no announced threshold, analysts often judge whether or not a given treatment is cost-effective by comparing its ICER with those of other, already funded, interventions in the jurisdiction concerned. This, of course, assumes that the decisions to fund these existing interventions were made appropriately. Also, the World Health Organization has suggested a threshold range, of between 1 and 3 times GDP per capita (Tan-Torres Edejer et al. 2003; WHO 2014).

Fig. 3.3 Breast cancer treatments on the cost-effectiveness plane.

3.2.9 Was uncertainty in the estimates of costs and consequences adequately characterized?

Every evaluation will contain some degree of uncertainty, imprecision, or methodological controversy. What if the compliance rate for influenza vaccination was 10% higher than considered for the analysis? What if the *per diem* hospital cost understated the true resource cost of a treatment programme by $100? What if a discount rate of 6% per annum was used instead of 3%? Or what if productivity changes had been excluded from the analysis? Users of economic evaluations will often ask these and similar questions; therefore, careful analysts will identify critical methodological assumptions or areas of uncertainty.

Briggs et al. (2012) distinguish among several different types or sources of uncertainty, relating to the data used in the analysis or the main methodological assumptions (see Table 3.2). The methods for handling uncertainty differ according to its source and the type of economic evaluation being performed. For example, in an economic evaluation conducted concurrently with a clinical trial, data (say) on length of hospital stay will be stochastic (i.e. have a mean and variance). Therefore in the analysis of *patient-level data* it is possible to conduct statistical analyses (see Chapter 8). In the case of studies employing decision-analytic modelling, data on key model parameters are drawn from a number of sources. Here the approach for dealing with parameter uncertainty is called *sensitivity analysis*, where the various parameters in the model are varied in order to assess how this impacts upon study results. Sensitivity analysis is also used to handle other types of uncertainty, such as that relating to methodological assumptions. (This is discussed further in Chapter 11.)

Sensitivity analysis is an important feature of economic evaluations and study results can be sensitive to the values taken by key parameters. In a review of economic evaluations, Schackman et al. (2004) found that quantitatively important changes in results

Table 3.2 Uncertainty: concepts and terminology

Preferred term	Concept	Other terms sometimes employed
Stochastic uncertainty	Random variability in outcomes between identical patients	Variability Monte Carlo error First-order uncertainty
Parameter uncertainty	The uncertainty in estimation of the parameter of interest	Second-order uncertainty
Heterogeneity	The variability between patients that can be attributed to characteristics of those patients	Variability Observed or explained heterogeneity
Structural uncertainty	The assumptions inherent in the decision model	Model uncertainty

were obtained in 31% of sensitivity analyses of health-related quality of life estimates, 20% of those of cost estimates, and 15% of those using different discount rates. In judging the quality of a sensitivity analysis conducted in an economic evaluation, readers should consider (1) how the uncertain parameters were identified, (2) how the plausible ranges for the parameters were specified, and (3) whether an appropriate form of sensitivity analysis was used.

1 *Identifying the uncertain parameters*. It is difficult to specify firm guidelines for this step, beyond the fact that, in principle, all parameters in the analysis are potential candidates for sensitivity analysis. One approach might be for the analyst to give the reasons why particular variables had *not* been included. Possible reasons for exclusion could be that parameter estimates are known with absolute certainty (e.g. the unit cost of a resource in a given location), or that they represent policy choices (e.g. the discount rate).

2 *Specifying the level of uncertainty*. A frequent weakness in published economic evaluations is that, while they include a sensitivity analysis, the reasons for specifying the plausible ranges for the variables are not given. Frequently estimates are doubled or halved with no justification. When judging published studies the user should assess the justification given for plausible ranges or distributions assigned, in conjunction with the statements authors make about their analyses. Sometimes the author's conclusion is that the result is very robust, although the ranges chosen for varying key estimates are unjustifiably small. The moral appears to be that if you do not shake your study too hard it is unlikely to fall apart!

3 *Deciding on the form of sensitivity analysis*. In the past, the most common form of sensitivity analysis was to undertake a *one-way analysis*. Here estimates for each parameter are varied one at a time in order to investigate the impact on study results. A common way to present the results of a one-way analysis is in a 'tornado diagram'. The impact that variation in each parameter has on the study result is shown by the width of the respective band. These diagrams are normally arranged so that the parameter in which variation has the biggest impact on the study result is at the top (see Chapter 11). Although one-way sensitivity analysis is one of the most common forms of sensitivity analysis in the literature, it is not now regarded as a comprehensive approach for handling parameter uncertainty, because the overall uncertainty in the cost-effectiveness ratio depends on the *combined* variability in several parameters. A variant of one-way sensitivity analysis is to undertake a *threshold analysis*. Here the critical value(s) of a parameter or parameters central to the decision are identified. For example, a decision-maker might specify an increase in cost, or an ICER, above which the programme would not be acceptable. Then the analyst could assess which combinations of parameter estimates could cause the threshold to be exceeded. Alternatively, the threshold values for key parameters that would cause the programme to be too costly or not cost-effective could be defined. The decision-makers could then make a judgement about whether particular thresholds were likely to be breached or not (see Box 3.3).

A more sophisticated approach is to undertake a *multiway analysis*. This recognizes that more than one parameter is uncertain and that each could vary within its specified

Box 3.3 Cost-effectiveness of hip prostheses: a two-way threshold analysis

The newer hip prostheses are more expensive than the existing, well-established ones. However, they may offer several advantages, one of which is a lower rate of revision (i.e. reoperation) due to failure of the prosthesis. Although new prostheses may require fewer revisions, this reduction is not known with certainty. Therefore, in their economic evaluation, Briggs et al. (1998) conducted a two-way threshold analysis, on price and revision rate. This is shown in Figure 3.4.

The interpretation is as follows. If, as a decision-maker, your requirement was neutrality in overall treatment costs, a prosthesis cost of 150% (of the cost of existing prostheses) would be justified if the reduction in the revision rate was round 35%. If, on the other hand, you also valued the increased benefits, in QALYs, that new prostheses may confer (in improved quality of life or reduced mortality from the reoperations), a prosthesis cost of around 230% (of existing prostheses) would be justified at a willingness-to-pay threshold of £10 000 per QALY.

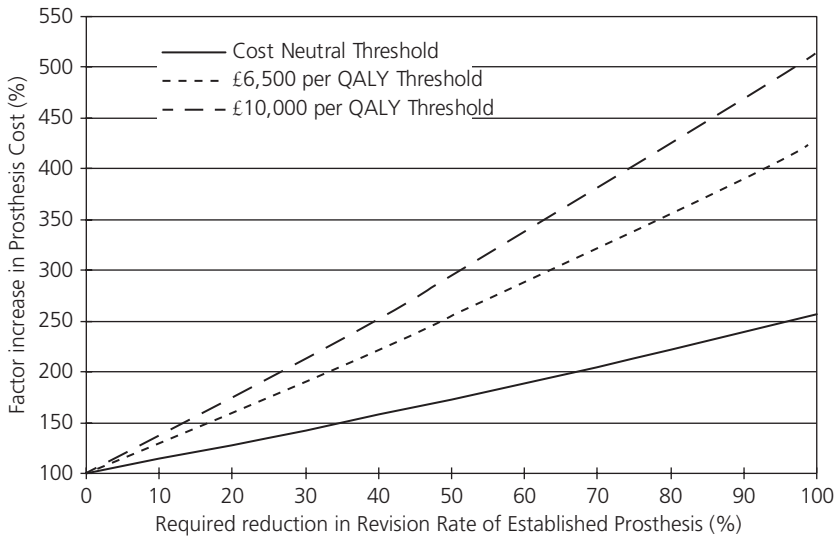


Fig. 3.4 Example of a threshold analysis.

From Andrew Briggs et al., The costs and benefits of primary total hip replacement: how likely are new prostheses to be cost-effective? *International Journal of Technology Assessment in Health Care*, Volume 14, Issue 4, pp. 743–61, Copyright © Cambridge University Press 1998, reproduced with permission.

range. Overall, this approach is more realistic but, unless there are only a few uncertain parameters, the number of potential combinations becomes very large.

Another approach is to use *scenario analysis*. Here a series of scenarios is constructed representing a subset of the potential multiway analyses. Typically, the scenarios will include a base case (best guess) scenario and the most optimistic (best case) and most

pessimistic (worst case) scenarios. Alternatively, they may include scenarios that the analyst or user of the study feel could probably apply. Scenario analysis is often used to explore the impact of structural assumptions in a decision-analytic model (see Chapter 9).

However, a limitation of multiway and threshold analyses is that they become impossible to undertake if there are more than a few parameters of interest. A final form of sensitivity analysis, *probabilistic sensitivity analysis*, is now becoming widely used in economic evaluations. Here probability distributions are applied to the specified ranges for the key parameters and samples drawn at random from these distributions to generate empirical distributions of the costs and consequences.

The main advantage of this approach is that it is possible to characterize the combined effect of all parameter uncertainty in the analysis and report on their implications for a decision based on mean costs and consequences. It also provides a foundation for assessing the value of additional information, the type of evidence that might be required and the implications of that need for evidence might have for the decision to adopt the new intervention or technology. (See Chapter 11 for a full discussion of uncertainty in economic evaluations.)

3.2.10 Did the presentation and discussion of study results include all issues of concern to users?

It will be clear from the foregoing discussion that the economic analyst has to make many methodological judgements when undertaking a study. Faced with users who may be mainly interested in the 'bottom line'—for example, 'should we buy a CT scanner?'—how should the analyst present the results? Decision indices such as cost-effectiveness and cost–benefit ratios are a useful way of summarizing study results. However, they should be used with care, as in interpreting them, the user may not be completely clear on what has gone into their construction. For example, does the estimate of 'cost' include losses in productivity or not?

Another point raised of interest to decision-makers is that the cost-effectiveness of the programme may vary by subgroups of the patient population. For example, in the study of chemotherapy for early breast cancer discussed under in Section 3.2.8 above, Campbell et al. (2011) also presented ICERs for patients of different age and risk of recurrence. For women considered high risk for recurrence, FEC-D was the most cost-effective treatment strategy even though it was the most expensive. This remained the case when patient age was increased from 40 to 60 years. However, for younger, low-risk women, the ICER for FEC-D exceeded £70 000 per QALY and E-CMF was the most cost-effective strategy.

This is the issue of heterogeneity mentioned earlier. This information is useful in situations where decision-makers feel it would be acceptable to limit the use of a particular technology to the patient groups for which it is most cost-effective. However, these decisions sometimes raise other considerations, such as fairness or equity in the use of resources, which would need to be considered alongside cost-effectiveness.

Finally, a good study should begin to help users interpret the results in the context of their own particular situation. This can be done by being explicit about the perspective

for the analysis (an earlier point) and by indicating how particular costs and benefits might vary by location. For example, the costs of instituting day-care surgery may vary, depending on whether a purpose-built day-care unit already exists or whether wards have to be converted. Similarly, the benefits of day-care surgery may vary depending on whether, in a particular location, there is pressure on beds and whether beds will be closed or left empty. Obviously it is impossible for the analyst to anticipate every possibility in every location, but it is useful to explore the factors, varying from place to place, that might impact upon the likely cost-effectiveness of programmes.

The issue of whether the results of economic evaluations can be interpreted in, or 'transferred' to, other jurisdictions has been widely discussed, since some key elements of data, such as unit costs (prices) and resource use (being related to clinical practice patterns) are likely to be context-specific. A good introduction to the topic can be found in the ISPOR Good Practices Task Force Report on Economic Data Transferability (Drummond et al. 2009).

3.3 **Reporting guidelines for economic evaluation**

The presentation, interpretation, and use of economic evaluation results raise a number of practical issues. For example, can guidelines for good practice in the reporting of studies be developed? Can the results (e.g. cost-effectiveness ratios) from different studies be meaningfully compared? Can results of studies be generalized from one setting, or country, to another?

There have been several attempts to develop reporting guidelines for economic evaluation. A recent example of such guidelines, along with references to earlier attempts, is given in the report of the Consolidated Health Economic Evaluation Reporting Standards (CHEERS) Task Force (Husereau et al. 2013). These guidelines, which were developed by an approach consistent with the CONSORT methodology used to develop reporting standards for clinical trials (Schulz et al. 2010), cover all the main steps in conducting an economic evaluation, as outlined in our critical appraisal checklist shown in Box 3.1. However, the CHEERS guidelines are more detailed and make several finer distinctions surrounding what should be reported in economic evaluations conducted alongside a single clinical study, such as a clinical trial, as opposed to an economic evaluation conducted using a decision-analytic model.

Although it makes sense to be fairly prescriptive about standards for reporting, in order to be able to compare the results of different studies in a meaningful fashion one would have to be confident that the studies adopted a similar methodological approach. Therefore, some sets of guidelines for economic evaluation go beyond merely specifying reporting standards, to advising on preferred methods. Since there is much more debate about the appropriateness of different methods for economic evaluation, one approach to achieving some measure of standardization is to specify a 'reference' case. Then, advice can be given to analysts to present their results in a manner consistent with the reference case, while also allowing them to present them (in addition) using alternative methods (see Box 3.4).

In those jurisdictions where there is a formal requirement for economic evaluations as part of the reimbursement process, it is usual (although not universal) to develop

Box 3.4 Developing guidelines for the presentation of results: specifying a 'reference case'

The notion of a 'reference case' was first proposed by the Public Health Service Panel on Cost-Effectiveness in Health and Medicine (Gold et al. 1996). In considering the methodology and practice of economic evaluation in health care, the panel recognized that many methodological issues were unresolved. On the other hand, they also recognized the need to develop a standardized approach for the conduct and reporting of studies, so that the results from different studies could be compared.

Therefore, the 'reference case' is a preferred set of methodological principles that should be used for the one of the analyses undertaken. Then, if the analyst prefers, they can also report other results, applying different methods. However, if the reference case analysis is always reported, reliable comparisons of studies can be made.

The reference case proposed by the panel embodied most of the good methodological principles of economic evaluation existing at the time. The main features were as follows:

1 The societal perspective should be adopted.
2 Effectiveness estimates should incorporate benefits and harms.
3 Mortality and morbidity consequences should be combined using QALYs.
4 Effectiveness estimates from best-designed and least-biased sources should be used.
5 Costs should include health care services, patient and caregiver time, and costs of non-health impacts.
6 Comparison should be made with existing practice and (if necessary) a viable low-cost alternative.
7 Discounting of costs and health outcomes should be undertaken at a real rate of 3% per annum (plus 5% for comparison with existing studies).
8 One-way and multiway sensitivity analysis (for important parameters) should be undertaken.
9 Comparison of the ICER should be made with those for other relevant interventions.

The notion of a reference case now underpins several sets of methodological guidelines for economic evaluation.

Source: data from Gold, M.R, et al. (ed.) *Cost-effectiveness in health and medicine*, Oxford University Press, New York, USA, Copyright © 1996 Oxford University Press USA.

methods guidelines for conducting studies. See, for example, the methods guidelines proposed by the National Institute for Health and Care Excellence (NICE) in the United Kingdom (NICE 2013).These guidelines are intended both for manufacturers making submissions of evidence and for those that assess them. The primary purposes of these guidelines are both to encourage minimum methodological standards and to

make the various submissions more comparable. It is hoped that, taken together, these factors will lead to a more consistent decision-making process.

The various national guidelines vary in their level of detail and scope. However, the general principles behind the different sets of guidelines are very similar, although there are several differences in areas such as the perspective for the analysis (health care costs or broader), the acceptability of various types of clinical data (randomized controlled trials only, or including observational studies), the assessment of health gain (QALYs or other approaches), and the characterization of uncertainty.

The most comprehensive source of information on national methods guides for economic evaluation is the review and classification maintained on the website of the International Society for Pharmacoeconomics and Outcomes Research (ISPOR) (<http://www.ispor.org/PEguidelines/index.asp>). Currently, this includes details of more than 30 sets of guidelines and is updated as new ones are produced. Most of the guidelines in the classification are 'official' guidelines developed by decision-makers to assist reimbursement decisions, although some are produced by academic groups with the more general purpose of improving the quality of economic evaluations undertaken within a given jurisdiction of internationally.

In low and middle income countries, much of the use of economic evaluation to date has been driven by international agencies, such as the World Health Organization (WHO) and the World Bank, where economic evaluations have been conducted alongside development programmes. WHO has produced its own set of methods guidelines that have been extensively used in regions of the world outside Europe and North America (Tan-Torres Edejer et al. 2003).

Since these methods guidelines are usually developed in the context of requiring economic evaluations as part of the process for decision-making on pricing or reimbursement of coverage for health technologies and programmes, they are often specific to the decision problems in the jurisdiction concerned. In a recent reference case that has been developed for the Bill and Melinda Gates Foundation (Bill and Melinda Gates Foundation 2013), the authors acknowledge that, in making proposals that are intended to serve the needs of economic analysts and decision-makers in several jurisdictions, it is possible to specify several key principles underlying a well-conducted economic evaluation, but that the precise methods will inevitably depend on how the decision problem is defined in the jurisdiction concerned. This in turn will inevitably depend on several broader aspects of the structure, organization, and financing of that jurisdiction's health care system.

3.4 Limitations of economic evaluation techniques

Our main purpose in this chapter is to make the user of economic evaluation results more aware of the methodological judgements involved in undertaking an economic evaluation in the health care field. In Box 3.1 we have consolidated the points made in this chapter into a suggested checklist of questions to ask when critically assessing economic evaluation results. In addition, there are several other limitations of which users should be aware.

Of primary concern from a policy perspective is the fact that economic evaluations do not usually incorporate into the analysis the importance of the distribution of

costs and consequences among different patient or population groups. Yet, in some cases, the identity of the recipient group (e.g. the poor, the elderly, working mothers, or a geographically remote community) may be an important factor in assessing the social desirability of a service or programme. Indeed, it may be the motivation for the programme in the first place. Although it is sometimes suggested that differential weights be attached to the value of outcomes accruing to special recipient groups, this is not normally done within an economic evaluation. Rather, an equitable distribution of costs and consequences across socio-economic or other defined groups in society is viewed as a competing dimension upon which decisions are made, in addition to that of cost-effective deployment of resources (see Chapter 2). One possibility is to undertake an equity analysis, to complement the economic evaluation (Cookson et al. 2009).

A more subtle, yet important, point is that the various forms of analysis discussed above embody different normative judgements, as discussed in Chapter 2. For example, if an economic evaluation values health outcomes in terms of individuals' willingness to pay, this may be constrained by ability to pay and therefore valuations are dependent on the existing income distribution. On the other hand, the simple aggregation of QALYs in an economic evaluation implies that a QALY is being valued the same no matter to whom it accrues. Therefore, in reality it is difficult to divorce equity considerations from the economic evaluation and analysts should be aware of this when selecting a particular analytic technique.

Finally, evaluation of any sort is in itself a costly activity. Bearing in mind that *even economic evaluations should be subject to economic evaluation*, it seems reasonable to suggest that economic evaluation techniques will prove most useful in situations where programme objectives require clarification, the competing alternatives are significantly different in nature, or large resource commitments are under consideration.

3.5 Conclusions

In this chapter, we have tried to assist users of economic evaluations in interpreting evaluation studies and assessing their usefulness for health care decisions, or for planning further analyses. We have identified and discussed ten questions that readers of economic evaluations can ask in order to critically assess a particular study; a checklist of these questions is given in Box 3.1.

Our intent in offering a checklist is not to create hypercritical users who will be satisfied only by superlative studies. It is important to realize, as emphasized at the outset, that for a variety of reasons it is unlikely that every study will satisfy all criteria. However, the use of these criteria as screening devices should help users of economic evaluations to identify quickly the strengths and weaknesses of studies. Moreover, in assessing any particular study, users should ask themselves one final question, '*How does this evaluation compare with our normal basis for decision-making?*' They may find that the method of organizing thoughts embodied in the evaluation compares well with alternative approaches, even bearing in mind the possible deficiencies in the study.

3.6 Critical appraisal of published articles

In this section we undertake critical appraisals of two papers. In order to understand better the key methodological issues in undertaking economic evaluations, you should obtain copies of the two studies and work through the critical appraisal checklist outlined in Box 3.1, answering each question as 'yes', 'no', or 'can't tell'.

The paper studied in Section 3.6.1, by Blomström et al. (2008), is an example of an economic evaluation conducted primarily alongside a single clinical study. In such studies, often called 'trial-based' studies, it is possible to use the opportunity of conducting the clinical trial, or other prospective clinical study, to collect detailed data on resource use and to have access to the individual patient data. (The issues in conducting economic evaluations using individual patient data are discussed in Chapter 8.)

The paper by McKenna et al. (2010) evaluated in Section 3.6.2 is an example of a study conducted using a decision-analytic model. In such studies, often called 'modelling studies', the model is used as a framework for synthesizing data from a variety of sources. Typically, the clinical data for the model will come from a systematic review of all the available literature, as opposed to a single clinical study, although McKenna et al. found that the available clinical literature was limited to a single RCT. (The issues in conducting economic evaluations using decision-analytic models are discussed in Chapter 9.)

Although the distinction between 'modelling' and 'trial-based' studies is useful for explaining the key methodological issues, the differences between the two approaches are not often that large in practice. For example, in common with many 'trial-based' studies, the study by Blomström et al. uses a statistical model to extrapolate survival beyond the end of the trial, so as to estimate the QALYs gained over lifetime. On the other hand, 'modelling' studies often use individual patient data as a basis for estimating the key clinical parameters in the model.

Our answers to the critical appraisal questions for the two studies are given below.

3.6.1 Cost effectiveness of cardiac resynchronization therapy in the Nordic region: an analysis based on the CARE-HF trial (Blomström et al. 2008)

1. Was a well-defined question posed in an answerable form?

YES ☒ NO ☐ CAN'T TELL ☐

The aim of the study is to investigate the cost-effectiveness of cardiac resynchronization therapy (CRT) in three Nordic countries (Denmark, Finland, and Sweden), using results from the CARE-HF trial alongside additional country-specific parameters (Blomström et al. 2008, p.870). The authors state that they conducted a cost-effectiveness analysis (CEA) to evaluate the associated costs and QALYs of CRT as an addition to standard pharmacological treatments compared with standard treatment alone, as defined by the CARE-HF trial (p.869).

The perspective taken for the analysis is not stated; however, the nature of the included costs and effects would suggest a health care perspective as the analysis includes intervention, hospitalization, and cardiac day care and outpatient costs.

2. Was a comprehensive description of the competing alternatives given? (i.e. can you tell who? did what? to whom? where? and how often?)

YES ☒ NO ☐ CAN'T TELL ☐

Although the authors devote a section of the publication to discussing the CARE-HF trial (p.870), several questions are not answered about the approach used in the trial. The major question unanswered is where the CARE-HF trial was performed, since cost-effectiveness results are presented for three different countries. Secondly, few details are given on the interventions being compared. For example, what constitutes standard and optimized pharmacological treatments, and how is CRT performed. Thirdly, although the average follow-up of the CARE-HF trial is reported (29.4 months, p.870) and Figure 1 of the paper provides an indication of the duration of the trial, the duration of the trial is not explicitly stated.

No relevant alternatives appear to have been omitted from the analysis, as the trial is only referred to as having two treatment groups (pharmacological therapy alone or in combination with CRT, p.870). However, the reader should consider whether the trial excluded any potentially relevant alternative treatment options, which is not discussed in the publication. For example, such treatment options may include different definitions of pharmacological therapy, since it is unclear if such therapy is consistent between countries. In addition, the authors could have considered including the possibility of replacing the implant's battery, rather than assuming only that the implant became non-functional. The exclusion of a 'do nothing' alternative from the analysis appears reasonable due to both its exclusion from the trial and the significant risks associated with not providing treatment for patients with heart failure; however, this is not discussed in the publication.

Although a significant amount of detail is excluded from the publication, several references are provided that can be expected to contain details of the CARE-HF trial. However, these are not sufficiently well referenced at suitable points of discussion to provide the reader with complete information.

3. Was there evidence that the programme's effectiveness had been established?

YES ☒ NO ☐ CAN'T TELL ☐

The CARE-HF randomized trial compared CRT in combination to pharmacological therapy with pharmacological therapy alone. The primary end point was death from any cause or an unplanned hospitalization for a document major cardiovascular event, such as worsening heart failure or myocardial infarction. The hazard ratio for the primary end point was HR 0.63; 95% CI 0.51 to 0.77 (p.870). The main secondary end point was death from any cause, and its HR was 0.64; 95% CI 0.48 to 0.85.

No specific discussion is made as to how representative is the CARE-HF protocol of the clinical practice in all three countries being analysed.

As the trial did not run for the patient's lifetime or for the expected lifetime of the device, the authors extrapolate the trial results over a lifetime horizon. The authors are very clear on the approach to estimate the extrapolated survival of patients in each

randomized group. They discuss their approach to separate the extrapolated survival curves (represented as Kaplan–Meir curves) into two (pp.870–1), one from the end of the trial until the end of the battery's expected lifetime and from that point until the patient's death. The authors are also clear about the base case assumptions associated with the survival extrapolation approach, including that the survival rate is the same for both groups beyond the end of the implant's lifetime.

4. Were all the important and relevant costs and consequences for each alternative identified?

YES ☒ NO ☐ CAN'T TELL ☐

The authors appear to have considered a complete set of the costs associated with each of the alternative treatments, and a good discussion is presented on the estimation of relevant costs for each of the three countries (p.871). Resource use was obtained from the CARE-HF trial, to which country-specific unit costs were applied. A discussion is also presented as to the relative merits of considering the initial implant costs either as an investment to be considered over its lifetime or as a one-off cost, as applied in a similar analysis taken from the UK perspective (p.875).

The consequences were expressed in terms of QALYs using the survival and the health-related quality of life (HRQoL) weights estimated from data collected during the CARE-HF trial. In the CARE-HF trial, EQ-5D was collected at baseline and 3 months and the MLWHF instrument was collected at baseline, 3 months, 18 months, and at the end of the trial. Therefore, the HRQoL weights at 18 months and at the end of the trial follow-up are estimated using a mixed model of EQ-5D and MLWHF. Patients are assumed to experience the HRQoL estimated for the end of the study throughout their lifetime (p.870).

The approach taken to HRQoL has a number of issues. First, no details are given on the MLWHF instrument, namely which dimensions of HRQoL it considers and whether it is a condition-specific or a general instrument, such as EQ-5D. Secondly, no details are given on the model used to map between EQ-5D and MLWHF, its appropriateness, and the goodness of fit of the estimates. Thirdly, the limitations of assuming that patients experience constant HRQoL throughout their lifetime are acknowledged (p.876) but this is not subjected to sensitivity analysis. Fourthly, while the CARE-HF trial recoded a set of primary outcomes associated with cardiovascular disease, the HRQoL weights associated with these do not appear to be considered in the study, relying only on the final overall average (a secondary outcome of the trial). While this may not have a significant impact on the analysis, it is an important caveat as the study may overlook significant differences in some of the primary outcomes between the two trial arms that may have large impacts on health utilities (this may have an impact on the cost-effectiveness decision if the distribution of primary outcomes is significantly different between the two arms).

In addition, it is not clear what costs or health utilities occur once the patient's implant reaches the end of its battery life. There is no discussion of whether a further surgery is required to remove/replace the battery and the associated costs and consequences.

5. Were costs and consequences measured accurately in appropriate physical units? (e.g. hours of nursing time, number of physician visits, lost work-days, gained life-years)

YES ☒ NO ☐ CAN'T TELL ☐

As mentioned above, resource use was measured in the CARE-HF trial; these were then applied to country-specific costs (using a range of hospital price lists, p.871).

Pharmaceutical costs were removed from the analysis because there were no significant differences between the two trial arms in pharmaceutical consumption (p.871). Costs appear to have been considered from the health care provider perspective; this appears to be consistent with the rest of the analysis.

No discussion is made to the role of joint resource use, e.g. operating theatres or specialist surgeons. This is likely to be a significant cost consideration, as the alternative of pharmacological treatment alone requires no such costs.

As mentioned in the previous question, the initial device implantation costs are treated as an annuity. The reader should be aware that considering the costing of the hospitalization and procedural costs associated with the implant as an investment spread over the lifetime of the implant may not be consistent with the approach of the health care provider.

6. Were costs and consequences valued credibly?

YES ☒ NO ☐ CAN'T TELL ☐

The majority of the consequences appear to be valued credibly, coming from the CARE-HF trial; however, there is a lack of clarity about how health utility data was collected. It is reported that a mix of MLWHF and EuroQol EQ-5D was used to determine health utility (p.870); however, no details are provided about how these instruments were used, the completeness of response rates, or the mixed model used to combine the results of the two instruments.

The costs included in the annual total for each arm are provided in Table 1 of the publication. They are based on values extracted from hospital price lists, and appear to have been suitable selected. However, it is unclear if there is consistency in the definition of these unit costs. It should be identified that, for example, Sweden's cost associated with heart transplants may be significantly higher than that in Denmark or Finland (€109 151 compared to €67 955 and €51 216 respectively) due to a better level of care being provided, having impacts on health utility, or a difference in the definition of what is included in the DRG.

7. Were costs and consequences adjusted for differential timing?

YES ☒ NO ☐ CAN'T TELL ☐

Both costs and health effects were discounted at a rate of 3% (p.871). This is justified with reference to previous publications and is subject to scenario analysis (both varied from 0% to 5%).

8. Was an incremental analysis of costs and consequences of alternatives performed?

YES ☒ NO ☐ CAN'T TELL ☐

In addition to a net monetary benefit analysis for the within trial analysis (see Table 3 of the paper) incremental analysis was performed for both within trial and for the main extrapolation analysis (Table 3 and Table 2 respectively). Results are presented for the median incremental cost per QALY and cost per life year gained as well as at the 95% confidence intervals. The reader should be aware that the use of confidence intervals to represent uncertainty in the ICERs can be misleading and that an alternative approach such as the estimation of the probability of cost-effectiveness for all alternative comparators may be more accurate and informative, as presented in Figure 2.

9. Was uncertainty in the estimates of costs and consequences adequately characterized?

YES ☒ NO ☐ CAN'T TELL ☐

In addition to the reporting of 95% confidence interval results for the analysis results (see previous answer), several sensitivity analyses were conducted for both the within-trial analysis and the extrapolated main analysis. These were as follows (one-way analyses unless stated):

- Uncertainty in the cost-effectiveness threshold was considered through the production of a cost-effectiveness acceptability curve (see Figure 2 of the paper).
- The impact of the assumptions around the average additional lifetime at the end of the follow-up period was tested by implementing a range of different survival curves for the two groups as well as assuming the same survival curve for both.
- Uncertainty in the discount rate was considered by varying the applied discount rates together from 0% (undiscounted) to 5%. In addition, a two-way sensitivity analysis was conducted by setting the discount rate associated with the health effects to 0%, while keeping the discount rate for costs at 3%. The inclusion of discount rates in the analysis is unclear as it represents a policy decision made by the health care provider rather than a parameter in the analysis.
- The assumption around differences in survival between the CRT group and the control group was tested by assuming the same mortality rate at the end of follow-up (both truncated and not at 6 years after follow-up), and assuming differences in mortality rate persisted after follow-up.
- Uncertainty around device lifetime was also tested by varying the lifetime to 5 or 7 years (this was only conducted on the within-trial analysis).

It is unclear what the decision rule for choosing which parameters were judged to be critical for the cost-effectiveness result in the main analysis as is stated on p.872, or how the respective scenarios were selected.

Results for these sensitivity analyses are produced in Table 4 of the paper. The results of the analysis were relatively insensitive to many of the scenarios; however, varying the survival assumptions after the end of follow-up had a larger impact on results for all three countries. The authors observe that in no cases did the sensitivity analyses force the ICER values beyond what could be considered cost-effective (p.874). However, the authors noted that the results of the analysis are highly sensitive to the longevity of the device, as shown in their Table 4.

10. Did the presentation and discussion of study results include all issues of concern to users?

YES ☐ NO ☐ CAN'T TELL ☒

The authors produce an extensive discussion of the results of the analysis as well as of the structural approach adopted (pp.874–6). Areas of discussion include the rationale for using an extended analysis (p.874), comparison to previous studies (p.875), a consideration of the role of the appropriate cost-effectiveness threshold to the analysis (p.875), and a brief discussion on the limitations of the study (p.876) including the extrapolation of survival from the clinical trial, the use of the 'last value carried forward' approach to health utility values from the trial, and the potential for bias in the exclusion of pharmaceutical costs. Although differences in pharmaceutical costs were not observed during the trial, as patients age and experience different risks of cardiovascular events it is likely that pharmaceutical costs could differ by intervention. Therefore, excluding pharmaceutical costs could make CRT appear less cost-effective.

However, throughout the publication a significant set of limitations is evident that result from the scale of the analysis attempted. In attempting to conclude a cost-effectiveness result for three Nordic countries the study may have overlooked several important factors within each country; these may include relevant patient factors that could result in different cost-effectiveness results for different subpopulations (e.g. by age or gender). The authors also continually assume the transferability of results from the CARE-HF trial and perfect representation of the trial population to each of the three countries, without explicitly stating this assumption. Clearly this is highly unlikely, and may have significant impacts on the suitability of the conclusions derived from the analysis. It is highly likely that there are reasons why the CARE-HF trial results are not suitable to be used for the analysis; this is a significant factor that is consistently overlooked by the authors.

3.6.2 Cost-effectiveness of enhanced external counterpulsation (EECP) for the treatment of stable angina in the United Kingdom (McKenna et al. 2010)

1. Was a well-defined question posed in an answerable form?

YES ☒ NO ☐ CAN'T TELL ☐

The authors state that their aim was to 'develop a UK-specific cost-effectiveness model of EECP compared with no treatment as additional therapy to usual care for the treatment of chronic stable angina' (p.176 of the paper).

The authors explain the context of their study, which is that 'EECP results in upfront costs but the potential quality of life benefits through improved symptoms and long-term relief from symptoms may outweigh the costs when compared with not giving the therapy' (p.176).

They state that 'a probabilistic decision analytic model was developed to assess the cost-effectiveness of EECP in the UK NHS' (p.176).

The authors state that 'the model evaluates costs from the perspective of the National Health Service and Personal Social Services (NHS & PSS), expressed in UK £ sterling at a 2008 price base. Outcomes in the model were expressed in terms of QALYs, with costs and benefits discounted at 3.5 percent per year' (p.176).

2. Was a comprehensive description of the competing alternatives given? (i.e. can you tell who? did what? to whom? where? and how often?)

YES ☒ NO ☐ CAN'T TELL ☐

EECP as an additional therapy to usual care is compared with usual care alone. The authors state that '[the model] evaluates a strategy of EECP treatment compared with no treatment on the assumption that angina patients would receive EECP treatment over and above standard current clinical practice care' (p.176).

The choice of competing alternatives follows the MUST-EECP study, a randomized clinical trial which compared EECP treatment (n = 72) with sham-EECP (n = 67). The authors describe EECP as 'a non-invasive technique used in the treatment of angina to increase blood flow to the heart. Three pairs of pressure cuffs are wrapped around the patient's calves, lower thighs, and upper thighs and are sequentially inflated during diastole. All pressure is released at the onset of systole by simultaneously deflating the cuff. . . . An EECP treatment course conventionally consists of thirty-five 1-hour sessions over a period of 4 to 7 (can do two sessions per day) weeks' (p.176).

The base-case population refers to the baseline characteristics of the trial population, which is assumed to be representative of the angina patients typically presenting for EECP in the UK clinical setting.

3. Was there evidence that the programme's effectiveness had been established?

YES ☒ NO ☐ CAN'T TELL ☐

A systematic review on the clinical effectiveness of EECP identified a single RCT (the MUST-EECP trial) comparing EECP with an alternative treatment, in this case, sham-EECP. The MUST-EECP study recorded a number of outcomes: (i) exercise treadmill duration; time to ≥1 mm ST-segment depression; (iii) angina counts; (iv) nitroglycerin use; and (v) HRQoL (p.177).

The authors do not present the results of the MUST-EECP trial and refer to the original publications. Since no evidence was found to link the four intermediate outcomes (i–iv) to final health outcomes, the outcome used in the model is the improvement in HRQoL at 12 months after the end of treatment (p.177). HRQoL was measured during the trial using the SF-36 instrument. In order to estimate QALYs, the SF-36 scores are mapped into EQ-5D health state preference values using a published algorithm. Table 1 of the paper (p.178) presents the improvement in EQ-5D values for EECP relative to sham-EECP at 1 year. On average, patients randomized to EECP experienced a EQ-5D improvement at 1 year of 0.1068 compared with 0.0351 for those randomized to sham-EECP, a difference of 0.0717.

The duration of the benefits from EECP is informed by expert elicitation due to lack of both experimental and observational evidence. Expert elicitation 'involved asking clinical experts to report their beliefs about the duration of HRQoL benefits with some estimate of their uncertainty' (p.177). The authors report that 'Five experts with experience and knowledge of EECP in the UK completed the exercise independently giving their own belief about the unknown quantities with estimates

of uncertainty [the proportion of patients that they would expect to sustain the average HRQoL benefits observed at one year].' Full details of the elicitation exercise are reported in another publication. Table 2 of the paper presents the mean and standard deviations for the probability of sustaining HRQoL benefits in each subsequent year. The mean probability of sustaining HRQoL benefits in subsequent years varied from 0.605 to 0.908 for the second year after treatment, 0.600 to 0.905 for the third year after treatment, and from 0.526 to 0.898 for the subsequent years. The average probability across all experts is assumed to be representative of the beliefs of relevant clinical experts.

Two key assumptions are employed. Firstly, since the authors found no evidence to suggest that EECP treatment compared with placebo has a differential impact on the risk of cardiovascular events and death, it was assumed that EECP has only a palliative benefit on angina patients. Therefore, EECP treatment improves HRQoL improvement compared to no treatment (p.177). Secondly, although the authors found no trial evidence to indicate the degree to which improvement in HRQoL from EECP is sustained over time and, given the clinical expert belief that the benefits of EECP are sustained beyond one year, the duration of the HRQoL benefits was informed from the results of the elicitation exercise (pp.177–178). Both assumptions seem reasonable in light of the available evidence.

4. Were all the important and relevant costs and consequences for each alternative identified?

YES ☒ NO ☐ CAN'T TELL ☐

The costs are evaluated from the perspective of the NHS and PSS. Outcomes are expressed in terms of QALYs (p.176). The perspective is consistent with the guidelines in place for economic evaluations informing NHS decisions (NICE 2013).

Given that EECP is assumed to have only a palliative benefit on angina patients, no additional costs other than the costs of EECP itself are considered. The costs of EECP per patient are based on the annual costs associated with EECP and a throughput of 12 patients per year. The costs of EECP include: (1) capital cost of new EECP machine annuitized over a useful life of 10 years, using an interest rate of 3.5% per annum, (2) equipment replacement costs, (3) consumables, (4) overheads, and (5) staffing costs (p.178).

5. Were costs and consequences measured accurately in appropriate physical units?

YES ☒ NO ☐ CAN'T TELL ☐

The costs are expressed in UK £ sterling at a 2008 price base (p.176). All relevant items are included in the costs. The outcomes are expressed in terms of QALYs. The health state preference values are obtained by mapping SF-36 summary scores reported in the MUST-EECP trial into EQ-5D utility scores. Life expectancy is estimated using standard UK age- and sex-specific mortality and including a competing mortality risk due to cardiovascular events. Mortality due to cardiovascular events is informed by the risk equations applied in the EUROPA trial (pp.177–178).

6. Were costs and consequences valued credibly?

YES ☒ NO ☐ CAN'T TELL ☐

The sources of all values are clearly identified. Health state preference values are obtained by mapping SF-36 summary scores collected during the MUST-EECP trial into EQ-5D utility scores using a published algorithm (p.177). Although the authors do not give any details of the collection of the SF-36 details or the algorithm used to map to EQ-5D it can be reasonably assumed that these were suitable and not subject to any significant biases; however, the reader should be aware that mapping of utility scores is often associated with lower statistical power, which is discussed briefly by the authors (p.181).

Unit costs and per patient costs of each resource item included in the model are presented in Table 3 of the paper (p.179). The capital costs, equipment replacement costs, staffing costs, and overhead costs are based on personal communications and on the EECP manufacturer's price list.

7. Were costs and consequences adjusted for differential timing?

YES ☒ NO ☐ CAN'T TELL ☐

Both costs and health outcomes are discounted at 3.5% per year (p.177). This is in line with guidelines for economic evaluations informing NHS decisions (see NICE 2013).

8. Was an incremental analysis of costs and consequences of alternatives performed?

YES ☒ NO ☐ CAN'T TELL ☐

The authors present mean lifetime costs and QALYs of both treatment strategies, as well as ICERs. Table 4 of the paper (p.179) presents the estimates of cost-effectiveness for the base case, together with best- and worst-case scenarios for duration of HRQoL benefits. For the base-case analysis, the ICER associated with EECP is £18 643 per additional QALY. For the worst-case scenario, in which HRQoL benefits are assumed to last 1 year, the ICER is £63 072 per additional QALY. For the best-case scenario, in which HRQoL benefits are assumed to be sustained throughout lifetime, the ICER is £5831 per additional QALY.

9. Was uncertainty in the estimates of costs and consequences adequately characterized?

YES ☒ NO ☐ CAN'T TELL ☐

Uncertainty around the cost-effectiveness estimates is analysed with probabilistic sensitivity analysis and a number of alternative scenarios.

Probabilistic analyses are conducted using Monte Carlo simulation (repeated random sampling from the joint probability distribution of parameters). Under base-case assumptions, the probability that EECP as an add-on to usual care is more cost-effective than usual care alone is 0.444 for a threshold of £20 000 per QALY gained and 0.698 for a threshold of £30 000 QALY gained (p.180).

The cost-effectiveness results for a number of alternative scenarios are presented in Supplementary Table 1 of the paper. The scenarios varied: (1) the probability of sustaining HRQoL benefits, (2) the cost of EECP sessions, and (3) the probability of repeat EECP sessions within 2 years of treatment. First, the probability of sustaining HRQoL

benefits over time is tested over the range of values elicited from the clinical experts, in addition to the worst- and best-case scenarios. The worst-case scenario assumed that HRQoL benefits from EECP are only maintained in the first year after treatment, and lost in subsequent years. The best-case scenario assumes that HRQoL benefits last over a patient's lifetime. Second, the cost of EECP is varied by £500 and £1000. Third, the probability of repeat EECP sessions is varied from 10% to 30%. The range of values used in the scenario analysis of the cost of EECP and of the probability of repeat EECP sessions is not justified (pp.179–180).

The results indicated the cost-effectiveness of EECP is highly sensitive to the probability of sustaining HRQoL benefits over time. In terms of the scenario on the costs of EECP, only if costs are expected to be £3000 more than the base-case estimate of £4347 does the cost-effectiveness of EECP become unlikely under the thresholds of £20 000 and £30 000 per QALY gained. The cost-effectiveness of EECP is robust to the likelihood of patients requiring repeat EECP sessions.

10. Did the presentation and discussion of study results include all issues of concern to users?

YES ☐ NO ☐ CAN'T TELL ☒

The authors include a fairly complete discussion of the results. No previously published studies examining the cost-effectiveness of EECP were found as part of the systematic literature review conducted by the authors. The ICER was compared to the conventional thresholds used in the UK NHS. The uncertainty around the cost-effectiveness estimates is expressed by the probability that EECP is cost-effective. A number of scenarios explored the key assumptions and parameter inputs employed in the decision model. The authors conclude that the results 'demonstrate that long-term maintenance of HRQoL benefits from EECP is central to the estimate of cost-effectiveness' (p.180). However, no evidence was found on the duration of benefits and the experts did not share similar beliefs. Therefore, whether EECP is a cost-effective technology remains uncertain.

The authors discuss the limitations of the analysis (p.181). First, the HRQoL estimates and the duration of HRQoL benefits are highly uncertain due to the limited evidence base. A mapping algorithm is used to convert SF-36 summary scores into EQ-5D values. Expert elicitation is used to obtain estimates of the probability of sustaining HRQoL benefits from treatment over time. Second, since no evidence was found to support the potential impact of EECP on health outcomes, the model only considers the impact of EECP on HRQoL. Therefore, the results can be considered conservative if EECP leads to a reduction in adverse health outcomes. Third, there is uncertainty regarding the need for repeat EECP sessions and on the treatment costs. The authors state 'if the number of patients undergoing the therapy were to increase . . . , the cost per patient would fall yet further, mean cost-effectiveness improve, and the uncertainty in cost-effectiveness decline' (p.181).

The authors mention that 'the generalisability of the findings to a broader range of patients who could potentially benefit from EECP should be viewed with due caution' (p.181). In addition, the results may not be generalizable to settings outside the UK. However, no further discussion is offered on the issue.

Although the NHS and PSS perspective is taken, the authors hypothesize that a societal perspective, which includes patient-borne costs, may increase the ICER associated

with EECP. Although no rationale is presented, patient-borne costs are likely to increase the ICER due to the travel required to and from the centre providing EECP treatment.

A final issue, not explored in the paper, is whether there are any implementation constraints, such as the feasibility of implementing EECP nationwide given the costs and training requirements of EECP.

Acknowledgements

The exercises in Section 3.6 are reproduced courtesy of Rita Faria and Sebastian Hinde. We are grateful to them for preparing these exercises and allowing us to use them here.

Text extracts in section 3.6.2 from McKenna, C. et al., Cost-effectiveness of enhanced external counterpulsation (EECP) for the treatment of stable angina in the United Kingdom, *International Journal of Technology Assessment in Health Care*, Volume 26, Issue 2, pp. 175–82, Copyright © Cambridge University Press 2010, reproduced with permission.

References

Bill and Melinda Gates Foundation (2013). *Methods for economic evaluation: reference case*. Seattle, WA: Bill and Melinda Gates Foundation.

Black, W.C. (1990). The cost-effectiveness plane: a graphic representation of cost-effectiveness. *Medical Decision Making*, **10**, 212–15.

Blomstrom, P., Ekman, M., Blomstrom Lundqvist, C., Calvet, M.J., Freemantle, N., et al. (2008). Cost effectiveness of cardiac resynchronization therapy in the Nordic region: An analysis based on the CARE-HF trial. *European Journal of Heart Failure*, **10**, 869–77.

Brazier, J., Roberts, J., and Deverill, M. (2002). The estimation of a preference-based single index from the SF-36. *Journal of Health Economics*, **21**, 271–92.

Briggs, A.H., Sculpher, M.J., Britton, A., Murray, D., and Fitzpatrick, R. (1998). The costs and benefits of primary total hip replacement. *International Journal of Technology Assessment in Health Care*, **14**, 743–61.

Briggs, A.H., Weinstein, M.C., Fenwick, E.A.L., Karnon, J., Sculpher, M.J., Paltiel, D., on behalf of the ISPOR-SMDM Modeling Good Practices Task Force. (2012). Model parameter estimation and uncertainty: a report of the ISPOR-SMDN Modeling Good Practices Task Force-6. *Value in Health*, **15**, 835–42.

Byford, S., Knapp, M., Greenshields, J., et al. (2003). Cost-effectiveness of brief cognitive behaviour therapy versus treatment as usual in recurrent deliberate self-harm: a decision-making approach. *Psychological Medicine*, **33**, 977–86.

Campbell, H.E., Epstein, D., Bloomfield, D., Griffin, S., Manca, A., Yarnold, J., Bliss, J., et al. (2011). The cost-effectiveness of adjuvant chemotherapy for early breast cancer: a comparison of no chemotherapy and first, second, and third generation regimens for patients with differing prognoses. *European Journal of Cancer*, **47**, 2517–30.

Centre for Reviews and Dissemination (2009). *Systematic reviews: CRD's guidance for undertaking reviews in health care*. Available at: <http://www.york.ac.uk/inst/crd/pdf/Systematic_Reviews.pdf> (accessed 26 November 2013).

Cookson, R., Drummond, M.F., and Weatherly, H. (2009). Explicit incorporation of equity considerations into economic evaluation of public health interventions. *Health Economics, Policy and Law*, **4**, 231–45.

Drummond, M.F., Barbieri, M., Cook, J., et al. (2009). Transferability of economic evaluations across jurisdictions: ISPOR Good Research Practices Task Force report. *Value in Health*, **12**, 409–18.

EuroQoL Group (1990). EuroQoL—a new facility for the measurement of health-related quality of life. *Health Policy*, **16**, 199–208.

Feeny, D., Furlong, W., Boyle, M., and Torrance, G. (1995). Multi-attribute health status classifications systems: health utilities index. *PharmacoEconomics*, **7**, 490–502.

Gold, M.R., Siegel, J.E., Russell, L.B., and Weinstein, M. (ed.) (1996). *Cost-effectiveness in health and medicine*. New York: Oxford University Press.

Guyatt, G., Rennie, D., Meade, M.O., and Cook D.J. (ed.) (2008). *Users' guides to the medical literature*, 2nd edition. Chicago: American Medical Association.

Husereau, D., Drummond, M.F., Petrou, S., et al. (2013). Consolidated Health Economic Evaluation Reporting Standards (CHEERS): Explanation and Elaboration: A Report of the ISPOR Health Economic Evaluation Publication Guidelines Good Reporting Practices Task Force. *Value in Health*, **16**, 231–50.

Mark, D.B., Hlatky, M.A., Califf, R.M., et al. (1995). Cost-effectiveness of thrombolytic therapy with tissue plasminogen activator as compared with streptokinase for acute myocardial infarction. *New England Journal of Medicine*, **332**, 1418–24.

McKenna, C., Hawkins, N., Claxton, K., McDaid, C., Suekarron, S., and Light, K. (2010). Cost-effectiveness of enhanced external counterpulsation (EECP) for the treatment of stable angina in the United Kingdom. *International Journal of Technology Assessment in Health Care*, **26**, 175–82.

Moher, D., Liberati, A., Tetzlaff, J., and Altman, D.G. (2009). Preferred reporting items for systematic reviews and meta-analyses: the PRISMA statement. *Open Medicine*, **3**, e123–30.

Neuhauser, D. and Lewicki, A.M. (1975). What do we gain from the sixth stool guaiac? *New England Journal of Medicine*, **293**, 226–8.

NICE [National Institute for Health and Care Excellence] (2013). *A guide to the methods of technology appraisal*. London, July. Available at: <http://www.nice.org.uk/aboutnice/howwework/devnicetech/guidetothemethodsoftechnologyappraisal.jsp> (accessed 5 December 2013).

Rawlins, M.D. and Culyer, A.J. (2004). National Institute of Clinical Excellence and its value judgments. *BMJ*, **329**, 224–7.

Schackman, B.R., Taffet Gold, H., Stone, P.W., and Neumann, P.J. (2004). How often do sensitivity analyses for economic parameters change cost–utility analysis conclusions? *PharmacoEconomics*, **22**, 293–300.

Schulz, K.F., Altman, D.G., and Moher, D. (2010). CONSORT 2010 Statement: updated guidelines for reporting parallel group randomized trials. *Open Medicine*, **4**, e60–8.

Sculpher, M.J., Poole, L., Cleland, J., et al. (2000). Low doses versus high doses of angiotensin converting enzyme inhibitor lisinopril in chronic heart failure: a cost-effectiveness analysis based on the Assessment of Treatment with Lisinopril and Survival Analysis (ATLAS) study. *European Journal of Heart Failure*, **2**, 447–54.

Tan-Torres Edejer, T., Baltussen, R., et al. (2003). *Making choices in health: WHO guide to cost-effectiveness analysis*. Geneva: World Health Organization.

Weatherly, H., Drummond, M.F., Claxton, K., Cookson, R., Ferguson, B., et al. (2009). Methods for assessing the cost-effectiveness of public health interventions: key challenges and recommendations. *Health Policy*, **93**, 86–92.

WHO [World Health Organization] (2014). WHO-CHOICE <http://www.who.int/choice/costs/CER_thresholds/en/> (accessed 29 January 2014).

Chapter 4

Principles of economic evaluation

4.1 Alternatives, costs, and benefits: some basics

The purpose of any type of economic evaluation is to inform decisions about which of the alternative courses of action available ought to be recommended, approved for widespread use, or reimbursed for specific groups of patients. The same analysis can also inform pricing as well as decisions about access and 'coverage'. For example, economic evaluation can identify whether a new drug ought to be approved for widespread use at its existing price, but it can also be used to identify the maximum price that the health care system can afford to pay and how this price will differ across the subgroups of patients that could benefit from it (see Section 4.2.2).

Informing decisions is the primary role of any type of economic evaluation, irrespective of the type of health care system or alternative views about which values ought to guide health care decisions (see Section 2.4 for a discussion of alternative views). Of course, alternative views about the purpose of health care and whether or not the health care system faces restrictions on the growth in health care expenditure will change the way an economic evaluation is conducted and how the results should be interpreted to inform decisions. To be useful for decision-making, any economic evaluation must consider four key issues. These are important irrespective of the values adopted or the type of health care system in which the analysis is to be used.

4.1.1 What are the alternatives?

Informing a particular decision requires the identification of the possible alternative courses of action that could be taken to improve the health of patients who find themselves in a particular situation and facing a choice between mutually exclusive alternatives: this is, 'either/or' decisions (see Section 2.3.1). Therefore, the alternatives may include different combinations or sequences of treatment and different ways in which an intervention can be used (e.g. what dose, or when to start and when to stop). The same principles apply for other types of interventions like diagnostic tests and public health programmes.

For simplicity we first explain how the results of economic evaluation can inform decisions when considering only two alternatives in Section 4.2. In most circumstances, however, more than two alternatives are available and in some circumstances there may be very many indeed. The implications for how economic evaluation should be reported and interpreted for decisions involving multiple alternatives is discussed in Sections 4.4.1 and 4.4.2.

Sometimes the choices faced are not mutually exclusive. For example, different decisions can be made about which interventions to offer to different subgroups of patients

with the same indication and within the same disease area. Equally, health care systems face choices about which interventions to make available across different areas of disease, relevant to different patient populations. These are not either/or decisions. How economic evaluation can be used to inform choices between non-mutually-exclusive alternatives is discussed in Section 4.4.3.

4.1.2 Which measure of benefit?

Improving health is often the stated objective of many health policies and is the objective that underpins most published economic evaluations. The alternative measures of health that are available were introduced in Section 2.4.1 and are discussed in detail in Chapter 5. There are a number of reasons why a generic measure of health-related quality of life (HRQoL) has advantages. First, it allows comparison of the health effects of alternatives that affect different aspects of health or have effects in a number of diseases. Secondly, it provides consistency with how other decisions relevant to other groups of patients with different diseases are made. Finally, it allows a comparison of health gained with the health expected to be lost elsewhere as a consequence of additional health care costs. Restricting attention to health and conducting cost-effectiveness analysis (CEA) based on a generic measure of health outcome (e.g. QALYs (quality-adjusted life-years)) could be justified by taking a social decision-making approach to the role of economic evaluation in health care (see Section 2.4.3). How decisions can be made based on the results of this type of CEA are set out in Sections 4.2 and 4.4.

A broader view is that the purpose of health care is to improve welfare. This suggests that health care decisions should be judged in the same way as any other public or private choice. Traditionally this view of welfare is founded on individuals' preferences and the monetary compensation that individuals are willing to offer to gain benefits or accept to incur losses. It is this view that underpins traditional welfarist cost–benefit analysis (CBA) (see Section 2.4.2 and Chapter 6). It requires the benefits of health care to be expressed as the equivalent amount of consumption; that is, the amount of money that an individual would be willing to pay (or to receive) in return for the (dis)benefits offered (see Section 4.3.3). The real rather than apparent distinction between CEA and CBA and the circumstances when they are equivalent are discussed in Section 4.3.4.

4.1.3 How can the costs and benefits of each alternative be estimated?

No single study is likely to provide all, or the only, evidence required to estimate the costs and health effects of the alternatives available, over the period of time during which they are likely to differ, or across the range of different subgroups of patients that could be identified (heterogeneity and subgroups are discussed in Chapters 8, 9, and 11). The evidence from relevant studies needs to be sought systematically, extracted, interpreted and then, where appropriate, combined, or synthesized to provide estimates of the key parameters (see Chapter 10). Decision-analytic models are commonly used as the structure within which evidence from different sources can inform the parameters which, in combination with explicit assumptions and judgments, provide estimates of costs and health effects (see Chapter 9).

4.1.4 What will be given up as a consequence of additional costs?

To decide whether the additional benefits offered by one alternative (compared to the other courses of action available) is sufficient to justify any additional costs depends critically on the value of what is given up by others as a consequence—the opportunity costs. Therefore, unless there is some assessment of the likely opportunity costs, the result of economic evaluation cannot be put to use to inform decisions—it remains only a description of costs and consequences. What is likely to be given up (whether health care for other patients or consumption opportunities in the rest of the economy) depends to some extent on the nature of the health care system, but is a key question irrespective of social values adopted and whether or not benefits are measured in health or the equivalent consumption. This question of how opportunity costs might be assessed, what a cost-effectiveness 'threshold' ought to represent and how it might be estimated is discussed in Section 4.3.

These issues were first introduced and discussed in Chapters 1 and 2. Addressing them in greater detail and identifying appropriate methods of analysis is the subject of subsequent chapters. This chapter presumes that they have been addressed using appropriate methods to estimate the expected costs and benefits of the alternatives courses of action available. The primary question that will be addressed in this chapter is how decisions can be informed using the results of this type of analysis.

4.2 Making decisions about health care

The concept of cost-effectiveness and how it might be represented was introduced in Chapter 3 (see Box 3.2). When faced with a choice between mutually exclusive alternatives, the questions is whether the additional or incremental health benefits of choosing one intervention rather than another are sufficient to justify the additional or incremental costs. In some circumstances the choice between the alternatives is clear. For example, if an intervention offers additional health benefits but at lower costs compared to the other alternatives available, it should be regarded as cost-effective and can be said to *dominate* (choosing this alternative would improve health outcome and reduce health care costs). Similarly, if the intervention is less effective (the incremental health benefits are negative) and has additional health care costs than one or more alternatives, then it can be said to be *dominated*. It is clearly not cost-effective because its use would reduce health outcomes and increase health care costs. However, if an alternative offers incremental health benefits but at some additional health care costs, the questions of whether it should be regarded as cost-effective and be approved for widespread use will depend on the value of what will be give up as a consequence.

4.2.1 Incremental cost-effectiveness ratios, thresholds, and net benefit

These situations are illustrated in Figure 4.1, which represents the cost-effectiveness plane for a choice between two alternatives, A and B (the concept of a cost-effectiveness plane was first introduced in Box 3.2). Alternative A, which might represent current

clinical practice, is at the origin, so the x-axis represents the incremental health benefit of alternative B and the y-axis the incremental costs of B. The situation in which B dominates A because it is more effective and less costly (positive incremental health benefits at negative incremental cost) is represented by points to the south-east of A (lower right). Similarly, the situation in which B is dominated by A because it is less effective and more costly (negative incremental health benefits at positive incremental cost) is represented by points to the north-west (upper left).

Figure 4.1 illustrates the situation in which B is estimated to offer incremental health benefits (Δh = 2 QALYs per patient treated) at some positive incremental cost. The incremental cost is the difference between the expected costs of using B rather than A. These incremental costs include the additional acquisition costs of B and any other costs associated with its use. They also include any expected resource savings such as from early recovery and discharge or avoiding subsequent costly clinical events (see Section 4.5.1 and Chapter 7). Consequently, the incremental costs might be greater or less then the acquisition cost of the intervention. For example, the incremental cost of a particular pharmaceutical intervention may differ from the price charged for the drug. Indeed, it is possible that an expensive drug might have negative incremental costs (reduce health care costs overall) if the subsequent resource savings are sufficient to offset the initial prescribing costs.

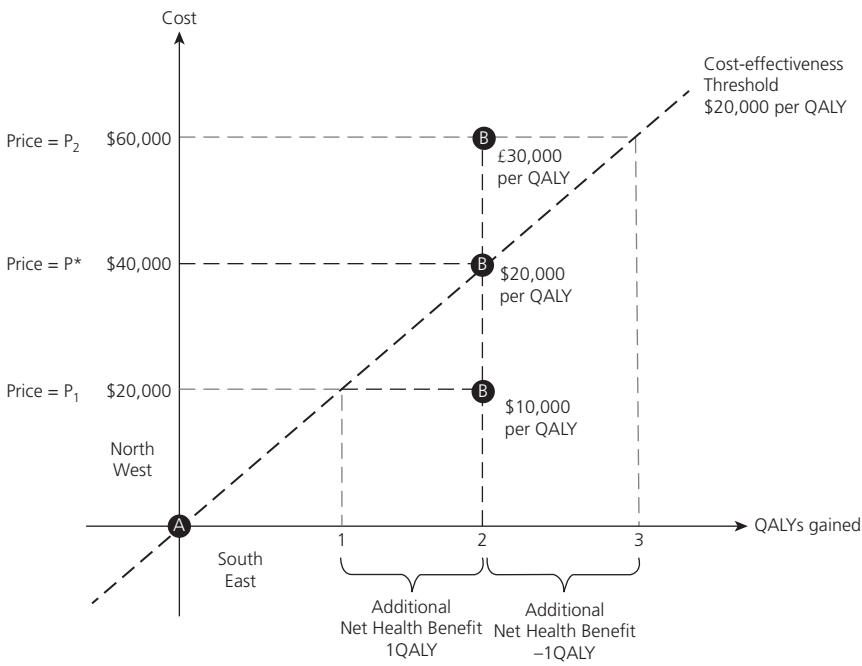


Fig. 4.1 ICERs, decisions, and net benefit.

Adapted by permission from BMJ Publishing Group Limited. *BMJ*, Claxton K. et al., Value based pricing for NHS drugs: an opportunity not to be missed?, Volume 336, pp. 251–4, Copyright © 2008, British Medical Journal Publishing Group.

At an acquisition cost (price) of P1 in Figure 4.1, the incremental costs of B (Δc) are estimated to be $20 000 per patient treated. These additional costs generate incremental benefits of 2 QALYs so we can say that B offers one QALY gained for every additional $10 000 spent. In other words the ratio of incremental cost to incremental effect, or the incremental cost-effectiveness ratio (ICER), is equal to $10 000 per QALY gained ($\Delta c/\Delta h$ = $20 000/2 QALYs). Although this is a useful summary of the cost-effectiveness of B compared to A, the question remains whether or not 2 QALYs gained justifies the additional $20 000; or equivalently whether an ICER of $10 000 per QALY is acceptable and B ought to be regarded as cost-effective.

To inform this decision some assessment must be made of what is likely to be given up as a consequence of the additional costs, and the value of what is forgone relative to the health benefits (i.e. the opportunity costs). In a health care system which faces some restrictions on the growth in health care expenditure these opportunity costs fall on health, because the additional costs are resources that will no longer be available to offer effective health care that would benefit other patients. That is, the opportunity cost is the health expected to be given up as a consequence of the incremental costs. The assessment of this type of opportunity cost is commonly described as a cost-effectiveness threshold (k) which can be compared to the ICER for B (Weinstein 2013; Weinstein and Zeckhauser 1973).

A cost-effectiveness threshold of $20 000 per QALY is represented by the rising diagonal in Figure 4.1. It means that every $20 000 of health care resources is expected to displace one QALY elsewhere in the health care system (see Section 4.3.2 for discussion of how thresholds have been used and might be estimated). Now it is possible to compare the health expected to be gained by using B (2 QALYs) with the health expected to be lost as a consequence of the additional costs of $20 000. At a threshold of $20 000 per QALY, these incremental costs are expected to displace one QALY elsewhere ($\Delta c/k$ = $20 000/$20 000). Therefore, at a price of P1, alternative B is expected to offer incremental net health benefits of 1 QALY ($\Delta h - \Delta c/k = 2$ QALYs gained minus 1 QALY displaced elsewhere), so it should be regarded as cost-effective. Asking whether B offers positive incremental net health benefits is entirely equivalent to asking whether the ICER for alternative B is less than the threshold. Since the ICER for B is less than the threshold ($\Delta c/\Delta h < k$) it indicates that B generates more health for a given amount of resource than the health care that is likely to be given up.

Cost-effectiveness can also be expressed as the equivalent health care system resources or incremental net monetary benefit. Instead of transforming health care costs into their health equivalent, the threshold can be used to establish how much additional health care resource would be required to generate the same health benefits elsewhere (i.e. $40 000 would be required to generate the 2 QALYs that are offered by B). The difference between this valuation of the incremental health benefits of B and the incremental costs is the incremental net monetary benefit. Therefore, once some assessment of a cost-effectiveness threshold is made (whether explicit or implicit) this can be used in two possible ways. One is to transform incremental health care costs into their health equivalent ($\Delta c/k$) so they can be compared to the incremental health benefits. The second is to transform incremental health benefits into their resource

equivalent ($\Delta h.k$) so they can be compared to incremental health care costs (Laska et al. 1999; Phelps and Mushlin 1991; Stinnett and Mullahy 1998).

We have introduced three equivalent ways of deciding whether or not an intervention is expected to be cost-effective, all of which can be illustrated in Figure 4.1:

1 At price P1, alternative B has an ICER of \$10 000 per QALY which is less than the cost-effectiveness threshold of \$20 000 per QALY ($\Delta c/\Delta h < k$).
2 The incremental net health benefit of B ($\Delta h - \Delta c/k$) is positive because the health gained (2 QALYs) exceeds the health equivalent of the additional health care costs (\$20 000/\$20 000 = 1 QALY).
3 The incremental net monetary benefit of B ($\Delta h.k - \Delta c$) is positive because the resources required to provide 2 QALYs elsewhere ($2 \times \$20\,000 = \$40\,000$) exceeds the incremental costs (\$20 000).

Therefore, the incremental net benefit to the health care system of alternative B at price P1 can be expressed as either 1 QALY (on the x-axis of Figure 4.1) or \$20 000 (on the y-axis of Figure 4.1) per patient treated. These three equivalent ways to consider cost-effectiveness are also summarized in Table 4.1. They will be revisited when considering the choice between multiple alternatives in Section 4.4.

There are some important implications from this material. First, it is not possible to make any statements about what is and what is not cost-effective without reference to a cost-effectiveness threshold that represents an assessment of opportunity costs (Johannesson and Weinstein 1993; Weinstein and Zeckhauser 1973). Secondly, some implicit or explicit assessment of the threshold is unavoidable because, when any decision is made, it implies a value for the threshold (Phelps and Mushlin 1991). Thirdly, in a resource-constrained health care system, health care costs really represent the health outcomes for other patients with competing claims on health care resources; therefore, decisions based on economic evaluation are really about identifying the alternative which offers the greatest net health benefits overall (Culyer et al. 2007; McCabe et al. 2008).

4.2.2 What price for a new technology?

It should be apparent that health care costs matter because they represent the opportunity to improve the health of other patients with legitimate claims on the health care system. Therefore, the cost of a health care intervention is just as important as how effective it might be. Although the health gains for the beneficiaries of an effective intervention

Table 4.1 A summary of decision rules

	Constraints on health expenditure		No constraints on health expenditure
	Health	Consumption	
ICER	$\Delta c/\Delta h < k$	$\Delta c/(v.\Delta h) < k/v$	$\Delta c/\Delta h < v$
Incremental net health benefit	$\Delta h - \Delta c/k > 0$	$\Delta h - \Delta c/k > 0$	Δh—$\Delta c/v > 0$
Incremental net money benefit	$k.\Delta h - \Delta c > 0$	$v(\Delta h - \Delta c/k) > 0$	$v.\Delta h$—$\Delta c > 0$

might be more readily identifiable compared to the health likely to be displaced elsewhere as a consequence of the additional costs, there seems little reason to treat those that are known or more easily identifiable differently to those that are not (see Section 2.4.3.2). Therefore, the price charged for a health technology (e.g. a new branded pharmaceutical) becomes just as important as how effective it might be (see Box 4.1).

Economic evaluation and the type of explicit assessment of cost-effectiveness discussed in Section 4.2.1 can also be used to identify the maximum price the health care systems can afford to pay for health care inputs and how this will differ across the subgroups of patients that could benefit from it (Claxton et al. 2008; Danzon and Economics 2014; Drummond et al. 2011). The link between price and cost-effectiveness is also illustrated in Figure 4.1. At a price of P1, alternative B is expected to be cost-effective and offers incremental net benefits to the health care system. Therefore, P1 is not the maximum price that the health care system could afford to pay for this technology. If the price was increased to P* so that the incremental costs are now $40 000, the ICER would be just equal to the threshold ($20 000 per QALY gained). At this point the health benefits of 2 QALYs are just offset by the health that is expected to be forgone elsewhere ($40 000/$20 000 = 2 QALY) and the incremental net health benefits would be zero. Similarly, the incremental costs of B are now just equal to the additional health care resources that would be required to generate the same additional health benefits elsewhere (2 × $20 000 = $40 000); that is, the incremental net monetary benefit is also zero.

At a higher price of P3, the incremental costs of B would be $60 000 and the ICER would be $30 000 per QALY gained, which is greater than the threshold. At this higher price, B is not cost-effective because the health that is likely to be displaced as a consequence ($60 000/$20 000 = 3 QALYs) exceeds the additional benefits offered (2 QALYs); that is, the incremental net health benefits are negative. Accepting this technology at this price would reduce overall health outcomes for the health care system by 1 QALY for every patient treated. Incremental net monetary benefit is also negative because the higher incremental costs of B are now greater than the additional health care resource required to generate the same additional health benefits elsewhere ($40 000 − $60 000 = −$20 000). This means that the resources required for B could be used to generate more health elsewhere in the health care system so approving B at this price would be equivalent to discarding $20 000 for every patient treated.

Therefore, the price at which the ICER is just equal to the cost-effectiveness threshold, or where the additional net benefits offered falls to zero, is the maximum the health care system can afford to pay. It is the price at which health expected to be lost as a consequence of the additional cost is just offset by the health expected to be gained. This makes clear that the amount a health care system that can afford to pay for a new technology will depend on the expected health benefits it offers to patients but also on a threshold that represents an estimate of the health impact of health care costs, as well as other costs and/or cost savings associated with its use.

4.3 **The cost-effectiveness threshold**

It should be quite clear that estimating the costs as well as the health effects of alternative interventions is essential to adequately inform decisions about their use. However,

Box 4.1 Pharmaceutical pricing and incentives for research and development

An assessment of cost-effectiveness can inform what price ought to be paid for new drugs. Furthermore, it can do so in a way that aligns the incentives manufacturers face when making investment decisions with the needs and constraints faced by the health care system. For example, the total value of a new drug indicated for a patient population of Q* each year, which is priced at P* in Figure 4.1, will be $P^* \times Q^*$. This total value will be received by the manufacturer in sales revenue each year over the remaining period of patent protection (at P* the health care system receives no net benefit; it is the maximum or *value-based price*). This predictable signal allows manufacturers to consider whether the expected benefits likely to be offered by their product will command a price that will provide a return on the investment required (Claxton et al. 2008).

The health care system will gain net benefit in the longer run when the patent expires, but only if cheaper generics enter the market and as long as prescribing switches to cheaper generic versions of the brand. However, new branded pharmaceuticals may well be developed and launched in the future as well. Therefore, it is important that they are compared to the cheaper generic versions of previously branded drugs when assessing their cost-effectiveness and in identifying the maximum value-based price. Whatever consumption value of health is deemed appropriate, it will not change how much the health care system can afford to pay for a new drug when there are constraints of health care expenditure (see Section 4.3.3) (Claxton et al. 2011a; Jayadev and Stiglitz 2009).

The value-based price for a new drug is likely to differ by subgroups within an indication and for the same drug used to treat different conditions (e.g. P* will be higher where it offers greater incremental health benefits). Therefore, prices based on an assessment of cost-effectiveness can also inform price and volume agreements. This also provides a framework to consider some of the issues around the global pricing of pharmaceuticals, since value is likely to differ between different health care systems that face different resource constraints (i.e. have different cost-effectiveness thresholds; see Section 4.3). Being able to charge different prices in different health care systems reflecting differences in value has a number of advantages. It means that lower-income countries do not necessarily face a choice of either paying global prices that exceed the value to their health care system or not having access to the technology (Danzon et al. 2011). Manufacturers also have an interest in being able to maintain differential prices as this means they will be able to maximize the returns to their investment (Danzon et al. 2012). The difficulty is being able to maintain price differentials when many health care systems base pricing decisions not on an assessment of value but on the prices charged in other markets. Nonetheless, economic evaluation provides the analytic framework within which a more rational, evidenced-based approach to domestic and global pricing can be explored.

reporting appropriate ICERs is not sufficient. Any statement about what is and what is not cost-effective rests on some assessment of an appropriate threshold, whether its value and evidential foundation is made explicit or not. Indeed, any decision to adopt or to reject an intervention which offers health benefits but imposes additional costs implies possible values for a threshold. In other words, some implicit or explicit assessment of a cost-effectiveness threshold is unavoidable. The only question is whether this should be done explicitly and be informed by such evidence that is available. If decisions which have impact on others should be evidence based, coherent, and accountable, it seems that an explicit and evidence-based approach to establishing an appropriate threshold has much to commend it (Culyer et al. 2007).

The importance of the concept of a cost-effectiveness threshold has been recognized for some considerable time (Neumann et al. 2014; Weinstein and Stason 1977; Weinstein and Zeckhauser 1973), and a number of possible values have been suggested in different contexts (Laupacis et al. 1992; Neumann et al. 2014; Newall et al. 2014). One problem with proposed values has been a lack of clarity about what they ought to represent, which turns on separate questions of fact and value (McCabe et al. 2008). The other problem has been that, until recently, there has been limited empirical evidence on which to found an assessment of the health consequences of additional health care costs (see Section 4.3.2).

4.3.1 Questions of fact and questions of value

Before considering what evidence might be available to inform an appropriate threshold, it is essential to have clarity about what the threshold ought to represent and, therefore, what type of evidence to consider. What it ought to represent will depend on what type of effects additional costs are expected to have. In other words, the key consideration is where the opportunity costs are expected to fall (a question of fact) and how the effects should be valued (a question of value). Importantly, this requires consistency in the way the benefits offered are measured and valued and how opportunity costs are to be identified, measured, and valued. For example, which of the alternative broad approaches to social value discussed in Section 2.4 is adopted also implies how the opportunity costs should be valued (also see Sections 4.3.2.4 and 4.5.3).

In Section 4.3.2 we examine the implications of restrictions on the growth in health care expenditure (when opportunity costs fall on health); and Section 4.3.3 considers situations when there are no restrictions on the growth in health care expenditure (when opportunity costs fall outside health care). In Section 4.3.4 we examine the implications of restrictions on health care expenditure but when a broader welfarist view (see Section 2.4.2) is taken. In doing so, we illustrate why the distinction between questions of fact and value is so important for the interpretation of different types of economic evaluation. The more common situation, in which some opportunity costs fall on health and some fall outside the health care system, is examined in Section 4.5.3.

4.3.2 Opportunity costs fall on health

Most health care systems face some constraints on the growth in health care expenditure. These constraints might be in the form of an explicit administrative budget

(e.g. the NHS in the United Kingdom or Veterans Administration in the United States); limits on increasing funding for health care through additional taxation; some limits placed on the growth in premiums/payment in a social insurance system; or some process of reviewing the benefits package and copayments as health care costs rise. Therefore, a cost-effectiveness threshold representing opportunity costs in terms of health is relevant in most health care systems. A few health care systems have revealed something about the type of threshold values likely to be used when making decisions. However, it is only in the UK that the National Institute for Health and Care Excellence (NICE) has made clear that the threshold ought to represent the health consequences of additional NHS costs. Furthermore, since 2004, NICE has published a range for the threshold used in its deliberative decision-making process (NICE 2013; Rawlins and Culyer 2004).

Importantly, however, the NICE threshold range (currently £20 000–£30 000 per QALY gained) was founded on the values implied by the decisions that had been previously made rather than on evidence about the likely health consequences of decisions which impose additional costs on the NHS. Over more recent years the evidence suggests that NICE does not reject technologies with ICERs below its upper bound of £30 000 per QALY and some analysis suggests that NICE is approving technologies with ICERs substantially higher than this (Dakin et al. 2014; Devlin and Parkin 2004).

4.3.2.1 What evidence is available?

There is little evidential foundation to any of the cost-effectiveness thresholds, whether these are explicit, implied, or suggested. However, the expected health effects of making decisions that impose addition costs on the health care system is an empirical question that can be addressed. Only a few studies have attempted to undertake such research. One approach is to ask what other health care was actually displaced and what the health effects of these 'disinvestment' decisions were (Appleby et al. 2009). There are several problems with this approach. First, it is a challenge to identify which specific changes took place, where and who to ask, and how to be sure that any changes were actually caused by the additional costs. Secondly, it is difficult to estimate the health effects of the changes that actually took place as a consequence of having to find additional resources (with little and only observational data available). The challenges of conducting this type of detailed local analysis for a health care system are immense but, thankfully, unnecessary once it is recognized that it is only the expected health effects of a change in available resources that is needed, rather than what specific changes took place, or where, by and for whom, with what specific health effects. In other words, the problem of estimating a cost-effectiveness threshold is the same as estimating the relationship between changes in health care expenditure and health outcome (Bokhari et al. 2007; Martin et al. 2008, 2012; Moreno-Serra and Smith 2012, 2015). This is the approach that was taken in research conducted in the United Kingdom (see Box 4.2 and Table 4.2) (Claxton et al. 2015b).

The estimates reported in Table 4.2 suggest that approving new technologies with ICERs of £30 000 or even £20 000 is likely to do more harm than good since more health is expected to be lost elsewhere than is gained by approving the intervention. For example, an intervention with an ICER of £30 000 that would cost £10 m per year to implement fully would be expected to generate up to 334 additional QALYs each

Box 4.2 Estimating the cost-effectiveness threshold

The problem of estimating a cost-effectiveness threshold is the same as estimating the relationship between changes in health care expenditure and health outcome. This is the approach that was taken in research conducted in the United Kingdom. The study used national data on expenditure and outcomes in different areas of disease (programme budget categories) reported at a local level (Claxton et al. 2015b; Martin et al. 2008). By exploiting the variation in expenditure and mortality outcomes, the relationship between changes in spending and mortality was estimated while accounting for endogeneity.

The cost per death averted was estimated to be £114 272 (see Table 4.2). With additional information about age and gender of the patient population, however, these mortality effects were expressed as a cost per life-year threshold (£25 241 per life-year). These life-year effects were adjusted for HRQoL with additional information about HRQoL norms by age and gender, as well as the HRQoL impacts of different types of disease (£30 270 per QALY based on population norms, see Table 4.2). By using the effect of expenditure on the mortality and life-year burden of disease as a surrogate for the effects on a more complete measure of health burden (i.e. that also includes HRQoL burden), a cost per QALY threshold that reflects the likely impact of expenditure on both mortality and morbidity was estimated (£12 936 per QALY, see Table 4.2) (Claxton et al. 2015b).

The estimates of the threshold reported in Table 4.2 are founded on estimating health effects (on deaths, life-years, and HRQoL) in different disease areas (see Table 4.3). These 23 programme budget categories (PBCs) are made up of International Classification of Diseases (ICD) codes. Therefore, with information about the age and gender distribution in these different diseases as well as estimates of incidence and duration of disease, it was possible to estimate the severity of disease (QALY loss) associated with the average of the displaced QALY effects (a QALY burden of 2.07 QALYs on average). Other work by the UK Department of Health estimated the wider social benefits or net production effects of changes in length and HRQoL by age, gender, and type of disease. This facilitated an estimate of the net production impact associated with the average of displaced QALY effects (£11 611 per QALY including marketed and non-marketed activities) (Claxton et al. 2015a).

year. However, finding the £10 m required from existing resources would be expected to lead to a loss of 773 QALYs each year (see Table 4.3). Table 4.3 indicates how these 773 QALYs are likely to be made up including (1) the type of health effects (e.g. 51 additional deaths and 233 life-years); and (2) where these different types of health effects are likely to occur. For example, Table 4.3 suggests there are greater life-year effects in cancer and circulatory diseases and greater quality of life effects in mental health and respiratory and neurological diseases.

These results, and any other estimates that might be possible in the future, are subject to uncertainty. The policy question, however, is whether an appropriate threshold

Table 4.2 Cost-effectiveness thresholds for the UK NHS (2008–09)

	Cost per death averted	Cost per life-year	Cost per QALY (life-year effects only)	Cost per QALY
Life-years per death averted	–	*4.5*	*4.5*	*4.5*
QALYs per death averted	–	–	*3.8*	*12.7*
11 PBCs (with mortality)	£105 872	£23 360	£28 045	£8308
All 23 PBCs	£114 272	£25 214	£30 270	**£12 936**

PBCs, programme budget categories; QALYs, quality-adjusted life-years.

Source: data from Claxton, K. et al., *Methods for the estimation of the NICE cost effectiveness threshold*, Center for Health Economics Research Paper 81, University of York, UK, Copyright © 2015.

Table 4.3 The health impact of £10 m

Totals	**Change in spend**	**Additional deaths**	**LY lost**	**Total QALY lost**	**Due to premature death**	**Quality of life effects**
	10 (£m)	**51**	**233**	**773**	**150**	**623**
Cancer	0.45	3.74	37.5	26.3	24.4	1.9
Circulatory	0.76	22.78	116.0	107.8	73.7	34.1
Respiratory	0.46	13.37	16.1	229.4	10.1	219.3
Gastro-intestinal	0.32	2.62	24.7	43.9	16.2	27.7
Infectious diseases	0.33	0.72	5.3	15.7	3.6	12.1
Endocrine	0.19	0.67	5.0	60.6	3.2	57.3
Neurological	0.60	1.21	6.5	109.1	4.3	104.8
Genito-urinary	0.46	2.25	3.3	10.6	2.1	8.5
Trauma and injuries*	0.77	0.00	0.0	0.0	0.0	0.0
Maternity & neonates*	0.68	0.01	0.4	0.2	0.2	0.1
Disorders of blood	0.21	0.36	1.7	21.8	1.1	20.7
Mental health	1.79	2.83	12.8	95.3	8.3	87.0
Learning disability	0.10	0.04	0.2	0.7	0.1	0.6
Problems of vision	0.19	0.05	0.2	4.2	0.2	4.1
Problems of hearing	0.09	0.03	0.1	14.0	0.1	13.9
Dental problems	0.29	0.00	0.0	6.8	0.0	6.8
Skin	0.20	0.24	1.1	1.9	0.7	1.2
Musculo-skeletal	0.36	0.39	1.8	23.2	1.2	22.1

Table 4.3 (continued) The health impact of £10 m

Totals	Change in spend	Additional deaths	LY lost	Total QALY lost	Due to premature death	Quality of life effects
	10 (£m)	**51**	**233**	**773**	**150**	**623**
Poisoning and A&E	0.09	0.04	0.2	0.8	0.1	0.7
Healthy individuals	0.35	0.03	0.2	0.7	0.1	0.6
Social care needs	0.30	0.00	0.0	0.0	0.0	0.0
Other	1.01	0.00	0.0	0.0	0.0	0.0

A&E, accident and emergency; LY, life-years; QALYs, quality-adjusted life-years.

Source: data from Claxton, K. et al., *Methods for the estimation of the NICE cost effectiveness threshold*, Center for Health Economics Research Paper 81, University of York, UK, Copyright © 2015.

should be based on an assessment of the balance of existing evidence. Maintaining a cost-effectiveness threshold that is too high will be expected to damage health outcomes. Adopting a threshold that is too low will also reduce health because access to some health care will be restricted when it need not be. Therefore, there seems little reason to maintain a threshold that is higher than the balance of evidence suggests, despite the obvious stakeholder interest in doing so.

So far we have described a threshold based on what the health care system currently does when responding to increases or decreases in available resources rather than what, in principle, it could do. For example, a decision-maker with the remit to reorganize the entire health care system (or a specific part of it) rather than make decisions about the choice between specific mutually exclusive alternative interventions would need much more information (see Section 4.5.3) (Eckermann and Pekarsky 2014). Similarly, some resources might currently be devoted to activities that are not particularly effective and could be (but are currently not) subject to disinvestment with little impact on health outcomes. Consequently, as the health care system changes (resources, productivity, and prices of inputs change), including the quality of other investment and disinvestment decisions, the threshold will also change.

4.3.2.2 Is the threshold likely to change over time?

To make a decision about an intervention that has costs and health benefits only in the current period (t), an estimate of the threshold relevant to this period is required. This reflects current expenditure, costs of health care inputs and productivity (the nominal threshold relevant to period t). The question is whether estimates of the threshold, necessarily from retrospective data, are a useful guide to what the current threshold is likely to be? It might be tempting to assume that the threshold will necessarily increase with growth in health care expenditure because, other things equal, the investments that become possible with more resources will tend to be less valuable (i.e. the better investments with lower cost per QALY should have already been made). It might also be tempting to assume that the nominal threshold

will increase at a similar rate to the average prices of health care inputs or other goods and services.

The problem with these assumptions is that other things are not equal. For example, the productivity of health care (the health gained for a unit of health care resource) is likely to improve with improvements in medicine, the ways in which services are delivered, and innovation in health technologies as well as cheaper prices for some inputs (e.g. cheaper generic versions of branded drugs). Changes in the determinants of health outside health care will also influence how the threshold might change over time. For example, healthier lifestyles and a better environment that improves life expectancy will mean that any reduction in mortality from health care will gain more life-years. On the other hand, a reduction in the baseline risk of events (e.g. myocardial infarction) will mean less absolute health effects from health care that reduces the relative risk of these events or the mortality and morbidity associated with them.

For all these reasons, whether or not the threshold has fallen or is likely to rise, fall, or stay constant with rising expenditure and prices is an empirical question. There is some limited empirical evidence from the UK,where estimates of the threshold for different years of expenditure showed no evidence of any growth (in either nominal or real terms) at a time when total expenditure increased in real terms and so did NHS prices (Claxton et al. 2015b). Consequently, it would be useful to re-estimate the threshold periodically to ensure that the most recent estimate does indeed reflect the likely opportunity costs given current resources, productivity and the nature of local decisions.

Decisions made now often impose costs in future periods (see Section 4.5.1). Therefore, some view about how the threshold is likely to change in the future is also needed. In most economic evaluations the threshold is often assumed (implicitly) to be constant in real terms. This might be reasonable when considering a choice between alternatives which offer similar distributions of costs and health benefits over time. However, when comparing interventions that offer future benefits but require an investment of resources today (e.g. a public health intervention) with one that imposes costs in future periods (e.g. long-term drug treatment) the question of how the threshold is likely to change becomes important (see Section 4.5.2).

4.3.2.3 Affordability and cost-effectiveness

The cost-effectiveness threshold should represent what is expected to be given up in order to be able to afford the implementation of an alternative that imposes additional costs. It is also evident that it is impossible to make meaningful statements about what is and is not cost-effective without specifying the threshold. Therefore, to say that an alternative is cost-effective but not affordable is really saying that the wrong threshold is being used to judge cost-effectiveness because it does not reflect the scale and value of what must be given up to afford to implement the alternative. This is maybe because the threshold is simply too high. For example, a reimbursement authority or national or global advisory body may recommend an alternative as cost-effective, but the local health care providers who face the task of disinvestment to afford it may believe that what must be given up exceeds the benefits offered.

In other circumstances the notion of affordability (low total additional cost of implementing an alternative) is sometimes cited as a criterion that might offset a lack

of cost-effectiveness. This may be considered the case in the context of rare diseases. Although the ICERs of some interventions (e.g. some new drugs) for rare diseases can be very high, as the total cost is low because of the rarity of the disease, it may be felt that they should be approved nonetheless, i.e. that 'they might not be cost-effective but they are affordable'. Although drugs for rare diseases pose particular challenges (Drummond et al. 2007), the problem is that a modest budget impact just changes the scale of *both* health that is displaced *and* health that is gained. For example, an intervention with an ICER of £50 000 with a total additional cost of £1 m to implement fully would be expected to offer health benefits of 20 QALYs each year. However, although $1 m might be regarded as modest and affordable compared to fully implementing a similar intervention in a more common disease, it would nonetheless be expected to lead to a loss of 77.3 QALYs elsewhere (see Table 4.3).

The net effect of accepting affordability as a reason to approve this intervention would be a net loss of over 57 QALYs falling on other patients. This might be acceptable if there are aspects of outcome that are regarded as especially important for the beneficiaries, such as severity and burden of disease. However, if there are additional considerations that ought to apply, they should apply whether the total cost is small or large, or necessarily whether the disease is rare or common (McCabe et al. 2005). The real question is not affordability, but whether these additional considerations for the beneficiaries are worth the net loss of 57.3 QALYs for other patients.

The scale of the total additional cost required to fully implement an intervention for its target population is, however, relevant to the proper assessment of the value of what is likely to be given up. The previous discussion of the threshold and the evidence available represents the health consequences of small (or marginal) changes in expenditure relative to total spending. This is the case whether these are increases (cost savings) or decreases (additional costs) in resources available for other health care activities. An informed decision-maker concerned with value for money is likely to displace the least valuable of those activities available for disinvestment. Therefore, to accommodate greater budget impact, more valuable activities would need to be displaced. In other words implementing an intervention with greater net budget impact is likely to displace not just more health but proportionately more health per $. That is, the threshold for interventions with greater total additional costs is likely to be lower than for small or marginal net budget impacts (Birch and Gafni 1992, 2013). There is some limited evidence that supports what might be expected, with lower thresholds in geographic areas of the NHS that were under greater budgetary pressure compared to those that were better resourced, which tended to be investing (Claxton et al. 2015b).

4.3.2.4 Other aspects of value

Consistency in the way benefits and opportunity costs are identified, measured, and valued is important. This is because, without such consistency, decisions may be self-defeating as the valuable attributes of benefit offered by an intervention may be more than offset by the same attributes of benefit that are given up as consequence of the additional costs. For example, improvements in health in a disease which is regarded as particularly severe or is associated with greater burden (i.e. a greater loss of QALYs) might be regarded as more important and carry greater weight than the same health

gain in a disease that imposes less of a burden. Similarly, health care interventions also have non-health effects on the wider economy which might include patients' and their carers' ability to contribute through unpaid and paid activities (see Section 4.5.3). Including such additional aspects of value may not lead to decisions that improve overall value unless it is also possible to assess the same type of attributes that are likely to be lost as a consequence of an intervention imposing additional costs.

Table 4.4 provides an example of this issue. The table compares the attributes of a proposed investment (ranibizumab (Lucentis) for the treatment of diabetic macular oedema) with the attributes of the disinvestments that are expected to occur elsewhere (see Box 4.2). The ICER for the subgroup of patients where the drug was likely to be most cost-effective was approximately £25 000 per QALY gained, prior to the commercial in confidence discounts that were offered during NICE's appraisal of this technology (NICE 2011). Estimates of the incidence of the disease suggested that the total additional cost to the NHS of approval restricted to this subgroup would be just over £80.6 m per year, and that 3225 QALYs would be gained in the relevant ICD code with a disease burden of 2.68 QALYs (see Box 4.2) (Claxton et al. 2015b). These health effects in this patient population were associated with net production benefits of £88.4 m (£27 421 per QALY gained on average for this ICD code). These attributes of benefit can be compared with the losses associated with the effects of the disinvestments that are likely to take place in Table 4.4. In this example the question is whether the expected net loss of almost 3000 QALYs, which is made up of over 400 additional deaths and a loss of over 1800 life-years (see Table 4.3), is worth the additional £16.6 m of net production gains in the wider economy, or whether the type of health gained is worth almost twice as much as the health lost because the burden of disease is almost 30% higher in this group (see Box 4.2) (Claxton et al. 2015a).

In principle, explicit weights could be assigned to these attributes. For example, weights could be used that represent how much quality-adjusted life expectancy might be given up in low-burden conditions in exchange for smaller QALY gains in diseases that have a more significant burden (see Chapter 6). The weights that

Table 4.4 Attributes of investment and expected disinvestment

Attributes	Investment	Disinvestment	Net effects
	Lucentis for diabetic macular oedema	Expected effects	
Deaths	0	–411	–411
Life-years	0	–1864	–1864
QALYs	3225	–6184	–2959
Severity of disease QALY loss per patient	2.68	2.07	0.61
Net production benefits	£88.4 m	–£71.8 m	£16.6 m

might be attached to non-health effects could be based on the consumption gains (compensation) that would be regarded as equivalent to the loss of 1 QALY (see Section 4.3.3). If such weights are explicitly specified, then a weighted QALY threshold could be estimated and used to compare to an ICER based on weighted QALY gains, which could also include the health equivalent of net consumption effects. By rearranging terms this is equivalent to adjusting the 'basic' cost per QALY threshold that can be compared to an ICER based on unweighted QALYs (the adjustment would be the ratio of weights attached to QALY gained to weights attached to the QALYs lost). Therefore, even if decision-makers are unwilling to specify explicit weights, consistent and accountable deliberation still requires an assessment of the other aspects of benefit that are likely to be forgone and a representation the trade-offs at stake when decisions are made (see Box 4.3) (Peacock and Mitton 2013; Peacock et al. 2007).

Box 4.3 Multi-criteria decision analysis

Multi-criteria decision analysis (MCDA) has been proposed as a means of taking account of a number of different aspects or attributes of benefit (Baltussen and Niessen 2006; Devlin and Sussex 2011). In fact, measures of HRQoL are a form of MCDA. For example, the EQ-5D is a form of MCDA with six criteria (length of life plus five different attributes of quality) with three performance scores for each. The attributes of investment illustrated in Table 4.4 also represent how MCDA can be conducted with three criteria (QALYs, severity, and wider social benefit), with proper account taken of the attributes of benefit that are likely to be forgone due to the additional costs. Three questions are important when considering how MCDA should be conducted:

- *What criteria should be used in MCDA?* Criteria should represent attributes of benefit that are valued alongside health gain. Each should make an independent contribution to benefit to avoid double counting. The means of measuring performance against each attribute should be prespecified, including the evidence that would support particular performance scores. The additional cost of an intervention is not a criterion (it is not an attribute of benefit). Rather, these costs are resources required to achieve an improvement in the composite measure of benefit. The costs indicate the scale of benefit that will be given up elsewhere.
- *How should weights be assigned to performance against each criterion?* The performance scores are not weights that represent the relative value of attributes, so should not be simply added up. Some have suggested that weights might emerge during the decision-making process, but this might not offer predictability or consistency, and risks strategic behaviour. The weights ought to represent how much one is willing to give up of one attribute to achieve an improvement on another. These types of weights can be estimated using choice-based methods to elicit preference (see Chapters 5 and 6).

Box 4.3 Multi-criteria decision analysis (continued)

- *What attributes of benefit are lost due to additional costs?* Including cost-effectiveness as a criterion cannot do this. Without a proper assessment of the other attributes of benefit forgone, decisions may reduce both health and the other attributes of benefit that originally motivated the use of MCDA. For example, in Table 4.4 one might conclude that the investment offered wider social benefits of £88.4 m when, in fact, the net wider social benefits were £16.6 m.

Therefore, the task of conducting MCDA correctly is considerable and it should not be regarded as a simple alternative to CEA as the same issues and methods apply except that, for MCDA, other attributes of benefit are considered in addition to health outcome (Peacock et al. 2007). If not done properly, instead of making decisions that improve a composite measure of benefit, which better represents society's preferences, it may actually reduce it (Claxton et al. 2015a).

4.3.3 Opportunity costs fall on consumption

If there is no explicit budget constraint or there are no other restrictions placed on the growth in health care expenditures in the health care system, then any additional costs associated with a new intervention to be adopted by that system will simply increase the total amount of health care expenditure. In these circumstances the additional costs will not (directly) displace health care for other patients. Nonetheless, additional costs will displace other socially valuable activities, so even if the primary purpose of health care is regarded as improving health, an intervention that only imposes costs outside the health care system cannot be treated as if it is 'free'. The additional costs will mean that the resources required will not be available for use elsewhere outside health care. In other words, the opportunity costs fall on the consumption of other goods and services rather than health. These reduced consumption opportunities might fall entirely on the patients who can benefit from the use of the intervention if they must pay the additional costs themselves. In other circumstances, however, it will be others who will bear these opportunity costs; for example, though higher health insurance premiums paid by other patients.

Now a decision about the use of the intervention will turn on whether the health benefits are regarded as more valuable than the consumption losses that will be incurred. This requires some assessment of how much additional consumption would be required compensate for a reduction in health, or what reduction in consumption would be regarded as equivalent to an improvement in health. Therefore, with no constraints on health care expenditure, some assessment of a consumption value of health (v) is required, rather than an estimate of the health opportunity costs (k). In these circumstances, there are three equivalent ways of deciding whether or not an intervention is expected to be worthwhile (see Table 4.1 and the similarities with Section 4.2.1):

- The ICER for an intervention is less than v. Here the health gained is more valuable than the incremental costs, which represents the consumption opportunities rather than the health that will be lost, ($\Delta c/\Delta h < v$).
- The equivalent consumption value of the health benefits ($v.\Delta h$) is greater than the consumption costs (Δc), so the incremental net consumption benefits are positive, ($v.\Delta h - \Delta c > 0$). This is sometimes described as net present value in cost–benefit analysis and can also be expressed as a benefit:cost ratio (($v.\Delta h)/\Delta c$) or a cost:benefit ($\Delta c/(v.\Delta h)$) ratio (Phelps and Mushlin 1991; Sugden and Williams 1979).
- If the health equivalent of the consumption losses ($\Delta c/v$) is less than the health benefits then the incremental net health benefit is positive ($\Delta h - \Delta c/v > 0$).

Estimates of a social consumption value of health (v) can be revealed in situations where people have made actual choices in which health is valued implicitly as an attribute in other markets, such as the trade-off between risk and wages in the labour market. Alternatively, such estimates can be derived from experiments which offer people hypothetical choices to try to find how much they are willing to pay for health or the collection of benefits offered by an alternative; or what trade-offs they would be willing to make between health and a range of other attributes, one of which can be valued on money terms (see Chapter 6). There is a range of estimates of values that might be used in different circumstances based on different methods and founded on different views about what values ought to inform social rather than individual choices (Ryen and Svensson 2014). These methods are discussed at greater length in Chapter 6.

4.3.3.1 What about the health effects of increased health care costs?

Even when there are no restrictions on the growth in health care expenditure, reimbursing or covering a higher-cost intervention will increase the costs of private insurance and/or out-of-pocket expenditure. For instance, there can be increases in direct costs, premiums, or higher copayments and deductibles. This will inevitably reduce access and health outcomes for some. For example, individuals may be unable to afford higher premiums, copayments, or the direct costs of care, and employers may be unwilling to offer health insurance at all or may select a more restricted benefits package. Therefore, even when there are no restrictions on the growth in health care expenditure, opportunity costs will fall on both consumption (those who are able and willing to pay the higher costs) and indirectly on health as well (those unable or unwilling to pay).

So, when would it be appropriate to only focus on the consumption opportunity costs (represented by v) and disregard any indirect health opportunity costs (represented by k)? This depends on whether one believes that the value of the health effects is fully reflected in the choices individuals make in health care and insurance markets. It will depend on one believing that the health lost by those who are unwilling or unable to pay the higher costs is less valuable than the increase in costs, but that the health gained by those willing and able to pay the higher costs is more valuable (Pauly 1995). This turns on the question of whether patients are making fully informed choices in an

undistorted market (see Section 2.4.2) and, even if they are, whether such individual choices are consistent with other social objectives and values (e.g. see Section 2.4.3) (Brouwer et al. 2008). Since neither of these conditions is likely to be met, the opportunity costs that fall on health, even when there are no constraints on health care expenditure, cannot be disregarded. Therefore, some assessment of a cost-effectiveness threshold (k) is required even in these circumstances.

4.3.4 What are the real distinctions between CEA and CBA?

It should already be evident that the distinction between CEA and cost–benefit analysis (CBA) seems more apparent than real (Garber and Phelps 1997; Phelps and Mushlin 1991; Weinstein and Manning 1997). Although CBA measures benefits in terms of their consumption equivalent and CEA represents health benefits in natural units or measures of HRQoL, these are valued in monetary terms when cost-effectiveness is assessed. For example, comparing an ICER to a threshold that represents health displaced when there are constraints on health care expenditure ($\Delta c/\Delta h < k$) is equivalent to comparing the resource equivalent of the health benefits ($k.\Delta h$) with the health care costs (Δc) (see Table 4.1). Health is inevitably valued in monetary terms; in this case it is valued in health care resources rather than consumption.

In these circumstances, a CBA that valued the health benefits using a consumption value of health would lead to the same decision as a CEA if the health opportunity costs are properly accounted for. Not only must the health gained be valued at its equivalent consumption value ($v.\Delta h$), the health displaced must also be valued using the same consumption value of health ($v.(\Delta c/k)$). Whatever value of v might be chosen and how much greater than k it might be, the same decision would be made because v simply scales benefits and the health opportunity costs to the same extent.

This can be illustrated by reconsidering the example in Figure 4.1 which was discussed in Sections 4.2.1 and 4.2.2 and Box 4.1. For example, if \$60 000 per QALY was deemed to be an appropriate consumption value of health (v) then the 2 QALYs gained by the intervention would be valued at \$120 000. However, at a price of P* the additional costs of \$40 000 fall on limited health care resources and will still displace 2 QALYs (k = \$20 000). Therefore, the opportunity costs fall on health, not consumption, so they should also be valued at \$120 000 (2 × \$60 000) not \$40 000. The fact that v might be greater than k makes no difference to whether or not the technology should be regarded as worthwhile at each price, or the maximum the health care system can afford to pay for this technology.

Applying a consumption value of health to inform this choice without accounting for the fact that the opportunity costs fall on health rather than consumption would lead to a decision that would not only reduce health but also net consumption value too. In this example, ignoring the threshold, k, would lead to conclusion that the technology is worthwhile at P3, because the ICER of \$30 000 is less than v (\$60 000). This would wrongly suggest that it would offer positive net consumption benefits of \$60 000 (2 × \$60 000 − \$60 000) per patient treated. However, this ignores k and the real opportunity costs. It would, in fact, result in a net loss of 1 QALY (2 − (\$60 000/\$20 000)) and a net consumption loss of \$60 000 ((2 × \$60 000) − (3 × \$60 000)) per patient.

Ignoring the impact of constraints on health care expenditure also leads to paying too much for health technologies. For example, one would conclude that $120 000 is the maximum the health care system can afford to pay for this technology (2 × $60 000) when in fact the maximum is only $40 000. The consequences of paying for this technology based on an estimate of v rather than k would be a loss of 4 QALYs per patient treated (2-($120 000/20 000)), which is equivalent to a net consumption loss of $240 000 (see Box 4.1) (Claxton et al. 2011a; Jayadev and Stiglitz 2009; Jena and Philipson 2007).

As discussed in Section 2.4.2 and developed in more detail in Chapter 6, there may be aspects of health or other non-health benefits that might not be fully reflected in measures of HRQoL, but that might be captured in a measure of benefit (B) based on consumption value or willingness to pay (i.e. $B \neq v.\Delta h$). In these circumstances, however, the consumption value of those aspects of health and non-health benefit that will be displaced as a result of constraints in the growth of health care expenditure also needs to be considered (see *other aspects of value* in Section 4.3.2.4 and Box 4.3). Therefore, some assessment of a threshold that reflects the opportunity costs associated with any restrictions on health care expenditure is still required even if measures of health benefit are rejected in favour of measures based on compensation and willingness to pay (see Chapter 6). Adopting CBA in favour of CEA does not avoid the question of what an appropriate threshold might be (Sculpher and Claxton 2012).

In the context of CBA, a threshold representing opportunity cost can be represented as critical benefit:cost ratio (v/k) or cost:benefit ratio (k/v) that must be achieved before an intervention can be regarded as worthwhile (see Table 4.1). For this reason, establishing an appropriate willingness to pay for health or a consumption value of a QALY will not be sufficient if there are restrictions on the growth in health care expenditure. Even if there are no restrictions it will not be sufficient if the value of the indirect health and other effects of increased health care costs are not fully reflected in individual choices in the market for health care.

4.3.4.1 Is health care spending 'optimal'?

Therefore, some assessment of a threshold that represents health opportunity costs is always necessary when there are restrictions on health care expenditure, and will also be sufficient if all costs fall on the health care system. However, the threshold cannot directly inform the broader question of whether current restrictions are too tight and whether expenditure on health care ought to be increased. Observing that an estimate of the consumption value of health is higher than the amount of health care resource required to improve health ($v > k$) would suggest that the health care system is not meeting individuals' preferences. This is because individuals would be willing to give up more of the resources available to them to improve their own health than the health care system would require. It would suggest that more resources could be transferred from other consumption opportunities to health care; that is, that the current budget or level of health care expenditure is too low.

Simply assuming that $v = k$, however, would be inappropriate because both are ultimately separate empirical questions. Indeed, there are good reasons to expect $v > k$. This is because there are costs associated with socially acceptable ways to finance health care systems, so society would not choose to increase expenditure to the point where

$v = k$. In addition, v represents how much an individual would be willing to give up of their own consumption to improve their own health, whereas k represents how much of collectively pooled resources is currently required to improve health. Furthermore, k reveals something about how much society is willing to pay for improvement in the public's health given other objectives, constraints, and competing claims on public funds. So v and k are not necessarily valuing the same thing. Therefore, unless one believes that the purpose of collectively funded health care is to satisfy individuals' preferences, that these are fully reflected in social decision-making and political processes, and that there are no welfare losses associated with financing the health care system, it is not at all clear that health care expenditure will be, or ought be, set to equalize v and k.

Appropriate assessment of v and k are ultimately empirical matters. A review of 383 estimates of the consumption value of a QALY suggests that v (mean estimate of 74 000 euros per QALY by Ryen and Svensson 2014) may well be greater than k (15 000 euros per QALY, based on estimates reported in Table 4.2). This would suggest that the scale of pooled or public funds available for health care is falling short of individual expectations and preferences, which seems to reflect the political reality in many health care systems. It also implies a ratio of $v/k > 1$, which represents the value of collective compared to private resource or the shadow price of public expenditure. A ratio of $v/k > 1$ suggests that public expenditure is scarce (more valuable) relative to private consumption, which seems to reflect the fiscal reality in many economies.

Failing to assess k or ignoring evidence that $k < v$ and simply using v to inform decisions on the grounds that expenditure ought to be increased so that $v = k$ would be inappropriate. This is because it assumes that the change in health care resources that would be required would be made available immediately. Of course, if expenditure on health care does increase in the future then the threshold is likely to change, in which case the decision can be reconsidered once more resources are actually available rather than only assumed to be available (see Sections 4.3.2 and 4.5.2 for a discussion of the implications of changes in the threshold over time).

The broader question of whether expenditure on health care should be increased, and to what extent, depends on social choices that are mediated through a political process. This process determines how public resources could be used for other purposes (education, transport, defence, etc.), and whether public expenditure should be increased, for example by raising additional taxation or running budget deficits. Economic evaluation cannot claim to prescribe this choice, but estimates of the health that could be gained through additional expenditure (k) and how this compares to individuals' preferences (v) can inform this debate.

4.4 Making decisions with multiple alternatives

Informing a particular decision requires identifying the possible alternative courses of action that could be taken to improve the health of patients who face a choice between mutually exclusive alternatives (i.e. either/or decisions). So far in this chapter we have discussed how the results of economic evaluation can inform decisions when there are only two alternative courses of action available. In these circumstances there is a single ICER that summarizes the cost-effectiveness of choosing the more effective but more

costly alternative since there is only one possible comparison of incremental cost (Δc) and incremental effect (Δh). In most circumstances, however, more than two alternatives are available and in some circumstances there may be very many indeed. This is because alternative strategies include the different combinations or sequences of treatment and different ways in which interventions can be used (what dose, when to start, when to stop, etc.).

4.4.1 Which ICER?

When there are multiple mutually exclusive alternatives available there are multiple pairwise comparisons that can be made, each providing different incremental costs and benefits, resulting in many different ICERs that could be reported. Three questions have to be addressed (Johannesson and Weinstein 1993; Weinstein 2013):

- How should alternatives be compared?
- Which comparisons are relevant when calculating and reporting ICERs?
- How should cost-effectiveness be judged in these circumstances?

These common circumstances are illustrated in Table 4.5 which reports the expected costs and expected QALYs associated with four alternative courses of action which could be taken (A–D). These costs and effects are also illustrated graphically on the cost-effectiveness plane in Figure 4.2. There are many different comparisons that could be made, each resulting in different incremental costs, health effects, and ICERs. For example, B, C, and D could be compared to the lowest-cost and least effective alternative, A. Alternative D has an ICER of $29 885 when compared to A, so using a threshold of $30 000 per QALY, D might appear to be cost-effective based on this comparison. This would only be the correct conclusion if A and D were the only courses of action available. It is clear, however, that B and C are also possible choices that could be made. A comparison of the costs and effects of these alternatives indicates that D is more costly and less effective than C. Therefore, D is strongly dominated by C (point D lies to

Table 4.5 ICERs and net benefit with multiple alternatives

	Cost	QALYs	ICERs compared to			Net benefit	
			Lowest cost (A)	Next lowest cost	Relevant alternative	$20 000 per QALY	$30 000 per QALY
A	$4147	0.593	-	-	-	$7713	$13 643
B	$8363	0.658	$64 862	$64 862	ED	$4797	$11 377
C	$8907	0.787	$24 536	$4217	$24 536	$6833	$14 703
D	$9078	0.758	$29 885	SD	SD	$6082	$13 662

ED, extended dominance; ICERs, incremental cost-effectiveness ratios; QALYs, quality-adjusted life-years; SD, strong dominance.

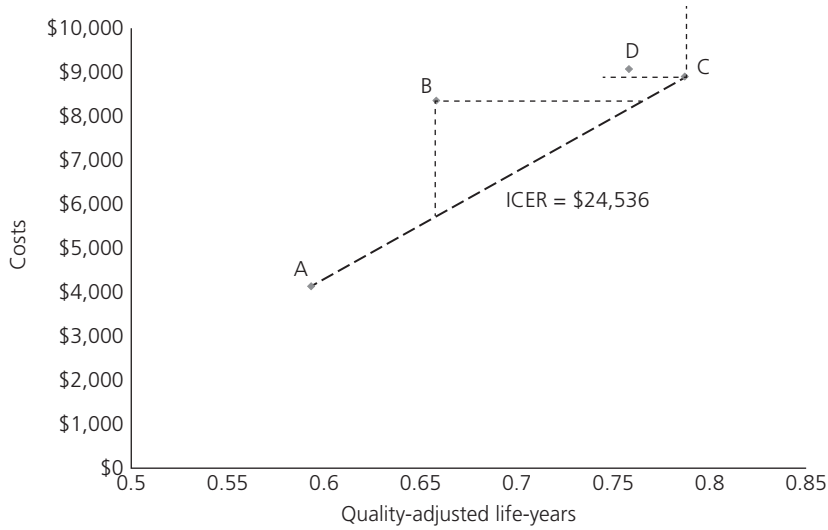


Fig. 4.2 The cost-effectiveness plane with multiple alternatives.

the north-west of point C in Figure 4.2) so should never be chosen irrespective of any threshold that might be applied.

This illustrates the importance of considering all the available alternatives in an economic evaluation. Failure to do so by, for example, excluding C from an analysis, is likely to mean that the cost-effectiveness of an intervention being considered (e.g. D) can be seriously overestimated. This is because it has not been compared to more cost-effective alternatives that might be available. In other words, any alternative can look attractive if it can be compared to something sufficiently bad! For this reason ICERs should not be based on comparisons with strongly dominated alternatives.

4.4.1.1 Extendedly dominated alternatives

Once strongly dominated alternatives have been ruled out, ICERs can be calculated based on comparisons of moving from a lower cost to the next more costly and effective alternative (neither A, B, or C is strongly dominated). This suggests that the ICER of alternative C when compared to B (the lower-cost alternative) is $4 794 per QALY in Table 4.5. One might be tempted to conclude that C appears very cost-effective at a threshold of $20 000 per QALY. However, the question is whether or not B is the appropriate alternative for this comparison.

The problem is that the ICER of B, when compared to A (the lower-cost alternative to B), is $64 862 per QALY, so would not be regarded as cost-effective at a threshold of $20 000 per QALY. In fact, B would never be chosen irrespective of the threshold so long as C is an available option. For example, if the threshold was $70 000 per QALY, B would be regarded as cost-effective compared to A but a decision-maker would not be satisfied with B because moving from B to C offers greater health improvements at a cost per QALY that would also be regarded as worthwhile; that is, moving from B to C offers greater net benefits (see Section 4.4.2). Importantly, not only is the ICER of

moving from B to C lower than the threshold, it is also lower than the ICER of moving from A to B. In other words, if a decision-maker were willing to pay enough for health outcome to make B seem worthwhile then they will also be willing to pay the additional costs to move to C because the ICER is lower. Therefore, alternative B will never be chosen and is described as being *extendedly dominated* by A and C.

Once B is ruled out as an alternative that will never be chosen, the ICER for C should be based on a comparison with A (the lower-cost non-dominated alternative). Therefore, the appropriate ICER for C is not $4 794 but $24 536 per QALY, so at a threshold of $20 000 per QALY it would not be considered cost-effective. It is A that should be regarded as cost-effective and has the highest net benefit (see Section 4.4.2). This illustrates the importance of identifying and ruling out extendedly dominated alternatives when calculating and reporting ICERs. Failure to do so will mean that ICERs based on comparisons with extendedly or strongly dominated alternatives will be underestimated and can lead to the acceptance of alternatives that are not cost-effective (Weinstein 2013).

Extendedly dominated alternatives can be identified in the way illustrated in Table 4.5:

- Rule out strongly dominated alternatives.
- Calculate ICERs based on the comparisons of moving to increasingly costly and increasingly effective alternatives.
- If the ICER associated with moving to more costly alternative falls, then the lower-cost alternative used to calculate the ICER is extendedly dominated and should be ruled out.
- Recalculate ICERs based on comparisons of moving to increasingly costly but increasingly effective alternatives that are neither strongly nor extendedly dominated.

This concept of extended dominance can also be illustrated graphically in Figure 4.2. Point B lies to the north-west of the part of the line that joins A and C. Therefore, alternative B can be thought of as being *strongly dominated* by some combinations of A and C, that is, the costs and effects of offering A to some proportion of the patient population and C to the others. If this type of 'mixture' was a practical possibility there would be more alternatives to consider, some of which would strongly dominate B. As in this example, this type of mixture is often not regarded as feasible (e.g. due to equity constraints, see Sections 4.4.3 and 4.6), so B remains extendedly rather than strongly dominated and will never offer higher net benefits than the other alternatives (see Section 4.4.2).

4.4.1.2 What are relevant alternatives?

This section has illustrated the importance of considering all the available alternatives. Failure to do so may mean that the cost-effectiveness of an intervention maybe seriously overestimated (its ICER is underestimated) because it has not been compared to more cost-effective alternatives that might available. Including every possible alternative strategy can be challenging (see Chapters 9 and 10). However, the guiding principle should be to include all those alternatives that have some possibility of being cost-effective, whether or not they are currently part of clinical practice. Indeed, including a 'do nothing' alternative with zero costs can be useful because if an intervention is

not cost-effective when compared to 'do nothing' then it will not be cost-effective when compared to any other alternative that is not dominated. In other words, a comparison with 'do nothing' provides a necessary but not sufficient condition for cost-effectiveness. The example illustrated in Table 4.5 is loosely based on the results of an evaluation of interventions for advanced ovarian cancer. Offering no health care whatsoever for this condition was not regarded as a feasible policy option and there is very limited evidence to estimate the QALYs associated with such a decision. Therefore, any conclusions based on the ICERs reported in Table 4.5 rest on the assumption that the lowest-cost alternative (A) is itself worthwhile compared to offering no health care at all.

4.4.2 Net benefit and multiple alternatives

When there are more than two alternatives, summarizing cost-effectiveness using ICERs requires consideration of which pairwise comparisons are appropriate when calculating incremental costs and effects. The discussion in Section 4.4.1 demonstrates the importance of excluding dominated and extendedly dominated alternatives before ICERs are calculated and reported. Comparing an ICER to a threshold ($\Delta c/\Delta h < k$) is equivalent to asking whether the incremental net health benefits offered by the intervention are positive ($\Delta h - \Delta c/k$), or whether the incremental net monetary benefits (the equivalent health care resources) are positive ($k.\Delta h - \Delta c$) (see Table 4.1 and Section 4.2) (Laska et al. 1999; Phelps and Mushlin 1991).

Expressing cost-effectiveness in terms of net benefit (whether expressed as health or its health care resource equivalent) is particularly useful when there are multiple alternatives. This is because it is not necessary to make only pairwise comparisons, calculate increments, and identify which are the appropriate comparisons that should be made. Calculating *incremental* net benefit is not necessary because the net benefit of each alternative can be calculated and directly compared. The net benefit of the four alternatives, for thresholds of $20 000 or $30 000 per QALY, is reported in Table 4.5. The alternative that should be considered cost-effective is simply the one that provides the highest net benefit. For example, at a threshold of $20 000 per QALY, A offers the highest net benefit of $7713, but at a threshold of $30 000 per QALY, C offers the highest net benefit of $14 703. Exactly the same conclusions are reached about cost-effectiveness as calculating ICERs and comparing them to the thresholds once dominated alternatives have been ruled out (Laska et al. 1999).

The reason why net benefit can be used in this way is that comparing the net benefit of any two of the alternatives is exactly the same as calculating the incremental net benefit of this comparison. In this example, the incremental health benefit of C compared to A is 0.194 and the incremental cost of C compared to A is $4760. So the incremental net monetary benefit at a threshold of $30 000 per QALY is $1060. This is exactly the same as calculating the difference between the net benefit of C ($14 703) and the net benefit of A ($13 643).

Using net benefit rather than ICERs to report cost-effectiveness requires a threshold to be specified. However, it does not require pairwise comparisons to be made, so strongly or extendedly dominated alternatives do not need to be ruled out because any alternative that is either dominated or strongly dominated will never offer the highest net benefit and will never be considered cost-effective. The equivalence of representing cost-effectiveness in terms of net benefit or ICERs is illustrated in Figure 4.3, where the net benefit of each of

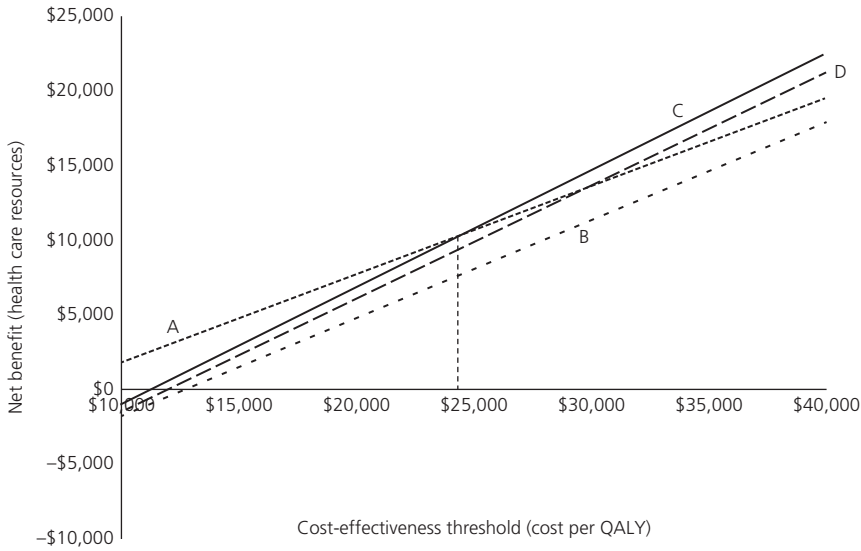


Fig. 4.3 Net benefit and the threshold.

the four alternatives is reported for a range of possible values for the threshold. At thresholds less than $24 536 per QALY (equal to the ICER of C compared to A), alternative A offers the highest net benefit and should be considered the cost-effective alternative. At a threshold greater than $24 536 per QALY, C has the highest net benefit ($ICER_{C,A} < k$) and is cost-effective. Importantly, neither D nor B will ever have the highest net benefit—they will never be on the outer envelope of these net benefit lines.

A few other things are interesting to note in Figure 4.3. Although D and B are dominated, so will never be considered cost-effective, they may be 'better' than other non-dominated (but not cost-effective) alternatives; for example, D has higher net benefit than A at thresholds greater than $30 000. Also the extendedly dominated alternative (B) has lower net benefit than the alternative that is strongly dominated (D). This demonstrates that the alternatives that need to be ruled out when calculating appropriate ICERs might nonetheless represent the next best choice and should not be excluded from the analysis or disregarded, especially when exploring the uncertainty associated with decisions based on cost-effectiveness (see Chapter 11).

4.4.3 Non-mutually-exclusive alternatives

In previous sections of this chapter we have considered the choice between the mutually exclusive courses of action that are available to improve the health of a patient (or a group of similar patients). Many of the choices faced in health care, however, are not mutually exclusive—these are not either/or decisions.

4.4.3.1 Subgroups and patient characteristics

For example, choosing which interventions should be offered to different subgroups of patients with the same condition are not mutually exclusive choices. Different decisions about which alternative to offer can be made for each subgroup of patients who

have different characteristics that directly affect cost-effectiveness. These characteristics could include, for example, patients' comorbidities, previous responses to treatment, or disease severity (see Chapters 8 and 10). Once subgroups have been identified and the costs and health effects of the alternatives have been estimated for each subgroup, cost-effectiveness can be assessed in the way described in Sections 4.4.1 and 4.4.2. In other words, each subgroup can be considered separately; identifying which of the mutually exclusive alternatives available to each subgroup should be regarded as cost-effective by either comparing the ICERs of non-dominated strategies to a threshold (see Section 4.4.1) or calculating the subgroup specific net benefit of each alternative (see Section 4.4.2). A particular intervention may offer the highest net benefit in all subgroups, only in some, or in none. Therefore, the calculation of appropriate ICERs or net benefits specific to each subgroup can inform any decision about which interventions can be regarded as cost-effective and should be offered to different types of patients.

The total cost and health effects of each alternative for the whole patient population are the sum of the costs and health effects across each of the subgroups. The average per-patient costs and effects is the weighted average across the subgroups, with weights reflecting the relative size of each subpopulation. This may be useful if decision-makers are concerned that approving or reimbursing an intervention for use in some subgroups but not in others will not be sustainable. For example, it might be difficult to monitor and control appropriate use once an intervention is made available for some (see Section 11.5) (Espinoza et al. 2014). Alternatively, making access conditional on certain characteristics (e.g. age and gender) might be regarded as discriminatory or in conflict with other social values and equity concerns. Here it is useful to compare the sum of the net benefits that could be achieved if different decisions for different subgroups can be made and sustained with the total net benefit if the same decision must be made for all. This indicates the value of being able to monitor and sustain differential access or the opportunity costs of any equity concerns that might be at play (Epstein et al. 2007; Stinnett and Paltiel 1996).

4.4.3.2 Priority-setting and defining a benefit package

Health care systems also face choices about which interventions to make available across different areas of disease, relevant to very different patient populations. For example, systems can decide which interventions to make available for which conditions and patient populations, what should be added to an existing benefit package, or what collection of health care interventions should be selected to form a new benefit package. The principles of how to judge whether or not a particular intervention for a specific patient population ought to be included are the same as those outlined in Sections 4.4.1 and 4.4.2. In a similar way to patient subgroups, each of the different interventions being considered for inclusion in the package can be considered separately. If an intervention is regarded as cost-effective when compared with the other mutually exclusive alternatives available for the treatment of patients with the specific condition, then it should be included in the collection of interventions available to treat the range of diseases relevant to different patient populations.

An important consideration, however, is how to assess opportunity costs and identify an appropriate threshold to assess cost-effectiveness in these circumstances. This

will depend on context: whether the system is considering an amendment to an existing package, with either the same resources or with additional resources; whether it is selecting a new package; or completely reorganizing an existing one. If considering amendments to an existing package but using the same resources, then the question is what other parts of the existing benefit package will need to be removed and what the health effects likely to be. This might be based on an assessment of what is likely to be forgone elsewhere (see Section 4.3.2) or on explicitly identifying the specific matching disinvestments that would need to take place. Identifying what might be suitable matching disinvestment still requires an assessment of how it compares to what would be likely to be forgone elsewhere if a matching disinvestment was not specified (i.e. an assessment of a threshold). Without this there is a danger that an intervention would be accepted and a matching disinvestment made when in fact both should be rejected (the ICER of the investment and proposed disinvestment are greater than the threshold). Similarly, an intervention might be rejected when compared to a specific disinvestment when both should be accepted (the ICER of the intervention and proposed disinvestment are less than the threshold).

When considering how best to use additional resources to expand a package, ideally one would wish to have information about the costs and effects of all the other potential investments that could be made, selecting those which offer the cheapest ways to improve health first (those with the lowest ICERs). The task is to identify the collection of investments that offer the greatest improvement in health given the additional resources that have been made available. This is similar to the task faced when considering how to select a new package or completely reorganizing an existing one: that is, selecting the collection of interventions that will provide the greatest improvements in health given the resources that are made available (Evans et al. 2013; Weinstein 2013). The information required is considerable. In principle it requires information, not only about the costs and effects of all the types of care currently offered to treat patient (sub)populations with different diseases, but also about the costs and effects of all the interventions that could be offered. With different sets of multiple (mutually exclusive) alternatives to choose from within a wide range different of conditions and patient populations, this is a complex task: there are multiple mutually exclusive alternatives within each of the many non-mutually-exclusive choices that are faced (Weinstein 2013).

Mathematical programming (linear and integer programming) provides a useful analytical framework within which this type of constrained optimization, across mutually and non-mutually exclusive choices, can be undertaken (Earnshaw and Dennett 2003; Stinnett and Paltiel 1996). Due to the considerable informational requirements, the examples of its application tend to be rather stylized (Epstein et al. 2007). Nonetheless, thinking about these questions as a mathematical programming problem is useful. It demonstrates that reporting simple 'league tables' of ICERs (one ICER for each non-mutually-exclusive programme of care) and implementing each in turn starting with the lowest ICER until available resources are spent is unlikely to be sufficient. This is because there are many mutually exclusive alternatives so potentially many ICERs within each of the non-mutually-exclusive programmes (Drummond et al. 1993; Gerard and Mooney 1993).

One of the results of applying mathematical programming to this type of problem is an estimate of the health effects of relaxing or tightening the constraint on health care

expenditure—the cost-effectiveness threshold or the shadow price of this constraint. For example, when considering reducing health care expenditure or including a more costly intervention, the threshold will be the ICER of the least cost-effective intervention that is currently included, because it is this that would be given up first. When considering increasing expenditure it will be the ICER of the most cost-effective intervention that was not included, because it is this that would be included first as more resources are made available. Therefore, setting priorities and selecting those interventions that should be included in a benefits package also reveals or implies the appropriate threshold for a given level of expenditure (Epstein et al. 2007; Stinnett and Paltiel 1996). With an appropriate estimate of the threshold, it is possible to decide whether or not a new intervention (or a collection of interventions) should be included without having to reconsider the entire benefits package every time these types of choices are faced. It is the threshold that provides the link between the type of decisions discussed in Section 4.4.1 and these broader choices about what collection of health care interventions should be included.

4.5 Some methodological implications

Previous sections have set out how estimates of the costs and effects of the available alternatives can be used to assess cost-effectiveness. However, there are important questions about which costs and other non-health effects ought to be included. For example the following questions are important:

- What future health care costs should be included (see Section 4.5.1)?
- What account should be taken of the timing of future costs and health benefits (see Section 4.5.2)?
- Should attention be restricted to health care costs or include non-health care costs and other non-health benefits (see Section 4.5.3).

4.5.1 Future costs and benefits

A decision to provide an effective intervention for current patients will have costs not just in the current period but in future periods as well. These future costs will include the costs of the intervention itself if continued treatment is required but, even when this is not the case, there will be an effect on the future costs of treating these patients. In some circumstances an effective intervention may reduce future costs by avoiding future events and the costs associated with them. In other circumstances it might increase health care costs if patients survive for longer but in a health state that requires health care (see Chapters 7 and 9).

An important question is whether these future costs should be restricted to the resources required to treat the particular condition (or conditions) that the intervention directly affects. These costs are sometimes referred to as future *related* health care costs. Or should health care costs that are likely to be incurred as a consequence of increasing life expectancy also be included? For example, an intervention that prevents, or reduces mortality from myocardial infarction will mean that more patients avoid this event or survive it. As a consequence, the life expectancy of these patients will increase, so some

will ultimately be diagnosed and treated for other conditions that are unrelated to the direct effect of the intervention. For example, some patients will be diagnosed with prostate cancer and, if treated for that condition, will incur the costs associated with it. The question is whether these 'future unrelated health care costs' should be considered as part of the cost of preventing or reducing the mortality associated with myocardial infarction (Garber and Phelps 1997; Weinstein and Manning 1997).

If these future costs will necessarily be incurred by deciding to provide an intervention that reduces mortality and extends survival, then they are part of the opportunity costs of this intervention (Feenstra et al. 2008; Lee 2008; van Baal et al. 2013, 2014). However, the decision to offer an intervention today does not necessarily commit to treating these unrelated conditions in the future and incurring the costs associated with them. For example, one would expect that a decision to diagnose and treat prostate cancer in the future would also be based on an assessment of which of the alternatives available at that time is cost-effective. If treatment is offered at this point it is because the benefits it offers are justified by the costs. In other words, the longer survival preserves the option to make future decisions which offer positive net health benefits which, with innovation in medicine and health technologies, might be better (offer greater net health benefit) than those available today.

Therefore, how this question is dealt with depends in part on what commitments are being made when making a decision about a specific intervention today. If a decision-making body has responsibilities for decisions about other interventions across different conditions now and in the future, then future unrelated costs should not be regarded as an irreversible commitment. Rather, they should be considered as a decision that can, in principle at least, be addressed in the future based on an assessment of whether incurring these costs are worthwhile. However, if there is no irreversible commitment to incurring future costs, then the estimate of the benefits of an intervention being considered today should not include the benefits associated with future treatments either. That is, an analysis that excludes future unrelated health care costs should also exclude the benefits that relevant future interventions offer as well. This can be difficult if published estimates of survival and life expectancy are used in the analysis. This is because such estimate will include some of the effects of other unrelated health care. In other words, the benefits of only committing to the future related costs of the intervention are likely to be overestimated. However, such analysis will also have excluded the possible positive net health benefits that future decisions about unrelated health care might be able to offer. So extending survival preserves an 'option value' of using more effective and cost-effective health care in the future. Therefore, restricting attention to the costs and benefits of future related health care will not necessarily always overestimate the cost-effectiveness of an intervention.

The problem of including future unrelated health care costs and the benefits associated with the relevant interventions is that it runs the risk of compounding poor decisions. For example, an intervention that would be considered cost-effective when excluding future unrelated health care costs and the benefits that they offer might not be cost-effective when unrelated health care costs (and its benefits) are included. This would occur if the future health care is not itself cost-effective. In other words, a cost-effective intervention might be rejected because cost-ineffective care is being (and it is

assumed will continue to be) provided elsewhere in the health care system. This seems inappropriate if there is a possibility of addressing cost-ineffective care in some future decisions. For these reasons most economic evaluations tend to restrict attention to future related health care costs, although it should be recognized that, in principle, the assessment of benefit should also be restricted to this related health care too (van Baal et al. 2014).

4.5.2 Discounting future costs and benefits

A decision to provide an effective intervention for the current patient population may offer some immediate health benefits but, in many circumstances, the health benefits will occur in future periods. For example, the life-years and QALYs gained from an intervention that reduces mortality will occur in future periods, even if the reduction in mortality is restricted to the current period. Of course other interventions are intended to reduce the risk of future events (e.g. primary prevention of myocardial infarction or stroke), so the health benefits they offer will not be realized for many years. Similarly, interventions will not just impose costs and offer cost savings in the current period but in future periods as well. The question is how account should be taken of when costs are incurred and health benefits are received.

4.5.2.1 The rationale for discounting

Different opportunity costs are imposed depending on when costs are incurred. These opportunity costs reflect the fact that the resources required for health care could, instead, have been invested elsewhere in the economy which would provide a positive rate of return (Paulden 2014). This means that costs that are incurred in the near future are more important than those incurred in the more distant future. This is illustrated in Table 4.6. At a real rate of return of 3.5% per annum, health care costs of $1 million incurred today ($t = 0$) could instead have been saved (invested elsewhere in the economy) and would provide more than $1 million in real terms in future periods, i.e. the value of $1 m invested at $t = 0$ is $1 410 599 in 10 years' time ($t = 10$). Therefore, a cost of $1 m which will be incurred in the next period ($t = 1$) requires less than $1 m of current resources to be set aside to cover this future liability because the funds can be invested with a positive rate of return. The amount of current resource required ($966 184) is described as the present value of $1 m incurred in $t = 1$. The further in the future the same real cost is incurred the less of current resources is required to meet it; that is, the present value of $1 m incurred in $t = 10$ is lower ($708 919); see Table 4.7. The opportunities to invest resources elsewhere in the economy with a real rate of return, rather than incur health care costs, makes it clear that costs need to be expressed in terms of their value in a common period. This is usually the present period, with their values discounted to present values based on when they are incurred, and a discount rate (r) that reflects real rates of return (see Section 7.2).

An important question is whether health should be discounted in the same way. On the one hand, health is unlike resources because there is no opportunity directly to invest health elsewhere at some real rate of return. However, health care transforms resources into health, so if resources can be traded over time then so can health. For example, we could choose to reduce expenditure and health now, invest the resources

Table 4.6 Why discount costs?

Year 0	1	2	3	10
£1 m	£1 m(1 + *r*) = £1 035 000	£1 m(1 + *r*)(1 + *r*) = £1 071 225	£1 m(1 + *r*)(1 + *r*)(1 + *r*) = £1 108 718	£1 m(1 + *r*)^{10} = £1 410 599
£1 m/(1 + *r*) = £966 184	£1 m			
£1 m/(1 + *r*)^{10} = £708 919				£1 m

Table 4.7 Why discount health?

Year 0	1	2	3	10
£1 m				**100 QALYs**
			Cost per QALY = £14 106	£1 410 599 (100 QALYs)
£1 m (70.9 QALYs)	Cost per QALY = £14 106			

that are released at a positive rate of return so that more resource will be available in the future, which can be used to generate more health, and vice versa. Indeed, many of the decisions in health care relate to these types of choices: whether to commit resources now or later with consequent impacts on current and future health (Claxton et al. 2006). The relative value of current compared to future health is implied by how resources are allocated to health care over time. So the way in which government sets health care budgets, or how other commitments to health care expenditure are made, reveals something about a the time preference, or discount rate, for health (Paulden and Claxton 2012).

Since health care resources are ultimately health, if we discount health care resources we also discount their health effects. This is illustrated in Table 4.7 where an intervention costs an additional $1 m now and generates 100 additional QALYs in $t = 10$. Rather than discount health to its present value, the costs could be expressed as their equivalent value in $t = 10$ when the health benefits occur, i.e. to their terminal rather than their present value. The opportunity costs of committing $1 m now is $1 410 599 in $t = 10$, so the cost per QALY of this intervention is $14 106 per QALY. This is exactly the same as discounting the health benefits occurring in $t = 10$ back to their present value (70.9 QALYs) and comparing them to the costs occurring in that period (Nord 2011).

4.5.2.2 Should costs and health be discounted at the same rate?

Health should be discounted, but an important and disputed question is whether health care costs and health should be discounted at the same rate (Brouwer et al. 2005; Claxton et al. 2006; Nord 2011). This depends on: (1) whether there are constraints

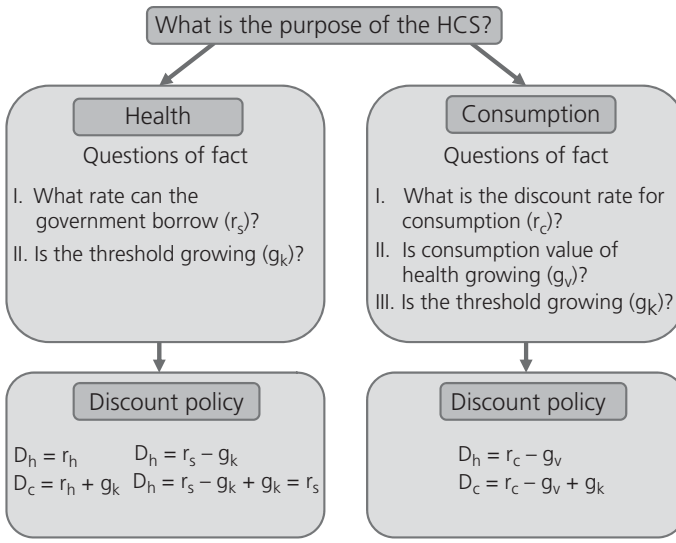


Fig. 4.4 Selecting a discounting policy.

on health care expenditure; (2) whether the cost-effectiveness threshold is expected to grow over time; and (3) whether the consumption value of health is expected to grow. How the answers to these questions influence discounting policy are summarized in Figure 4.4 and are discussed below (Claxton et al. 2011b).

If there are constraints on the growth in health expenditure, then health care costs are health that is expected to be forgone at a rate represented by the cost-effectiveness threshold, k (see Section 4.3.2). So future health care costs are simply future health that is likely to be forgone. Cost-effectiveness can be expressed by calculating net health benefits in each period and discounting future net health benefits using a discount rate for health, $D_h = r_h$. Equivalently, if the threshold is expected to be constant over time, then the ICER can be compared to the current threshold using incremental health benefits and incremental costs discounted at the same discount rate for health ($D_c = D_h = r_h$). A revealed time preference or discount rate for health (r_h) can be based on the rate at which government or the funders of health care can borrow or save (r_s) and whether the threshold is expected to grow (g_k) because this indicates the relative value (in terms of health care resources) of current compared to future health ($r_h = r_s - g_k$) (Paulden and Claxton 2012).

If the threshold is expected to grow in real terms ($g_k > 0$), then any future costs are less important than current costs because they will be expected to displace less health. Alternatively, if the threshold is expected to decline ($g_k < 0$), then future costs are more important than current costs because they will be expected to displace more health. The relative importance of future and current costs needs to be reflected in estimates of cost-effectiveness either by calculating net benefits in each period or through discounting costs differently when calculating an ICER:

- Net health benefits can be calculated in each period (t) by applying the threshold relevant to that period, k_t, to the costs that occur that period (i.e. $\Delta h_t - \Delta c_t/k_t$, see Section 4.2.1). These future net health benefits can then be discounted to the present period using a discount rate for health, $D_h = r_h$.
- If an ICER is compared to the current threshold, some account must be taken of whether future costs are expected to be less ($g_k > 0$) or more important ($g_k < 0$). This can be achieved by discounting the incremental health benefits using discount rate for health ($D_h = r_h$) but discounting the incremental costs at a rate that reflects any growth in the threshold and the relative importance of future costs ($D_c = r_h + g_k$). For example, if a discount rate for health is 1.5% and the threshold is expected to grow at 2%, future costs are less important and one way to reflect this is to discount them at a higher rate ($D_c = 3.5\%$) compared to health benefits ($D_h = 1.5\%$)

The purpose of health care might be regarded as improving a broader notion of welfare or consumption value (see Section 2.4.2) rather than health itself. In these circumstances it is also necessary to consider whether the consumption value of health (v_t, see Section 4.3.3) is likely to grow ($g_v > 0$). There are good reasons to believe that it will increase with economic growth and consumption (Smith and Gravelle 2001). However, if there are constraints on health care expenditure, then future costs do not displace consumption but health (see Section 4.4.3). Therefore, any growth in the consumption value of health will mean that future health and future health care costs will both be more valuable. This can be reflected in estimates of cost-effectiveness by calculating the net consumption value of the net health effects in each period using the consumption value of health relevant to that period (i.e. $v_t\,(\Delta h_t - \Delta c_t/k_t)$, see Table 4.1 and Section 4.3.4) and then discounting these net consumption benefits at a discount rate for consumption (r_c). An appropriate discount rate for consumption effects might be based on either the long-run real rates of return to a safe asset, like a government bond, or on estimates of social rates of time preference which reflect individual time preference and expected growth in consumption (Paulden 2014). Alternatively, if an ICER is compared to the current threshold, then incremental costs and health benefits can be discounted at a lower rate than consumption ($D_h = D_c = r_c - g_v$) to reflect growth in the consumption value of health and the greater importance of both future health benefits and future health care costs (Claxton et al. 2011b).

The important thing to note is that it is growth in the threshold that leads to discounting health care costs differently from health benefits (a higher discount rate for health care costs if the threshold is expected to grow) and only when comparing an ICER to the current threshold. Growth in the consumption value of health leads to lower discount rate for both health and health care costs. However, if there are no constraints on health care expenditure then future cost will displace future consumption rather than health (see Table 4.1 and Section 4.3.3). In these circumstances the value of the health effects in each period can be calculated using the consumption value of health relevant to that period. The future consumption benefits ($v_t.\Delta h_t$) and consumption costs (Δc_t) can then be discounted at the same rate for consumption (r_c). Alternatively, if an ICER is compared to the current consumption value of health, then incremental health

benefits can be discounted at a lower rate than consumption ($D_h = r_c - g_v$), to reflect the fact that future health is more valuable in terms of consumption, but with costs discounted at the higher rate for consumption ($D_c = r_c$) (Claxton et al. 2011b; Smith and Gravelle 2001). In short, discounting health benefits at a lower rate than health care costs will be appropriate if there are no constraints on health care expenditure and the consumption value of health is expected to grow.

4.5.3 Perspective for costs and benefits?

The type of CEA described in Sections 4.2 and 4.4 compares the health benefits expected to be gained from using an intervention with the health that is likely to be displaced due to the additional health care costs. This will lead to reasonable decisions if the measure of health captures the important and valuable aspects of health and any other non-health effects are small or are of limited value compared to the effects on health. However, interventions will often have other non-health effects, which fall into two broad types: direct costs of care that do not fall on the health care system and the indirect effects on the rest of the economy.

Some direct costs of health care are borne by patients, such as out-of-pocket costs as well as their time in accessing care. Other direct costs include the time and resources devoted to caring for patients outside the health care system. So these costs might include marketed and non-marketed activities (e.g. time and informal care) which will need to be valued in some way (see Section 7.1). An effective intervention may reduce these costs (e.g. a quicker recovery) or increase them (e.g. prolong survival in a chronic state).

The indirect effects on the wider economy are external to the patients, their family, or informal carers but are valued by the rest of society. For example, enabling a patient to return to productive activity in the labour market (as well as other non-marketed activities in the community and the household) will, in many circumstances, add to production in the economy (see Section 7.3). So there will be a net benefit to rest of society if the value of the additional (marketed and non-marketed) production exceeds the individual's additional consumption. Again, an effective intervention might provide external benefits by reducing mortality in economically active groups. However, it may also impose external costs if it reduces mortality in populations where consumption exceeds the value of production (Johannesson et al. 1997; Kruse et al. 2012; Meltzer 1997, 2013).

This poses disputed questions of which non-health effects should count and how they should be measured and valued relative to each other. Any attempt to aggregate all these different effects by specifying what should count and how they should be valued either requires or implies a particular definition of social welfare (see Sections 2.4.2 and 2.4.3).

4.5.3.1 A single societal perspective

Traditionally economics addresses this problem by founding a view of welfare on individuals' preferences which are revealed by the choices people make. In particular, these choices are made in markets which show the compensation (additional consumption) individuals require to give up something of value or the consumption they would be

willing to offer to gain something of value (see Section 2.4.2). These principles enable all the non-health effects to be valued in a common numeraire of the equivalent consumption gains and losses (Tsuchiya and Williams 2001). The equivalent consumption value of non-health effects can be based on market prices where they are believed to be undistorted or, in the absence of such markets, by estimating shadow prices that would reflect the outcome of an undistorted market (Boadway and Bruce 1984; Sugden and Williams 1979). Where aspects of non-health effects are not marketed (e.g. patient and carer time), shadow prices can be based on how these are implicitly valued in other markets or in hypothetical choices (see Chapters 6 and 7). Therefore, a traditional welfarist approach provides guidance on what should count (any effects where individuals require compensation or are willing to pay) and how they should be measured and valued (using market prices or shadow prices). This means that all the non-health effects can be aggregated and expressed as a net consumption benefit or net consumption cost associated with the intervention.

4.5.3.2 Taking account of restrictions on health care expenditure

Aside from questions about whether this implied definition of social welfare is acceptable (see Section 2.4.3), there are issues regarding its implementation: in particular, how non-health benefits and costs should be taken into account alongside the health effects, and the health care costs of an intervention when there are restrictions on the growth in health care expenditure.

An intervention which would not be regarded as cost-effective, on the basis of a comparison of the health benefits with the health forgone as a consequence of its additional health care costs, might offer considerable non-health consumption benefits. These benefits may come, for instance, from reducing mortality and improving quality of life in younger, economically active, populations. For example, in Table 4.4 the use of Lucentis for the treatment of diabetic macular oedema is not cost-effective, with negative net health benefits of −2959 QALYs each year, but offers consumption benefits of £88.4 m. Equally, an intervention that might be regarded as cost-effective (health gains exceed health forgone for the health care system), but might impose considerable consumption costs (e.g. reducing mortality in older, less economically active, populations). Simply combining the health care costs (e.g. £80.6 m in Table 4.4) and consumption cost and benefits (e.g. £88.4 m in Table 4.4) would be inappropriate as it would ignore the reality of existing constraints on health care expenditure (see Section 4.3.4). It would treat consumption benefits as if they are health care resources that could be used to offer health care, and it would treat consumption costs as if they were health care costs and would displace other health care. For example, in Table 4.4 one would wrongly conclude that Lucentis reduces total costs (£80.6 m − £88.4 m = −£7.8 m) and offers net health benefits of 3225 QALYs so should be regarded as worthwhile.

Therefore, it is important to distinguish where the opportunity costs fall and value them appropriately. Additional health care costs (Δc_h) will impose opportunity costs on health so a cost-effectiveness threshold that reflects these health opportunity costs is required to calculate the net health benefits within the health care system ($\Delta c_h/k$). Any non-health consumption cost, Δc_c, (or benefit) will displace (or offer) consumption opportunities in the wider economy. So to compare the net health benefits (within

the health care system) to the consumption costs or benefits (in the wider economy), an estimate of the consumption value of health is also required (see Sections 4.3.3 and 4.3.4). The net consumption effects can then be expressed as their health equivalent ($\Delta c_c/v$) and can be compared to net health benefits ($\Delta h - \Delta c_h/k$). Alternatively, the net health benefits can be expressed as their consumption equivalent ($v(\Delta h - \Delta c_h/k)$) and compared to the external net consumption costs (Δc_c). In so far as k is less than v, then less weight ($k/v < 1$) ought to be placed on consumption costs compared to health care costs (see Table 4.1 and Section 4.3.4) (Claxton et al. 2010).

For example, in Table 4.4, if v = £60 000 per QALY then the consumption benefits of £88.4 m would be equivalent to a gain of 1473 QALYs (£85 200 000/£60 000), which is less than the loss of net health benefit within the health care system. Alternatively this loss of net health benefit (2959 QALYs) can be valued as £177.54 m of equivalent consumption (2959 × £60 000), which is greater than the consumption benefits of £88.4. Therefore, failing to take account of where opportunity costs fall, and valuing cost appropriately by reflecting the impact of any constraint on health care expenditure, risks concluding that an intervention is worthwhile when it is not.

Adopting a single societal perspective still requires an estimate of a threshold that represents health opportunity cost. Simply assuming that $v = k$ would be inappropriate as there are good reasons to expect $v > k$ because there are costs associated with socially acceptable ways to finance health care systems. This seems to be supported by the evidence that is available (see Section 4.3.4) (Claxton et al. 2015b; Ryen and Svensson 2014). It suggests that publicly funded health care is falling short of individual expectations and that public expenditure is relatively scarce compared to private consumption. This also seems to reflect the political and fiscal realities in many health care systems and economies. Appropriate assessment of v and k are ultimately separate empirical matters. Failing to assess k or ignoring evidence that $k < v$ and simply using v to inform decisions on the grounds that expenditure ought to be increased so that $v = k$ would be inappropriate because it assumes that the change in health care resources that would be required would be made available immediately (see Section 4.3.4).

Therefore, adopting a single societal perspective still requires the impact of constraints on health care expenditure to be accounted for. Indeed, there are likely to be other constraints which, in principle, ought to be taken into account as well, such as costs that fall on other sectors where public expenditure is also restricted (e.g. education or criminal justice).

4.5.3.3 Accounting for displaced non-health benefits

Including non-health benefits for patients, carers, and the wider economy, may not improve decisions unless it is also possible to assess the same non-health benefits that are also likely to be displaced as a consequence of additional health care costs (see Section 4.3.2.4 and Box 4.3). An intervention which would not be judged to be cost-effective when considering only its health effects (net health benefits are negative), might offer consumption benefits to patients, carers, and the wider economy that might compensate and make it worthwhile nonetheless.

However, the health that is expected to be lost as consequence of a new intervention's additional health care costs might also be associated with benefits to patients,

carers, and the wider economy that will be displaced as well. Therefore, whether or not the non-health benefits offered by an intervention can compensate for the loss of net health benefit requires some assessment of the non-health benefits that are also likely to be displaced. For example, although the investment illustrated in Table 4.4 might offer consumption benefits of £88.4 m in addition to the 3225 QALYs gained, the 6184 QALYs expected to be lost as a consequence of the additional £80 m of health care costs will also be associated with a loss of £71.8 m of consumption benefits (see Box 4.2). Therefore, the net consumption benefits are £16.6 m rather than £88.4 m; equivalent to 273 rather than 1473 QALYs if v = £60 000 per QALY.

In other circumstances an intervention might offer consumption benefits but they might be less than the consumption benefits that are likely to be displaced. Therefore, net consumption losses would be imposed: the intervention is less cost-effective when non-health benefits are considered. For this reason, adopting a single societal perspective that includes non-health benefits maybe self-defeating unless some assessment of the non-health benefits that are likely to be displaced as a consequence of additional health care costs is also possible (see Section 4.3.2 and Box 4.3) (Claxton et al. 2015a).

It should be recognized that restricting attention to a health care system perspective does not necessarily disadvantage the assessment of the cost-effectiveness of a new technology. Although this narrower perspective excludes some potential aspects of benefit, it also excludes the opportunity costs associated with them too. Indeed, 'on average', a narrower health care perspective may lead to very similar decisions once account has been taken of the restrictions on health care expenditure and non-health benefits that are likely to be displaced. For example, if non-health care benefits are associated with improvements in health, then those technologies that would be judged to be cost-effective from a health care system perspective (as they increase net health) will also offer net non-health benefits outside the health care system. Those not judged cost-effective will reduce overall health (their net health benefits will be negative) so will also impose net non-health costs outside the health care system too.

4.5.3.4 A multi-sectoral perspective

Adopting a single societal perspective requires or implies a particular definition of social welfare. The problem is that there is no consensus about what should count and how all the effects that ought to count should be valued. Expressing them in a single common numeraire requires a particular definition of social welfare which will be disputed. This will be the case whether it is based on the type of welfarist or extra-welfarist principles discussed in Sections 2.4.2 and 2.4.3. In addition, there may be other important arguments and social objectives that are difficult to articulate precisely, let alone specify, measure, and value. These might include, for example, different types of concern for equity (e.g. equity of health, income, or income-related health), or more nebulous but potentially no less important objectives, such as the role that access to health care might play in society (e.g. effects on social cohesion and a sense of community) (Olsen and Richardson 1999). As a consequence, economic evaluation using a single societal perspective is unlikely to provide a prescription for social choice. This is because it is likely to be an incomplete and disputed description of all the effects that are regarded as socially valuable.

A more modest role of informing, rather than prescribing, social decisions was described in Section 2.4.3. It is this role that economic evaluation has tended to play in health policy and underpins much of the economic evaluations that have been conducted (Coast et al. 2008; Williams 1993). In part, this explains why many economic evaluations adopt a narrower health care system perspective. This reflects the way a principal (e.g. government) allocates resources and devolves powers to an agent (the health care system or a reimbursement agency). The principal gives responsibilities to pursue explicit, specific, and measureable objectives that are generally agreed to be the primary purpose of health care (e.g. improving health). This is not to say, however, that the other non-health effects are unimportant or can be disregarded. Rather, the purpose of economic evaluation is to set out the scale of the expected effects within and outside health care and the inevitable trade-offs required when decisions are made. In other words, a *multi-sectoral perspective* might be more useful than a single societal one. This would set out the net effects on health within the health care system, the net effects on private consumption opportunities in the wider economy, and the effects on other public sectors where expenditure might also be constrained.

This type of multi-sectoral perspective has already been illustrated in Table 4.4. For example, in Table 4.4 the new technology would not be judged to be cost-effective when considering only its health effects because its approval would lead to a net loss of −2959 QALYs per year (net health benefits are negative). The health benefits it offers are also associated with consumption benefits to the wider economy of £88.4 m. However, the health that is expected to be displaced will also displace £71.8 m of consumption benefits to the wider economy. Therefore, the trade-off that needs to be considered when making a decision about this technology is whether the net loss of 2959 QALYs is worth the net gain of £16.6 m for the wider economy. If the consumption value of a QALY is greater than £5610 then these consumption benefits cannot compensate for loss of health benefits. This type of disaggregated approach enables comparison of the impact of the proposed investment across the different sectors. It presents the aspects of benefit matched with estimates of where the opportunity costs fall, in which sector what aspects of benefit are likely to be displaced. Economic evaluation might not prescribe how these trade-offs ought to be made but, by making them explicit, it can contribute to consistency in decision-making and to a process of accountability, scrutiny and change.

4.6 **Concluding remarks**

Decisions informed by the type of CEA described in Sections 4.2 and 4.4 will ensure health is improved overall so long as the cost-effectiveness threshold used reflects how much health is likely to be displaced elsewhere as a consequence of additional costs. Whether or not these types of 'decision rule' improve the performance of the health care system depends on whether the opportunity costs have been properly taken into account.

There are a number of circumstances where this is likely to be more challenging. For example, if the total additional cost represents a significant net budget impact, then proportionally more health is likely to be displaced; that is, more health is likely to be displaced per $ (see Section 4.3.2.3). In these circumstances, comparing an ICER to a

threshold that represents the health effects of a marginal change in expenditure may not lead to decisions that improve health. This is more likely to be an important consideration in health care systems with particularly limited resources and/or when interventions require investments in infrastructure to deliver them effectively: that is, when an investment is indivisible (all or nothing) and where the additional costs represent a significant net budget impact (Birch and Gafni 2013). As well as infrastructure, equity considerations might mean that an intervention must be implemented for all patients who might benefit, or not offered at all. In other words, horizontal equity considerations can be regarded as constraints that make approval of an intervention 'lumpy' or indivisible so the total additional costs will have a greater net budget impact (Stinnett and Paltiel 1996). In these circumstances, consideration needs to be given to the scale of what is likely to be displaced. This could be based on evidence of how the threshold changes with the scale of changes in expenditure (see Section 4.3.2.3).

Mathematical programming provides a useful analytic framework within which these types of additional constraint and indivisibilities can be considered (see Section 4.4.3.2). As well as reflecting indivisibilities using integer programming (Earnshaw and Dennett 2003), mathematical programming can also be used to reflect other types of constraints such as horizontal equity or the need not to discriminate between certain groups or geographical areas. In many circumstances, there is not a single overall constraint on total health care expenditure, but multiple constraints. For example, health care expenditure might be constrained in different budgetary periods (Epstein et al. 2007). This becomes important when interventions impose additional costs in future periods. One way to reflect whether expenditure is likely to be more or less constrained in the future is to take account of expected changes in the threshold over time (see Section 4.5.2).

Once uncertainty and variability in costs and outcomes are considered, formulating simple 'decision rules' becomes even more challenging. This is especially so when research might be conducted to reduce these uncertainties in the future, but adopting the intervention will commit irrecoverable opportunity costs (see Chapter 11). In principle, uncertainty, variability and different type of budgetary policies can be reflected in stochastic mathematical programming solutions to this complex allocation problem (Al et al. 2005; McKenna et al. 2010; Sendi et al. 2003). Implementing a mathematical programming solution can account for indivisibilities and a range of different types of constraints. As well as ensuring that a decision to adopt a particular technology improves health outcomes overall, it can also identify how to reorganize completely what is provided within a health care system. That is, it can help identify the collection of interventions that will provide the greatest improvements in health given the resources that are made available. The difficulty is the considerable informational requirement necessary to implement such modelling. It would require information not only about the costs and effects of all the types of care currently offered to treat patent (sub)-populations with different diseases, but also about the costs and effects of all the interventions that could be offered.

Unfortunately, these informational requirements are beyond what are available in most circumstances. However, understanding the limitations of what can be said, with the type of information that is available (estimates of the costs and effects of the mutually

exclusive alternatives available to improve the health of a specific patient (sub)-population with a particular condition) is useful in two respects. First, it indicates that an assessment of how opportunity costs are likely to differ is critical: for example, whether these opportunity costs are due to significant net budget impact, indivisibilities, a range of other constraints on expenditure, as well as other types of constraint. Secondly, it suggests that, even if the problem of defining social welfare is set aside (see Sections 4.5.3 and 2.4.3), economic evaluation is unlikely to provide a prescription for health care decisions because it is unlikely to capture the opportunity costs of all the constrains likely to be faced, or to consider all the ways in which the health care system could in principle be reorganized. Moreover, this will be the case whether welfarist or extra-welfarist principles are being used. Rather, the purpose of economic evaluation might be better viewed as informing health care decisions by identifying the best estimates of the likely opportunity cost given the resources available and the current arrangement of the health care system. By doing so, it can also identify the potential implications of other constraints that might be faced and how things might be improved by relaxing them.

References

Al, M.J., Feenstra, T.L., and Van Hout, B.A. (2005). Optimal allocation of resources over healthcare programmes: dealing with decreasing marginal utility and uncertainty. *Health Economics*, **14**, 655–67.

Appleby, J., Devlin, N., Parkin, D., et al. (2009). Searching for cost effectiveness thresholds in the NHS. *Health Policy*, **91**, 239–45.

Baltussen, R. and Niessen, L. (2006). Priority setting of health interventions: the need for multi-criteria decision analysis. *Cost Effectiveness and Resource Allocation*, **4**, 14.

Birch, S. and Gafni, A. (1992). Cost effectiveness/utility analyses: do current decision rules lead us to where we want to be? *Journal of Health Economics*, **11**, 279–96.

Birch, S. and Gafni, A. (2013). Decision rules in economic evaluation revisited, in A. Jones (ed.), *The Elgar companion to health economics*, pp. 528–38. Cheltenham: Edward Elgar.

Boadway, R.W. and Bruce, N. (1984). *Welfare economics*. Oxford: Blackwell.

Bokhari, F.A., Gai, Y., and Gottret, P. (2007). Government health expenditures and health outcomes. *Health Economics*, **16**, 257–73.

Brouwer, W.B.F., Culyer, A.J., van Exel, N.J.A., et al. (2008). Welfarism vs extra-welfarism. *Journal of Health Economics*, **27**, 325–38.

Brouwer, W.B.F., Niessen, L.W., Postma, M.J., et al. (2005). The need for differential discounting of costs and effects in cost-effectiveness analyses. *BMJ*, **331**, 446–8.

Claxton, K., Briggs, A., Buxton, M., et al. (2008). Value based pricing for NHS drugs: an opportunity not to be missed? *BMJ*, **336**, 251–4.

Claxton, K., Martin, S., Soares, M., et al. (2015b). Methods for the estimation of the NICE cost effectiveness threshold. *Health Technology Assessment*, **19**(14), doi10.3310/hta19140.

Claxton, K., Paulden, M., Gravelle, H., et al. (2011b). Discounting and decision making in the economic evaluation of health-care technologies. *Health Economics*, **20**, 2–15.

Claxton, K., Sculpher, M., and Carroll, S. (2011a). *Value-based pricing for pharmaceuticals: Its role, specification and prospects in a newly devolved NHS. CHE Research Paper 60*. <http://www.york.ac.uk/media/che/documents/papers/researchpapers/CHERP60.pdf>. York: Centre for Health Economics, University of York.

Claxton, K., Sculpher, M.J., Culyer, A.J., et al. (2006). Discounting and cost-effectiveness in NICE—stepping back to sort out a confusion. *Health Economics*, **15**, 1–4.

Claxton, K., Sculpher, M., Palmer, S., et al. (2015a). Causes for concern: is NICE failing to uphold its responsibilities to all NHS patients? *Health Economics*, **24**, 1–7.

Claxton, K., Walker, S., Palmer, S., et al. (2010). *Appropriate perspectives for health care decisions. CHE Research Paper* 54. <http://www.york.ac.uk/inst/che/pdf/rp54.pdf>. York: Centre for Health Economics, University of York.

Coast, J., Smith, R., and Lorgelly, P. (2008). Welfarism, extra-welfarism and capability: the spread of ideas in health economics. *Social Science and Medicine*, **67**, 1190–98.

Culyer, A., McCabe, C., Briggs, A., et al. (2007). Searching for a threshold, not setting one: the role of the National Institute for Health and Clinical Excellence. *Journal of Health Services Research and Policy*, **12**, 56–58.

Dakin, H., Devlin, N., Feng, Y., et al. (2014). The influence of cost-effectiveness and other factors on NICE decisions. *Health Economics*, DOI: 10.1002/hec.3086.

Danzon, P.M. (2014). Pricing and reimbursement of biopharmaceuticals and medical devices in the USA, in A.J. Culyer (ed.), *Encyclopedia of health economics*, pp. 127–35. Amsterdam: Elsevier.

Danzon, P.M., Towse, A., and Mestre-Ferrandiz, J.M. (2012). *Value-based differential pricing: efficient prices for drugs in a global context. National Bureau of Economic Research Working Paper w18593*. Cambridge, MA: National Bureau of Economic Research.

Danzon, P.M., Towse, A.K., and Mulcahy, A. (2011). Setting cost-effectiveness thresholds as a means to achieve appropriate drug prices in rich and poor countries. *Health Affairs*, **30**, 1529–38.

Devlin, N. and Parkin, D. (2004). Does NICE have a cost-effectiveness threshold and what other factors influence its decisions? A binary choice analysis. *Health Economics*, **13**, 437–52.

Devlin, N.J. and Sussex, J. (2011). *Incorporating multiple criteria in HTA. Methods and processes*. London: Office of Health Economics.

Drummond, M., Jönsson, B., Rutten, F., et al. (2011). Reimbursement of pharmaceuticals: reference pricing versus health technology assessment. *European Journal of Health Economics*, **12**, 263–71.

Drummond, M., Torrance, G., and Mason, J. (1993). Cost-effectiveness league tables: more harm than good? *Social Science Medical*, **37**, 33–40.

Drummond, M.F., Wilson, D.A., Kanovos, P., et al. (2007). Assessing the economic challenges posed by orphan drugs. *International Journal of Technology Assessment in Health Care*, **23**, 36–42.

Earnshaw, S.R. and Dennett, S.L. (2003). Integer/linear mathematical programming models—a tool for allocating healthcare resources. *PharmacoEconomics*, 21, 839–51.

Eckermann, S. and Pekarsky, B. (2014). Can the real opportunity cost stand up: displaced services, the straw man outside the room. *PharmacoEconomics*, **32**, 319–25.

Epstein, D.M., Chalabi, Z., Claxton, K., et al. (2007). Efficiency, equity and budgetary policies: informing decisions using mathematical programming. *Medical Decision Making*, **27**, 128–37.

Espinoza, M.A., Manca, M., Claxton, K., et al. (2014). The value of heterogeneity for cost-effectiveness subgroup analysis: conceptual framework and application. *Medical Decision Making*, **34**, 951–64.

Evans, D.B., Chisholm, D., and Tan-Torres Edejer, T. (2013). Generalized cost-effectiveness analysis: principles and practice. *The Elgar companion to health economics*, pp. 479–91. Cheltenham: Edward Elgar.

Feenstra, T.L., van Baal, P.H., Gandjour, A., et al. (2008). Future costs in economic evaluation. A comment on Lee. *Journal of Health Economics*, **27**, 1645–9.

Garber, A.M. and Phelps, C.E. (1997). Economic foundations of cost-effectiveness analysis. *Journal of Health Economics*, **16**, 1–31.

Gerard, K. and Mooney, G. (1993). QALY league tables: handle with care. *Health Economics*, **2**, 59–64.

Jayadev, A. and Stiglitz, J. (2009). Two ideas to increase innovation and reduce pharmaceutical costs and prices. *Health Affairs*, **28**, w165–8.

Jena, A. and Philipson, T. (2007). Cost-effectiveness as a price control. *Health Affairs*, **26**, 696–703.

Johannesson, M., Meltzer, D., and O'Conor, R.M. (1997). Incorporating future costs in medical cost-effectiveness analysis: implications for the cost-effectiveness of the treatment of hypertension. *Medical Decision Making*, **17**, 382–89.

Johannesson, M. and Weinstein, S. (1993). On the decision rules of cost-effectiveness analysis. *Journal of Health Economics*, **12**, 459–67.

Kruse, M., Sørensen, J., and Gyrd-Hansen, D. (2012). Future costs in cost-effectiveness analysis: an empirical assessment. *European Journal of Health Economics*, **13**, 63–70.

Laska, E.M., Meisner, M., Siegel, C., et al. (1999). Ratio-based and net benefit based approaches to health care resource allocation: proofs of optimality and equivalence. *Health Economics*, **8**, 171–4.

Laupacis, A., Feeny, D., Detsky, A.S., et al. (1992). How attractive does a new technology have to be to warrant adoption and utilization? Tentative guidelines for using clinical and economic evaluations. *Canadian Medical Association Journal*, **146**, 473–81.

Lee, R.H. (2008). Future costs in cost effectiveness analysis. *Journal of Health Economics*, **27**, 809–18.

Martin, S., Rice, N., and Smith, P.C. (2008). Does health care spending improve health outcomes? Evidence from English programme budgeting data. *Journal of Health Economics*, **27**, 826–42.

Martin, S., Rice, N., and Smith, P.C. (2012). Comparing costs and outcomes across programmes of health care. *Health Economics*, **21**, 316–37.

McCabe, C., Claxton, K., and Culyer, A.J. (2008). The NICE cost-effectiveness threshold: what it is and what that means. *PharmacoEconomics*, **26**, 733–44.

McCabe, C., Claxton, K., and Tsuchiya, A. (2005). Orphan drugs and the NHS: should we value rarity? *BMJ*, **331**, 1016–19.

McKenna, C., Claxton, K., Chalabi, Z., et al. (2010). Budgetary policies and available actions: a generalisation of decision rules for allocation and research decisions. *Journal of Health Economics*, **29**, 170–81.

Meltzer, D. (1997). Accounting for future costs in medical cost-effectiveness analysis. *Journal of Health Economics*, **16**, 33–64.

Meltzer, D.O. (2013). Future costs in medical cost-effectiveness analysis, in A. Jones (ed.), *The Elgar companion to health economics*, pp. 481–9. Cheltenham: Edward Elgar.

Moreno-Serra, R. and Smith, P.C. (2012). Does progress towards universal health coverage improve population health? *Lancet*, **380**, 917–23.

Moreno-Serra, R. and Smith, P.C. (2015). Broader health coverage is good for the nation's health: evidence from country level panel data. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, **178**, 101–24.

Neumann, P.J., Cohen, J.T., and Weinstein, M.C. (2014). Updating cost-effectiveness—the curious resilience of the $50 000-per-QALY threshold. *New England Journal of Medicine*, **371**, 796–7.

Newall, A.T., Jit, M., and Hutubessy, R. (2014). Are current cost-effectiveness thresholds for low- and middle-income countries useful? Examples from the world of vaccines. *PharmacoEconomics*, **32**, 525–31.

NICE [National Institute for Health and Care Excellence] (2011). *Ranibizumab for the treatment of diabetic macular oedema* (TA237). London: NICE.

NICE [National Institute for Health and Care Excellence] (2013). *Guide to the methods of technology appraisal*. London: NICE.

Nord, E. (2011). Discounting future health benefits: the poverty of consistency arguments. *Health Economics*, **20**, 16–26.

Olsen, J.A. and Richardson, J. (1999). Production gains from health care: what should be included in cost-effectiveness analysis. *Social Science and Medicine*, **49**, 17–26.

Paulden, M. (2014). Time preference and discounting, in A.J. Culyer (ed.), *Encyclopedia of health economics*, pp. 395–403. Amsterdam: Elsevier.

Paulden, M. and Claxton, K. (2012). Budget allocation and the revealed social rate of time preference for health. *Health Economics*, **21**, 612–18.

Pauly, M.V. (1995). Valuing health benefits in monetary terms, in F.A. Sloan (ed.), *Valuing health care. Costs, benefits and effectiveness of pharmaceuticals and other medical technologies*. Cambridge: Cambridge University Press.

Peacock, S. and Mitton, C. (2013). Priority setting methods in health services, in A. Jones (ed.), *The Elgar companion to health economics*, pp. 576–85. Cheltenham: Edward Elgar.

Peacock, S., Richardson, J., Carter, R., et al. (2007). Priority setting in health care using multi-attribute utility theory and programme budgeting and marginal analysis. *Social Science and Medicine*, **64**, 897–910.

Phelps, C.E. and Mushlin, A.I. (1991). On the (near) equivalence of cost-effectiveness and cost-benefit analyses. *International Journal of Technology Assessment in HeaIth Care*, **7**, 12–21.

Rawlins, M.D. and Culyer, A.J. (2004). National Institute for Clinical Excellence and its value judgments. *BMJ*, **329**, 224–27.

Ryen, L. and Svensson, M. (2014). The willingness to pay for a quality-adjusted life year: a review of the empirical literature. *Health Economics*, DOI: 10.1002/hec.3085.

Sculpher, M.J. and Claxton, K. (2012). Real economics needs to reflect real decisions: a response to Johnson. *PharmacoEconomics*, **30**, 133–6.

Sendi, P., Al, M.J., Gafni, A., et al. (2003). Optimizing a portfolio of health care programs in the presence of uncertainty and constrained resources. *Social Science and Medicine*, **57**, 2207–15.

Smith, D. and Gravelle, H. (2001). The practice of discounting in economic evaluations of healthcare interventions. *International Journal of Technology Assessment in Health Care*, **17**, 236–43.

Stinnett, A. and Mullahy, J. (1998). Net health benefits: a new framework for the analysis of uncertainty in cost-effectiveness analyses. *Medical Decision Making*, **18**, S68–S80.

Stinnett, A.A. and Paltiel, A.D. (1996). Mathematical programming for the efficient allocation of health care resources. *Journal of Health Economics*, **15**, 641–53.

Sugden, R. and Williams, A.H. (1979). *The principles of practical cost-benefit analysis.* Oxford: Oxford University Press.

Tsuchiya, A. and Williams, A. (2001). Welfare economics and economic evaluation, in M. Drummond and A. McGuire (ed.), *Economic evaluation in health care: merging theory with practice*. Oxford: Oxford University Press.

van Baal, P., Meltzer, D., and Brouwer, W.B.F. (2013). Pharmacoeconomic guidelines should prescribe inclusion of indirect medical costs! A response to Grima et al. *PharmacoEconomics*, **31**, 375–76.

van Baal, P., Meltzer, D.O., and Brouwer, W.B. (2014). Future costs, fixed health care budgets and the decision rules of cost effectiveness analysis. *Health Economics*, doi: 10.1002/hec.3138. [Epub ahead of print.]

Weinstein, M.C. (2013). Decision rules for incremental cost-effectiveness analysis, in A.M. Jones (ed.), *The Elgar companion to health economics*, pp. 505–14. Cheltenham: Edward Elgar.

Weinstein, M.C. and Manning, W.G. (1997). Theoretical issues in cost-effectiveness analysis. *Journal of Health Economics*, **16**, 121–28.

Weinstein, M.C. and Stason, W.B. (1977). Foundations of cost-effectiveness analysis for health and medical practices. *New England Journal of Medicine*, **296**, 716–21.

Weinstein, M.C. and Zeckhauser, R. (1973). Critical ratios and efficient allocation. *Journal of Public Economics*, **2**, 147–57.

Williams, A. (1993). Priorities and research strategy in health economics for the 1990s. *Health Economics*, **2**, 295–302.

Chapter 5

Measuring and valuing effects: health gain

5.1 Some basics

The starting point for the assessment of health gain in economic evaluations is the measurement of health effects. These can be improvements in the main health outcome of interest (e.g. survival) or other effects, such as the side effects of therapy, which could impact on health-related quality of life either positively or negatively. These data come from both randomized controlled trials and observational data sets, such as registries, administrative databases, clinical series and long-term epidemiological studies.

Randomized controlled trials are the best source for estimates of relative clinical effect (comparing two or more therapies), although these are sometimes obtained from observational studies when no relevant trials are available. Estimates of some other effects, particularly rare side effects or effects that only become apparent in the long term, are usually obtained from observational studies, since these studies tend to have long-term follow-up.

Health effects can be incorporated in economic evaluations in two main ways. First, an economic evaluation can be conducted alongside a single clinical study, using individual patient data. This is discussed further in Chapter 8. Alternatively, an economic evaluation can be conducted using decision-analytic modelling. In this case, models usually incorporate a synthesis of data on health effects from a number of clinical studies. Modelling studies and issues in the synthesis of data are discussed further in Chapters 9 and 10 respectively.

The main focus in this chapter is on the ways in which health effects can be used to assess health gain in the context of economic evaluations, but it should be remembered that the identification and measurement of health effects can also be important in the estimation of costs. For example, a bleeding complication not only has important impacts on the health of the patient, but will also require health care resources to treat it. Indeed, economic evaluations conducted using individual patient data and those using decision-analytic models view health effects as events potentially resulting in both costs and changes in health status.

In this chapter we explore how health effects can be used as measures of health gain in economic evaluations and the circumstances under which particular measurement approaches are most suitable. We begin with consideration of clinical data alone, moving on to descriptive (health-related) quality-of-life (QoL) data and finally generic measures of health gain. The latter measures involve the formal elicitation of preferences for

health states. The main methods for doing this are also explained, along with the key methodological and practical issues that arise.

5.2 Using health effects in economic evaluation

As discussed in Chapter 2, there are several good reasons for using generic measures of outcome in economic evaluations, even when there is a single clinical outcome from therapy, such as increasing survival. This is especially true in budget-constrained systems, where it is important to compare the health gain achieved by the treatment with that which will be forgone in the alternative treatments that the new treatment displaces. However, the literature on economic evaluations contains studies using several types of outcome measures. Therefore, these are discussed before the chapter focuses on the issues surrounding the derivation and use of the generic measures, such as the quality-adjusted life-year (QALY).

5.2.1 Clinical outcomes

Some economic evaluations use the outcomes, as reported in the relevant clinical study, as the measure of health gain. This approach can only be considered appropriate when there is one major objective of therapy. For example, therapies having the primary objective of extending life, such as renal dialysis or treatment for advanced stages of cancer, could be assessed in terms of their cost per year of life gained, as compared with the relevant alternative(s). Therapies could be compared both within the particular medical field of interest and with life-extending therapies in other fields.

Of course, the usefulness of this approach rests on whether there is truly one major objective of therapy. In a review of economic evaluations of cancer therapies, Tengs et al. (2000) established that life extension accounted for 90% of the total gain in health. However, it is likely that patients would also be interested in the quality of life during the extra years gained. Indeed, in the case of cancer therapies with high toxicity, the patient may face a trade-off between length of life and quality of life. It is also known that different modalities of dialysis (e.g. hospital dialysis, home dialysis, continuous ambulatory peritoneal dialysis) have differing impacts on quality of life. Therefore, an economic evaluation based on the cost per life-year gained is at best incomplete and could also be misleading.

Comparisons of life-saving therapies using life-years gained as the measure of benefit at least has the advantage that this clinical outcome measure is generic across all life-extending therapies. The problem with using most clinical outcome measures is that they are specific to the clinical field concerned (e.g. seizure reduction for epilepsy, symptom-free days for asthma, change in ACR 20 for rheumatoid arthritis). Economic evaluations could still be conducted to examine some constrained choices. This may be helpful for a decision-maker choosing among therapies for the same condition, but still poses difficulties in interpretation. For example, if one therapy is more expensive than another, but delivers more clinical improvement, how do decision-makers assess whether a given improvement is worth the extra cost? Perhaps one could compare the incremental cost per unit of improvement with that obtained from other interventions in the same clinical field, although this assumes

that the measure being used encapsulates all the relevant health changes resulting from the various treatments.

This is the approach currently advised by the Institute for Quality and Efficiency in Health Care (IQWiG) in Germany. The institute suggests that, in assessing the cost-effectiveness of new technologies, analysts should construct an 'efficiency frontier' expressing the value for money from existing interventions within the same therapeutic area in terms of their cost per unit of clinical improvement. Then it can be assessed whether a new intervention is on the frontier (i.e. of comparable cost-effectiveness to the last new therapy in the therapeutic area concerned) (IQWiG 2011). However, the authors acknowledge that this approach may be complicated if there are several relevant clinical outcomes in the therapeutic area concerned. In such cases, a composite measure of outcome would need to be constructed.

Nevertheless, Sculpher and Claxton (2010) point out that the proposed efficiency frontier requires the apparently broad terms of 'therapeutic' or 'disease area' to be interpreted very narrowly, defining sufficiently similar groups of patients to ensure there is a common choice between (mutually exclusive) alternatives. This is much narrower than disease areas (e.g. breast cancer, whether early or metastatic disease) or even diagnostic results (e.g. HER 2 positive). Indeed, the licensed indications for medicines commonly include a number of such therapeutic areas. Furthermore, they argue that only by comparing the health gains from the new treatment with the other activities displaced can the question of value be addressed. It is unlikely that a new treatment with additional costs will only displace activities within a narrowly defined 'therapeutic area'. Even, if as is sometimes argued, there is no formal constraint on health expenditure in a given jurisdiction, an increase in health care costs will displace activities outside the health care sector—for example, private consumption if taxation or social insurance payments are increased.

Another important issue in using clinical outcomes is that the end points measured in the clinical studies may not be final end points. Survival, or life-years gained, is a final end point in that it relates directly to the patient's health status. However, because of the challenges, such as length of follow-up and sample size, required to assess final end points, many clinical trials assess intermediate end points, such as change in blood pressure, LDL cholesterol, bone mineral density, or time to progression. Such measures are only useful to the extent that they are good predictors of the relevant final end point.

In some medical fields, epidemiological studies have been used to establish the link between intermediate and final end points. For example, Ciani et al. (2013) explored the validity of complete cytogenetic response and major molecular response at 12 months as predictors of overall survival in first-line treatment of chronic myelogenous leukemia by undertaking a systematic review and meta-analysis of the available observational studies. In this case, policy-makers were willing to accept the observational association, but this may not be the case for all therapeutic areas, as there may not be convincing evidence on the relationship between the intermediate and final end points.

Therefore, although it would be possible to compare treatment strategies in terms of their cost per mmHg blood pressure reduction or cost per unit of LDL cholesterol reduction, most economic evaluations take advantage of the epidemiological data, where it exists, to model the changes in final outcomes. (Examples of such studies are given in

Chapter 9.) It goes without saying that in clinical fields where the links between intermediate and final outcomes have not been established, an economic evaluation using an intermediate outcome could be misleading.

Finally, intermediate outcomes may also be used in clinical studies of diagnostic and preventive interventions, with measurements being made of 'patients diagnosed', 'cases detected', or 'cases averted'. An economic evaluation based on these intermediate measures rests on the proposition that achieving the intermediate outcome leads to cost-effectiveness in the long run. For example, in the case of a diagnostic test or screening strategy, the necessary requirement is that there is a clinical and cost-effective therapy to treat the cases diagnosed or detected.

Even if there is such a therapeutic intervention, it is still important to proceed with caution. A more expensive diagnostic strategy may detect more patients suffering from the condition, but the additional cases detected may be in individuals with less serious forms of the disease. Therefore, treating the additional cases may be less valuable than treating those more serious cases that are also detected by the existing strategy. Thus, as with many of the other clinical outcomes discussed above, an economic evaluation using cases detected, or cases averted, as the measure of benefit is at best partial and potentially misleading.

5.2.2 Quality-of-life measures

There is increasing interest in including patient-reported outcome measures (PROs) in clinical studies. PROs include measures of patient satisfaction and health-related quality of life and capture aspects of treatment effect that may not be captured in the main clinical outcomes. Since quality-of-life measures focus on treatment effects that, in principle, impact on the patient's well-being, it is worth considering their relevance for economic evaluation.

There are two categories of QoL measures.

- *Disease-specific, or condition-specific* measures concentrate on the main quality-of-life impacts of a particular disease. Examples include the St George's Asthma Quality of Life Scale (Jones et al. 2002) and the EORTC instrument (Fayers and Bottomly 2002) in cancer. Although these measures may be useful in assessing the efficacy of treatment, their use in economic evaluation is limited to comparisons of treatments within the disease in question. In this respect they suffer from the same limitations as the clinical outcomes, but may have the additional complication that they are normally multidimensional. Therefore, unless the measure has an accompanying algorithm for generating a summary measure, it will be difficult to compare two treatments that perform differently in the different dimensions of quality of life, unless one treatment dominates the other in all respects.
- *Generic measures* do not focus on the impacts of a particular disease. Rather, they consider a broad range of dimensions of quality of life that, in principle, could be impacted by any disease, including physical function, mental well-being, social function, and pain. The most widely used measure of this type is the Short Form (SF)-36 (Ware et al. 1993). In principle, generic QoL measures have broader

application, but, as is the case of disease-specific measures, comparisons of treatments will usually require an algorithm to generate a summary score. Some measures have such an algorithm, others do not.

However, when we consider the perspective of a decision-maker wishing to allocate health care resources, two other issues need to be considered. First, the impact of treatments on survival needs to be considered alongside their impacts on quality of life. Indeed, some treatments (e.g. chemotherapy for cancer) may involve a trade-off between length and quality of life. Secondly, it is important to consider how such trade-offs are made; for example, whose preferences are involved and how are they elicited? Some of the algorithms used to generate a summary outcome in quality-of-life measures are based on a simple scoring system, not a formal consideration of preferences for different health effects or health states.

5.2.3 Generic measures of health gain

It should be clear from the discussion above that the ideal measure of health effect for assessing benefit in economic evaluations would be a generic measure of health gain that (1) encompasses the major elements of changes in length and quality of life, and (2) is based on the formal consideration of preferences for health states. For a decision-maker facing constraints on the overall growth in health care expenditure, such a measure would also have the advantage that not only can the health gained from adoption of the new intervention be assessed, but also the loss in health resulting from abandoning interventions that can no longer be funded as a result. There are a number of generic measures of health gain, the most widely used being the QALY. A widely used measure in economic evaluations conducted in developing countries is the disability-adjusted life-year or DALY (Tan-Torres Edejer et al. 2003).

A taxonomy of the alternative measures of health effect is given in Figure 5.1. Since the generic measures of health gain have the widest application in economic evaluation, these are discussed in more detail below.

5.2.3.1 The quality-adjusted life-year (QALY)

The concept of the QALY was first introduced in 1968 by Herbert Klarman and colleagues in a study on chronic renal failure (Klarman et al. 1968). They noted that the quality of life with a kidney transplant was better than that with dialysis, and estimated that it was 25% better. The cost per life-year gained by the different treatment options was calculated with and without this quality adjustment. Although they did not use the term 'quality-adjusted life-year', the concept was identical.

As was mentioned above, the advantage of the QALY as a measure of health outcome is that it can simultaneously capture gains from reduced morbidity (quality gains) and reduced mortality (quantity gains), and combine these into a single measure. Moreover, the combination is based on the relative desirability of the different outcomes. A simple example is displayed in Figure 5.2. Without the intervention the individual's health-related quality of life would deteriorate according to the lower path and the person would die at time Death 1. With the intervention the person would deteriorate more slowly, would live longer, and would die at time Death 2. The area between the

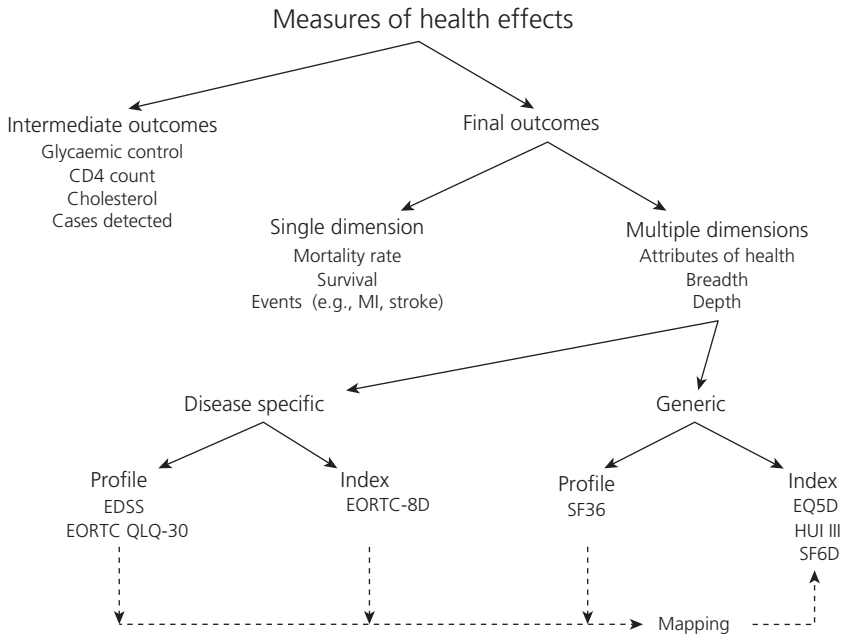


Fig. 5.1 A taxonomy of measures of health effects.

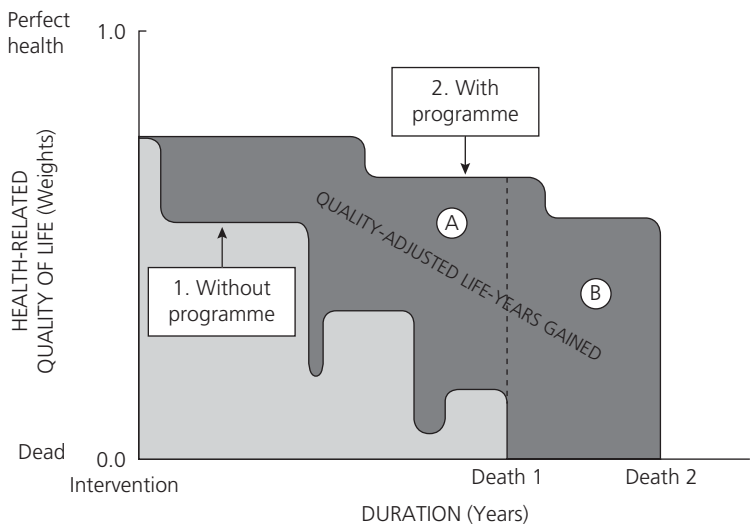


Fig. 5.2 Quality-adjusted life-years gained from an intervention.

Reproduced from Gold, M.R. et al. (ed.), *Cost-effectiveness in health and medicine*, Figure 4.2, pp. 92, Oxford University Press, New York, USA, Copyright © 1996, with permission of Oxford University Press, USA. Source: data from Torrance, G.W., Designing and conducting cost–utility analyses, pp. 1105–11, in B. Spilker (ed.), *Quality of life and pharmacoeconomics in clinical trials*, 2nd edition, Lippincott-Raven, Philadelphia, USA, Copyright © 1996.

two curves is the QALYs gained by the intervention. For instructional purposes the area can be divided into two parts, A and B, as shown. Part A is the amount of QALY gained due to quality improvement (the gain in health-related quality of life during the time that the person would have otherwise been alive anyhow) and part B is the amount of QALY gained due to quantity improvement (the amount of life extension but factored by the quality of that life extension).

Much more complicated cases can be handled. The paths may cross each other. For example, many cancer treatments cause a QALY loss in the short term in order to achieve a QALY gain in the longer term. The paths may be identical for a long time after the intervention and only diverge in the distant future. An example of this pattern could be a hypertension drug that is well tolerated and has no side effects but eventually averts serious cardiovascular events. The paths may be uncertain, reflecting the variability between apparently similar patients in terms of their prognoses. This uncertainty can be characterized by including a series of alternative paths, with the likelihood of a given patient following each being reflected by a probability. With each path having a QALY value associated with it, the expected (or mean) QALY is calculated as sum of the QALY for each pathway weighted by its respective probability. This is an example of a simple decision-analytic model, which is described in more detail in Chapter 9.

To operationalize the QALY concept, as described above, one needs quality weights that represent the health-related quality of life (HRQoL) of the health states under consideration. These quality weights are the scale for the vertical axis in Figure 5.2. The instruments that are used to obtain the required weights are discussed in Section 5.3.

The QALY weights for health states should be based on preferences for the health states. This way the more desirable (more preferred) health states receive greater weight and will be favoured in the analysis. The scale of QALY weights may contain many points, but two points that must be on the scale are perfect health and death. These two are required because they will both occur in programmes being evaluated with the QALY model, and weights will be required for them. Because these two must always be on the scale, and because they are well specified and understood, they are usually selected to be the two anchor points (actually, a better term would be reference points) for the interval scale of QALY weights. This is akin to selecting the freezing point and the boiling point of water to be the anchor points for the interval scale of temperature.

To define an interval scale of QALY weights, death and perfect health can be given any two arbitrary values as long as the value for death is smaller than the value for perfect health. The pair of values could be (32, 212), (0, 100), (−5.9, 2.3), (0, 1), or whatever, and the resulting scale would be an interval scale of QALY weights. However, one pair of scores stands out as particularly convenient (death = 0 and perfect health = 1), and this has become the conventional scale for QALY weights. Note that this still allows states worse than death, which would have scores less than 0, and indeed states better than perfect health, if they exist, which would have scores greater than 1.

Scales of measurement can be nominal (e.g. colours—red, blue, green), ordinal (e.g. size—small, medium, large, extra large, extra extra large), or cardinal (e.g. length in metres, or temperature in °C). Cardinal scales can be interval (e.g. temperature) or ratio (e.g. length). The difference between these two is that the ratio scale has an unambiguous zero point that indicates there is absolutely none of the phenomenon being

measured. For example, if something has a length of zero, it has no length. However, if something has a temperature of zero, it still has a temperature. A convenient memory aid that lists the types of scales in increasing order of their mathematical properties is the French word for black, *noir*, standing for *n*ominal, *o*rdinal, *i*nterval, *r*atio.

Because it has an absolute zero, a ratio scale is unique under a positive multiplicative transformation. This means that any ratio scale can be multiplied by any positive constant and the result is still a ratio scale of the same phenomenon, just in different units. This property is used, for example, to convert feet to yards, or metres to miles. Because it has no natural zero, an interval scale is unique under a positive linear transformation. This means that any interval scale x can be transformed to a scale y using a function $y = a + bx$, where a can be any constant and b can be any positive constant. The result will still be an interval scale of the same phenomenon, but in different units and with a different zero. This property is used to convert °F to °C.

An interval scale has the property that ratios of intervals have meaning, but ratios of scale quantities do not. In a ratio scale both types of ratios have meaning. For example, with temperature, the interval scale property means that it is correct to state that the gain in temperature in going from 40°F to 80°F is twice as much as the gain in going from 40°F to 60°F, but it is incorrect to state that 80°F is twice as hot as 40°F. The former statement holds true whether the temperature is measured in °F or °C, while the latter does not. Conversely, in length, it is both correct to state that the gain in length from 40 metres to 80 metres is twice as much as the gain from 40 metres to 60 metres, and that 80 metres is twice as long as 40 metres. Both statements remain true whether the lengths are measured in metres, inches, miles, fathoms, light years, or any other unit of length.

At first glance it may seem that a scale of HRQoL does have a natural zero at death. After all, death represents no HRQoL. However, the problem is that there could be states worse than death (Patrick et al. 1994; Torrance et al. 1982; Torrance 1984), and these states require a score for their HRQoL. Thus, death is not the bottom of the scale. In fact, there is no well-defined bottom of the scale. Conventionally, as discussed above, death is assigned a zero score and states worse than death take on negative scores. This is akin to the temperature at which water freezes being assigned 0°C and temperatures colder than that being negative.

Finally, it is useful to note that for economic evaluation an interval scale is required for the QALY weights, but an interval scale is all that is required. First, an interval scale is required because it is important that intervals of equal length on the scale have equal interpretation, and this is the fundamental nature of an interval scale. That is, it is important that a gain from 0.2 to 0.4 on the scale represents the same increase in desirability as a gain from 0.6 to 0.8. This is required because in the QALY calculations those two types of gains will appear equal.

Second, an interval scale is all that is required; there is no need to have a ratio scale. There are two reasons. First, because an interval scale is a type of cardinal scale, all parametric statistical calculations are allowed; for example, mean, standard deviation, *t*-test, analysis of variance, and so on. Second, because all economic evaluations are comparative, the analysis is always dealing with differences between the programme and the comparator, and all mathematical manipulations on differences (intervals) are

valid with an interval scale. That is, it is valid to take ratios of differences (the incremental QALYs gained in programme A compared to its comparator are twice those of programme B compared to its comparator), and to use the differences in other ratios (the incremental cost per incremental QALY for programme A is one-third of that for programme B), as well as to perform the statistical tests (the incremental QALYs gained in programme A are not statistically significantly different from those gained in programme B at the 5% level).

Conceptually, the QALY calculation is very straightforward. If both groups start with exactly the same baseline health state, as is the case in Figure 5.2, the incremental QALYs gained is simply the area under path 2 less the area under path 1. If, on the other hand, there is a difference in baseline health state between the two groups, the area between the two curves must be adjusted to account for this difference. The recommended adjustment method uses multiple regression to estimate the incremental QALY and an associated measure of sampling variability (Manca et al. 2005). The area under a path can be thought of as the sum of the areas under each component health state on the path, where the area under a health state is the duration of the health state in years, or fraction of a year, multiplied by the quality weight for the health state. This is the QALYs gained without discounting.

Because individuals, and society, generally prefer gains of all types, including health gains, to occur earlier rather than later, future amounts are multiplied by a discount factor to account for this time preference. The technique of discounting, as applied to costs, is described in detail in Chapter 7. The logic is the same when applied to QALYs (see Chapter 4). Essentially the method consists of taking the amounts that will occur in future years and moving them year by year back to the present, reducing the amount each year by *r*% of the remaining amount, where *r*% is the annual discount rate.

5.2.3.2 The disability-adjusted life-year (DALY)

The concept of DALYs was originally developed by the World Health Organization (WHO) initially for their Global Burden of Disease and Injury study (Murray and Lopez 1996). Subsequently they have been recommended by WHO for use in generalized cost-effectiveness analysis (Tan-Torres Edejer et al. 2003). See Box 5.1.

Since its introduction in 1993 the DALY approach has been heavily debated. For example, Arnesen and Kapiriri (2004) show that the value choices built into the DALY, notably the age weights, the discount rate, and the disability weights, have a major influence on the rankings of programmes, and they argue that these value choices are, in part, arbitrary and are far from transparent. They also conclude that the disability weights are of doubtful validity. Some of the issues in calculating and presenting DALYs are discussed by Rushby and Hanson (2001), who suggest a set of minimum reporting criteria. Comparisons between QALYs and DALYs and the implications for health policy decisions are discussed in the papers by Airoldi and Morton (2009) and Robberstad (2005).

In response to these criticisms and the need to make use of more recent data, WHO undertook a major re-estimation of DALYs for its Global Burden of Disease (GBD) study in 2010. An earlier expert consultation addressed the conceptual, ethical, and measurement issues in undertaking a comprehensive revision of the disability weights

Box 5.1 Disability-adjusted life-years (DALYs)

DALYs, as originally developed, were conceptually similar to QALYs but differed in the following important ways:

- The life expectancy used in the QALY depends on the situation. The life expectancy used in the DALY was constant and was set at the greatest reported national life expectancy, that of Japanese women.
- The disability weights in the QALY are based on preferences, either those of the general public or those of the patients in the study. The disability weights in the DALY were not preferences but were person trade-off scores from a panel of health care workers who met in Geneva in August 1995.
- Although both sets of disability weights are on the same scale where death has a score of 0 and full health has a score of 1, the QALY weights can take on any value depending upon the health state while the DALY weights, in contrast, can only take on one of seven discrete values. That is, in the original DALY system there were only seven health states in addition to dead and healthy.
- The QALY does not use age weights. The initial DALY used age weights that give lower weight to years of young and elderly people.

WHO originally suggested that DALY users should use the age weights in their base case analysis but that a sensitivity analysis could be undertaken without the age weights. This is because the age weights may raise some equity concerns (Tan-Torres Edejer et al. 2003).

(Salomon 2008). In addition, the GBD study undertook a comprehensive re-estimation of disability weights through a large-scale empirical investigation with a major emphasis on surveying respondents from the general population, in which judgements about health losses associated with many causes of disease and injury were elicited through a new standardized approach. The GBD 2010 study estimated disability weights for 220 health states using a method involving paired comparisons of health states described using lay descriptions consisting of a brief summary of the health state of an average or modal case in 30 words or less (Salomon et al. 2012).

The original GBD 1990 study and subsequent WHO updates also incorporated age-weighting in the standard DALYs used in most publications and analyses. The standard age weights gave less weight to years of healthy life lost at young ages and older ages. With the clearer conceptualization of DALYs as purely a measure of population health loss rather than broader aspects of social welfare, it is difficult to justify the inclusion of age weights, and the GBD 2010 study dropped them. The result was a simplified DALY, where the calculation of years of healthy life lost to disability was merely the prevalence of each sequela multiplied by the relevant disability weight (WHO 2013).

Although the primary focus of DALYs remains the estimation of the GBD, they continue to be used in economic evaluations of health care programmes and treatments.

They are most likely to be used when required by the international agency, such as WHO, commissioning the study, or when the economic evaluaton is being conducted in a country for which no local health state preference values exist. The latest information on the development of the DALY estimates can be found at <http://www.who.int/healthinfo/global_burden_disease/en/>.

5.3 Measuring preferences for health states

5.3.1 A note on terminology

Many people use the terms 'utility', 'value', and 'preference' interchangeably, but in fact there are differences. Preference is the umbrella term that describes the overall concept; utilities and values are different types of preferences. What you get depends on how you do the measurements (see Table 5.1). There are two key aspects of the measurement process. One is the way in which the question is framed, specifically whether the outcomes in the question are certain or uncertain. The other is the way in which the subject is asked to respond, specifically whether the subject is asked to perform a scaling task based on introspection or to make a choice.

The term 'utility' is particularly problematic. It has been around for several centuries, has been used by a variety of disciplines, and has a number of related but different meanings (Cooper and Rappoport 1984; Miyamoto 1988; Sen 1991). Thus, it creates a significant potential for confusion and for people to talk past each other. In economics the term has tended to be synonymous with preference; the more preferable an outcome, the more utility associated with it. The differences in meaning arise when approaches are developed to define the concept more precisely and especially when attempts are made to measure it.

In the literature on health state preference measurement reviewed here, the term 'utility' has its origins in the theory of decision-making under uncertainty, as developed

Table 5.1 Methods of measuring preferences

Response method	**Question framing**	
	Certainty (values)	**Uncertainty (vNM utilities)**
	1	2
Scaling	Rating scale	
	Category scaling	
	Visual analogue scale	
	Ratio scale	
	3	4
Choice	Time trade-off	Standard gamble
	Paired comparison	
	Equivalence	
	Person trade-off	

by von Neumann and Morgenstern (1944), often referred to as 'vNM utilities'. Because of the potential confusion over the use of the term 'utility', we will use the more general term 'preferences', unless we are directly referring to literature, or measurement instruments, that use the term 'utility'.

5.3.2 General measurement concepts

Consider a subject being asked preference questions for health outcomes, where each outcome is a specific lifetime path for the subject. That is, each outcome describes a path from now to death consisting of one or more health states for specified time periods. This, in fact, is the most general case of measuring preferences for health outcomes, and all health state preference measurement uses, or should use, this format. Even measuring preferences for single temporary states, such as 1 week of hospitalization for an acute episode of some disease, cannot be done in isolation from what will follow. What will follow should always be described explicitly, else the subject will implicitly assume something and it will affect the measurement in unknown ways.

A question framed under certainty would ask the subject to compare two or more outcomes and to choose between them or to scale them. In thinking about each outcome, the subject is asked to assume that the outcome would occur with certainty. There are no unknowns and no probabilities in the way the various futures are described. A question framed under uncertainty would ask the subject to compare two alternatives, where at least one of the alternatives contained uncertainty; that is, it contained probabilities. The conventional standard gamble question, described in Section 5.4.2, is a common example. The difference between these two forms of questioning is that the certainty method does not capture the subject's risk attitude, while the uncertainty method does.

Risk attitude is a well-known concept in preference measurements and utility theory (Gafni and Torrance 1984; Holloway 1979; Keeney and Raiffa 1976). The intuitive notion is that if a person shies away from more risky alternatives in favour of less risky alternatives, they are risk averse. If they are indifferent, they are risk neutral, and if they prefer risky situations, they are risk seeking. Mathematically, the concept can only be operationalized when measuring preferences over outcomes that are themselves defined on an interval scale. Then, the definition is that if the subject prefers the expected value of an uncertain alternative to the uncertain alternative itself, the subject is risk averse; indifference between the two represents risk neutrality; and a preference for the gamble indicates a risk-seeking attitude.

For example, a subject who prefers $100 for sure to a 50/50 gamble of receiving $0 or $200 would be said to be risk averse with respect to money. On the other hand, if the subject was indifferent between the two, they would be risk neutral; and if the gamble was preferred, they would be risk-seeking. Similarly, a subject who rated three health outcomes, A, B, and C, on a visual analogue scale (VAS) (see Section 5.4.1) as valued at 0.4, 0.6, and 0.8, who then preferred outcome B for sure to a 50/50 gamble of receiving outcome A or C, would be said to be risk averse with respect to value. As in the case for money, if the subject had been indifferent, they would have been classed as risk neutral; and if they had preferred the gamble, they would be called risk seeking. Risk attitude

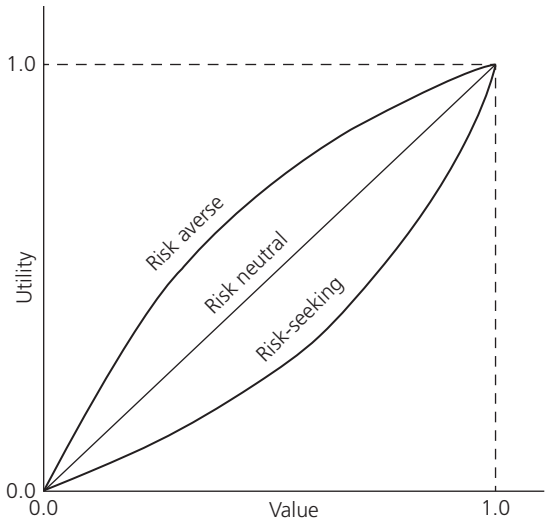


Fig. 5.3 The three generic types of relative risk attitude.

with respect to values is sometimes called relative risk attitude (Dyer and Sarin 1979, 1982; Torrance et al. 1995) to differentiate it from risk attitudes with respect to fundamental consequences like dollars or years of healthy life.

Note that the risk attitude really only pertains to a specific question. There is no requirement that a person have a consistent risk attitude over multiple questions. However, the existence of a consistent risk attitude that can be modelled mathematically is often assumed for practical convenience. For example, the three generic types of relative risk attitude are shown in Figure 5.3. As the figure shows, a person whose relative risk attitude is consistently risk averse over the length of the scale will have utilities (preferences adjusted for risk) that exceed their values (riskless preferences). Empirically, this is the common finding.

The second dimension of Table 5.1 refers to the response method. A subject can be asked to determine a strength of preference by introspection and to indicate the result on a numerical scale. Alternatively, a subject can be asked to choose between two alternatives, thus revealing their preference indirectly. The advantage of scaling is that it takes the respondent less time. The advantage of the choice-based methods is that choosing, unlike scaling, is a natural human task at which we all have considerable experience, and furthermore it is observable and verifiable. Thus, many analysts prefer the choice-based methods in designing studies.

Table 5.1 is divided into four cells. Cell 1 contains instruments that require the subject to think introspectively about outcomes presented with certainty and to provide a rating or a score. Rating scales (assign a number), category scales (assign a category), and VASs (mark a line) are all variations on the same theme. Ratio scaling, as used by Rosser and colleagues (Rosser and Kind 1978; Rosser and Watts 1978), also belongs in this category. In ratio scaling, subjects were asked to indicate how many times worse one outcome was compared to the next best outcome. The outcomes were defined with

certainty and the task was one of introspection. There are no instruments that fall in cell 2 to our knowledge, although presumably one could ask subjects to rate their preferences for gamble alternatives. Cell 3 contains the time trade-off (TTO) approach (see Section 5.4.3), the paired comparison approach (Hadorn et al. 1992; Hadorn and Uebersax 1995; Streiner and Norman 1989), and the old equivalence approach (Patrick et al. 1973; Patrick and Erickson 1993), now renamed the person trade-off (PTO) approach (Green 2001; Nord 1995, 1996, 1999; Nord et al. 1993). Finally, cell 4 contains the well-known standard gamble in all its variations (Bennett and Torrance 1996; Furlong et al. 1990; O'Brien et al. 1994; Torrance 1986; Torrance and Feeny 1989; Torrance et al. 2002).

To summarize, all of the methods in Table 5.1 measure preferences. Those in cells 1 and 3 measure values; those in cell 4 measure what, in the decision analysis literature, are called 'utilities'. Because the task is different in each cell, one should not be surprised that the resulting preference scores will differ. Indeed, the common finding is that, for states preferred to death, standard gamble scores are greater than TTO scores, which in turn are greater than visual analogue scores (Bass et al. 1994; Bennett and Torrance 1996; Churchill et al. 1987; O'Leary et al. 1995; Read et al. 1984; Rutten-van Molken et al. 1995; Stiggelbout et al. 1994; Torrance 1976; Wolfson et al. 1982). However, one study produced the contrary finding of TTO scores exceeding standard gamble scores (Dolan et al. 1996a). The reason given for the differences between cells 3 and 4 is risk attitude, which is only captured in cell 4. The reason for the difference between cells 1 and 3 presumably lies in the difference between choosing and scaling.

Which method is best? As indicated earlier, other things being equal, most health economists prefer choice-based methods over scaling methods. In practice, other things are not equal, notably the time required to use the different approaches. In addition, preferences estimated under conditions of uncertainty, incorporating individuals' risk attitude, are most relevant to the majority of decision-making situations, which typically involve uncertainty. Therefore, many analysts (Gold et al. 1996; Mehrez and Gafni 1991) argue that because future health outcomes are clearly uncertain in the real world, the preferences measured under uncertainty are the more appropriate. It should be noted, however, that these theoretical arguments are technically only valid at the individual level. Von Neumann–Morgenstern utility theory only covers individual decision-making, and once we aggregate the preferences across the respondents and use the results to inform societal decision-making, the theory no longer directly applies. On the other hand, the theory would apply if we assume that society is a single individual with preferences equal to the mean preferences of the community.

As a final caveat, users of economic evaluation studies and preference-scored health status classification systems should be aware that all of these methods are in use. Users should check carefully to determine what method was used in studies or pre-scored instruments of interest to them, and to ensure that the method suits their purpose.

5.4 **Methods for measuring preferences**

The various methods for measuring preferences are summarized briefly in this section. Further descriptions of most of the methods are available in the literature. A detailed

technical manual describing how to build and use standard gamble boards, TTO boards, and feeling thermometers (VASs) is available (Furlong et al. 1990). A video demonstrating an interview using these instruments can also be obtained (O'Brien et al. 1994). The book by Spilker (1996) contains descriptions of the standard gamble, the TTO, and VASs in Chapters 12 and 27. The book by Gold et al. (1996) contains a brief summary of a variety of measurement approaches in Chapter 6. Journal articles covering the three main techniques are also available (Torrance 1986; Torrance et al. 2002).

The three most widely used techniques to measure directly the preferences of individuals for health outcomes are the rating scale and its variants, the standard gamble, and the TTO. These three are summarized below.

5.4.1 Rating scale, category scaling, and visual analogue scale (VAS)

The simplest approach to measuring preferences is to ask subjects first to rank health outcomes from most preferred to least preferred, and second, to place the outcomes on a scale such that the intervals or spacing between placements correspond to the differences in preference as perceived by the subject. That is, outcomes that are almost equally desirable would be placed close together while outcomes that are very different in desirability would be placed far apart. The subject should be instructed to concentrate on these intervals and comparisons of one interval to another, rather than on the scores themselves. The purpose is to encourage the subject to produce an interval scale of preferences. Note that because ratios of scale values are meaningless in an interval scale, it is inappropriate for subjects to make comparisons like 'outcome A is twice as desirable as outcome B and so I will place it twice as high on the scale'. The correct comparisons are ones like, 'the difference in desirability between outcomes A and B is twice as great as the difference between C and D, hence I will make the interval between A and B twice as large'.

There are a number of variations on the rating scale approach. The scale can have numbers (e.g. 0–100), categories (e.g. 0–10), or just consist of a 10 cm line on a page. The different variations often have different names. Rating scale usually refers to a scale of numbers, often 0–100. Category rating or category scaling is the variation that consists of a small number of categories, often 10 or 11, that the subject is to assume to be equally spaced. A VAS consists of a line on a page, often 10 cm in length, with clearly defined end points and with or without other marks along the line.

Preferences for chronic states can be measured on a rating scale. The chronic states are described to the subject as irreversible; that is, they are to be considered permanent from age of onset until death. The subject must be provided with the duration of time for which the state will be experienced, and this should be the same for all states that are measured together relative to each other in one batch. States with different ages of onset and/or ages of death can be handled by using multiple batches. Two additional chronic states are added to each batch as reference states for the scale—healthy (from age of onset to age of death) and death (at age of onset).

The subject is asked to select the best health state of the batch, which presumably would be 'normal healthy life' and the worst state, which may or may not be 'death at

age of onset' and to place these at the ends of the scale. They are then asked to locate the other states on the rating scale relative to each other such that the distances between the locations are proportional to their preference differences. The rating scale is measured between 0 at one end and 1 at the other end. If death is judged to be the worst state and placed at 0 on the rating scale, the preference value for each of the other states is simply the scale value of its placement. If death is not judged to be the worst state but is placed at some intermediate point on the scale, say d, the preference values for the other states are given by the formula $(x - d)/(1 - d)$, where x is the scale placement of the health state.

Note that if the respondent places the best and/or the worst state near, but not at, the ends of the scale, the formula above must be modified. A simple way to handle this situation is to linearly rescale the interval between the worst and best states on to 0–1, and then to proceed with the formulae as shown above. This approach, for example, is needed when using the VAS included in the EuroQoL Group's EQ-5D instrument, assuming the researcher wants the scores on the conventional dead–healthy 0–1 scale. The approach is needed because the VAS in the EQ-5D has ends labelled 'best imaginable health state' and 'worst imaginable health state', which encourages respondents to place actual states not at the ends (one can always imagine something better or worse).

Preferences for temporary health states can also be measured on a rating scale. Temporary states are described to the subject as lasting for a specified duration of time at the end of which the person returns to normal health. As with chronic states, temporary states of the same duration and same age of onset should be batched together for measurement. Each batch should have one additional state, 'healthy', added to it. The subject is then asked to place the best state (healthy) at one end of the scale and the worst temporary state at the other end. The remaining temporary states are located on the scale such that the distances between the locations are proportional to the subject's preference differences.

If the programmes being evaluated involve only morbidity and not mortality and if there is no need to compare the findings to programmes that do involve mortality, the procedure described above for temporary health states is sufficient. However, if this is not the case, the interval preference values for the temporary states must be transformed on to the standard 0–1 health preference scale. This can be done by redefining the worst temporary health state as a chronic state of the same duration, and measuring its preference value by the technique described for chronic states. The values for the other temporary health states can then be transformed on to the standard 0–1 dead–healthy scale by a positive linear transformation (just like converting °F to °C).

Scores from a rating scale give the investigator a firm indication of the ordinal rankings of the health outcomes, and some information on the intensity of those preferences. However, rating scales are subject to measurement biases, and the empirical findings are that when compared to preferences measured by the standard gamble or the TTO, the rating scale scores are not an interval scale of preferences (Bleichrodt and Johannesson 1997; Robinson et al. 2001; Torrance 1976; Torrance et al. 1982, 1996, 2001). Notable biases that seem to be at work are the end-of-scale bias, in which subjects tend to shy away from using the ends of the scale, and the context bias, in which subjects tend to space out the outcomes over the scale regardless of how good or bad the states are

(Bleichrodt and Johannesson 1997; Torrance et al. 2001). However, Parkin and Devlin (2006) have shown that in some circumstances the VAS performs quite well.

Empirical findings indicate that rating scale scores can be converted to standard gamble or TTO scores by using a power curve conversion (Torrance 1976; Torrance et al. 1982, 1996, 2001). Thus, one approach is to use the rating scale method, which is quick and efficient, and to convert the resulting scores to utilities by a suitable power curve conversion. However, in a review of seven studies exploring the relationship between VAS and one of the choice-based approaches, Brazier et al. (2003) found that there was not a stable relationship and no evidence of a power function performing better, in statistical terms, than the linear form. Therefore, they conclude that obtaining values for health states and then mapping then onto one of the choice-based methods for eliciting preferences can only ever be second best compared with the direct use of a choice-based technique.

A second approach, which is not mutually exclusive, is to use the rating scale task primarily as a warm-up for subjects, to familiarize them with the descriptions of the outcomes, and to have them begin to think hard about their preferences before measuring the important preferences by some other technique.

5.4.2 Standard gamble

The standard gamble is the classical method of measuring cardinal preferences. It is based directly on the fundamental axioms of utility theory, first presented by von Neumann and Morgenstern (1944) (see Box 5.2). In fact, the standard gamble method is a direct application of the third axiom in Box 5.2. The method has been used extensively in the field of decision analysis, and good descriptions of the methods are available in books in this field (see e.g. Holloway 1979).

The method can be used to measure preferences for chronic states but the approach varies somewhat depending upon whether or not the chronic state is preferred to death or considered worse than death. For chronic states preferred to death the method is displayed in Figure 5.4. The subject is offered two alternatives. Alternative 1 is a treatment with two possible outcomes: either the patient is returned to perfect health and lives for an additional t years (probability P), or the patient dies immediately (probability $1 - P$). Alternative 2 has the certain outcome of chronic state i for life (t years). Probability P is varied until the respondent is indifferent between the two alternatives, at which point the required preference score for state i for time t is simply P; that is, $h_i = P$. Here, h_i is measured on a scale where perfect health for t years is 1.0 and immediate death is 0.0.

Because most subjects cannot readily relate to probabilities, the standard gamble is often supplemented with the use of visual aids, particularly a probability wheel (Furlong et al. 1990; Torrance 1976). This is an adjustable disc with two sectors, each of different colour, and constructed so that the relative size of the two sectors can be readily changed. The alternatives are displayed to the subject on cards, and the two outcomes of the gamble alternative are colour-keyed to the two sectors of the probability wheel. The subject is told that the chance of each outcome is proportional to the similarly coloured area of the disc.

Preferences for temporary health states can be measured relative to each other using the standard gamble method as shown in Figure 5.5. Here intermediate states i are

Box 5.2 Axioms of von Neumann–Morgenstern utility theory

The original axioms of von Neumann and Morgenstern have been refined and restated over the years by various authors. Bell and Farquhar (1986) present the axioms as follows.

1 *Preference exist and are transitive*. For any pair of risky prospects y and y' either y is preferred to y', y' is preferred to y, or the individual is indifferent between y and y'. In addition, for any three risky prospects, y, y', and y'', if y is preferred to y', and y' is preferred to y'', then y is preferred to y''; similarly, if y is indifferent to y', and y' is indifferent to y'', then y is indifferent to y''.

2 *Independence*. An individual should be indifferent between a two-stage risky prospect and its probabilistically equivalent one-stage counterpart derived using the ordinary laws of probability. For example, consider two risky prospects y and y' where y is made up of outcome x_1 with probability p_1 and outcome x_2 with probability $(1 - p_1)$, indicated symbolically as $y = \{p_1, x_1, x_2\}$, and $y' = \{p_2, x_1, x_2\}$. This axiom implies that an individual would be indifferent between the two-stage risky prospect (p, y, y'), and its probabilistically equivalent one-stage counterpart $\{pp_1 + (1 - p)p_2, x_1, x_2\}$.

3 *Continuity of preferences*. If there are three outcomes such that x_1 is preferred to x_2, which is preferred to x_3, there is some probability p at which the individual is indifferent between outcome x_2 with certainty or receiving the risky prospect made up of outcome x_1 with probability p and outcome x_3 with probability $1 - p$.

measured relative to the best state (healthy) and the worst state (temporary state j). Note that all states must last for the same duration, say t, followed by a common state, usually healthy. In this format the formula for the preference value of state i for time t is $h_i = P + (1 - P)h_j$, where i is the state being measured and j is the worst state. Here h_i is measured on a scale where perfect health for duration t is 1.0. If death is not a

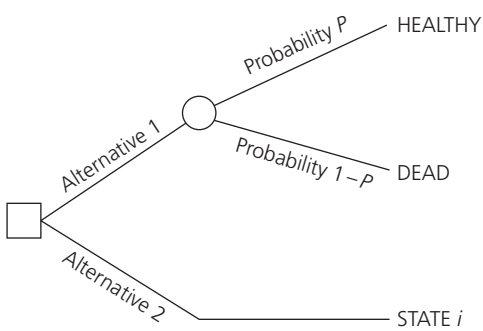


Fig. 5.4 Standard gamble for a chronic health state preferred to death.

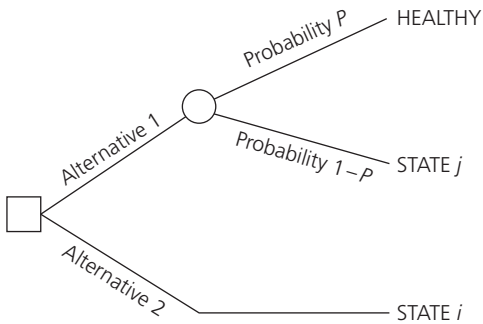


Fig. 5.5 Standard gamble for a temporary health state.

consideration in the use of the preference values, h_j can be set equal to zero and the h_i values determined from the formula, which then reduces to $h_i = P$. However, if it is desired to relate these values to the 0–1 dead–healthy scale, the worst of the temporary states (state *j*) must be redefined as a short-duration chronic state for time *t* followed by death and measured on the 0–1 scale by the technique described above for chronic states. This gives the value for h_j for time *t* which can then, in turn, be used in the above formula to find the value for h_i for time *t*.

Variations on this method are also possible. For example, in Figure 5.5 state *j* can be the state considered next best compared to state *i*, rather than being the worst state. This does not change the formula $h_i = P + (1 - P)h_j$ but it does mean that the *h* values for the states have to be solved in sequence, starting with the worst state.

The traditional method of obtaining standard gamble measurements is through individual face-to-face interviews with the subjects, complete with carefully scripted interviews and helpful visual aids (Furlong et al. 1990). Other, more efficient techniques are, however, being developed. These include interactive computer approaches (Lenert 2001), paper-based approaches (Ross et al. 2003), and group interviews with paper-based response (Gorber 2003).

5.4.3 Time trade-off

The TTO method was developed specifically for use in health care by Torrance et al. (1972). It was originally developed as a simple, easy-to-administer instrument that gave comparable scores to the standard gamble (Torrance 1976). Subsequently, its theoretical properties have been explored (Bleichrodt 2002; Mehrez and Gafni 1990), and further empirical work indicates that TTO scores require adjustment before they can be used as vNM utilities (Martin et al. 2000).

The application of the TTO technique to a chronic state considered better than death is shown in Figure 5.6. The subject is offered two alternatives:

- state *i* for time *t* (life expectancy of an individual with the chronic condition) followed by death
- healthy for time $x < t$ followed by death.

Time *x* is varied until the respondent is indifferent between the two alternatives, at which point the required preference score for state *i* is given, $h_i = x/t$.

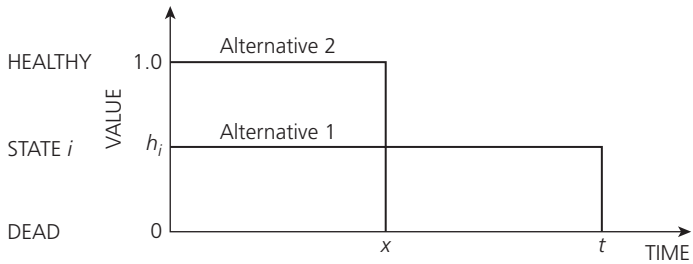


Fig. 5.6 Time trade-off for a chronic health state preferred to death.

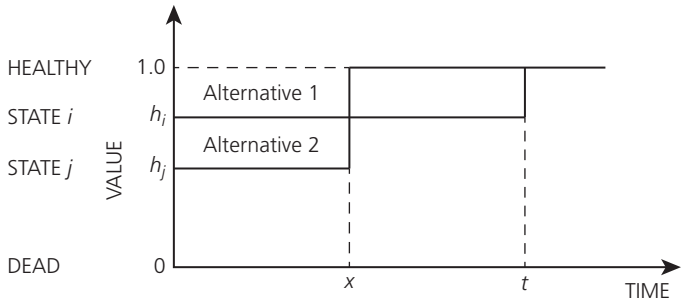


Fig. 5.7 Time trade-off for a temporary health state.

Preferences for temporary health states can be measured relative to each other using the TTO method as shown in Figure 5.7. As with the rating scale and the standard gamble, intermediate states *i* are measured relative to the best state (healthy) and the worst state (temporary state *j*). The subject is offered two alternatives:

- temporary state *i* for time *t* (the time duration specified for the temporary states), followed by healthy state
- temporary state *j* for $x < t$, followed by healthy state.

Time x is varied until the respondent is indifferent between the two alternatives, at which point the required preference score for state *i* is $h_i = 1 - (1 - h_j)x/t$. If we set $h_j = 0$, this reduces to $h_i = 1 - x/t$. Figure 5.6 shows the basic format, but other variations are possible. State *j* need not be the worst state as long as it is any state worse than *i*. In using variations, however, care must be taken to ensure that all preference values can be calculated. In one systematic variation that has been used (Sackett and Torrance 1978; Torrance 1976; Torrance et al. 1972), state *j* is always the next worse state to state *i*. Although the formula is still the same, $h_i = 1 - (1 - h_j)x/t$, the states must now be solved in sequence from worst to best.

Finally, as with the rating scale and the standard gamble, if the preference scores for the temporary states are to be transformed to the 0–1 dead–healthy scale, the worst of the temporary states must be redefined as a short-duration chronic state and measured by the method for chronic states described above.

The methods described above represent the conventional approach to TTO as developed by Torrance and colleagues. Variations have been suggested by others. Buckingham et al. (1996) experimented with three approaches to trading off time: conventional TTO where the respondent trades against unwanted premature death, annual TTO where the trade is against unwanted convalescence, and daily TTO with a trade against unwanted sleep. Based on ease of use and relationship to independent variables, they recommended daily TTO. However, one potential problem with this recommendation is that if the TTO scores are used for calculating QALYs, they are in fact being used to represent trade-offs between living states and death, and it would seem that scores based on trades against death would be more appropriate for the task.

Cook et al. (1994) investigated the second stage of using TTO for temporary states. This is the stage where the worst temporary state is redefined as a short-term chronic state followed by death and measured using the method for chronic states. They were concerned that the imminence of death in such a scenario would inappropriately distort the result. Accordingly, in an application where the short duration was 12 weeks, they chose to present the state at two longer durations, 12 months and 12 years, in part to determine if the duration would affect the results. To their surprise there was no effect of duration on the TTO score. However, other evidence, reviewed in Brazier et al. (2003), suggests that the prospect for poor health states, particularly severe ones, the longer they are specified to last in the valuation task, the worse they seem to become. This would result in TTO values declining with time. Conversely, people may believe that with time they will adjust to the state and hence the duration effect may raise the TTO health state values. This is an empirical issue, but we can observe that the fact that the duration in a state might affect its valuation is problematic for the construction of QALY profiles, since this is done by multiplying the given health state value by the time spent in the states. (We return to this issue in Section 5.8.1.)

There have been several developments to the standard approach for conducting the TTO discussed above (Rowen and Brazier 2011). One of the most important is the lead time TTO, developed by Devlin et al. (2011). This counters a problem experienced with the standard TTO protocol, whereby states worse than death are valued by a different measurement task and this may be apparent to the respondent, especially as a different prop is used. The standard TTO task for states worse than death provides respondents with a choice between (1) health state h for x years, followed by full health for y years after which they will die, or (2) immediate death. Another problem with this approach is that respondents may not believe that they could return to full health after experiencing a very severe state.

In the lead time TTO, a period of full health (i.e. 'lead time') is added to the start of the normal TTO, meaning that states worse than death can be valued by cutting into the lead time. This approach means that the same method, props, and formula to calculate the TTO value are used for all states. This new procedure has been the subject of further methodological work undertaken by the EuroQoL Group concerning the EQ-5D measure (Devlin and Krabbe 2013). (The EQ-5D is discussed in Section 5.5.2.)

5.5 Multi-attribute health status classification systems with preference scores

Measuring preferences for health outcomes, as described in the previous section, is a very time consuming and complex task. An alternative that is very attractive and widely used is to bypass the measurement task by using one of the pre-scored multi-attribute health status classification systems that exist. The three most widely used systems, described here in some detail, are the Health Utilities Index (HUI), EQ-5D from the EuroQoL Group, and Short Form 6D (SF-6D). Other systems include the 15D (Sintonen 2001) and the Assessment of Quality of Life (AQoL) (Hawthorne et al. 2001).

In this section we first describe the applicable theory, multi-attribute utility theory, and then the three main systems.

5.5.1 Multi-attribute utility theory

Traditional vNM utility theory was extended to cover multi-attribute outcomes by Keeney and Raiffa (1976). To accommodate the extension they had to add one additional assumption to the three axioms of utility theory. This assumption is that the utility independence among the attributes can be represented by at least first-order utility independence, and perhaps by stronger utility independence (mutual utility independence, additive independence). This is best explained by example. Consider the Health Utilities Index Mark 2 (HUI2) which is a multi-attribute health status classification system consisting of the following six core attributes: sensation, mobility, emotion, cognition, self-care, pain. Each attribute in turn consists of four or five levels of specified impairment from no impairment to full impairment. (See Section 5.5.4 for a full description of the system.)

First-order utility independence implies that there is no interaction (synergism or antagonism) between preferences among levels on any one attribute and the fixed levels for the other attributes. An example would be the case where level 3 mobility has a utility of 0.6 on the mobility subscale, regardless of the health status levels on the other attributes. The mobility subscale is the single-attribute utility function for mobility, scaled such that the best level of mobility is 1.0 and the worst level of mobility is 0. Note that the overall weight for mobility could change on the basis of health status on the other attributes, and thus the overall effect of changes in mobility could change without violating first-order utility independence. For example, a change from level 1 mobility to level 3 mobility could reduce overall utility by 0.2 if that were the only health status deficit, but by less than 0.2 if the individual already had other major health status deficits. All that is required for first-order utility independence is that the relative scaling *within* the mobility subscale stays constant.

Mutual utility independence is a stronger assumption. It requires that there be no interaction between preferences for levels on *some* attributes and the fixed levels for other attributes. This characteristic must hold for all possible subsets of attributes. An example of mutual utility independence would be the case where level 2 on sensation coupled with level 3 on mobility has a utility of 0.7 on the sensation–mobility subscale, regardless of the health status levels on the other attributes. The sensation–mobility subscale is the subscale for these two attributes combined, such that the worst level on

sensation coupled with the worst level of mobility is 0 and the best level on sensation coupled with the best level on mobility is 1. Note that the weight of this subscale for sensation and mobility could change given different health status on other attributes, so that the overall impact of changes within sensation and mobility could differ without violating mutual utility independence. For example, a change from level 1 on sensation and level 1 on mobility to level 2 on sensation and level 3 on mobility could reduce overall utility by 0.25 if those were the only deficits, but by less than 0.25 if the individual already had other major deficits. What is required for mutual utility independence is that the relative scaling *within* the sensation–mobility subscale stays constant.

Additive utility independence implies that there is no interaction for preferences among attributes at all. That is, the overall preference depends only on the individual levels of the attributes and not on the manner in which the levels of the different attributes are combined. An example of additive independence would be the case where a change from level 1 mobility to level 3 mobility would reduce the overall utility by 0.2 regardless of the levels on the other attributes.

The three independence assumptions lead to three different multi-attribute functions. The simplest assumption, first-order utility independence, leads to the most complex mathematical function, the multilinear function. The second possible assumption, mutual utility independence, leads to the multiplicative function. The strongest assumption (most difficult to fulfil), additive independence, leads to the simplest function, the additive function. See Box 5.3 for the three multi-attribute utility functions.

5.5.2 **EQ-5D**

The EuroQoL Group, a consortium of investigators in western Europe, initially developed a system with six attributes: mobility, self-care, main activity, social relationships, pain, and mood (EuroQoL Group 1990). Subsequently it was revised to include five attributes: mobility, self-care, usual activity, pain/discomfort, and anxiety/depression (Brooks 1996; Essink-Bot et al. 1993; Kind 1996). Each attribute has three levels: no problem, some problems, and major problems, thus defining 243 possible health states, to which have been added 'unconscious' and 'dead' for a total of 245 in all. Preferences for the scoring function were measured with the TTO technique on a random sample of approximately 3000 members of the adult population of the United Kingdom (Dolan et al. 1995, 1996b). The scoring function was developed using econometric modelling as opposed to multi-attribute utility theory. The scores fall on the 0.0 (dead) to 1.0 (perfect health) value scale.

The full system and the original scoring function are shown in Box 5.4 and Table 5.2 (Dolan et al. 1995). There have been several developments in the EQ-5D since it was originally developed and surveys have been conducted in several countries, including the United States of America (Shaw et al. 2005). Comprehensive details on the EQ-5D can be obtained from the web site of the EuroQoL Group (<http://www.euroqol.org>).

In 2011, the EuroQoL Group developed and tested a new five-level version of the EQ-5D, the EQ-5D-5L (Herdman et al. 2011). (At the same time, the original EQ-5D instrument was renamed the EQ-5D-3L.) The motivation for developing the new instrument was the growing evidence from use of the three-level EQ-5D that it can suffer from ceiling effects, particularly when used in general population surveys but also in some patient populations. As a result there might be issues in its ability to detect small changes in health, especially in patients with milder conditions (Herdman et al. 2011).

Box 5.3 Types of multi-attribute utility functions

Additive:

$$u(x) = \sum_{j=1}^{n} k_j u_j(x_j)$$

$$\text{where } \sum_{j=1}^{n} k_j = 1.$$

Multiplicative:

$$u(x) = (1/k)\left[\prod_{j=1}^{n}(1 + kk_j u_j(x_j)) - 1\right]$$

$$\text{where } (1+k) = \prod_{j=1}^{n}(1 + kk_j).$$

The multiplicative model contains the additive model as a special case. In fitting the multiplicative model, if the measured k_j sum to 1, then $k = 0$ and the additive model holds.

Multilinear:

$$\begin{aligned} u(x) = {} & k_1 u_1(x_1) + k_2 u_2(x_2) + \ldots \\ & + k_{12} u_1(x_1) u_2(x_2) + k_{13} u_1(x_1) u_3(x_3) + \ldots \\ & + k_{123} u_1(x_1) u_2(x_2) u_3(x_3) + \ldots \\ & + \ldots \end{aligned}$$

where the sum of all ks equals 1.

Hybrid: Various hybrid models are possible, based on hierarchically nested subsets of attributes.

Notation: $u_j(x_j)$ is the single-attribute utility function for attribute *j*. $u(x)$ is the utility for health state x, represented by an n-element vector. k and k_j are model parameters. Σ is the summation sign. $\prod$ is the multiplication sign.

In the first phase of the development, a pool of potential labels for the new levels was identified and provisional labels chosen after a response scaling task carried out in face-to-face interviews with members of the general public. In the second phase, face and content validity of two alternative five-level systems were tested in focus group sessions with healthy individuals and those with chronic illness. The development work was conducted simultaneously in English and Spanish, since they are two if the most widely spoken languages. The language to describe the dimensions of health states the new 5L version is similar to that in the original EQ-5D, except that the labels used for the five levels (e.g. for mobility, self-care, and usual activities) are 'no problems', 'slight problems', 'moderate problems', 'severe problems', and 'unable'. In the case of pain/discomfort

Box 5.4 EQ-5D classification system

Mobility

1 No problems walking
2 Some problems walking about
3 Confined to bed

Self-care

1 No problems with self-care
2 Some problems washing or dressing self
3 Unable to wash or dress self

Usual activities

1 No problems with performing usual activities (e.g. work, study, housework, family or leisure activities)
2 Some problems with performing usual activities
3 Unable to perform usual activities

Pain/discomfort

1 No pain or discomfort
2 Moderate pain or discomfort
3 Extreme pain or discomfort

Anxiety/depression

1 Not anxious or depressed
2 Moderately anxious or depressed
3 Extremely anxious or depressed

Note: For convenience each composite health state has a five-digit code number relating to the relevant level of each dimension, with the dimensions always listed in the order given above. Thus 11223 means

1 No problems walking about
1 No problems with self-care
2 Some problems with performing usual activities
2 Moderate pain or discomfort
3 Extremely anxious or depressed

Table 5.2 EQ-5D scoring formula

Coefficients for TTO tariffs	
Dimension	**Coefficient**
Constant	0.081
Mobility	
level 2	0.069
level 3	0.314
Self-care	
level 2	0.104
level 3	0.214
Usual activity	
level 2	0.036
level 3	0.094
Pain/discomfort	
level 2	0.123
level 3	0.386
Anxiety/depression	
level 2	0.071
level 3	0.236

Reproduced from Dolan, P. et al, *A social tariff for EuroQoL: Results from a UK general population survey*, Discussion Paper Number 138, Centre for Health Economics, University of York, UK, Copyright © 1995.

EuroQol time trade-off scores are calculated by subtracting the relevant coefficients from 1.000. The constant term is used if there is any dysfunction at all. The N3 term is used if any dimension is at level 3. The term for each dimension is selected based on the level of that dimension. The algorithm for computing the tariff is quite straightforward. For example, consider the state 11223:

Full health	= 1.000
Constant term (for any dysfunctional state)	–0.081
Mobility (level 1)	–0
Self-care (level 1)	–0
Usual activities (level 2)	–0.036
Pain or discomfort (level 2)	–0.123
Anxiety or depression (level 3)	–0.236
N3 (level 3 occurs within at least one dimension)	–0.269
Therefore, the estimated value for 11223	= 0.255

and anxiety/depression, the label 'extreme' or 'extremely' is used instead of 'unable'. (See Herdman et al. (2011) for a full description.)

Once the new instrument had been agreed upon, the immediate task was to develop a new value set for the 3125 states in the new classification system. Therefore, an interim value set was developed by mapping from the EQ-5D-3L (van Hout et al. 2012).

Both instruments were coadministered to 3691 respondents in 6 countries with conditions of varying severity, covering a broad range of levels of health. Then four models were used to generate value sets for the EQ-5D-5L. A non-parametric model was chosen because if its simplicity while performing similarly to the other models. The results of the mapping exercise are shown in van Hout et al. (2012).

Although the interim value set allows values for EQ-5D-5L states to be assigned from existing EQ-3D value sets, the approach has its limitations. Therefore, the EuroQoL Group has embarked on an international research programme to generate a series of value sets for a range of countries. This research also has the objectives of overcoming some of the problems of the TTO approach in valuing health states considered to be worse than dead and of exploring the use of discrete choice methods. (Discrete choice methods are discussed in Chapter 6.) The results of this research are reported in a supplement of the *European Journal of Health Economics*, along with an accompanying editorial (Devlin and Krabbe 2013). An international protocol for studies to produce value sets for the EQ-5D-5L has been developed and it is anticipated that value sets for a large range of countries will eventually be available. Up-to-date information can be obtained from the EuroQoL Group website (<www.euroqol.org>).

5.5.3 Short Form 6D

The SF-6D is a preference-based instrument based on the popular HRQoL questionnaire, the Short Form 36 (SF-36) (Brazier et al. 2002). The instrument was developed, in part, because the SF-36 has been widely used in a large number of studies, and it would be useful to be able to convert the study results to health state preference values and hence to QALYs. The SF-6D consists of a multi-attribute health status classification system with six attributes (Box 5.5) and a scoring table (Table 5.3). The classification system was developed from the information collected on the SF-36 questionnaire. It uses 11 items from the SF-36 (8 from the SF-12, which is an abbreviated version of the SF-36, and 3 others from the SF-36 itself). The classification system consists of 4–6 levels on each of the 6 attributes for a total of 18 000 unique health states.

The scoring model for the SF-6D was developed based on standard gamble measurements on a random sample ($n = 836$) of the general population of the United Kingdom. Each subject provided estimates for 6 states. A total of 249 different health states were valued. Using econometric modelling on these data the developers investigated a number of different scoring models and recommended a particular one, which is shown in Table 5.5.

To use the SF-6D system, you first must use the SF-36 questionnaire or the SF-12 questionnaire plus the three additional questions to collect the data to classify the patients into the SF-6D classification system. Then you use the scoring table to compute the health state preference values. These fall on the conventional scale for health state preference values, where dead is 0.0 and healthy is 1.0. The worst state in the SF-6D system has a value of 0.30.

There is currently a study to develop a version 2 of the SF-6D to overcome some of the problems identified with version 1, namely: 'floor effects' from the scores only going

Box 5.5 SF-6D classification system

Physical functioning

1 Your health does not limit you in *vigorous activities.*
2 Your health limits you a little in *vigorous activities.*
3 Your health limits you a little in *moderate activities.*
4 Your health limits you a lot in *moderate activities.*
5 Your health limits you a little in *bathing and dressing.*
6 Your health limits you a lot in *bathing and dressing.*

Role limitations

1 You have no problems with your work or other regular daily activities as a result of your physical health or any emotional problems.
2 You are limited in the kind of work or other activities as a result of your physical health.
3 You accomplish less than you would like as a result of emotional problems.
4 You are limited in the kind of work or other activities as a result of your physical health and accomplish less than you would like as a result of emotional problems.

Social functioning

1 Your health limits your social activities *none of the time.*
2 Your health limits your social activities *a little of the time.*
3 Your health limits your social activities *some of the time.*
4 Your health limits your social activities *most of the time.*
5 Your health limits your social activities *all of the time.*

Pain

1 You have no pain.
2 You have pain but it does not interfere with your normal work (both outside the home and housework).
3 You have pain that interferes with your normal work (both outside the home and housework) *a little bit.*
4 You have pain that interferes with your normal work (both outside the home and housework) *moderately.*
5 You have pain that interferes with your normal work (both outside the home and housework) *quite a little bit.*
6 You have pain that interferes with your normal work (both outside the home and housework) *extremely.*

Box 5.5 SF-6D classification system (continued)

Mental health

1 You feel tense or downhearted and low *none of the time.*
2 You feel tense or downhearted and low *a little bit of the time.*
3 You feel tense or downhearted and low *some of the time.*
4 You feel tense or downhearted and low *most of the time.*
5 You feel tense or downhearted and low *all of the time.*

Vitality

1 You have a lot of energy *all of the time.*
2 You have a lot of energy *most of the time.*
3 You have a lot of energy *some of the time.*
4 You have a lot of energy *none of the time.*

down to 0.3, inconsistencies in some levels, and confusion from mixing positively and negatively worded items. The new SF-6D will be valued by a variant of TTO using DCE with duration. (For further information see <https://www.sheffield.ac.uk/scharr/sections/heds/mvh/sf-6d>.)

5.5.4 Health Utilities Index (HUI)

The HUI currently consists of two systems, HUI2 and HUI3 (Furlong et al. 2001; Horsman et al. 2003). Each includes a health status classification system and a scoring formula. In both cases the scoring formula is based on standard gamble utilities measured on the general public, and the scores are on the conventional dead–healthy 0–1 scale.

For most applications, HUI3 should be used as the primary analysis. It has the more detailed descriptive system, it has full structural independence, and population norms are available. HUI2 can be used in a secondary role to provide additional insight. HUI2 has some additional attributes not in the HUI3 that may be useful in specific studies: self-care, emotion with a focus on worry/anxiety, and fertility. HUI2 can also be used as a sensitivity analysis.

Preferences for the HUI2 scoring function were measured on a random sample of parents of schoolchildren in the City of Hamilton, Canada and surrounding district using both a visual analogue technique and a standard gamble instrument. Thus, both value and utility functions are available, although the utility function is the one recommended for most applications. States worse than death were identified, but were scored as equal to death. The scoring formula is a multiplicative multi-attribute utility function, with scores that fall on the 0.0 (dead) to 1.0 (perfect health) scale.

Table 5.3 SF-6D scoring model

General terms		Physical functioning		Role limitations		Social functioning		Pain		Mental health		Vitality	
Term	**Score**	**Level**	**Score**	**Level**	**Score**	**Level**	**Score**	**Level**	**Score**	**Level**	**Score**	**Level**	**Score**
C	1.000	PF1	–0.000	RL1	–0.000	SF1	–0.000	PAIN1	–0.000	MH1	–0.000	VIT1	–0.000
MOST	–0.070	PF2	–0.053	RL2	–0.053	SF2	–0.055	PAIN2	–0.047	MH2	–0.049	VIT2	–0.086
		PF3	–0.011	RL3	–0.055	SF3	–0.067	PAIN3	–0.025	MH3	–0.042	VIT3	–0.061
		PF4	–0.040	RL4	–0.050	SF4	–0.070	PAIN4	–0.056	MH4	0.109	VIT4	–0.054
		PF5	–0.054			SF5	–0.087	PAIN5	–0.091	MH5	–0.128	VIT5	–0.091
		PF6	–0.111					PAIN6	–0.167				

Preference value = C + PF + RL + SF + PAIN + MH + VIT + MOST

Where C is a constant term, PF*x* denotes level *x* on the physical functioning dimension (same for other dimensions), and MOST is a term to be used if any dimension is at its most severe level.

The HUI3 classification system was based closely on that of the HUI2. The application-specific attribute, fertility, was dropped. The sensory attribute of HUI2 was expanded in HUI3 into the three attributes: vision, hearing, and speech. The remaining changes were made to increase the structural independence (orthogonality) of the attributes. An attribute is structurally independent of other attributes if it is conceivable for an individual to function at any level on that attribute, regardless of the levels on the other attributes. If all attributes are structurally independent of each other, all combinations of levels in the system are possible. This goal has been achieved in the HUI3. Structural independence is not only useful for the descriptive classification system, but also greatly simplifies the estimation of the scoring function.

Preferences for the HUI3 were measured on a random sample of general population adults living in the City of Hamilton, Canada using both a visual analogue technique and a standard gamble instrument. States worse than death were measured as negative scores on the 0.0 (dead) to 1.0 (perfect health) scale. Both a multiplicative model and a multilinear model have been estimated. The multiplicative model is the one recommended and is the one described below (Feeny et al. 2002).

Shown in Tables 5.4–5.7 are the HUI2 classification system (Table 5.4), the HUI2 scoring formula (Table 5.5), the HUI3 classification system (Table 5.6), and the HUI3 scoring formula (Table 5.7). An exercise on calculating HUI scores is provided in Box 5.6.

Table 5.4 Health Utilities Index mark 2 classification system

Attribute	Level	Level description
Sensation	1	Ability to see, hear, and speak normally for age
	2	Requires equipment to see or hear or speak
	3	Sees, hears, or speaks with limitations even with equipment
	4	Blind, deaf, or mute
Mobility	1	Able to walk, bend, lift, jump, and run normally for age
	2	Walks, bends, lifts, jumps, or runs with some limitations but does not require help
	3	Requires mechanical equipment (such as cane, crutches, braces, or wheelchair) to walk or get around independently
	4	Requires the help of another person to walk or get around and requires mechanical equipment as well
	5	Unable to control or use arms and legs
Emotion	1	Generally happy and free from worry
	2	Occasionally fretful, angry, irritable, anxious, depressed, or suffering 'night terrors'
	3	Often fretful, angry, irritable, anxious, depressed, or suffering 'night terrors'
	4	Almost always fretful, angry, irritable, anxious, depressed
	5	Extremely fretful, angry, irritable, anxious, or depressed; usually requiring hospitalization or psychiatric institutional care

(*continued*)

Table 5.4 (continued) Health Utilities Index mark 2 classification system

Attribute	Level	Level description
Cognition	1	Learns and remembers schoolwork normally for age
	2	Learns and remembers schoolwork more slowly than classmates as judged by parents and/or teachers
	3	Learns and remembers very slowly and usually requires special educational assistance
	4	Unable to learn and remember
Self-care	1	Eats, bathes, dresses, and uses the toilet normally for age
	2	Eats, bathes, dresses, or uses the toilet independently with difficulty
	3	Requires mechanical equipment to eat, bathe, dress, or use the toilet independently
	4	Requires the help of another person to eat, bathe, dress, or use the toilet
Pain	1	Free of pain and discomfort
	2	Occasional pain. Discomfort relieved by non-prescription drugs or self-control activity without disruption of normal activities
	3	Frequent pain. Discomfort relieved by oral medicines with occasional disruption of normal activities
	4	Frequent pain, frequent disruption of normal activities. Discomfort requires prescription narcotics for relief
	5	Severe pain. Pain not relieved by drugs and constantly disrupts normal activities
Fertility[a]	1	Able to have children with a fertile spouse
	2	Difficulty in having children with a fertile spouse
	3	Unable to have children with a fertile spouse

[a]Fertility attribute can be deleted if not required. Contact developers for details (<http://www.healthutilities.com>).

To use the system, researchers must describe the health states of subjects according to an HUI classification system, and then use the corresponding scoring formula. For clinical studies or population studies, questionnaires have been developed for self-administration or interviewer administration to collect sufficient data to classify the patient or subject into both the HUI2 and the HUI3 systems. The questionnaire takes less than 10 minutes for self-administration and only 2–3 minutes for interviewer administration.

The HUI system has been widely used throughout the world and, accordingly, has been translated and culturally adapted using established guidelines (Guillemin et al. 1993) into a large and growing number of languages, currently 36. In addition to clinical

Table 5.5 Health Utilities Index mark 2 scoring formula

Sensation		Mobility		Emotion		Cognition		Self-care		Pain		Fertility	
x_1	b_1	x_2	b_2	x_3	b_3	x_4	b_4	x_5	b_5	x_6	b_6	x_7	b_7
1	1.00	1	1.00	1	1.00	1	1.00	1	1.00	1	1.00	1	1.00
2	0.95	2	0.97	2	0.93	2	0.95	2	0.97	2	0.97	2	0.97
3	0.86	3	0.84	3	0.81	3	0.88	3	0.91	3	0.85	3	0.88
4	0.61	4	0.73	4	0.70	4	0.65	4	0.80	4	0.64	4	n/a
5	n/a	5	0.58	5	0.53	5	n/a	5	n/a	5	0.38	5	n/a

Formula: $u^* = 1.06(b_1 \times b_2 \times b_3 \times b_4 \times b_5 \times b_6 \times b_7) - 0.06$, where u^* is on a health state preference scale where dead has a value of 0.00 and healthy has a value of 1.00. Because the worst possible health state was judged by respondents as worse than death, it has a negative value of –0.03. The standard error of u^* for estimating validation states within the sample is 0.015 for measurement error and sampling error, and 0.06 if model error is also included. x_i is attribute level code for attribute i; b_i is level score for attribute i.

Table 5.6 Health Utilities Index mark 3 classification system

Attribute	Level	Level description
Vision	1	Able to see well enough to read ordinary newsprint and recognize a friend on the other side of the street, without glasses or contact lenses
	2	Able to see well enough to read ordinary newsprint and recognize a friend on the other side of the street, but with glasses
	3	Able to read ordinary newsprint with or without glasses but unable to recognize a friend on the other side of the street, even with glasses
	4	Able to recognize a friend on the other side of the street with or without glasses but unable to read ordinary newsprint, even with glasses
	5	Unable to read ordinary newsprint and unable to recognize a friend on the other side of the street, even with glasses
	6	Unable to see at all
Hearing	1	Able to hear what is said in a group conversation with at least three other people, without a hearing aid
	2	Able to hear what is said in a conversation with one other person in a quiet room without a hearing aid, but require a hearing aid to hear what is said in a group conversation with at least three other people
	3	Able to hear what is said in a conversation with one other person in a quiet room with a hearing aid, and able to hear what is said in a group conversation with at least three other people with a hearing aid

Table 5.6 (continued) Health Utilities Index mark 3 classification system

Attribute	Level	Level description
	4	Able to hear what is said in a conversation with one other person in a quiet room without a hearing aid, but unable to hear what is said in a group conversation with at least three other people even with a hearing aid
	5	Able to hear what is said in a conversation with one other person in a quiet room with a hearing aid, but unable to hear what is said in a group conversation with at least three other people even with a hearing aid
	6	Unable to hear at all
Speech	1	Able to be understood completely when speaking with strangers or friends
	2	Able to be understood partially when speaking with strangers but able to be understood completely when speaking with people who know me well
	3	Able to be understood partially when speaking with strangers or people who know me well
	4	Unable to be understood when speaking with strangers but able to be understood partially by people who know me well
	5	Unable to be understood when speaking to other people (or unable to speak at all)
Ambulation	1	Able to walk around the neighborhood without difficulty, and without walking equipment
	2	Able to walk around the neighborhood with difficulty; but does not require walking equipment or the help of another person
	3	Able to walk around the neighborhood with walking equipment, but without the help of another person
	4	Able to walk only short distances with walking equipment, and requires a wheelchair to get around the neighborhood
	5	Unable to walk alone, even with walking equipment. Able to walk short distances with the help of another person, and requires a wheelchair to get around the neighborhood
	6	Cannot walk at all
Dexterity	1	Full use of two hands and ten fingers
	2	Limitations in the use of hands or fingers, but does not require special tools or help of another person
	3	Limitations in the use of hands or fingers, is independent with use of special tools (does not require the help of another person)
	4	Limitations in the use of hands or fingers, requires the help of another person for some tasks (not independent even with use of special tools)

Table 5.6 (continued) Health Utilities Index mark 3 classification system

Attribute	Level	Level description
	5	Limitations in use of hands or fingers, requires the help of another person for most tasks (not independent even with use of special tools)
	6	Limitations in use of hands or fingers, requires the help of another person for all tasks (not independent even with use of special tools)
Emotion	1	Happy and interested in life
	2	Somewhat happy
	3	Somewhat unhappy
	4	Very unhappy
	5	So unhappy that life is not worthwhile
Cognition	1	Able to remember most things, think clearly, and solve day-to-day problems
	2	Able to remember most things, but has a little difficulty when trying to think and solve day-to-day problems
	3	Somewhat forgetful, but able to think clearly and solve day-to-day problems
	4	Somewhat forgetful, and has a little difficulty when trying to think or solve day-to-day problems
	5	Very forgetful, and has great difficulty when trying to think or solve day-to-day problems
	6	Unable to remember anything at all, and unable to think or solve day-to-day problems
Pain	1	Free of pain and discomfort
	2	Mild to moderate pain that prevents no activities
	3	Moderate pain that prevents a few activities
	4	Moderate to severe pain that prevents some activities
	5	Severe pain that prevents most activities

studies, the HUI has been used in a number of population health surveys. Thus, population norm data are available for comparative purposes. (For further information on the HUI and the availability of questionnaires and support services, contact Health Utilities Incorporated, Dundas, Canada <www.healthutilities.com>. The website also has a useful list of references to all HUI methodological studies, clinical and evaluative studies, and population health applications.)

5.5.5 Other generic preference-based instruments

In this chapter we have focused on the three most widely used instruments or classification systems, the EQ-5D, HUI, and SF-6D. However, other instruments exist. The most

Table 5.7 Health Utilities Index mark 3 scoring formula

Vision		Hearing		Speech		Ambulation		Dexterity		Emotion		Cognition		Pain	
x_1	b_1	x_2	b_2	x_3	b_3	x_4	b_4	x_5	b_5	x_6	b_6	x_7	b_7	x_8	b_8
1	1.00	1	1.00	1	1.00	1	1.00	1	1.00	1	1.00	1	1.00	1	1.00
2	0.98	2	0.95	2	0.94	2	0.93	2	0.95	2	0.95	2	0.92	2	0.96
3	0.89	3	0.89	3	0.89	3	0.86	3	0.88	3	0.85	3	0.95	3	0.90
4	0.84	4	0.80	4	0.81	4	0.73	4	0.76	4	0.64	4	0.83	4	0.77
5	0.75	5	0.74	5	0.68	5	0.65	5	0.65	5	0.46	5	0.60	5	0.55
6	0.61	6	0.61	6	n/a	6	0.58	6	0.56	6	n/a	6	0.42	6	n/a

Formula (dead–perfect health scale): $u^* = 1.371(b_1 \times b_2 \times b_3 \times b_4 \times b_5 \times b_6 \times b_7 \times b_8) - 0.371$, where u^* is the health state preference value on a scale where dead has a vlaue of 0.00 and healthy has a value of 1.00. States worse than dead have negative values. x_i is attribute level code for attribute i; b_i is level score for attribute i. For the attribute 'Cognition', the score for level 3 is greater than the score for level 2. This is not a typo, but reflects that level 3 was seen as preferable to level 2.

The standard error of u^*, including model error, is 0.08 for estimating validation states within the sample. For estimating validation states (based on $n = 73$) from an independent sample, the standard error is 0.10 if the states are unweighted, 0.006 if the states are weighted by prevalence excluding the state of perfect health, and 0.004 if the states are weighted by prevalence including the state of perfect health.Reproduced with permission from Lippincott Williams and Wilkins/Wolters Kluwer Health: Feeny, D. et al., Multiattribute and single-attribute utility functions for the Health Utilities Index Mark 3 system, *Medical Care*, Volume 40, Issue 2, Table 3, pp. 113–28, Copyright © 2002, Lippincott-Raven Publishers.

notable are the 15D, developed in Finland (Sintonen 2001; Sintonen and Pekurinen 1993) and the AQoL, developed in Australia (Hawthorne et al. 1999). Both these have had extensive use, mainly within their own geographical region, although the AQoL has also been compared with the other instruments discussed here (Hawthorne et al. 2001; Richardson et al. 2014). Further details on these measures can be found in Brazier et al. (2007).

5.5.6 Which system to use?

Having decided to use a preference-based multi-attribute health status system in a study, a researcher must then decide which one to use. While we cannot answer that question for the researcher, we can give some guidance on the considerations.

First, the decision does matter. These systems are far from identical. They differ in the dimensions of health they cover, in the number of levels defined on each dimension, in the description of these levels, and in the severity of the most severe level. In addition, they differ in the population surveyed and in the instruments used to determine the preference-based scoring. Finally, they differ in the theoretical approach taken to modelling the preference data into a scoring formula. For example, although all are multi-attribute systems, only the HUI uses multi-attribute utility theory for the estimation of the scoring formula. EQ-5D and SF-6D use econometric modelling. (See Stevens et al. (2007) for a discussion of the pros and cons of each approach.) Because of these various differences, it is not surprising that comparative studies show that the same patient groups can score

Box 5.6 Exercise: Health Utilities Index

1. In the HU12 system (Table 5.4), the health state of an individual is described as a six- or seven-element vector with each element denoting the level on an attribute. For example, 1321221 would be an individual who was at level 1 sensation, level 3 mobility, level 2 emotion, level 1 cognition, level 2 self-care, level 2 pain, and level 1 fertility. Because the fertility attribute is optional in the system, if only six elements are specified they refer to the first six attributes.

The score for a health state is determined using the formula from Table 5.5. If only six elements are specified, b_7 is omitted from the formula (or equivalently, b_7 is set equal to 1).

Determine the HU12 scores for the following health states:

(a) 1321221

(b) 2132113

(c) 111111

(d) 112114

(e) 332325

Answers: 0.72, 0.62, 1.00, 0.57, 0.17.

2. In the HU13 system (Tables 5.6 and 5.7) there are more attributes, more levels, and a different scoring formula, but otherwise the notation and the method of calculation is the same.

Determine the HU13 scores for the following health states:

(a) 23112211

(b) 11121131

(c) 41131112

(d) 11111451

(e) 66566565

Answers: 0.68, 0.84, 0.58, 0.16, –0.36.

quite differently depending upon the instrument used (Conner-Spady and Suarez-Almazor 2003; Kopec and Willison 2003; Lubetkin and Gold 2003; O'Brien et al. 2003).

Fryback et al. (2010) compared five of the generic HRQoL indexes, including EQ-5D-3L, HUI2, HUI3, and SF-6D, by using them to assign health state preference values to a range of health states. Data were from the National Health Measurement Study (a telephone survey of 3844 US adults) and item response theory was used to estimate latent summary health score for each NHMS respondent based on the five index scores for that respondent. They found that simple linear functions may serve as 'cross-walks' between the indexes only for lower health states, albeit with low precision. However, ceiling effects make cross-walks among most of the indexes ill specified above a certain level of health.

Several other studies have compared the scores obtained for patients with different health conditions. In a study in schizophrenia, McCrone et al. (2009) found very similar mean scores at baseline and follow-up for the EQ-5D and SF-6D. In contrast, Heintz et al. (2012) found that different measures, including the EQ-5D-3L and HUI3, discriminated between different levels of eye health (in diabetic retinopathy) to a greater or lesser extent. Finally, in another study, Joore et al. (2010) investigated whether differences in utility scores based on the EQ-5D and SF-6D had an impact on incremental cost-effectiveness of treatments in five distinct patient groups. They found that there were small differences in incremental QALYs, but that these translated to large differences in incremental cost-effectiveness ratios, such that there was an important difference in whether a treatment intervention was cost-effective at a given threshold ratio. This study demonstrated that incremental cost-effectiveness ratios for different studies may not be comparable if the QALYs were estimated using different instruments. See Brazier et al. (2007) for a fuller discussion of the differences between the well-known instruments.

When selecting an instrument the researcher should consider a number of factors. In general, is the instrument seen as credible? That is, is it an established instrument, like those described above, which has demonstrated feasibility, reliability, validity, and responsiveness in a number of studies? There is a rapidly expanding literature describing applications and measurement characteristics of each of these instruments. For example, the web sites for EQ-5D and HUI each list hundreds of publications on the instrument. Does the health status classification system cover the attributes and the levels of these attributes that are likely to be important to the patient population under study? Has the instrument been used in similar patients and was it responsive? Is the instrument likely to be responsive to the changes expected in the study patients? Does the instrument have ceiling effects or floor effects that will reduce its sensitivity for the patients under study? Only some of the instruments measure states worse than death. Is this likely to be an important aspect of the study? Brazier and Deverill (1999) have produced a checklist for judging preference-based measures of HRQoL and this work is further reviewed in Brazier et al. (2007).

A major factor to consider is whether the intended audience for the study has any guidance or preference for a particular instrument? For example, the National Institute for Health and Care Excellence (NICE) in the United Kingdom specifies that:

> For the reference case, the measurement of changes in health-related quality of life should be reported directly from patients and the utility of these changes should be based on public preferences using a choice-based method. The EQ-5D is the preferred measure of health-related quality of life in adults. (NICE 2013)

However, NICE also recognizes that other approaches may need to be considered when directly measured EQ-5D data are not available, or in situations where EQ-5D has been shown to be an inappropriate measure for the health condition being studied. One approach would be to map from a descriptive quality instrument to the preferred preference-based instrument. (This is discussed in Section 5.6.) One advantage of the SF-6D is that algorithms exist to generate it from SF-36 or SF-12 data, if these have been collected as part of the clinical studies of the treatment of interest.

If clinicians are an intended audience, do the clinical opinion leaders have a preferred instrument for their field, although it is likely that clinicians would nearly always prefer a condition-specific measure, as opposed to one of the generic instruments? Is the instrument based on sound theory? How much time is required to complete the questionnaire and is the questionnaire clear and easy to follow—that is, is the patient burden acceptable? And finally, what is the overall cost of using the instrument, including licensing fees, data collection costs, scoring costs, and analysis costs?

If there is not a single obvious instrument for a study, the researcher may wish to consider a pilot study to test several contenders and determine which performs best in the type of patients being studied. The researcher may even wish to use several instruments in the study, designating one as the primary measure and the other(s) as secondary. This can not only provide additional insight into the study findings, but can also be of considerable interest in its own right as a head-to-head comparison of alternative instruments.

One issue that is sometimes raised in instrument selection is the fact that each instrument is scored based on preferences from a particular population, and those preferences may not apply to other populations. For example, the original HUI was scored based on preferences of residents of Hamilton, Canada, and the original EQ-5D and SF-6D were scored based on preferences of residents of the United Kingdom. The concern is that the scoring may not be appropriate when the instrument is used in other geographic locations. Therefore, the developers of the various instruments, especially the EQ-5D, have developed scores, or value sets, based on other populations.

There are now several studies that demonstrate differences in health state values by age, gender, marital, status, and own health (Dolan and Roberts 2002; Kharroubi et al. 2007). In another study comparing the different national value sets for the EQ-5D when applied to two hypothetical health states, Knies et al. (2009) found substantial differences, but attributed most of the variation to methodological differences in the various valuation studies. On the other hand, HUI scores seem to travel well when the measurements are done using the same methods. Wang et al. (2002) replicated the HUI2 scoring procedure on a separate sample of parents of childhood cancer patients and obtained similar results to the original scoring based on a sample of parents from the general population. Le Gales et al. (2002) replicated the HUI3 scoring procedure on a representative random sample of the French population and obtained results similar to the original scoring based on the Canadian data.

Although differences might exist due to geographic and other factors, these might be small compared with the differences that exist among instruments and methods for generating the value sets. Users of studies should probably be more concerned about the comparability of studies that use different indexes (EQ-5D, SF-6D, HUI) than about the appropriateness of using an EQ-5D instrument in the United States or a HUI instrument in Europe. Nevertheless, in a given jurisdiction using economic evaluation in reimbursement decisions, some standardization of the methods to generate health state preference values is advisable and it would obviously be preferable if health state preference values were available for the chosen instrument from the relevant local population.

5.6 Mapping between non-preference-based measures of health and generic preference-based measures

5.6.1 The purpose of mapping

In jurisdictions where economic evaluation is used in making decisions about the pricing and/or reimbursement of health technologies, decision-makers often want the health effects expressed as a generic measure of health gain, such as a QALY, since this allows them to make comparisons over a wide range of therapeutic areas. However, generic preference-based measures, such as the EQ-5D or HUI, are not often included in clinical trials of new therapies: it is much more common to include a non-preference-based measure such as the SF-36 or a disease-specific instrument for the health condition of interest. This can present economic analysis with a problem, as QALY estimates for the relevant health states within the disease of interest may not be available.

Mapping, sometimes called 'cross-walking', is one solution to this problem, since it enables health state preference values to be predicted when no preference-based measures have been included in the clinical study, or are not available in the literature. The approach involves estimating the relationship between a non-preference-based measure and a generic preference-based measure using statistical association, or estimating exchange rates between instruments (Brazier et al. 2010). It requires a degree of overlap between the descriptive systems of the two measures and that the two measures are administered on the same population. Two datasets are required, an 'estimation' dataset, where the non-preference-based and preference-based measures have been used, and a 'study' dataset containing only the non-preference measure. Regression techniques are used on the estimation dataset to determine the statistical association between the two measures and the results applied to the study dataset to obtain predicted health state preference values.

There are now numerous examples of mapping, in clinical areas as diverse as urinary incontinence (Brazier et al. 2008), obesity (Brazier et al. 2004), and cancer (Wu et al. 2007; McKenzie and van der Pol 2009). There has also been a cross-walk analysis of five of the main generic preference-based instruments (Fryback et al. 2010). Many studies have mapped non-preference-based measures to the EQ-5D, since this is the generic instrument preferred by NICE in the United Kingdom (NICE 2014). A review of studies mapping from quality-of-life or clinical measures to the EQ-5D has been conducted by Dakin (2013), who found 90 studies reporting 121 mapping algorithms. An online database has now been established and is available at <http://herc.ox.ac.uk/downloads/mappingdatabase>.

5.6.2 Methods of mapping

Given the growth in mapping studies, attention has been paid to the methods of mapping. Longworth and Rowen (2013), in reviewing the studies mapping to EQ-5D in NICE health technology appraisals, consider five elements of mapping: (1) defining the estimation data set; (2) model specification; (3) model type (e.g. ordinary least squares); (4) assessing performance (e.g. goodness of fit, predictive ability); (5) application (e.g. application to a validation sample, characterizing uncertainty in the estimates). Other

researchers have reviewed or compared different mapping methods (Brazier et al. 2010; Chuang and Kind 2009; Lu et al. 2013; Mortimer and Segal 2008).

Despite the growth in popularity of mapping, there are doubts about whether it should ever be the method of first choice. McCabe et al. (2013) argue that there has been very little discussion of the appropriate theoretical framework to guide the design and evaluation of mapping models, and that currently proposed quality standards are inadequate for producing robust or appropriate estimates of health state preference values. Using data from several clinical trials, Ades et al. (2013) considered the relative efficiency (estimate divided by its standard error) of treatment effects from the disease-specific QoL measure, the generic preference-based QoL measure, the generic measure indirectly estimated from the mapped disease-specific measure, and a pooled estimate combining the direct and indirect information on the generic QoL measure. They conclude that trials powered on disease-specific measures are likely to have sufficient (statistical) power to detect treatment effect on the generic QoL if a pooled estimate is used. Therefore, generic QoL instruments should be routinely included in randomized controlled trials. Also, in their review of the use of mapped health state preference values in NICE technology appraisals, Longworth and Rowen (2013) conclude that, in most cases, it is still advantageous to collect data directly by using the favoured preference-based instrument and that mapping should usually be viewed as a second-best solution.

5.6.3 Generic versus condition-specific measures

Longworth et al. (2014) have recently completed a review of the use of generic and condition-specific health-related QoL measures in the context of decisions made by NICE in the United Kingdom in four clinical areas: cancer, skin conditions, hearing, and vision disorders. They found that EQ-5D was valid and responsive for skin conditions and most cancers; in vision, its performance varied according to aetiology; and for hearing impairments its performance was poor. The HUI3 performed well for hearing and vision disorders. It also performed well in cancers, although evidence was limited, and there was no evidence in skin conditions. There were limited data for SF-6D in all four conditions and limited evidence on reliability of all instruments.

Brazier and colleagues (2014) undertook similar reviews in mental health. They found evidence to support validity in depression, and to some extent in anxiety and personality disorder. Results were more mixed in schizophrenia and bipolar disorder, with a suggestion that EQ-5D and SF-36 may be reflecting depression rather than other consequences of these conditions. Qualitative research undertaken by the same group suggests this is because these measures do not cover most of the concerns for patients with mental health problems (Connell et al. 2012).

The mapping algorithms studied by Longworth et al. (2014) were estimated to predict EQ-5D values from alternative cancer-specific measures of health. Response mapping using all the domain scores was the best-performing model for the EORTC QLQ-C30. An exploratory valuation study found that bolt-on items to EQ-5D for vision, hearing, and tiredness had a significant impact on values of the health states, but the direction and magnitude of differences depended on the severity of the health state. The vision

bolt-on item had a statistically significant impact on EQ-5D health state values and a full valuation model was estimated.

Therefore, they concluded that EQ-5D performs well in studies of cancer and skin conditions. Mapping techniques provide a solution to predict EQ-5D values where EQ-5D has not been administered. For conditions where EQ-5D was found to be inappropriate, including some vision disorders and hearing, bolt-ons provide a promising solution. More primary research into the psychometric properties of the generic preference-based measures is required, particularly in cancer and for the assessment of reliability. Further research is needed for the development and valuation of bolt-ons to EQ-5D.

Another approach to dealing with situations where generic measures are not considered sufficiently relevant or sensitive is to use one of the growing number of condition-specific preference-based measures. These can be developed *de novo* or based on existing measures, which may be more acceptable to clinicians. The process for constructing condition-specific measures is well developed (Brazier et al. 2012) and there are examples in areas as diverse as vision (Rentz et al. 2014), cancer, (Rowen et al. 2011), asthma (Revicki et al. 1998) and diabetes (Sundaram et al. 2010). However, there are concerns that QALYs calculated using condition-specific measures may not be comparable, even where they have been valued using the same methods, due to the impact of side effects and comorbidities (on health effects and valuation), condition labels, and focusing effects (from having narrower descriptive systems) (Brazier and Tsuchyia 2010). While a decision-specific measure may be more relevant and sensitive, this advantage must be weighed up against any potential loss of comparability for informing cross-condition resource allocation decisions.

5.7 Whose values should be used to value health states?

One important issue in valuing health states is that of whose values should be used. The early studies primarily used convenience samples, consisting mainly of patients, health professionals, or mixes of the two. Nowadays, most of the generic preference-based instruments for estimating QALYs use samples of the general public.

The source of values can matter. For example, Suarez-Alomar and Conner-Spady (2001) elicited preferences for two arthritis health states (mild and severe) using VASs, TTO, and standard gamble methods by interviewing 104 individuals from the general public, 51 patients with rheumatoid arthritis, and 43 health professionals. The health scenarios were based on attributes described in the EQ-5D. They then compared the ratings in their survey with those obtained for the same scenarios by the UK scoring algorithm used for the EQ-5D. Statistically significant differences were observed in the ratings of the health scenarios, mostly for the severe vignette. The cost-effectiveness ratio for a hypothetical intervention varied according to the method employed to determine the utility of the health states, from US$15 000 to US$111 000 per QALY.

Brazier et al. (2009) speculate that differences between patient preferences and those of the general population for life-saving interventions may be greater than those for interventions that predominantly affect quality of life. Indeed, due to adjustment to the condition, patient preferences for health states often tend to be higher than those

of the general public (Nord et al. 2009; Shaw 2011). Because of this, the use of patient preferences can result in a lower estimate of the QALYs gained from interventions that impact mainly on the quality, rather than the length, of life. Therefore, in the context of using economic evaluation to decide on which therapies should be provided, the use of patient preferences may not always be in the interests of patients!

The main debate centres on the choice between patients and the general public as the source of values for valuing health states. One way of making the choice would be purely on normative grounds. For example, one could argue that the values should come from patients because they are the potential recipients of the treatments that are being evaluated. Alternatively, one could argue that the values should come from the general public, since through the taxes they pay, they provide most of the funds for health care in many countries, particularly those with a publicly funded national health service or health insurance scheme. One might also argue that the choice of values may depend on the decision-making context. For example, in deciding whether or not to allocate public funds to a new intervention, it might be more appropriate to use the values of the general public, whereas in making more circumscribed treatment choices among therapies that are already funded, patient preferences may be more appropriate.

There may also be practical difficulties in eliciting the values. For example, asking patients suffering from a particular condition may raise ethical issues, particularly if the estimation method involves trading the length and quality of life, or the certainty of a given health state versus a gamble involving probabilities of full health or death. Alternatively, members of the general public may not be motivated to participate in studies, or may not take them seriously.

However, the most interesting aspect of the choice of source of values is that in the case of patients these represent experienced health states, whereas in the case of the general public the values will, in the most part, reflect *ex ante* preferences for states they have not experienced. Therefore, the key issue in eliciting health state preferences from the general public would be to explain how a state an individual has never experienced might impact on their quality of life. Even eliciting values from individuals who have experienced health states is not without its difficulties. Kahneman (2009) points out that measuring an experience is different from measuring a preference from an 'experienced' person. He cites a study where patients with colostomy are happy with their colostomy and expect to be happy without it. However, once the colostomy is removed, they remember their previous state as absolutely horrible and, in terms of preferences, would be willing to pay a great deal to get rid of it (Smith et al. 2006). He argues that a problem arising from this reversal in perception is that it is difficult to generate a number that decision-makers would take seriously.

Given that either patient preferences and general population preferences could be appropriate depending on the situation, some researchers have suggested a hybrid approach. Brazier et al. (2009) argue that, for economic evaluation to reflect practice, it must take account of the impact of patient preferences (e.g. on compliance with therapy), even if general population preferences are used to value the benefits of interventions. In trying to reach a consensus on this issue, Drummond et al. (2009) acknowledge that, depending on the situation, both *ex ante* preferences about a health state, and the preferences of those experiencing it, could be relevant. In some situations both sets of

preferences could be combined, in that members of the public could be informed about the views of individuals who have experienced the health states that are the subject of valuation.

On a practical level, it is important that those undertaking economic evaluations are aware of this debate, particularly if they are eliciting values for health states *de novo*. However, in practice, the source of values is often determined as a result of the choice of preference-based instrument. Many analysts chose to use a generic instrument, which usually implies the use of general population values.

5.8 Criticisms of QALYs

Despite the fact that QALYs are now widely used in economic evaluations, they have attracted several criticisms. (See Reed Johnson (2009) for a fairly broad critique.) The main criticisms are discussed below. In addition, some of the alternatives to QALYs are discussed here, while others are discussed in Chapter 6.

5.8.1 QALYs are not really 'utilities'

It was mentioned in Section 5.3.1 that QALYs are not 'utilities' as consistent with the conventional meaning of the term in economics. That is, in the health economic evaluation literature, the term 'utility' is used in the context of a vNM utility. So all QALYs that are formed from preferences measured in any way other than with a standard gamble, by definition, cannot be utilities. But what about QALYs formed from preferences for health states measured with a standard gamble? Can they not be utilities? It turns out they can be, but only under quite restrictive assumptions (Torrance and Feeny 1989; Weinstein and Fineberg 1980). The two attributes of quality and quantity (of life) must be mutually utility independent (preferences for gambles on the one attribute are independent of the amount of the other attribute); the trade-off of quantity for quality must exhibit the constant proportional trade-off property (the proportion of remaining life that one would trade off for a specified quality improvement is independent of the amount of remaining life); and the single-attribute utility function for additional healthy life-years must be linear with time (for a fixed quality level one's utilities are directly proportional to longevity, a property also referred to as risk neutrality with respect to time). These conditions, particularly the last one, may not hold in practice, and thus even a QALY weighted with preferences obtained via the standard gamble is generally not a utility, consistent with the strict use of term in economics.

By reference to practical clinical examples from coronary heart disease, Pliskin et al. (1980) demonstrated that certain plausible independence properties lead to a quasi-additive utility function for life-years and health status. In particular, remaining longevity can be shown to be utility independent of health status and the constant proportional trade-off may hold, as the amount individuals are willing to give up for an improvement in health status from any given level to another level does not depend on the absolute number of life-years remaining.

So could a QALY turn out to be a good approximation of a utility? Garber and Phelps (1995) argue that it could, and describe a set of assumptions under which decisions based on cost/QALY would be entirely consistent with welfare economic theory (see

also Garber et al. 1996). Empirically, it may turn out that QALYs are an adequate approximation of utilities, at least under most situations. So far, there are few data. One study has found that the approximation is not adequate at the individual level, but at the group level it looks promising to use a regression approach to predict path utilities from the utilities of the component states (Kuppermann et al. 1997). More research is clearly needed to determine the situations where a QALY is a good approximation of a utility and where it is not. (See Brazier et al. (2007) for a fuller discussion.)

However, does it really matter whether or not a QALY is a utility? Those following the 'extra-welfarist' approach (outlined in Chapter 2) take the view is that the QALY is a good basic definition of what we are trying to achieve in health care, and maximizing QALYs is quite an appropriate goal (Culyer 1989).

Two of the restrictive assumptions of QALYs that have most often been discussed are those of constant proportionality and additive independence. Namely, the value one places in a health state should not be dependent on the time spent in the state and the value placed on states should be independent of the order in which they are experienced. These are important because of the way (in the standard QALY approach) values for health states are multiplied by the time in each state to calculate the number of QALYs gained from treatment. However, there is some empirical evidence that the value individuals place on a sequence (or *profile*) of health states is different from that which would be inferred from the individual state values, calculating the value of the profile using the standard QALY approach (Richardson et al. 1996).

Healthy-year equivalents (HYEs) have been proposed as a theoretically superior alternative to QALYs, but one that is more challenging to execute (Mehrez and Gafni 1989, 1991, 1992). Essentially, the HYE approach, as proposed by Mehrez and Gafni, differs from the conventional approach to QALYs in two respects. First, it measures the preferences over the entire path (or profile) of health states through which the individual would pass, rather than for each state alone. Second, it measures the preferences using a two-stage standard gamble measurement procedure that first measures the conventional utility for the path and then measures the number of healthy years that would give the same utility.

There has been extensive discussion and debate on all aspects of HYEs. In addition to the references listed throughout this section, other articles on the HYE debate include Bleichrodt (1995), Culyer and Wagstaff (1993, 1995), Fryback (1993), Gafni and Birch (1993), and Mehrez and Gafni (1993). In this book we are not able to go into all the intricacies of the entire debate; interested readers can go to the original sources for that. However, a few of the key points are summarized below.

- Measuring preferences over a path of health states is theoretically attractive but more difficult in practice. It is theoretically attractive because it is a more general approach to preference measurement, and imposes fewer restrictive assumptions. Thus, it is more likely to capture more accurately the true preferences of the individuals. It is more difficult in practice for two reasons. First, each measurement task is more difficult for the respondent; the certain alternative in the standard gamble is a path of health states rather than a single health state. Because each health state often requires considerable detail to describe appropriately (examples include up to a half page of text, a description of the health status on eight

attributes, or even videos to describe a single health state), one may quickly run into cognitive overload with many subjects. Second, in many practical problems there may be a large number of health paths to be assessed. Indeed, a modest-sized Markov model could easily have 8 states and 20 cycles, which would give over 1018 unique health paths to be assessed—a daunting task.

- The concept of measuring preferences over a path of health states is not restricted to HYEs, but could also be used with QALYs, if desired. In such a case, the QALY for a path would be the utility of the path, as measured for example in a single standard gamble, multiplied by the duration of the path.
- All researchers who have independently evaluated the HYE have concluded that the two-stage standard gamble measurement procedure originally proposed for the HYE is theoretically equivalent to a one-stage TTO procedure (Buckingham 1993; Johannesson et al. 1993; Loomes 1995; Wakker 1996; Weinstein and Pliskin 1996; Williams 1995). This conclusion does not imply that the two measurement procedures will give identical results, but it does imply that if the two procedures differ empirically, there are no theoretical grounds for choosing one over the other. Presumably one would choose the procedure with the least potential for measurement error, which would seem to be the TTO method (Wakker 1996; Weinstein and Pliskin 1996). Note, also, that neither the HYE nor the TTO captures the individual's risk attitude, so neither captures fully the individual's preferences under risk.
- The HYE is not a utility and is not intended as an alternative to utility theory (Gafni 1996; Wakker 1996). Under vNM utility theory, the appropriate approach for an individual is to solve their decision problem using conventional utility theory. This would involve measuring utilities for each health path (the first standard gamble in the HYE approach), taking expected utilities, and selecting the alternative with the largest expected utility.

5.8.2 QALYs do not encapsulate all the relevant attributes of health care

By definition, QALYS focus on health gain in terms of improvements in length and quality of life. However, it is sometimes argued that individuals value other attributes associated with the provision of health care, such as increased convenience in accessing and using health care treatments and services (Higgins et al. 2014). This raises two issues: what value, if any, do individuals place on these attributes and is it legitimate to include them in the measure of benefit we use in evaluating health care treatments and programmes?

The first issue can be addressed by using methods of valuation that include attributes other than the two typically considered in estimating the QALYs gained. For example, one could revert to contingent valuation, where the gains from the consumption of health care are estimated in monetary terms. Alternatively, one can use an approach called 'discrete choice experiments', where several attributes of health care programmes and services can be specified and the trade-offs between them explored. These approaches are discussed in Chapter 6.

The second issue relates to the alternative approaches to the economic evaluation of health care discussed in Chapters 2 and 4. For example, if one adopted a welfarist approach, any value that individuals place on the various attributes of health care would be relevant to the analysis. Alternatively, if one adopted an extra-welfarist approach, one might question the relevance of considering attributes that do not directly contribute to increased health gain and might prefer to limit the consideration to improved length and quality of life. Of course, the question of relevance may not straightforward. For example, increased convenience might increase compliance with therapy, and this may translate into improvements in length and quality of life, although this would have to be demonstrated. Alternatively, some elements of the value from increases in the process utility of receiving care may be captured in the psychological dimension of the instrument used to estimate health state preference values.

5.8.3 QALYs may not reflect social values

In the standard approach for estimating the QALYs gained from health care interventions we elicit health state preference values from individuals and then aggregate the QALYs gained in order to estimate the total QALYs gained from a given treatment or programme. This approach treats all QALYs as if they are of equal value; however, it is sometimes argued that on a societal level we may value QALYs differently depending on the situation. For example, we may place a higher value on a QALY gained by someone suffering from a serious disease, as opposed to a mild one, or we may place a higher value on a QALY gained by a young person, as opposed to an elderly person. If this were true, a simple aggregation of the QALYs gained would not reflect the social value of health care programmes. There have been several studies exploring aspects of societal values. See Linley and Hughes (2013) for a recent example.

Researchers have found that when members of the general public are asked specific PTO questions, like 'how many patients of type A should be cured to be equivalent in social value to curing 10 patients of type B', the results do not match conventional QALYs (Nord 1995). The reasons often seem to relate to equity considerations (e.g. help the sicker people first, treat all equally regardless of capacity to benefit), and perhaps to the 'rule of rescue' (Hadorn 1991) in which life-saving is always given the highest priority.

Therefore, the PTO approach, to determine saved young life equivalents (SAVEs), has been suggested as an alternative to the conventional QALY (Green 2001; Nord 1995, 1996, 1999; Nord et al. 1993). Nord reports that PTO results do not match the results from traditional techniques like rating scale, standard gamble, and TTO and that the differences can be quite large. Moreover, Nord argues that the PTO scores are more appropriate for use in resource allocation, because they are based directly on the trade-offs that society considers appropriate.

The basic argument is that the weights for conventional QALYs, and thus the QALYs themselves, reflect an individualistic perspective and not a societal perspective, and thus the conventional QALY does not measure social value. Specifically, it is noted that the weights for conventional QALYs represent an aggregation of preferences and

trade-offs that individuals hold for their own health. That is, they represent the trade-offs among various living states and between living states and death that the individuals would want for themselves. In aggregating, all people's preferences are considered equal, although other weighting schemes are possible (Williams 1988), and so the resulting QALY reflects a particular equity position.

The SAVE is the common metric that can be used with the PTO approach. All programmes are converted through PTO measurements to their SAVE value, and programmes are compared on the basis of costs and SAVEs. Conventional QALYs and SAVEs take different approaches to the definition and measurement of preferences from a societal perspective. In the QALY approach, each member of society is asked what kinds of trade-offs they would like *for themselves*, and the societal decision-making is made consistent with these trade-offs. In the SAVE approach each member of society is asked what kinds of trade-offs they would like *for others*, and this forms the basis for the societal decision-making. It is not clear on theoretical or ethical grounds that one is better than the other. There would be no problem if the two approaches gave similar results, but it appears they do not. The SAVE approach appears to give more emphasis to quantity of life, and less to quality of life. That is, compared to the QALY approach the SAVE approach is less willing to take mortality risks to improve quality of life. This may represent the human tendency, also seen in other fields, of being more conservative when giving advice than when taking it.

The debate about whether QALYs reflect societal preferences raises issues about whether QALYs should be weighted other than equally, whether other factors (such as equity) should be considered in decisions about the allocation of health care resources and whether these factors (if relevant) should be introduced into the analysis itself or enter through a deliberative decision-making process.

5.9 **Further reading**

The literature on the measurement of HRQoL has expanded considerably in recent years and it is not possible to cover it in much depth in a book about economic evaluation in general. Therefore, this chapter has concentrated primarily on preference-based measures that generate a single index, because these measures have been used to estimate the QALYs gained in economic evaluations of health treatments and programmes.

Those wishing to study QoL measurement in more detail should consult a general book on the topic, such as that by Brazier et al. (2007). Examples of empirical studies and methodological articles can be found in journals such as *Quality of Life Research* (the official journal of the International Society for Quality of Life Research, ISOQOL) and *Value in Health* (the official journal of the International Society for Pharmacoeconomics and Outcomes Research, ISPOR). In addition, papers on QoL measurements in particular health conditions can be found in clinical journals in the relevant field. A useful paper on the reporting of QoL studies is the CONSORT statement produced by Calvert et al. (2013). In addition, several task force reports produced by ISPOR deal with aspects of QOL measurement (<http://www.ispor.org/workpaper/practices_index.asp>).

References

Ades, A.E., Lu, G., and Madan, J.J. (2013). Which health-related quality-of-life outcome when planning randomized trials: disease-specific or generic, or both? A common factor model. *Value in Health*, **14**, 185–94.

Airoldi, M. and Morton, A. (2009). Adjusting life for quality or disability: stylistic difference or substantial dispute? *Health Economics*, **18**, 1237–47.

Arnesen, T. and Kapiriri, L. (2004). Can the value choices in DALYs influence global priority-setting? *Health Policy*, **70**, 137–49.

Bass, E.B., Steinberg, E., Pitt, H., et al. (1994). Comparison of the rating scale and the standard gamble in measuring patient preferences for outcomes of gallstone disease. *Medical Decision Making*, **14**, 307–14.

Bell, D. and Farquhar, P. (1986). Perspectives on utility theory. *Operations Research*, **34**, 179–83.

Bennett, K.J. and Torrance, G.W. (1996). Measuring health state preferences and utilities: rating scale, time trade-off and standard gamble techniques, in B. Spilker (ed.), *Quality of life and pharmacoeconomics in clinical trials*, 2nd edition, pp. 253–65. Philadelphia: Lippincott-Raven.

Bleichrodt, H. (1995). QALYs and HYEs: Under what conditions are they equivalent? *Journal of Health Economics*, **14**, 17–37.

Bleichrodt, H. (2002). A new explanation for the difference between time trade-off utilities and standard gamble utilities. *Health Economics*, **11**, 447–56.

Bleichrodt, H. and Johannesson, M. (1997). An experimental test of a theoretical foundation for rating scale valuations. *Medical Decision Making*, **17**, 208–16.

Brazier, J.E. and Deverill, M. (1999). A checklist for judging preference-based measures of health related quality of life: learning from psychometrics. *Health Economics*, **8**, 41–51.

Brazier, J. and Tsuchiya, A. (2010). Preference-based condition specific measures of health: what happens to cross programme comparability. *Health Economics*, **19**, 125–9.

Brazier, J., Roberts, J., and Deverill, M. (2002). The estimation of a preference-based measure of health from the SF-36. *Journal of Health Economics*, **21**, 271–92.

Brazier, J.E., Green, C., McCabe, C., and Stevens, K. (2003). Use of visual analogue scales in economic evaluation. *Expert Review of Pharmacoeconomics and Outcomes Research*, **3**, 293–302.

Brazier, J.E., Koltkin, R.L., Crosby, R.D., and Rhys Williams, G. (2004). Estimating a preference-based single index for the Impact of Weight on Quality of Life-Lite (IWQOL-Lite) instrument from the SF-6D. *Value in Health*, **7**, 490–8.

Brazier, J.E., Ratcliffe, J., Salomon, J., and Tsuchiya, A. (2007). *Measuring and valuing health benefits for economic evaluation*. Oxford: Oxford University Press.

Brazier, J.E., Czoski-Murray, C., Roberts, J., Brown, M., Symonds, T., and Kelleher, C. (2008). Estimation of a preference-based index from a condition-specific measure: The King's Health Questionnaire. *Medical Decision Making*, **28**, 113–26.

Brazier, J.E., Dixon, S., and Ratcliffe, J. (2009). The role of patient preferences in cost-effectiveness analysis. *PharmacoEconomics*, **27**, 705–12.

Brazier, J.E., Yang, Y., Tsuchiya, A., and Rowen, D.L. (2010). A review of studies mapping (or cross walking) non-preference based measures of health to generic preference-based measures. *European Journal of Health Economics*, **11**, 215–25.

Brazier, J., Rowen, D., Mavranezouli, I., Tsuchiya, A., Young, T., and Yang, Y. (2012). Developing and testing methods for deriving preference-based measures of health from condition

specific measures (and other patient based measures of outcome). *Health Technology Assessment*, **16**(32).

Brazier, J., Connell, J., Papaioannou, D., et al. (2014). A systematic review, psychometric analysis and qualitative assessment of generic preference-based measures of health in mental health populations and the estimation of mapping functions from widely used specific measures. *Health Technology Assessment*, **18**(34), vii–viii, xiii–xxv, 1–188.

Brooks, R. with the EuroQol Group (1996). EuroQol: the current state of play. *Health Policy*, **37**, 53–72.

Buckingham, K. (1993). A note on HYE (healthy years equivalent). *Journal of Health Economics*, **11**, 301–9.

Buckingham, J.K., Birdsall, J., and Douglas, J.G. (1996). Comparing three versions of the time tradeoff: time for a change? *Medical Decision Making*, **16**, 335–47.

Calvert, M., Blazeby, J., Altman, D.G., Revicki, D.A., Moher, D., Brundage, M.D, for the CONSORT PRO Group (2013). Reporting of patient-reported outcomes in randomized trials. The CONSORT PRO extension. *JAMA*, **309**, 814–22.

Chuang, L-H. and Kind, P. (2009). Converting the SF-12 into the EQ-5D: an empirical comparison of methodologies. *PharmacoEconomics*, **27**, 491–505.

Churchill, D., Torrance, G., Taylor, D., et al. (1987). Measurement of quality of life in end-stage renal disease: The time trade-off approach. *Clinical and Investigative Medicine*, **10**, 14–20.

Ciani, O., Hoyle, M., Pavey, T., et al. (2013). Complete cytogenetic response and major molecular response as surrogate outcomes for overall survival in first-line treatment of chronic myelogenous leukemia: a case study for technology appraisal on the basis of surrogate outcomes evidence. *Value in Health*, **16**, 1081–90.

Connell, J., Brazier, J.E., O'Cathain, A., Lloyd-Jones, M., and Paisley, S. (2012). Quality of life of people with mental health problems: a synthesis of qualitative research. *Health and Quality of Life Outcomes*, **10**, 138.

Conner-Spady, B. and Suarez-Almazor, M.E. (2003). Variation in the estimation of quality-adjusted life-years by different preference-based instruments. *Medical Care*, **41**, 791–801.

Cook, J., Richardson, J., and Street, A. (1994). A cost–utility analysis of treatment options for gallstone disease: methodological issues and results. *Health Economics*, **3**, 157–68.

Cooper, R. and Rappoport, P. (1984). Were the ordinalists wrong about welfare economics? *Journal of Economic Literature*, **22**, 507–30.

Culyer, A. (1989). The normative economics of health care finance and provision. *Oxford Review of Economic Policy*, **5**, 34–58.

Culyer, A. and Wagstaff, A. (1993). QALYs versus HYEs. *Journal of Health Economics*, **11**, 311–23.

Culyer, A.J. and Wagstaff, A. (1995). QALYs versus HYEs: A reply to Gafni, Birch and Mehrez. *Journal of Health Economics*, **14**, 39–45.

Dakin, H. (2013). Review of studies mapping from quality of life or clinical measures to EQ-5D: an online database. *Health and Quality of Life Outcomes*, **11**, 151–7.

Devlin, N.J. and Krabbe, P.F.M. (2013). The development of new research methods for the valuation of EQ-5D-5L. *European Journal of Health Economics*, **14**(Suppl 1), S1–S3.

Devlin, N., Tsuchiya, A., Buckingham, K., and Tilling, C. (2011). A uniform time trade off method for states better and worse than dead: feasibility study of the 'lead time' approach. *Health Economics*, **20**, 348–61.

Dolan, P., Gudex, C., Kind, P., and Williams, A. (1995). *A social tariff for EuroQoL: Results from a UK general population survey*, Discussion Paper No. 138. York: Centre for Health Economics, University of York.

Dolan, P., Gudex, C., Kind, P., and Williams, A. (1996a). Valuing health states: a comparison of methods. *Journal of Health Economics*, **15**, 209–31.

Dolan, P., Gudex, C., Kind, P., and Williams, A. (1996b). The time trade-off method: results from a general population study. *Health Economics*, **5**, 141–54.

Dolan, P. and Roberts, J. (2002). To what extent can we explain time trade-off values from other information about respondents? *Social Science and Medicine*, **54**, 919–29.

Drummond, M.F., Brixner, D., Gold, M., Kind, P., McGuire, A., and Nord, E. (2009). Toward a consensus on the QALY. *Value in Health*, **12**(suppl. 1), S31–S35.

Dyer, J. and Sarin, R. (1979). Measurable multi-attribute value functions. *Operations Research*, **27**, 810–22.

Dyer, J. and Sarin, R. (1982). Relative risk aversion. *Management Science*, **28**, 875–86.

Essink-Bot, M., Stouthard, M., and Bonsel, G. (1993). Generalizability of valuations on health states collected with the EuroQol questionnaire. *Health Economics*, **2**, 237–46.

EuroQol Group (1990). EuroQol—a new facility for the measurement of health-related quality of life. *Health Policy*, **16**, 199–208.

Fayers, P. and Bottomley, A. (2002). Quality of life research within the EORTC—the EORTC QLQ-C30. *European Journal of Cancer*, **38**, 125–33.

Feeny, D., Furlong, W., Torrance, G., et al. (2002). Multiattribute and single-attribute utility functions for the Health Utilities Index Mark 3 system. *Medical Care*, **40**, 113–28.

Fryback, D.G. (1993). QALYs, HYEs, and the loss of innocence (editorial). *Medical Decision Making*, **13**, 271–2.

Fryback, D.G., Cherepanov, D., Bolt, D., and Kim, J-S. (2010). Comparison of 5 health-related quality of life indexes using item response theory analysis. *Medical Decision Making*, **30**, 5–15.

Furlong, W., Feeny, D., Torrance, G., Barr, R., and Horsman, J. (1990). *Guide to design and development of health-state utility instrumentation*, Working Paper No. 90–9. Hamilton, Ontario: McMaster University, Centre for Health Economics and Policy Analysis.

Furlong, W.J., Feeny, D.H., Torrance, G.W., and Barr, R.D. (2001). The Health Utilities Index (HUI) system for assessing health-related quality of life in clinical studies. *Annals of Medicine*, **33**, 375–84.

Gafni, A. (1996). HYEs: Do we need them and can they fulfil the promise? *Medical Decision Making*, **16**, 215–16.

Gafni, A. and Birch, S. (1993). Economics, health and health economics: HYEs versus QALYs. *Journal of Health Economics*, **11**, 325–39.

Gafni, A. and Torrance, G. (1984). Risk attitude and time preference in health. *Management Science*, **30**, 440–51.

Garber, A.M. and Phelps, C.E. (1995). *Economic foundations of cost-effectiveness analysis*. Stanford, CA: National Bureau of Economic Research.

Garber, A.M., Weinstein, M.C., Torrance, G.W., and Kamlet, M.S. (1996). Theoretical foundations of cost-effectiveness analysis, in M.R. Gold, J.E. Siegel, L.B. Russell, and M.C. Weinstein (ed.), *Cost-effectiveness in health and medicine*, pp. 25–53. New York: Oxford University Press.

Gold, M.R., Siegel, J.E., Russell, L.B., and Weinstein, M.C. (1996). *Cost-effectiveness in health and medicine*. New York: Oxford University Press.

Gorber, S. (2003). A new classification and measurement system of functional health. *Au Courant (Statistics Canada Catalogue 82-005-XIE)*, September, 2–3.

Green, C. (2001). On the societal value of health care: what do we know about the person trade-off technique? *Health Economics*, **10**, 233–43.

Guillemin, F., Bombardier, C., and Beaton, D. (1993). Cross-cultural adaptation of health-related quality of life measures: Literature review and proposed guidelines. *Journal of Clinical Epidemiology*, **46**, 1417–32.

Hadorn, D. (1991). Setting health care priorities in Oregon: cost-effectiveness meets the rule of rescue. *JAMA*, **265**, 2218–25.

Hadorn, D.C. and Uebersax, J. (1995). Large scale outcome evaluation: how should quality of life be measured? I. Calibration of a brief questionnaire and a search for preference subgroups. *Journal of Clinical Epidemiology*, **48**, 607–18.

Hadorn, D.C., Hays, R.D., and Hauber, T. (1992). Improving task comprehension in the measurement of health state preferences. *Journal of Clinical Epidemiology*, **45**, 233–43.

Hawthorne, G., Richardson, J., and Osborne, R. (1999). The assessment of quality of life (AQoL) instrument: a psychometric measure of health-related quality of life. *Quality of Life Research*, **8**, 209–24.

Hawthorne, G., Richardson, J., and Day, N.A. (2001). A comparison of the Assessment of Quality of Life (AQoL) with four other generic utility instruments. *Annals of Medicine*, **33**, 358–70.

Heintz, E., Wirehn, A-B., Bourghardt Peebo, B., Rosenqvist, U., and Levin, L-A. (2012). QALY weights for diabetic retinopathy—a comparison of health state valuations with HUI-3, ED-5D, EQ-VAS, and TTO. *Value in Health*, **15**, 475–84.

Herdman, M., Gudex, C., Lloyd, A., et al. (2011). Development and preliminary testing of the new five-level version of EQ-5D (EQ-5D-5L). *Quality of Life Research*, **20**, 1727–36.

Higgins, A., Barnett, J., Meads, C., Singh, J., and Longworth, L. (2014). Does convenience matter? A systematic review of convenience-based aspects of process utility. *Value in Health*, **17**, 877–87.

Holloway, C. (1979). *Decision making under uncertainty: models and choices*. Englewood Cliffs, NJ: Prentice-Hall.

Horsman, J., Furlong, W., Feeny, D., and Torrance, G. (2003). The Health Utilities Index (HUI): concepts, measurement properties and applications. *Health and Quality of Life Outcomes*, **1**, 54 (electronic version available free at <http://www.hqlo.com/content/1/1/54>).

IQWiG [Institute for Quality and Efficiency in Health Care] (2011). *General Methods* version 4.0. <https://www.iqwig.de/download/General_Methods_4–0.pdf>. (Accessed 2 February 2013.)

Johannesson, M., Pliskin, J., and Weinstein, M. (1993). Are healthy-years equivalents an improvement over quality-adjusted life-years? *Medical Decision Making*, **13**, 281–6.

Jones, P.W., Quirk, F.H., Baveystock, C.M., and Littlejohns, P. (2002). A self-complete measure for chronic airflow limitation—the St George's Respiratory Questionnaire. *American Review of Respiratory Disease*, **145**, 1321–7.

Joore, M., Brunenberg, D., Nelemans, P., et al. (2010). The impact of differences in EQ-5D and SF-6D utility scores on the acceptability of cost-utility ratios: results across five trial-based cost-utility studies. *Value in Health*, **13**, 222–9.

Kahneman, D. (2009) A different approach to health state valuation (summary of a presentation). *Value in Health*, **12**(suppl. 1), S16–S17.

Keeney, R. and Raiffa, H. (1976). *Decisions with multiple objectives: preferences and value trade-offs*. New York: Wiley.

Kharroubi, S.A., Brazier, J.E., and O'Hagan, A. (2007). Modelling covariates for the SF-6D standard gamble health state preference data using a nonparametric Bayesian method. *Social Science and Medicine*, **64**, 1242–52.

Kind, P. (1996). The EuroQol instrument: an index of health-related quality of life, in B. Spilker (ed.), *Quality of life and pharmacoeconomics in clinical trials*, 2nd edition, pp. 191–201. Philadelphia: Lippincott-Raven.

Klarman, H., Francis, J., and Rosenthal, G. (1968). Cost-effectiveness analysis applied to the treatment of chronic renal disease. *Medical Care*, **6**, 48–54.

Knies, S., Evers, S.M., Candel, M.J., Severens, J.L., and Ament, A.J. (2009). Utilities of the EQ-5D: transferable or not? *PharmacoEconomics*, **27**, 767–79.

Kopec, J.A. and Willison, K.D. (2003). A comparative review of four preference-weighted measures of health-related quality of life. *Journal of Clinical Epidemiology*, **56**, 317–25.

Kuppermann, M., Shiboski, S., Feeny, D., Elkin, E.P., and Washington, A.E. (1997). Can preference scores for discrete states be used to derive preference scores for an entire path of events? An application to prenatal diagnosis. *Medical Decision Making*, **17**, 4255.

Le Gales, C., Buron, C., Costet, N., Rosman, S., and Slama, G. (2002). Development of a preference-weighted health status classification system in France: the Health Utilities Index. *Health Care Management Science*, **5**, 41–51.

Lenert, L.A. (2001). The reliability and internal consistency of an Internet-capable computer program for measuring utilities. *Quality of Life Research*, **9**, 811–7.

Linley, W.G. and Hughes, D.A. (2013). Societal views on NICE, cancer drugs fund and value-based pricing criteria for prioritising medicines: a cross-sectoral survey of 4118 adults in Great Britain. *Health Economics*, **22**, 948–64.

Longworth, L. and Rowen, D. (2013). Mapping to obtain EQ-5D utility values for use in NICE health technology assessments. *Value in Health*, **16**, 202–10.

Longworth, L., Yang, Y., Young, T., et al. (2014). Use of generic and condition-specific measures of health-related quality of life in NICE decision-making: a systematic review, statistical modelling and survey. *Health Technology Assessment*, **18**, 1–224.

Loomes, G. (1995). The myth of the HYE. *Journal of Health Economics*, **14**, 1–7.

Lu, G., Brazier, J.E., and Ades, A.E. (2013). Mapping from disease-specific top generic health-related quality-of-life scales: a common factor model. *Value in Health*, **16**, 177–84.

Lubetkin, E.I. and Gold, M.R. (2003). Areas of decrement in health-related quality of life (HRQL): comparing the SF-12, EQ-5D, and HUI3. *Quality of Life Research*, **12**, 1059–67.

Manca, A., Hawkins, N., and Sculpher, M. (2005). Estimating mean QALYs in trial-based cost-effectiveness analysis: the importance of controlling for baseline utility. *Health Economics*, **14**, 487–96.

Martin, A.J., Glasziou, P.P., Simes, R.J., and Lumley, T.A. (2000). Comparison of standard gamble, time trade-off, and adjusted time trade-off scores. *International Journal of Technology Assessment in Health Care*, **16**, 137–47.

McCabe, C., Edlin, R., Meads, D., Brown, C., and Kharroubi, S. (2013). Constructing indirect utility models: some observations on the principles and practice of mapping to obtain health state utilities. *PharmacoEconomics*, **31**, 636–41.

McCrone, P., Patel, A., Knapp, M., et al. (2009). A comparison of SF-6D and EQ-5D utility scores in a study of patients with schizophrenia. *Journal of Mental Health Policy and Economics*, **12**, 27–31.

McKenzie, L. and van der Pol, M. (2009). Mapping the EORTC QLQ C-30 onto the EQ-5D instrument: the potential to estimate QALYs without generic preference data. *Value in Health*, **12**, 167–71.

Mehrez, A. and Gafni, A. (1989). Quality-adjusted life years, utility theory, and healthy-years equivalents. *Medical Decision Making*, **9**, 142–9.

Mehrez, A. and Gafni, A. (1990). Evaluating health-related quality of life: an indifference curve interpretation for the time trade-off technique. *Social Science in Medicine*, **31**, 1281–3.

Mehrez, A. and Gafni, A. (1991). The healthy-years equivalents: how to measure them using the standard gamble approach. *Medical Decision Making*, **11**, 140–6.

Mehrez, A. and Gafni, A. (1992). Preference based outcome measures for economic evaluation of drug interventions: quality adjusted life years (QALYs) versus healthy years equivalents (HYEs). *PharmacoEconomics*, **1**, 338–45.

Mehrez, A. and Gafni, A. (1993). Healthy-years equivalents versus quality-adjusted life years: in pursuit of progress. *Medical Decision Making*, **13**, 287–92.

Miyamoto, J.M. (1988). Generic utility theory: measurement foundations and applications in multiattribute utility theory. *Journal of Mathematical Psychology*, **32**, 357–404.

Mortimer, D. and Segal, L. (2008). Is the value of a life or life-year saved context specific? Further evidence from a discrete choice experiment. *Cost Effectiveness and Resource Allocation*, **6**, 8.

Murray, C.J.L. and Lopez, A.D. (1996). *The global burden of disease: a comprehensive assessment of mortality and disability from diseases, injuries, and risk factors in 1990 and projected to 2020*. Cambridge, MA: Harvard University Press.

NICE [National Institute for Health and Care Excellence] (2013). *Guide to the methods of technology appraisal*, April 2013. London: UK National Health Service, National Institute for Health and Care Excellence.

Nord, E. (1995). The person-trade-off approach to valuing health care programs. *Medical Decision Making*, **15**, 201–8.

Nord, E. (1996). Health status index models for use in resource allocation decisions: A critical review in the light of observed preferences for social choice. *International Journal of Technology Assessment in Health Care*, **12**, 31–44.

Nord, E. (1999). *Cost-value analysis in health care: making sense out of QALYs*. Cambridge, MA: Cambridge University Press.

Nord, E., Richardson, J., and Macarounas-Kirchmann, K. (1993). Social evaluation of health care versus personal evaluation of health states. *International Journal of Technology Assessment in Health Care*, **9**, 463–78.

Nord, E., Daniels, N., and Kamlet, M. (2009). QALYs: some challenges. *Value in Health*, **12** (suppl. 1), S10–S15.

O'Brien, B., Spath, M., Blackhouse, G., Severens, J.L., Dorian, P., and Brazier, J. (2003). A view from the bridge: agreement between the SF-6D utility algorithm and the Health Utilities Index. *Health Economics Letters*, **7**, 9–15.

O'Brien, B.J., Torrance, G.W., and Moran, L.A. (1994). *A Practical Guide to Health State Preference Measurement: a video introduction*, Working Paper No. 95–2. Hamilton, Ontario: Centre for Health Economics and Policy Analysis, McMaster University.

O'Leary, J.F., Fairclough, D.L., Jankowski, M.K., and Weeks, J.C. (1995). Comparison of time trade-off utilities and rating scale values of cancer patients and their relatives: evidence for a possible plateau relationship. *Medical Decision Making*, **15**, 132–7.

Parkin, D. and Devlin, N. (2006). Is there a case for using visual analogue scale valuations in cost-utility analysis? *Health Economics*, **15**, 653–64.

Patrick, D. and Erickson, P. (1993). *Health status and health policy: quality of life in health care evaluation and resource allocation*. New York: Oxford University Press.

Patrick, D., Bush, J., and Chen, M. (1973). Methods for measuring levels of well-being for a health status index. *Health Services Research*, **8**, 228–45.

Patrick, D., Starks, H., Cain, K., Uhlmann, R., and Pearlman, R. (1994). Measuring preferences for health states worse than death. *Medical Decision Making*, **14**, 9–18.

Pliskin, J.S., Shepherd, D.S., and Weinstein, M.C. (1980). Utility functions for life years and health status. *Operations Research*, **28**, 206–24.

Read, J., Quinn, R., Berwick, D., Fineberg, H., and Weinstein, M. (1984). Preferences for health outcomes—comparisons of assessment methods. *Medical Decision Making*, **4**, 315–29.

Reed Johnson, F. (2009). Editorial: Moving the QALY forward or just stuck in traffic? *Value in Health*, **12** (suppl 1), S38–9.

Rentz, A.M., Kowalski, J.W., Walt, J.G., et al. (2014). Development of a preference-based index from the National Eye Institute Visual Function Questionnaire-25. *JAMA Ophthalmology*, **132**, 310–18.

Revicki, D.A., Leidy, N.K., Brennan-Diemer, F., Sorensen, S., and Togias, A. (1998). Integrating patient preferences into health outcomes assessment: the multiattribute Asthma Symptom Utility Index. *Chest*, **114**, 998–1007.

Richardson, J., Hall, J., and Salkeld, G. (1996). The measurement of utility in multiphase health states. *International Journal of Technology Assessment in Health Care*, **12**, 151–62.

Richardson, J., McKie, J., and Bariola, E. (2014). Multiattribute utility instruments and their use, in A.J. Culyer (ed.), *Encyclopedia of Health Economics*, Vol. 2, pp. 341–57. San Diego, CA: Elsevier.

Robberstad, B. (2005). QALYs vs DALYs vs LYs gained: What are the differences, and what difference do they make for health care priority setting? *Norsk Epidemiologi*, **15**, 183–99.

Robinson, A., Loomes, G., and Jones-Lee, M. (2001). Visual analogue scales, standard gambles and relative risk aversion. *Medical Decision Making*, **21**, 17–27.

Ross, P.L., Littenberg, B., Fearn, P., Scardino, P.T., Karakiewicz, P.I., and Kattan, M.W. (2003). Paper standard gamble: a paper-based measure of standard gamble utility for current health. *International Journal of Technology Assessment in Health Care*, **19**, 135–47.

Rosser, R. and Kind, P. (1978). A scale of valuations of states of illness: is there a social consensus? *International Journal of Epidemiology*, **7**, 347–58.

Rosser, R. and Watts, V. (1978). The measurement of illness. *Journal of Operational Research Society*, **29**, 529–40.

Rowen, D. and Brazier, J. (2011). Health utility measurement, in S. Glied and P. Smith (ed.), *The Oxford handbook of health economics*, pp. 788–813. Oxford: Oxford University Press.

Rowen, D., Brazier, J., Young, T., et al. (2011). Deriving a preference-based measure for cancer using the EORTC QLQ-C30. *Value in Health*, **14**, 721–31.

Rushby, J.F. and Hanson, K. (2001). Calculating and presenting disability adjusted life years (DALYs) in cost-effectiveness analysis. *Health Policy and Planning*, **16**, 326–31.

Rutten-van Molken, M.P.M.H., Bakker, C.H., van Doorslaer, E.K.A., and van der Linden, S. (1995). Methodological issues of patient utility measurement: experience from two clinical trials. *Medical Care*, **33**, 9223–7.

Sackett, D. and Torrance, G. (1978). The utility of different health states as perceived by the general public. *Journal of Chronic Diseases*, **31**, 697–704.

Salomon, J. (2008). *Measurement of disability weights in the Global Burden of Disease 2005*. GBD Disability Weights Expert Consultation, Seattle, 4–5 September 2008. Seattle: Institute for Health Metrics and Evaluation, University of Washington.

Salomon, J.A., Vos, T., Hogan, D.R., et al. (2012). Common values in assessing health outcomes from disease and injury: disability weights measurement study for the Global Burden of Disease Study 2010. *Lancet*, **380**, 2129–43.

Sculpher, M.J. and Claxton, K. (2010). Sins of omission and obsfucation: IQWiG's guidelines on economic evaluation methods. *Health Economics*, **19**, 1132–6.

Sen, A. (1991). Utility: ideas and terminology. *Economics and Philosophy*, **7**, 277–83.

Shaw, J.W. (2011). Use of patient versus population preferences in economic evaluations of health care interventions. *Clinical Therapeutics*, **33**, 898–900.

Shaw, J.K., Johnson, J.A. and Coons, S.J. (2005). US validation of the EQ-5D states: development and testing of the D1 model. *Medical Care*, **43**, 203–20.

Sintonen, H. (2001). The 15D instrument of health-related quality of life: properties and applications. *Annals of Medicine*, **33**, 328–36.

Sintonen, H. and Pekurinen, M. (1993). A fifteen-dimensional measure of health-related quality of life (15D) and its applications. *Quality of life assessment: key issues in the 1990s*, pp. 185–95. Dordrecht, The Netherlands: Kluwer Academic Publishers.

Smith, D.M., Ubel, P., Sherriff, R.L., et al. (2006). Misremenbering colostomies? Former patients give lower utility ratings than do current patients. *Health Psychology*, **25**, 688–95.

Spilker, B. (1996). *Quality of life and pharmacoeconomics in clinical trials*, 2nd edition. Philadelphia: Lippincott-Raven.

Stevens, K., McCabe, C., Brazier, J.E., and Roberts, J. (2007). Multi-attribute utility function or statistical inference models: a comparison of health state valuation models using the HUI2 health state classification. *Journal of Health Economics*, **26**, 992–1002.

Stiggelbout, A., Kiebert, G., Kievit, J., Leer, J., Stoter, G., and de Haes, J. (1994). Utility assessment in cancer patients: adjustment of time tradeoff scores for the utility of life years and comparison with standard gamble scores. *Medical Decision Making*, **14**, 82–90.

Streiner, D.L. and Norman, G.R. (1989). *Health measurement scales: a practical guide to their development and use*. Oxford: Oxford University Press.

Suarez-Almazor, M.E. and Conner-Spady, B. (2001). Rating of arthritis health states by patients, physicians, and the general public. Implications for cost-utility analyses. *Journal of Rheumatology*, **28**, 648–56.

Sundaram, M., Smith, M.J., Revicki, D.A., Miller, L.A., Madhavan, S., and Hobbs, G. (2010). Estimation of a valuation function for a diabetes mellitus-specific preference-based measure of health: the Diabetes Utility Index. *PharmacoEconomics*, **28**, 201–16.

Tan-Torres Edejer, T., Baltussen, R., et al. (2003). *Making choices in health: WHO guide to cost-effectiveness analysis*. Geneva: World Health Organization.

Tengs, T. and Wallace, A. (2000). One thousand health-related quality-of-life estimates. *Medical Care*, **38**, 583–637.

Torrance, G.W. (1976). Social preferences for health states: An empirical evaluation of three measurement techniques. *Socio-Economic Planning Sciences*, **10**, 129–36.

Torrance, G.W. (1984). Health states worse than death, in W.V. Eimeren, R. Engelbrecht, and C.D. Flagle (ed.), *Proceedings of Third International Conference on Systems Science in Health Care*, pp. 1085–9. Berlin: Springer.

Torrance, G.W. (1986). Measurement of health-state utilities for economic appraisal: a review. *Journal of Health Economics*, **5**, 1–30.

Torrance, G.W. (1996). Designing and conducting cost–utility analyses, in B. Spilker (ed.), *Quality of life and pharmacoeconomics in clinical trials*, 2nd edition, pp. 1105–11. Philadelphia: Lippincott-Raven.

Torrance, G.W. and Feeny, D. (1989). Utilities and quality-adjusted life years. *International Journal of Technology Assessment in Health Care*, **5**, 559–75.

Torrance, G.W., Thomas, W., and Sackett, D. (1972). A utility maximization model for evaluation of health care programs. *Health Services Research*, **7**, 118–33.

Torrance, G.W., Boyle, M.H., and Horwood, S.P. (1982). Application of multi-attribute utility theory to measure social preferences for health states. *Operations Research*, **30**, 1043–9.

Torrance, G.W., Furlong, W.J., Feeny, D.H., and Boyle, M. (1995). Multi-attribute preference functions: health utilities index. *PharmacoEconomics*, **7**, 503–20.

Torrance, G.W., Feeny, D.H., Furlong, W.J., Barr, R.D., Zhang, Y., and Wang, Q. (1996). Multi-attribute utility function for a comprehensive health status classification system: health utilities index mark 2. *Medical Care*, **34**, 702–22.

Torrance, G.W., Feeny, D., and Furlong, W. (2001). Visual analog scales: do they have a role in the measurement of preferences for health states? *Medical Decision Making*, **21**, 329–34.

Torrance, G.W., Furlong, W., and Feeny, D. (2002). Health utility estimation. *Expert Review of Pharmacoeconomics and Outcomes Research*, **2**, 99–108.

van Hout, B., Janssen, M.F., Feng, Y-S., et al. (2012). Interim scoring for the EQ-5D-5L: mapping the EQ-5D-5L to EQ-5D-3L value sets. *Value in Health*, **15**, 708–15.

von Neumann, J. and Morgenstern, O. (1944). *Theory of games and economic behaviour*. Princeton, NJ: Princeton University Press.

Wakker, P. (1996). A criticism of healthy-years equivalents. *Medical Decision Making*, **16**, 207–14.

Wang, Q., Furlong, W., Feeny, D., Torrance, G., and Barr, R. (2002). How robust is the Health Utilities Index Mark 2 utility function? *Medical Decision Making*, **22**, 350–8.

Ware, J.E., Snow, K.K., Kolinski, M., and Gandeck, B. (1993). *SF-36 health survey manual and interpretation guide*. Boston, MA: The Health Institute, New England Medical Center.

Weinstein, M. and Fineberg, H.C. (1980). *Clinical decision analysis*. Philadelphia: W.B. Saunders.

Weinstein, M. and Pliskin, J. (1996). Perspectives on healthy-years equivalents (HYEs): What are the issues? *Medical Decision Making*, **16**, 205–6.

WHO [World Health Organization] (2013). *WHO methods and sources for global burden of disease estimates 2000–11*. Global Health Estimates Technical Paper WHO/HIS/GHE/2013.4. Department of Health Statistics and Information Systems. Geneva: World Health Organization.

Williams, A. (1988). Ethics and efficiency in the provision of health care, in M. Bell and S. Mendux (ed.), *Philosophy and medical welfare*, pp. 111–26. Cambridge: Cambridge University Press.

Williams, A. (1995). Economics, QALYs and medical ethics—a health economist's perspective. *Health Care Analysis: Journal of Health Philosophy and Policy*, **3**, 221–6.

Wolfson, A.D., Sinclair, A.J., Bombardier, C., and McGeer, A. (1982). Preference measurements for functional status in stroke patients: inter-rater and inter-technique comparisons, in R. Kane and R. Kane (ed.), *Values and long term care*, pp. 1912–14. Lexington, MA: D.C. Heath.

Wu, E.Q., Mulani, P., Farrell, M.H., and Sleep, D. (2007). Mapping FACT-P and EORTC QLQ-C30 to patient health status by EQ-5D in metastatic hormone-refractory prostate cancer patients. *Value in Health*, **10**, 408–14.

Chapter 6

Measuring and valuing effects: consumption benefits of health care

6.1 Some basics

Most of the economic evaluations in the published literature measure and value the benefits of health care interventions in terms of the measures of health gain described in Chapter 5. This is also the favoured approach in the majority of the published methods guidelines for economic evaluation specified by health care decision-makers (ISPOR 2014). However, there is a growing research literature discussing the development and application of alternative measures, such as willingness-to-pay (WTP) assessments and *discrete choice experiments* (DCEs).

Analysts may have one or more motivations for using these approaches. First, some analysts may be dissatisfied about some of the restrictive assumptions underlying the use of QALYs (quality-adjusted life-years) to reflect individuals' preferences, or the methods by which they are typically estimated. For example, in estimating the five-level version of the EQ-5D, Oppe et al. (2014) use DCEs to estimate individual's preferences for health states, in addition to the time trade-off (TTO). One of the factors favouring the DCE approach is that research has shown that it is important to administer the TTO in face-to-face interviews, which precludes studies with large numbers of respondents (Devlin and Krabbe 2013; Shah et al. 2013). In contrast, in the re-estimation of DALY weights for the WHO Global Burden of Disease study in 2010, Salomon et al. (2012), using paired comparisons, an approach similar to DCEs, were able to obtain responses from 13 902 individuals in household surveys in 5 countries, supplemented by an open-access web-based survey of 16 328 people. In addition, Brazier et al. (2013) used a discrete choice approach in a survey of 3669 individuals in the United Kingdom to explore societal values that might be used to generate weights for QALYs (Brazier et al. 2013).

Secondly, some analysts may feel that the measures of health gain discussed in Chapter 5 may not include, or poorly reflect, all the relevant benefits of health care. For example, these could include increased convenience, or process utility, resulting from a different mode of delivering the intervention concerned (e.g. oral administration of a drug, as opposed to an injection). (See Brennan and Dixon 2013 and Higgins et al. 2014 for recent reviews of studies of convenience-based aspects of process utility

in health care delivery.) However, if it could be shown that increased convenience increases adherence to therapy, the impact might be reflected in increased health gain. Alternatively, one of the major impacts of some health technologies, such as those concerned with screening or diagnosis, could be reassurance. Although increased piece of mind may be reflected in improved health-related quality of life (HRQoL), it may be inadequately captured by the methods used to estimate health gain. (See Lin et al. 2013 for a recent review of studies estimating the WTP for diagnostic technologies.)

Finally, some analysts may feel that economic evaluations in health care should be conducted using the welfarist approach discussed in Chapters 2 and 4, pointing out that there are many costs or effects of health care interventions that extend beyond the health care budget. To this end, they might question why economic evaluation in health care has departed from the use of cost–benefit analysis, measuring a wide range of costs and consequences, as is the case in the evaluation of environmental programmes or investments in the field of transport (Johnson 2012).

Therefore, this chapter discusses the various methods that have been used to measure and value the effects of health care programmes, in addition to those discussed in Chapter 5. Then, in Section 6.7, we discuss the issues associated with incorporating these methods within economic evaluations to inform health policy decisions.

6.2 Assigning money values to the outcomes of health care programmes

Traditionally economic evaluation in health has distinguished methods which assign monetary value to outcomes and those that keep outcomes in 'natural units'. However, it should be clear from the material in Chapters 2 and 4 that when the latter studies are used to make decisions, an explicit or implicit value is placed on the outcomes. Indeed, the explicit use of a cost-effectiveness threshold allows an incremental cost-effectiveness ratio to be expressed in terms of net monetary benefit (see Section 4). Therefore, as discussed in Chapter 4 (Section 4.3.4), the distinction between cost–benefit analysis and cost-effectiveness analysis is more subtle than often presented, and relates to more than whether or not outcomes are valued in monetary terms (Sculpher and Claxton 2012).

In the context of studies which have generally been termed 'cost–benefit studies', historically there have been three general approaches to the monetary valuation of health outcomes: (1) human capital, (2) revealed preferences, and (3) stated preferences of WTP. We discuss each of these approaches and review both theoretical and practical strengths and weaknesses. Attempting to assign money values to health outcomes explicitly has been, and remains, controversial. As one economist working in the field noted, 'To be trained in medicine, nursing or one of the other "sharp end" disciplines and then be faced with some hard-nosed, cold-blooded economist placing money values on human life and suffering is anathema to many' (Mooney 1992). What is often overlooked, however, is that such valuations occur—often implicitly—every day when decisions are made by both individuals and societies that trade off health objectives against other benefits.

6.2.1 The human capital approach

The utilization of a health care programme can be viewed as an investment in a person's human capital. In measuring the payback on this investment the value of the healthy time produced can be quantified in terms of the person's renewed or increased production in the marketplace. Hence the human capital method places monetary weights on healthy time using market wage rates and the value of the programme is assessed in terms of the present value of future earnings. This human capital method of valuing health status was used in the early years of economic evaluation, but is almost never used now (for an early application see Mushkin 1978). We can distinguish between two uses of the human capital concept: (1) as the *sole* basis for valuing all aspects of health improvements, and (2) as a method of valuing *part* of the benefits of health care interventions, using earnings data as a means of valuing productivity changes only. The approach does still have some application as a means of valuing productivity changes. This is discussed further in Chapter 7 (Section 7.3).

6.2.2 Revealed preference studies

Revealed preference is a common approach in economics for understanding the value that individuals place on goods and services. For example, if an individual pays $1 for an orange, we can infer that they value it at least that amount. However, in the health care field there are very few markets that allow preferences to be revealed by the purchases individuals make.

A number of wage–risk studies have been published, in which the goal is to examine the relationship between particular health risks associated with a hazardous job and wage rates that individuals require to accept the job (Marin and Psacharopoulos 1982). This approach is based on individual preferences regarding the value of increased (decreased) health risk, such as injury at work, as a trade-off against increased (decreased) income, which represents all other goods and services the person might consume. An example of the wage–risk approach is given in Box 6.1.

The strength of the wage–risk approach is that it is based on actual consumer choices involving health versus money, rather than hypothetical scenarios and preference

Box 6.1 Value of a statistical life

Wage–risk example

'Suppose jobs A and B are identical except that workers in job A have higher annual fatal injury risks such that, on average, there is one more job-related death per year for every 10 000 workers in job A than in job B, and workers in job A earn $500 more per year than those in job B. The implied value of statistical life is then $5 million for workers in job B who are each willing to forgo $500 per year for a 1-in-10 000 lower annual risk.' (Fisher et al. 1989)

statements. However, a weakness of the approach is that estimated values have varied widely and estimation seems to be very context and job specific. Using observed data there is always the problem of disentangling the many factors that will confound the relationship between wage and health risk. Furthermore, for use in a specific economic evaluation of a treatment programme it is necessary to observe an occupational choice where the relevant health outcome is the focus of compensation or payment. A more fundamental concern is that the observed risk–money trade-offs may not reflect the kind of rational choice-revealing preferences that economists believe, because of the many imperfections intervening in labour markets and limitations in how individuals perceive occupational risks. It is not possible in this chapter to comprehensively review the volume of work that has been done in this area, but the interested reader should consult Viscusi (1992) for a good review.

It is worth noting at this point that there is another valuation principle that might also be referred to as 'revealed preference' but is not based on individual consumers. This approach is a review of past decisions, such as court awards for injury compensation, to elicit the minimum value that society (or its elected representatives) places on health outcomes (Mooney 1977). In practice, however, many such legal awards are actually based on human capital calculation of discounted earnings streams.

Shifting the focus slightly, one might also be tempted to review previous government health care funding decisions as a source of revealed preference to determine dollar values assigned to health outcomes. But the danger here is one of circularity because we would use previous decisions in future analyses in the belief that some rational process had truthfully revealed societal values for health outcomes in the prior decision.

6.2.3 Stated preference studies—contingent valuation

As the name suggests, contingent valuation studies use survey methods to present respondents with hypothetical scenarios about the programme or problem under evaluation. It is a method for eliciting *stated* preferences. Respondents are required to think about the *contingency* of an actual market existing for a programme or health benefit and to reveal the *maximum* they would be willing to pay for such a programme or benefit. Why are we interested in the maximum WTP? Consider a simple consumer decision to buy a chocolate bar. A measure of how much the consumer values the chocolate bar is the maximum that they would be willing to pay. The difference between this value and the price they have to pay in the market is known as consumer surplus. Of course, for products like chocolate bars one does not need to hire high-priced economists to do a formal assessment; each consumer does this calculation in their own head. However, the logic carries over to contingent valuation studies for non-marketed goods such as a health care programme where we are trying to estimate value in relation to cost for purposes of collective funding. Hence in contingent valuation studies consumers are asked to consider what they would be willing to pay, and thereby sacrifice in terms of other commodities, for the programme benefits if they were in the marketplace.

Here we take health programme benefits to be broadly defined; some may be improvements in health status whereas others may be attributes such as the value of being better informed about one's health or the value associated with the process of care

(Donaldson and Shackley 1997). It is the aggregation of this consumer surplus—which can be large, small, positive, or negative—across individuals that forms the basis of the cost–benefit calculus. In many ways, therefore, economic evaluations based on contingent valuation and statements of WTP can be thought of as attempts to replace missing markets, albeit hypothetically, in an attempt to measure underlying consumer demand and valuation for non-marketed social goods such as health care programmes.

Before reviewing the use of contingent valuation methods in health care it is important to recognize that the need to value health gains and losses for inclusion in economic evaluations has arisen in other public sectors such as transport and environment. Indeed, much of the pioneering work on contingent valuation methods was undertaken in transport studies by economists such as Jones-Lee (1976). An example of a contingent valuation question to estimate a money value for loss of life in the context of road safety is given in Box 6.2. As can be seen, an important advantage of this example is its realism, because it involves an easily understood choice that many people have faced—albeit with a little less precision on the risk of death! Hence in this example the contingency is not difficult to imagine because actual markets for cars do exist where price is related to safety features (e.g. inclusion/exclusion of airbags). Indeed, with this example one can even compare stated preferences with revealed preferences from actual market data.

Box 6.2 Value of a statistical life: road safety contingent valuation example

'Suppose that you are buying a particular make of car. You can, if you want, choose to have a new kind of safety feature fitted to the car at an extra cost. The next few questions will ask about how much extra you would be prepared to pay for some different types of safety feature. You must bear in mind how much you personally can afford.

As we said earlier, the risk of a car driver being killed in an accident is 10 in 100 000. You could choose to have a safety feature fitted to your car which would halve the risk of the car driver being killed, down to 5 in 100 000. Taking into account how much you can personally afford, what is the most that you would be prepared to pay to have this safety feature fitted to the car?' (Jones-Lee et al. 1985)

Hypothetical example

Current risk of death without safety feature = 10 in 100 000
New risk with safety feature = 5 in 100 000
Reduction in risk (dR) = 5 in 100 000
Maximum (for example) premium willing to pay (dV) = £50
Implied value of life = dV/dR = £50/5 × 10^{-5} = £1 m

Table 6.1 Use of willingness-to-pay (WTP) and willingness-to-accept (WTA) questions in the contexts of compensating variation and equivalent variation

Temporal perspective and programme status		Does this consumer gain or lose in utility from before–after change?	Compensating variation (CV)	Equivalent variation (EV)
Before	After		$ +/– required *after* the change to make utility same as before the change	$ +/– required *before* the change to make utility the same as after the change
Project A			A_1	A_3
No programme	Programme	Gain	WTP: maximum amount that must be taken from gainer to maintain at current (before) level of utility	WTA: minimum amount that must be paid to *potential* gainers to forgo the gain and make utility equal to what it would have been after the change
			A_2	A_4
		Loss	WTA: minimum amount that must be paid to loser to maintain at current (before) level of utility	WTP: maximum amount that must be taken from *potential* loser to forgo the loss and make utility level equal to what it would have been after the change
Project B			B_1	B_3
Programme	No programme	Loss	WTA: minimum amount that must be paid to loser to maintain at current (before) level of utility	WTP: maximum amount that must be taken from *potential* loser to forgo the loss and make utility level equal to what it would have been after the change
			B_2	B_4
		Gain	WTP: maximum amount that must be taken from gainer to maintain at current (before) level of utility	WTA: minimum amount that must be paid to *potential* gainers to forgo the gain and make utility equal to what it would have been after the change

From O'Brien, B. and Gafni, A., When do the 'dollars' make sense? Toward a conceptual framework for contingent valuation studies in health care, *Medical Decision Making*, Volume 16, Issue 3, pp. 288–99, Copyright © 1996. Reprinted by permission of SAGE Publications.

In the transport literature, the estimates obtained by contingent valuation have been used in road investment decisions in some jurisdictions and are called the value of a prevented fatality (VPF). There has also been some used of these estimates in the health evaluation literature. For example, Caro et al. (2007) used estimates of VPF from France (2 million euro) and the United Kingdom (2.1 million euro) in a cost–benefit analysis of preventing sudden cardiac deaths with an implantable cardiac defibrillator versus amiodarone. Since there was a positive net monetary benefit when using these VPF estimates, they concluded that ICDs are a worthwhile investment compared with amiodarone in the countries studied.

Another set of conceptual distinctions is shown in Table 6.1. Studies can use either the utility concept of compensating or equivalent variation and can ask questions of WTP or willingness to accept (WTA), depending upon whether a programme is being introduced or removed. For example, in Table 6.1 under the concept of compensating variation and for the introduction of a programme for an individual who gains from this programme, we wish to find out the maximum amount that must be taken from the gainer to maintain them at the current (before-programme) level of utility. This is the maximum they would be willing to pay for the programme to go ahead. In contrast, equivalent variation for the same individual is the minimum amount that must be paid to this *potential* gainer to forgo the gain and to make their utility equal to what it would have been after the change. Hence the equivalent variation is the minimum the individual would be willing to accept in compensation to forgo the programme. A more rigorous derivation and discussion of these concepts can be found in Johansson (1995), and this text also gives a discussion of the circumstances under which these concepts yield the same money values. Reviews of studies conducting money valuations of health programme benefits indicate that the majority of studies use WTP in the context of programme introduction and compensating variation (Diener et al. 1998).

6.2.4 Willingness-to-pay (WTP) studies in health care

In recent years there has been rapid growth in the number of willingness-to-pay (WTP) studies published in the health care literature. Smith (2003) gives a comprehensive list of studies published between 1985 and 2001. For more recent reviews see Donaldson et al. (2012) and McIntosh et al. (2010). It should be noted, however, that most of the published health care WTP studies focus on exploring measurement feasibility issues rather than being full programme evaluations. This cautious embrace of the approach is partly due to some of the inherent difficulties in measuring WTP and partly to some ongoing conceptual debates concerning the ways in which questions should be asked, and of whom. To review and summarize some of these issues, the next section considers some of the theoretical and practical considerations that face the analyst seeking to design a WTP study to value the benefits of a health care programme.

6.3 What might we mean by willingness to pay (WTP)?

It is important to keep in mind that WTP is a measurement technique, and it is how and why this technique is applied that determines its usefulness for economic evaluation.

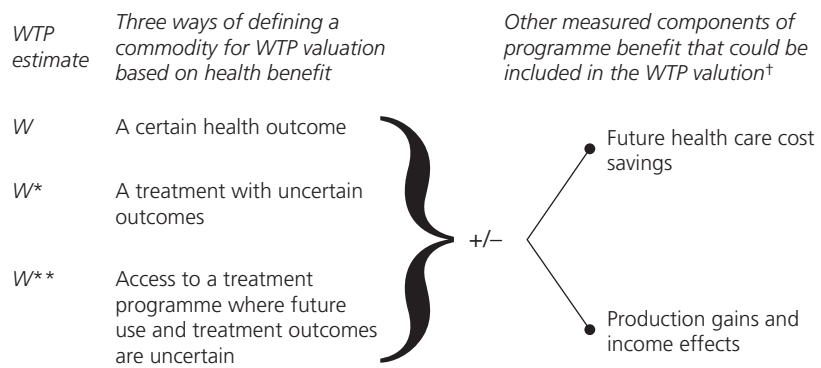
Reviews of WTP studies in health care have revealed wide variation in what questions are being asked, of whom, and how (O'Brien and Gafni 1996). There is disagreement, therefore, concerning how WTP should be measured and how such measures can be incorporated into the evaluation.

In this section we explore different ways in which the concept of WTP can be defined and measured for inclusion in a health care economic evaluation. A simple framework is presented in Figure 6.1, which distinguishes between improvements in health *per se* and other sources of benefit, all of which could, in theory, be valued by WTP. For the health component we also consider the nature of the commodity defined in a willingness-to-pay study with particular emphasis on the role of uncertainty. We describe and illustrate this framework in the following sections.

6.3.1 Global versus restricted WTP

Three broad categories of benefits can arise from a health care programme: (1) intangible benefits, which are the value of improved health *per se* to the individual consumer of a programme; (2) future health care costs avoided; (3) increased productive output due to improved health status. One 'restricted' perspective on WTP is that it would be used only to value those components of benefit for which no money values existed from other market sources. In this approach, WTP estimates are restricted to quantifying the money value of changes in health *per se*, with future health care cost savings and production gains being valued using market prices. (This distinction is discussed in Chapter 3; see Figure 3.1.)

An alternative 'global' perspective on this measurement task is to argue that the purpose of the WTP study is to learn about how the individual consumer would value a specific health care programme in a world where private markets and price signals for all goods and services were operational. However, in this free market scenario, consistency also calls for us to ask the respondent to consider in their valuation the future health care costs that they individually would



† The default money valuation method for these non-health benefits would be to use market prices (for example, wage rates for production). However, in theory, the WTP scenario could be a purely private market for all goods and services, requiring the respondent to state a global WTP based on all consequences of the programme.

Fig. 6.1 What might we mean by willingness to pay?

sustain in the private market world and also work-related income effects as a consequence of ill health or treatment. As an example, consider a decision to buy a more expensive but more effective cold medication over the counter from a pharmacy. A consumer's decision (and WTP) would be driven not only by anticipated health benefits but also, in part, by cost offsets from other medications they may no longer need to purchase if they bought the more expensive cold medications. They might also include the costs associated with work absence in deciding whether or not to buy the more expensive medication. Therefore in this simple private market consumer purchase example, an individual's WTP for a medication is a function not only of the health benefits but also of future out-of-pocket cost savings and income effects from work absence. By analogy, this thought process can be transferred to contingent markets for health care programmes that are covered by insurance or taxation.

As discussed earlier and illustrated in Figure 6.1, the concept of contingent markets is very powerful and can be used to assign money values to all aspects of benefit arising from a health care programme, not simply the value of the health gain itself. While these global and restricted strategies are alternative ways to proceed, great caution needs to be exercised in how respondents are being questioned and whether there is potential for double counting of some programme benefits. For example, when assessing an individual's WTP for a new antihypertensive medication the respondent needs to be told explicitly whether they should be considering income effects due to work absence arising from the disease or its treatment. Double counting would arise if the individual had considered income effects in answering the questions about WTP but the analyst also valued attributable production gains using wage rate data (i.e. a human capital calculation).

6.3.2 What good or service is being valued?

Even if we focus on the restricted form of WTP based on health benefits, as shown in Figure 6.1, there are at least three ways in which a good or service for valuation can be defined: (1) find the WTP for a certain health outcome (W); (2) find the WTP for a treatment with uncertain health outcomes (W^*); (3) find the WTP for access to a treatment programme where future use and treatment outcomes are both uncertain (W^{**}). Consistent with the welfare economics of health care market failure as outlined by Arrow (1963), the main distinction between these three definitions of the good or service being valued is uncertainty. The difference between W and W^* is the inclusion of uncertainty on the supply side with respect to outcomes for a given treatment. In moving to W^{**} we also include uncertainty on the demand side, because individuals are being asked about their WTP for a health care programme given they are uncertain whether they need or will demand this service in the future.

Here we review these three definitions of the goods or service being valued, as illustrated in Figure 6.1.

- *Valuing a certain health outcome (W).* Authors such as Pauly (1995) have suggested that finding the 'shadow price' for a QALY may be a useful bridge between CUA (cost utility analysis) and CBA (see Chapter 4, Section 4.3.4). Some empirical work on the relationship between health status measures and WTP has also

been undertaken (Reed Johnson et al. 1994). Studies that fall in this first use of WTP to value certain health outcomes would include the early work of Thompson (1986), where people with arthritis were asked open-ended questions for the maximum they would be willing to pay to achieve a cure of their arthritis.

- *Valuing a treatment with uncertain outcomes (W*).* As indicated by authors such as Gafni (1991), a limitation of basing estimates of WTP on certain health outcomes is that the consequences of health care programmes are inherently uncertain. Under *W**, therefore, the goal of measurement is to determine the maximum the respondent would be willing to pay to consume a treatment programme with outcomes that are not certainties but have specified probabilities. Although one can multiply certain health values (*h*) by probabilities to devise expected money values for the programme, the values collected directly on uncertain prospects (*W**) will only be the same as the expected values if individuals are risk neutral with respect to income and health.
- *Valuing access to a treatment programme (W**).* In most developed countries we observe that health care services are funded and delivered on the basis of insurance or tax contributions. This reflects an important characteristic of the health care market, which is that illness and the demand for health care is uncertain. The consequence of such insurance or tax arrangements is that persons do not bear the full cost (if any) of the service at the point of delivery. Hence it has been argued by Gafni (1991) that questions about WTP should be framed in a way that incorporates this demand-side uncertainty. Specifically, in Box 6.3 we characterize *W*** as being the maximum an individual would be willing to pay for access to a treatment programme where both future use and treatment outcomes are uncertain. For example, this hypothetical choice might use the payment vehicle of increased insurance premiums or taxation to ensure a programme is made available. A distinction is therefore made between an *ex post* perspective such as *W** where the individual undertaking the valuation knows that they are a consumer of the treatment and that the only uncertainty is on the probability of outcomes, versus an *ex ante* perspective such as *W*** where the individual's valuation needs to incorporate the probability of sustaining the illness and needing the service in question.

In Box 6.3 we illustrate the difference between the *ex post* and *ex ante* perspectives using an example of *in vitro* fertilization (IVF) from Neumann and Johannesson (1994). This was a population-based survey where the authors explored WTP for IVF services using both an *ex post* scenario (assuming infertility, what would you pay out of pocket) and an *ex ante* scenario (where the individuals are asked to assume they have a 10% chance of being infertile and they can buy insurance coverage for IVF). What is notable from Box 6.3 is that the implied value per statistical baby is much higher for the *ex ante* or insurance-based approach ($1.8 m) than the *ex post* or user-based approach ($0.17 m). This is because in the insurance-based setting persons are now also incorporating their risk aversion into the valuation of access to the programme.

As a further illustration of how such *ex ante* insurance-based questions can be asked in practice, Box 6.4, taken from a study by O'Brien et al. (1998), shows how respondents

Box 6.3 WTP for *in vitro* fertilization

This study illustrates two different approaches estimates to forming WTP questions. The *ex post* or user-based approach estimates how much you will pay at the point of consumption. The *ex ante* or insurance-based approach estimates how much you will pay for insurance coverage.

Ex post perspective (user based)

- Assume you are infertile and want children
- IVF has 10% chance of being successful if purchased
- Mean WTP:

  $17 730 (if 10% chance of success)

  $28 054 (if 25% chance of success)

  $43 576 (if 50% chance of success)

Ex ante perspective (insurance based)

- Assume you have 10% chance of being infertile
- IVF has 10% chance of success
- You can buy a one-time insurance premium for IVF coverage
- Mean WTP of $865

Implied WTP per statistical baby

- $177 730 (user based)
- $1.8 m (insurance based)

who were enrolled in a health maintenance organization (HMO) in the United States of America were asked whether they were willing to upgrade their insurance coverage to include a new supportive drug used in cancer chemotherapy known as GCSF. The benefit of this drug is that it reduces the risk of neutropenic fever following chemotherapy; in the example in Box 6.4 the reduction in risk is from 20% to 10% over the six cycles of chemotherapy. A bidding algorithm was used (see below) to find the maximum additional monthly premium persons would pay to have the new drug covered.

6.3.3 Connecting the Ws

The *W*s in Figure 6.1 are clearly connected, and the nature of the relationship depends, *inter alia*, upon the risk preferences of the respondents. For example, one could measure

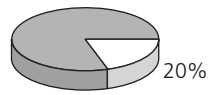
Box 6.4 Example of *ex ante* insurance-based WTP

Option A: your HMO plan covers chemotherapy

- Assume that your chance of getting cancer over the next 5 years is 1 in 100.
- You continue to pay your current monthly insurance premium for health care:

If you get cancer

- You get chemotherapy but not GCSF.
- Over six cycles of chemotherapy your chance of neutropenic fever is:



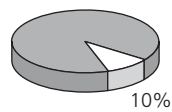
- You cannot buy GCSF or get it covered by another plan.

Option B: your HMO plan covers chemotherapy and GCSF

- Assume that your chance of getting cancer over the next 5 years is 1 in 100.
- You pay a monthly supplement to cover GCSF in addition to your current insurance premium for health care.

If you get cancer

- You get chemotherapy with GCSF.
- Over six cycles of chemotherapy your chance of neutropenic fever is:



- You cannot buy GCSF or get it covered by another plan.

If GCSF was not currently covered by your HMO plan (Option A), would you consider paying an increased premium for coverage of GCSF (Option B)?

money values for certain health outcomes (W) and multiply these by their probabilities of arising to calculate the expected money value of a treatment programme. However, this expected value would only correspond with the measured *ex ante* value W^* to the extent that individuals were risk neutral (with respect to income and health) in their preferences. As described in Chapter 5, we generally observe that individuals are risk averse and therefore the *ex ante* value would be less than the expected value. This risk

preference relationship is also true for W^* in relation to W^{**}, and this has been analysed by Johannesson (1996b).

We have already discussed the relationship between the restricted concept of WTP (W, W^*, or W^{**} in this framework), with its primary focus on the value of health benefits, and the more global concept of WTP where a respondent is required to value all health and non-health benefits in money terms. It is this concept of global or overall WTP that was labelled as W' in Figure 3.1. In practice it is unlikely that many studies will attempt to measure W', but in theory it could be done.

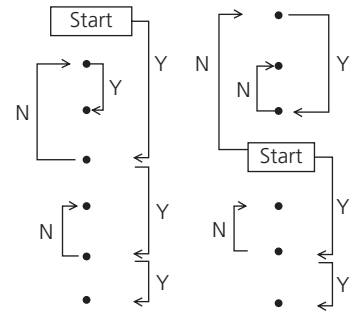
Finally, another important source of private health care market failure, identified by Arrow (1963), is spillovers or externalities. The concept of externality in the consumption of health care is best explained using the example of an infectious disease where one person might be willing to pay for another person to receive treatment so as to reduce the risks of disease transmission to themselves or others. More generally, there might also be humanitarian spillovers in that one person derives utility from the knowledge that others can gain access to needed health care services. The implications of externalities for willingness-to-pay studies is that the sampling frame for inquiry must extend to all persons whose utility is impacted by the introduction of the programme. For example, in assessing WTP for an (elective) new vaccination programme for an infectious disease one would need to draw survey samples from all persons who would benefit from the programme, including the direct benefit to those who vaccinate and the indirect benefit of those who do not vaccinate but are now at lower risk.

6.3.4 A simple example

To illustrate how WTP data might be used in an economic evaluation, consider the decision context of an HMO trying to decide whether to place a new drug on its formulary. Let us suppose a willingness-to-pay survey similar to that described in O'Brien et al. (1998)—the GCSF study described in Box 6.4 and Figure 6.2—has been undertaken on a sample of HMO enrollees. The willingness-to-pay scenario was the maximum additional insurance that respondents would pay to have the drug covered over a 5-year period. The first task would be to forecast the total WTP (for the HMO population) from the sample using multiple regression analysis based on known characteristics of the sample and population. Suppose the (discounted) total WTP over the 5 years is \$10 m. The cost of the programme is a function, in part, of how many people will receive the treatment, over the same time period, and this must also be forecast. Suppose the estimated cost is \$7 m with an estimated \$2 m in cost savings from health care resources not consumed by persons who receive the treatment for a net cost (discounted) of \$5 m. For simplicity, assume that there are no productivity losses or that gains and losses cancel out. Using these data the programme has a positive net benefit of \$5 m (i.e. \$10 m – \$7 m + \$2 m). How this information can be used to inform resource allocation depends on a number of things but particularly whether one is allocating resources within a fixed or non-fixed budget setting.

In a non-fixed budget scenario, the HMO might decide to add the new programme and actually raise insurance premiums, thus increasing its budget. In a competitive market, if these marginal adjustments to coverage, which could be up or down, do not

Bid level	Insurance-based bid scale ($)	Bid algorithm #1	Bid algorithm #2
1	1		
2	5		
3	10		
4	15		
5	25		
6	50		
7	100		



Y = willing to pay this bid; N = not willing to pay this bid.

Persons accepting bid level 7 were then asked an open-ended question for the maximum they were willing to pay.

Fig. 6.2 Bidding algorithms used in the GCSF willingness-to-pay study.

reflect consumers' values then consumers (or their employers, who choose the insurance plan) may elect to switch to other plans.

The second scenario assumes a fixed budget. In this circumstance, knowledge that the new programme has a positive net benefit is of partial value for resource allocation and prioritizing services. To implement the new programme without expanding the budget, efficiency criteria would argue for a rank ordering of existing programmes by the size of their net benefit to facilitate comparison with the new programme. Efficiency would require us to replace programmes with small net benefit by programmes with larger net benefits. The practical difficulty with this scenario is that it is data hungry and works by comparison of net benefit; having data on the benefits of the new programme is a necessary but not sufficient condition for making resource-allocation decisions because we need to know the opportunity cost of its adoption, in terms of the benefits of any existing programme(s) that will be displaced.

6.4 Pragmatic measurement issues in willingness to pay (WTP)

6.4.1 Issues of bias and precision

The goal of the measurement task is to obtain precise and unbiased estimates of WTP. To pose such a question in a way that is both believable and clear to a respondent is not a trivial undertaking and is at least as complex as the health state

preference measurement tasks described in Chapter 5 (probably more so). There are two types of general question format: open-ended and closed. Open-ended questions pose a difficult cognitive task for most respondents because we are typically not used to thinking about the *maximum* we would pay for something. Experience with this approach suggests that although it may produce unbiased estimates of WTP because the respondent is not prompted, it is very imprecise, with widely varying responses and many non-responses or protest responses (Johannesson 1996a); also see Donaldson et al. (1997) on evaluation of open-ended question formats.

Closed question formats have been used in health care WTP studies in two general formats: bidding games to find within-person maximum value and so-called 'take it or leave it' between-person surveys. Bidding games use a predetermined search algorithm to bid the respondent up or down, conditional on how they respond to a prompted monetary value. Much like an auction, if you say 'yes' to $50 we will ask you a higher amount; 'no', and we will ask you a lower amount; for example, see the study by O'Brien and Viramontes (1994). While the bidding game improves upon open-ended questions for the precision of the estimated maximum WTP, it may do so at the expense of introducing a bias in the form of starting point bias. This bias is a form of framing effect where the respondents' answers are influenced by the first numbers presented in the bidding game. Although a number of non-health and health care studies have found evidence of starting point bias (Stalhammer 1996), this result is not conclusive because others have used bidding games and found no evidence of starting point bias, even though it was explicitly tested for (O'Brien and Viramontes 1994; O'Brien et al. 1998).

To illustrate this concept, we show in Figure 6.2 the bidding game used in the same GCSF study mentioned in Box 6.4 from O'Brien et al. (1998). Respondents received one of two bid algorithms and analysis showed that the hypothesis of no starting point bias could not be rejected.

The second type of closed question format is an approach used widely in environmental economic evaluation where surveys of large numbers of persons are typically undertaken to elicit values for some environmental programme or problem. (A controversial environmental example where contingent valuation methods have been used is the valuation of natural resources destroyed by the *Exxon Valdez* oil spill in Prince William Sound, Alaska.) The essence of this approach is that each respondent is only asked one question ('take it or leave it'); for example, 'would you be willing to pay an extra $50 per month on your taxes for this programme—yes or no'? The money amount each person is asked is randomly selected from a range. So, for example, the next person might be asked if they are willing to pay $100, and so on. The data are then analysed using econometric techniques such as probit analysis to identify a bid curve—that is, the quantitative relationship between the proportion of persons accepting or rejecting the bid at different levels of the bid. By mathematically integrating for the area under this bid curve one can determine the mean WTP, or, alternatively, identify the median WTP. For a discussion of this approach in health care see Johannesson (1996a).

The 'take it or leave it' categorical approach has been used in health care WTP studies by Johannesson (1996b) with some success. The difficulties with this approach are in identifying the relevant range from which to sample bids and also in

the large sample size one needs for precise estimation. A variant of this approach, which increases precision, is to ask another (random) bid question of each respondent, but the direction being conditional on the answer to the first question. In the future it is likely that interviews will be computer based so that random bid selection (first or subsequent) is easy to achieve. However, there is still some residual risk of bias because the analyst must choose the *range* from which bids are sampled.

Some studies compare more than one method of eliciting WTP (Frew et al. 2003; 2004; Ryan et al. 2004; Whynes et al. 2003). Ryan et al. (2004) compared the payment card (bidding approach) with the dichotomous choice ('take it or leave it' approach). They found that the dichotomous choice method consistently gave higher estimates of WTP. Whynes et al. (2003) showed that range bias was prevalent in payment-scale willingness-to-pay formats. Although the bidding approach and dichotomous choice approach can be shown to produce different valuations, it is not clear that one produces 'better' valuations than the other.

Another measurement issue, discussed by Donaldson et al. (2012), is that, when valuing two alternatives, A and B, some studies have shown a lack of congruence between simple preferences and the magnitudes of willingness-to-pay responses, most likely because respondents compare the costs of the alternatives on offer and base their WTP on that rather than their strength of preference for each. They suggest that an alternative would be to adopt a marginal approach, whereby the respondent is asked for their maximum WTP to have their preferred option instead of that which is less preferred.

6.4.2 Validation of WTP by tests of scope

Is it possible to validate the findings of a willingness-to-pay study? The 'gold standard' against which we would like to compare predicted WTP from compensating variation (CV) surveys is what consumers would *actually* pay. Unfortunately, for most of the health programme benefits studied by CV methods, an actual market may not exist so *criterion validity* cannot easily be established. However, there are some useful tests of *construct validity* that can be examined in willingness-to-pay studies. The logic of construct validation in this setting is to determine whether the data are consistent with theoretical constructs that should be present if the willingness-to-pay responses are measuring the value we intend.

There are two simple propositions ('constructs') from economic theory that can be tested. First, most goods have what is known as a positive income elasticity: meaning that, other things being equal, higher respondent incomes should be associated with higher WTP. Second, the more of a positively valued good that is supplied by a hypothetical programme, the greater should be a persons' WTP, although the marginal utility of additional units of benefit is likely to decline.

This second principle was strongly endorsed by official guidelines for WTP studies for environmental damage assessment (NOAA 1993). The National Oceanic and Atmospheric Administration (NOAA) panel termed these validation techniques 'scope tests' because the proposition is that WTP should vary with the scope of the benefit (or damage) arising from the hypothetical programme. Scope tests are an important part of willingness-to-pay validation and have been recommended by European guidelines on

health care willingness-to-pay studies where, for example, the magnitude of a treatment effect or other health benefit can be varied in the survey (Johannesson et al. 1996). For examples of scope tests in health care willingness-to-pay studies see Kartman et al. (1996), O'Brien et al. (1998), and Stalhammer and Johannesson (1996). A review of 35 willingness-to-pay studies by Carson (1997) showed that the vast majority (31) were scope sensitive.

6.4.3 Relevance to the health care setting

One of the key difficulties with WTP studies is making the scenario realistic for the respondent. Even if we adopt an *ex ante* insurance-based perspective, most consumers will not be familiar with purchasing access to individual health care programmes. It is likely that these forms of payment scenarios will work better in health care systems, such as that of the United States, where consumers are used to paying more directly for health care, than in the United Kingdom, which has a system of social provision based on taxation contributions. In the environmental economic evaluation literature it has been customary to characterize the decision problem as whether to vote in favour of or against a proposal to have a programme implemented that would have an associated tax contribution. In this kind of format the respondent has their mind focused on the idea of a referendum rather than an actual purchasing decision. In some settings this may be more realistic for the respondent than to consider insurance contributions.

Donaldson et al. (2012) also consider the use of willingness-to-pay assessments in different decision-making situations, such as (1) a clinical decision-maker addressing the question of which type of care to provide for a given group of patients; (2) a health authority seeking to determine the community's strength of preference for which services to provide; and (3) a national decision-maker making a one-off decision about whether or not to fund a specific intervention. In the latter situation they discuss alternative approaches for estimating society's WTP for a QALY, a point we return to in Section 6.7.

6.4.4 Recent examples of willingness-to-pay (WTP) studies

In recent years there has been a rapid growth in the publication of contingent valuation studies of health care treatments and programmes. The recent literature contains studies on such diverse topics as smoking cessation (Heredia-Pi et al. 2012), prostate cancer treatment (Li et al. 2012), childhood asthma (Brandt et al. 2012) and dementia caregiving interventions (Jutkowitz et al. 2010). In addition, an increasing number of willingness-to-pay estimates are being obtained using DCEs (see Section 6.6).

6.5 Exercise: designing a willingness-to-pay (WTP) survey for a new treatment for ovarian cancer

6.5.1 Scenario

The government is trying to decide whether they should reimburse a new therapy for ovarian cancer. Design a willingness-to-pay survey for an economic evaluation of this new treatment for ovarian cancer. Assume that among women who receive this therapy there is a 5% rate of complete cure from the cancer, but the majority will sustain some side effects from the treatment. Data suggest there are productivity gains with more

women in the treated group being able to return to work. There are also cost offsets with treated women receiving fewer health care services in the future.

6.5.2 Specific questions

1 Which components of benefit arising from this treatment programme would you value using WTP? Consider the pros and cons of using a 'global' willingness-to-pay estimate for valuing all programme benefits versus a 'restricted' WTP for health benefits and market prices for other components of benefit.

2 How would you define the commodity that the respondent is being asked to pay for? Consider the alternative formulations of *W, W**, and *W*** discussed in this chapter. What kind of payment vehicle is 'believable' for each of these formulations?

3 The draft study proposal is to interview a sample of women with ovarian cancer. Would you include other subjects in the survey, and why?

6.5.3 Solutions/ideas

1 Using the 'global approach' discussed in this chapter we could frame a willingness-to-pay scenario for the individual for the contingent market for the new therapy, where future attributable costs and employment effects were a personal responsibility and to be met out of pocket. If we used this approach and made it explicit that the respondent should consider these attributes in their valuation then it would not be appropriate to use market prices to value future health care cost savings or wage rates to value productivity effects. To do so would be double counting, because the individual has been asked to consider these in a private market scenario where they are the responsibility of the individual. In practice this global approach may be a difficult cognitive task for the respondent. An easier route may be to use market prices for the future costs and productivity effects but use the willingness-to-pay approach only for the benefit of the health effects, that is, the 'restricted' approach to WTP. If this approach is adopted, however, it is still important to state explicitly to the respondent that they should not consider income effects associated with work absence or future costs associated with the disease in their valuation.

2 The key difference between the three *W*s is the incorporation of uncertainty into the valuation tasks. Perhaps the simplest task would be to estimate money values for the (certain) health states arising in the evaluation. More generally, however, it would be more desirable to include uncertainty with respect to outcomes such that individuals were being asked to value the treatment with probabilities of therapeutic benefit but also probabilities of harm. One of the difficulties here is the extent to which the multiple attributes of outcome and associated probabilities can be presented to respondents in a comprehensible manner. The payment vehicle one might adopt for this type of money valuation could be additional out-of-pocket expense at the point of consumption (e.g. a variable copayment on a medication). A difficulty with this payment vehicle format is that it may not be believable to respondents if the therapy is a major medical procedure that would normally be covered by a health care system at zero cost to the patient. The consequence of framing a question in such a way might be a number of protest responses from respondents.

Formulating the valuation question to estimate W^{**} is more complex yet. Now the respondents need to be presented with information both on the uncertain outcomes of therapy but also on probability of needing this therapy themselves in some future time period. How one frames a payment vehicle to address W^{**} is also complex and will be conditional upon the system of health care financing that the respondent is familiar with. For example, in a predominantly private insurance system it may be most meaningful to ask the respondent to consider additional insurance premiums that they would be willing to pay to gain coverage and access to the treatment programme. In a predominantly tax-financed health care system it may be necessary to frame the question in terms of additional tax contributions (either national or local) that would facilitate the availability of the new treatment programme.

3 While it may be of interest to interview women with ovarian cancer in the estimation of W or W^{*}, the total societal WTP will necessitate a broader sampling. For example, to estimate W^{**} one would need to also interview women who do not currently have ovarian cancer but are at risk of this disease. These individuals may be willing to preserve the option of having this programme available should they need it in the future. More generally, there is also the issue of externality or spillover benefits to other members of the population who are not at current or future risk of ovarian cancer (i.e. men). Men may be willing to pledge additional insurance or tax dollars to cover the ovarian cancer treatment programme either through a self-interested motivation (i.e. wives and daughters at risk) or through a general humanitarian or altruistic motivation where they are expressing a statement of value for women more generally.

6.6 Other stated preference approaches: discrete choice experiments (DCEs)

The survey method of data collection and analysis known as conjoint analysis was developed in mathematical psychology and marketing (Bradley and Terry 1952; Luce 1959). Conjoint analysis is another stated preference approach and is based on the premises that any good or service can be described by its characteristics (or attributes) and the extent to which an individual values a good or service depends on the levels of these characteristics.

There are several potential applications of conjoint analysis, not all of which relate directly to the valuation of health effects in money terms. For example, it can be used as a means to understand patient preferences for health states and as a means to value the various health states described by patient-reported outcomes and HRQoL scales, as discussed in Section 6.1. In addition, conjoint analysis can be used in a drug licensing context to assess patients' willingness to accept the risks associated with more effective treatments and also offers a mechanism for patients to participate in clinical decision-making (Bridges et al. 2011).

Good introductions to the use of ranking and rating conjoint analysis and DCEs are provided by Ryan and Farrar (2000) and Ryan and Gerard (2003). Because of its grounding in random utility theory, economists tend to prefer the methodology of DCEs rather than the ranking and rating methodologies of the conjoint analysis

method (Louviere and Fiebig 2010). They point out that the technique can be used to show individuals are willing to trade between the characteristics of treatments or services, to estimate the relative importance of different characteristics, to estimate whether a characteristic or attribute is important, and to predict the demand for a given good or service with given characteristics. Also, because of the underlying theoretical basis of DCEs, Carson and Louviere (2011) argue that they are distinct from the broader range of studies going under the more general name 'conjoint analysis'. The particular relevance in the context of economic evaluation is that the payment vehicle of cost (out of pocket) can be included as one of the attributes. Then, the ratio (or marginal rate of substitution) of any given attribute to the absolute parameter on the cost attribute shows how much money the individual is willing to pay for a unit change in that attribute.

There are several key steps in undertaking a DCE. First, the characteristics of the treatment or service must be identified. These may often be predefined, but can also be obtained from literature reviews or focus group discussions with health professionals or patients. In the case of a treatment, the characteristics are likely to relate to the main dimensions of efficacy and the major adverse events. A review of the literature (Ryan and Gerard 2003) suggested that a set of four to six attributes was acceptable, in relation to the respondents' ability to complete the choice task. However, there are no hard-and-fast rules on this. Bridges et al. (2011) argue that all attributes that potentially characterize the alternatives should be considered, although some may be excluded to ensure that the profiles are plausible to subjects.

Second, levels need to be assigned to the characteristics. These may be cardinal (e.g. cure after 1 day being twice as good as cure after 2 days, ordinal (e.g. severe pain being worse than moderate pain), or categorical, where there is no natural ordering (e.g. one adverse event versus another).

Third, scenarios need to be drawn up that describe all possible configurations of the characteristics and levels chosen. Clearly, the major issue here is that the potential number of scenarios is dependent on the number of characteristics and levels defined. Because it is usually impossible to include all the scenarios in any given survey (full factorial design), experimental designs are used to reduce the number to a manageable level. For a summary of these design criteria see Huber and Zwerina (1996) and Louviere et al. (2000). Also, Carlsson and Martinsson (2003) provide a concise review of design techniques for stated preference discrete choice methods in health economics including random designs and optimal design strategies.

Fourth, preferences for the scenarios need to be elicited using discrete choices. Respondents are presented with a number of choices and, for each, asked to choose their preferred one. Possible responses include stating that either A or B is preferred (where one option may also be a fixed 'status quo'), that A or B is preferred on a graded scale, that indifference is preferred, or that the 'non-participation or neither' option is preferred. These different elicitation formats each have advantages and disadvantages. However, it is important to identify the appropriate elicitation format for the question being addressed. For example, where a non-participation option (inclusion of a 'prefer neither' option) is incorrectly excluded, this may give rise to overestimation of any values obtained (Morey et al. 1993). (An example of a graded elicitation format with an opt-out option is shown in Table 6.2.)

One scaling approach worthy of particular mention is *best–worst scaling*. This approach was developed by Louviere and Woodworth (1990) and was first applied in the

Table 6.2 Options in the location and waiting time for orthodontic services

	(a) Current				(b) Alternative			Which option would you choose? (please tick one box for each choice)				
	First appointment	Second appointment	Waiting time (months)		First appointment	Second appointment	Waiting time (months)	Definitely (a) current	Probably (a) current	No preference	Probably (b) alternative	Definitely (b) alternative
Choice1	Hospital	Hospital	8	or	Local	Local	12					
Choice2	Hospital	Hospital	8	or	Hospital	Hospital	16					
Choice3	Hospital	Hospital	8	or	Local	Local	16					

case of food safety (Finn and Louviere 1992). A guide to its use in health care research was published by Flynn et al. (2007) and an important application in health was in valuing the ICECAP index for older people (Coast et al. 2008).

In best–worst scaling, instead of being asked to make a choice between two profiles of attributes as in a standard DCE, individuals choose the best and worst attribute displayed in the particular profile specified. It has been argued that this is important in situations where respondents do not have experience of making choices in the area of application and is simpler than keeping two or more profiles in mind at once (Potoglou et al. 2011). Also, with best–worst scaling it is possible to present a large number of attributes in a single profile, without being overly concerned about complicating the task.

Sometimes best–worst scaling is used in conjunction with standard DCEs in order to gain more information about respondents' preferences. However, an important question is whether the preference weights obtained by best–worst scaling differ from those obtained from those obtained from standard DCEs. An empirical comparison was made as part of the Outcomes of Social Care for Adults project. The findings showed that the preference weights from the two methods gave similar patterns in preferences and in the majority of cases the preference weights were not statistically different (Potoglou et al. 2011). However, the comparison of the advantages and disadvantages of the two methods will remain a topic for further research.

Finally, once collected, the data need to be analysed. Usually this is done using econometric techniques. For example, for a binary response format, the utility function is specified in additive form: $\Delta U = \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \ldots + \beta_n X_n$, where ΔU is the change in utility in moving from treatment A to B, X_j ($j = 1, 2, \ldots, n$) are the differences in attribute levels A and B, and β_j ($j = 1, 2, \ldots, n$) are the coefficients of the model to be estimated. The method of analysis depends on the elicitation format; for instance, when a non-participation option is included this may require analysis using a nested logit model. This is one area where several developments have taken place in recent years. Hauber et al. (2015 in press) describe the range of options available to researchers to analyse data generated from studies using conjoint analysis (including ranking, rating, and discrete-choice elicitation formats), provide researchers with an understanding of the implicit and explicit assumptions required to apply different analysis methods to data generated using different elicitation formats, and make recommendations for day-to-day use.

One feature of DCEs is that the range of attributes included can extend beyond the dimensions of length and quality of life that normally characterize the measures of health gain discussed in Chapter 5. The range can include aspects of the configuration of health services. For example, in an early study Ryan and Farrar (2000) explored individuals' preferences for various configurations of hospital location and waiting time in the provision of orthodontic services. The policy question determined that treatment location (i.e. local clinic or hospital) and waiting time were the main attributes individuals would be concerned about. In total 16 scenarios were possible, considering first and second appointments and four levels of waiting time. Fifteen discrete choices were constructed by comparing the current service to all alternatives. An example of one of these choices is given in Table 6.2.

A random effects ordered probit model was fitted thus:

$\Delta B = B_1 \text{LOC}_1 + B_2 \text{LOC}_2 + B_3 \text{WAIT}.$

Table 6.3 Model estimates

Variable	Coefficient	*P* value
LOC_1	$-0.77(B_1)$	<0.001
LOC_2	$-0.91(B_2)$	<0.001
WAIT	$-0.59(B_3)$	<0.001

Out of 160 individuals, 73 gave consistent responses. The results of the estimation are shown in Table 6.3. The interpretation is that the benefit of the service is significantly associated with lower waiting times and first and second appointments in a local clinic. Individuals are willing to wait an extra 1.3 months ($B_1/B_3 = 0.77/0.59$) to have their first appointment in a local clinic (Ryan and Farrar 2000).

DCEs are becoming increasingly popular in the health care field. Other early examples include examination of individuals' preferences for service provision (e.g. out-of-hours care provided by general practitioners) (Scott et al. 2003) and for treatment characteristics (e.g. therapies for osteoarthritis and prostate cancer) (Ratcliffe et al. 2004; Sculpher et al. 2004). (See Ryan and Gerard 2003 for a review of the literature.) This approach has many of the advantages of contingent valuation, in that it enables non-health characteristics, or attributes related to process utility, to be included. One major additional advantage is that, although WTP tells us about the valuation of the whole 'bundle' of characteristics, DCEs help us understand the relative valuations of, or the trade-offs between, various attributes. This would be of particular relevance to a health planner considering options for provision of a particular service, or a clinical researcher wanting to understand the relative valuation of the outcomes, side effects, or other characteristics of a particular therapy (e.g. how much is an oral formulation of a drug valued relative to administration by injection?).

Several situations exist where patients face trade-offs between the risks and benefits of alternative therapies. Sculpher et al. (2004) explored men's trade-offs in the field of non-metastatic prostate cancer, where different treatments, while increasing life expectancy, have various side effects including diarrhoea, hot flushes, breast swelling or tenderness, and loss of physical energy and sex drive. Also, in settings where patients have to pay for medication, or incur other expenses in obtaining care, there could also be impacts on out-of-pocket expenses.

Some results from their DCE are shown in Table 6.4. The coefficients for the attributes were all statistically significant from zero. Negative values indicate that the more severe the problem, the less likely the patient is to prefer that scenario. (The negative value for out-of-pocket expenses indicates that the higher the costs, the less likely the patient is to prefer that scenario.) The positive value for life expectancy indicates that the greater the life expectancy, the more likely the patient is to prefer that scenario.

Table 6.5 shows the marginal rates of substitution between life expectancy and the other attributes—that is, how much life expectancy the men were willing to trade off to

Table 6.4 Results of second part of discrete choice exercise

Variable	Coefficient (95% confidence interval)	Standard error	*P* value
Diarrhoea	–0.4193	0.0644	<0.001
	(–0.5454 to –0.2931)		
Hot flushes	–0.1225	0.0479	0.010
	(–0.2162 to –0.0287)		
Breast tenderness	–0.4329	0.0927	<0.001
	(–0.6147 to –0.2512)		
Out-of-pocket expenses	–0.0016	0.0004	0.001
	(–0.0025 to –0.0007)		
Life expectancy	0.2329	0.0256	<0.001
	(0.1827 to 0.2832)		
Constant	0.1278	0.0618	0.014
	(0.0262 to 0.2294)		
Number of observations	992; 164.35; $P < 0.0001$[a]		

[a] χ^2 test.

Source: data from *BMJ*, Sculpher, M.J. et al., Patients' preferences for the management of non-metastatic prostate cancer: discrete-choice experiment, Volume 328, pp. 382, Copyright © 2004, British Medical Journal Publishing Group.

achieve an improvement by one level in one of the other attributes. For example, men are willing to trade off 1.8 months of life expectancy to change diarrhoea from a moderate to mild level, or from mild to absent.

As with willingness-to-pay estimations, DCEs raise a number of methodological issues. (See Louviere et al. 2000 for a full discussion.) Some of these have already been discussed in the literature, including internal validity and consistency (McIntosh and Ryan 2002; Ryan et al. 1998) and test–retest reliability (Bryan et al. 2000). However, one issue of particular importance to economic evaluation deserves special mention, namely, the inclusion of a cost attribute in order to estimate WTP. Ratcliffe (2000) argues that this should be viewed with caution, as the level at which the cost attribute is set can influence the willingness-to-pay estimates for the levels of other attributes and hence the total WTP value inferred for that individual to receive their chosen intervention. Also, in an empirical study using a large dataset, Skjoldborg and Gyrd-Hansen (2003) found that the cost range applied in DCEs, and the inclusion of a dummy variable to represent the utility associated with payment *per se*, could affect the willingness-to-pay values. Mitchell and Carson (1989) and Morey et al. (1993) outline the importance of modelling the participation decision when estimating welfare estimates using DCEs. Morey et al. (1993) note that DCE willingness-to-pay values are sensitive to the valuation model employed and that modelling the participation decision is crucial in obtaining accurate estimates. Finally, weighting the values by

Table 6.5 Patients' marginal rates of substitution between life expectancy and other attributes

Attribute	Life expectancy willing to forgo (months)	Single-level improvement
Diarrhoea	1.8	From moderate to mild or from mild to absent
Hot flushes	0.5	From moderate to mild or from mild to absent
Breast swelling	1.9	From present to absent
Loss of libido	1.3	From present to absent
Problems in maintaining an erection		
Aged <70 years	1.8	From moderate to mild or from mild to absent
Aged 70 years	0.9	From moderate to mild or from mild to absent
Lack of energy or 'pep'	3.0	From present to absent

Source: data from *BMJ*, Sculpher, M.J. et al., Patients' preferences for the management of non-metastatic prostate cancer: discrete-choice experiment, Volume 328, pp. 382, Copyright © 2004, British Medical Journal Publishing Group.

the probability of choosing each alternative should also be carried out using a multiple alternative DCE study (Bennett and Blamey 2001; Lancsar and Savage 2004).

A review of DCEs in health care was conducted by de Bekker-Grob et al. (2012). They identified 114 DCEs published between 2001 and 2008, covering a wide range of policy questions. As compared with a baseline of studies conducted between 1990 and 2000, applications took place in a broader range of health care systems, and there has been a move to incorporating fewer attributes, more choices, and interview-based surveys. There has also been a shift towards statistically more efficient designs and flexible econometric models. The reporting of monetary values continues to be popular, the use of utility scores has not gained popularity, and there has been an increasing use of odds ratios and probabilities. The latter are likely to be useful at the policy level to investigate take-up and acceptability of new interventions. Incorporation of interaction terms in the design and analysis of DCEs, explanations of risk, tests of external validity, and incorporation of DCE results into a decision-making framework remain important areas for future research.

A later systematic review, covering the period 2009–2012, found a total of 179 health-related DCEs (Clark et al. 2014). There was a continuing trend towards conducting DCEs across a broader range of countries. However, the trend towards including fewer attributes was reversed, while the trend towards interview-based DCEs reversed because of increased computer administration. The trend towards using more flexible econometric models, including mixed logit and latent class, had continued. Reporting of monetary values had fallen compared with earlier periods, but the proportion of

studies estimating trade-offs between health outcomes and experience factors, or valuing outcomes in terms of utility scores, had increased, although use of odds ratios and probabilities had declined.

Clark et al. (2014) argue that the reassuring trend towards the use of more flexible and appropriate DCE designs and econometric methods has been reinforced by the increased use of qualitative methods to inform DCE processes and results. However, qualitative research methods are being used less often to inform attribute selection, which may make DCEs more susceptible to omitted variable bias if the decision framework is not known before the research is conducted.

The literature on DCEs of health care treatments and programmes continues to grow. The recent literature contains studies on such diverse topics as prophylactic granulocyte colony-stimulating factors in breast cancer (Johnson et al. 2014), symptom relief in gastroesophageal reflux disease (Deal et al. 2013), the effect of providing information about invasive follow-up testing in colorectal cancer screening (Benning et al. 2014), treatment for low back pain (Kløjgaard et al. 2014), hospital-at-home (Goosens et al. 2014), and genetic test information for treatable conditions (Kilambi et al. 2014).

As more studies have been carried out, methodological principles have been developed. Lanscar and Louviere (2008) provide a resource for current practitioners as well as those considering undertaking a DCE, using DCE results in a policy/commercial context, or reviewing a DCE. To aid in undertaking and assessing the quality of DCEs, they discuss the process of carrying out a choice study and have developed a checklist covering conceptualizing the choice process, selecting attributes and levels, experimental design, questionnaire design, pilot testing, sampling and sample size, data collection, coding of data, econometric analysis, validity, interpretation, derivation of welfare measures, and policy analysis.

The key methodological steps in conducting conjoint analyses, especially DCEs, have also recently been discussed by a task force convened by the International Society for Pharmacoeconomics and Outcomes Research (Bridges et al. 2011). This is shown in Box 6.5 and is organized under ten general headings: research question, selection of attributes and levels, construction of tasks, experimental design, preference elicitation, instrument design, data collection, statistical analyses, results and conclusions, and study presentation. The task force concludes that researchers conducting these studies in health care should always be clear about the approaches they are using and why these approaches are appropriate to a particular study.

6.7 **Valuation of health effects for health policy decisions**

Most of the research described in this chapter has focused solely on measuring the benefits of health care interventions and the measurements without formal integration into economic evaluations. One exception is the study by Haefeli et al. (2008), who undertook a feasibility study of using contingent valuation techniques to value benefits in a cost–benefit analysis (CBA) of spinal surgery. They used a contingent valuation survey with *ex post* WTP/WTA questions. Although it was possible to obtain estimates

Box 6.5 A checklist for conjoint analysis applications in health care

1 Was a well-defined research question stated and is conjoint analysis an appropriate method for answering it?

- 1.1 Were a well-defined research question and a testable hypothesis articulated?
- 1.2 Was the study perspective described, and was the study placed in a particular decision-making or policy context?
- 1.3 What is the rationale for using conjoint analysis to answer the research question?

2 Was the choice of attributes and levels supported by evidence?

- 2.1 Was attribute identification supported by evidence (literature reviews, focus groups, or other scientific methods)?
- 2.2 Was attribute selection justified and consistent with theory?
- 2.3 Was level selection for each attribute justified by the evidence and consistent with the study perspective and hypothesis?

3 Was the construction of tasks appropriate?

- 3.1 Was the number of attributes in each conjoint task justified (i.e. full or partial profile)?
- 3.2 Was the number of profiles in each conjoint task justified?
- 3.3 Was (should) an opt-out or a status-quo alternative (be) included?

4 Was the choice of experimental design justified and evaluated?

- 4.1 Was the choice of experimental design justified? Were alternative experimental designs considered?
- 4.2 Were the properties of the experimental design evaluated?
- 4.3 Was the number of conjoint tasks included in the data collection instrument appropriate?

5 Were preferences elicited appropriately, given the research question?

- 5.1 Was there sufficient motivation and explanation of conjoint tasks?
- 5.2 Was an appropriate elicitation format (i.e. rating, ranking, or choice) used? Did (should) the elicitation format allow for indifference?
- 5.3 In addition to preference elicitation, did the conjoint tasks include other qualifying questions (e.g. strength of preference, confidence in response, and other methods)?

6 Was the data collection instrument designed appropriately?

- 6.1 Was appropriate respondent information collected (e.g. sociodemographic, attitudinal, health history or status, and treatment experience)?

Box 6.5 A checklist for conjoint analysis applications in health care (continued)

6.2 Were the attributes and levels defined, and was any contextual information provided?

6.3 Was the level of burden of the data collection instrument appropriate? Were respondents encouraged and motivated?

7 Was the data collection plan appropriate?

7.1 Was the sampling strategy justified (e.g. sample size, stratification, and recruitment)?

7.2 Was the mode of administration justified and appropriate (e.g. face-to-face, pen-and-paper, web-based)?

7.3 Were ethical considerations addressed (e.g. recruitment, information and/or consent, compensation)?

8 Were statistical analyses and model estimations appropriate?

8.1 Were respondent characteristics examined and tested?

8.2 Was the quality of the responses examined (e.g. rationality, validity, reliability)?

8.3 Was model estimation conducted appropriately? Were issues of clustering and subgroups handled appropriately?

9 Were the results and conclusions valid?

9.1 Did study results reflect testable hypotheses and account for statistical uncertainty?

9.2 Were study conclusions supported by the evidence and compared with existing findings in the literature?

9.3 Were study limitations and generalizability adequately discussed?

10 Was the study presentation clear, concise, and complete?

10.1 Was study importance and research context adequately motivated?

10.2 Were the study data collection instrument and methods described?

10.3 Were the study implications clearly stated and understandable to a wide audience?

which suggested that surgery was cost-beneficial within a CBA framework, they recommend further studies to improve the reliability of the net-benefit estimates.

In another study, Brisson and Edmunds (2006) compared the economic desirability of varicella vaccination using CUA with that using WTP in a CBA. They found that while the cost-effectiveness of vaccination in the CUA was uncertain, it was highly beneficial in the CBA using WTP.

6.7.1 Potential uses of willingness-to-pay (WTP) studies and DCEs in health care decision-making

Donaldson et al. (2012) discuss the potential uses of willingness to pay in several decision-making scenarios. First, it might be useful in helping clinicians understand patients' preferences for one treatment over another. They recommend that this is best assessed by asking respondents about their WTP for each treatment, or by asking them for their additional WTP for their preferred treatment over the alternative. Of course, within the context of a comparative clinical study, these preferences could also be explored by asking patients about the value they place on the health state they are experiencing. However, depending on the health state preference measurement being used, willingness to pay may be more sensitive to differences in preferences. In this context, the tacit assumption is that there is no net budgetary impact of the treatment choice. For example, it could relate to a choice between two similarly priced drugs with slightly different profiles of side effects. If there is a net budgetary impact this would need to be considered alongside the net WTP, as discussed later in this section.

A second decision-making context where a WTP study could be useful is where a health authority is discussing priorities for services. In such a situation it would be useful to know if members of the public can compare disparate alternatives and express their preference for each in terms of WTP. This could include attributes that go beyond health (e.g. preferences for equity in access). However, unless the cost of the various programmes were identical there would be an opportunity cost if they were to be purchased from the same fixed budget. Olsen and Donaldson (1998) attempted to address this by asking respondents to consider three programmes that were competing for additional funding. They were then asked to give their WTP for each option in extra taxation for each programme. Donaldson et al. (2012) point out that this study raised several issues. First, the rankings implied by WTP did not often match the rankings given by respondents. Also, there could be ordering effects, in that if respondents reached some kind of budget constraint (despite being told that the options were competing with one another), their valuation for the last option valued could be deflated. The 1998 research led to further studies, to investigate whether the ordering of the presentation of the various options was important (Stewart et al. 2002).

A third situation where contingent valuation might be useful is in the context of a national body (e.g. a health technology assessment agency) making decisions about whether or not to fund particular interventions. As mentioned in Chapters 2, 3 and 4, these decisions are made with reference to a cost-effectiveness threshold, reflecting the opportunity cost in forgone health from the programmes that are displaced, given a fixed budget. It may be interesting to conduct surveys of the general public to elicit their maximum WTP for a unit of health gain (e.g. a QALY), so as to compare this with the value implied by the current budget constraint. (In Chapter 4 we called the former v and the latter k.) In principle this may provide some indication of whether the budget itself is set appropriately, given public preferences. However, as pointed out in Section 4.3.4, v represents how much an individual would be willing to give up of their own consumption to improve their own health, whereas k represents how much of collectively pooled resources (i.e. from a national health budget or private insurance plan)

is currently required to improve health. Economic evaluation cannot prescribe that k should equal v, but can inform that debate.

Mason et al. (2008) have conducted a review of studies of WTP for a QALY and found (at the time of the review) that the values estimated were very disparate and could not be recommended for use by policy makers. They conclude that 'We see the work reported here as being part of a process of reconciling values from surveys of the general public with those that may emerge from health service decision-making as each approach improves over time'. However, if decisions are being made that imply a cost to a constrained health care budget, the decision-maker must always consider k and knowing something about v does not inform those decisions.

One possibility is that the WTP for a QALY varies by context (e.g. disease type, age of respondent, and whether the health gains were primarily in length of life or quality of life). Several research studies conducted in the United Kingdom have begun to explore these issues (Brazier et al. 2014; Linley and Hughes 2013), but within the context of determining whether QALYs should be weighted other than equally, not to determine monetary values for a QALY in different health care contexts. A recent review by Ryen and Svensson (2014) of 24 studies containing 383 unique estimates of WTP for a QALY also found a wide range, with trimmed mean and median estimates of 74 159 and 24 226 euros repectively (2010 price level). Some of the differences related to study methods, but a regression analysis indicated that the WTP for a QALY is significantly higher if the QALY gain comes from life extension rather than quality-of-life improvements.

Turning to DCEs, Ryan et al. (2012) have discussed their use in policy analysis or as decision support tools. An obvious application is to gain a better understanding of patients' experiences and their trade-offs between different health outcomes and treatment attributes. This could be helpful in the design of future services, or in making choices between existing services, although, as in the context of contingent valuation, it would be important to recognize any opportunity costs arising from such decisions.

Ryan et al. argue that the greater use of DCEs within an economic evaluation framework would be critically dependent on decision-makers recognizing that elements of value beyond those typically included in the usual measures of health gain are relevant in health policy making. Louviere and Lancsar (2009) suggest that DCEs may be particularly important in the case of population health policies, or where there is a need to predict choice probabilities, which are useful for analysing the likely uptake of services. They argue that DCEs are complementary to other forms of preference elicitation and valuation in the health economist's toolkit and can be used in conjunction with other stated preference and evaluation methods.

6.7.2 Challenges in incorporating willingness to pay (WTP) and DCEs into economic evaluation

It was mentioned in Section 6.7.1 that the main challenge in incorporating WTP studies and DCEs into economic evaluation is to identify the opportunity costs of adopting the new programme. If a health care decision-maker operating with a fixed

budget only considers health gain and all the costs of adopting the new programme fall on the health care budget, the threshold (k) would provide an estimate of the health forgone through the displacement of other programmes. If some of the costs fall on other budgets (e.g. if the new programme were for elderly and mentally ill patients and consumed resources in social care), then consideration would have to be given to the opportunity costs in the social care sector, although these may not be a direct concern to the health care decision-maker.

If a health care decision-maker were convinced that other aspects of patient welfare, such as increased convenience or ease of access to services, were relevant, he or she would need to consider both the losses in health and of these additional benefits from any displaced programmes. Although possible in principle, this might be difficult in practice because there is currently no standardization in the range of attributes assessed in WTP studies or DCEs. In addition, the decision-maker would need to be sure that there was no double counting in considering these other process benefits alongside the health gains, due to the improvements in process also generating health gains though improved compliance (Brennan and Dixon 2013).

Finally, if a health care decision-maker took the view that these additional benefits (beyond health gain) were important, but that they should not be a charge to the health budget, one option would be to increase patient copayments. This implies that some of the additional cost would be in forgone consumption. In this case the estimate of WTP (v) may give an indication of individual (patient) or societal preferences and would be helpful in making that decision. However, the health care decision-maker would still need to consider k, as this would be an indication of the cost in forgone health in any displaced programmes. Also, as pointed out in Section 4.3.3, increasing patient copayments would have implications on health for individuals unable to afford the higher copayments.

Therefore, going beyond the consideration of health gain in health care decision-making raises several complexities, so analysts might want to consider carefully when this should be considered. A true welfarist would argue that it should be done as a matter of principle. Other analysts might make the decision based on the nature of the programmes being evaluated. For example, it has been argued that there might be a strong case for a broader consideration of benefits in the evaluation of complex public health programmes, since these have a broad range of costs and benefits falling on a wide range of public budgets and personal consumption (Weatherly et al. 2014).

6.8 Further reading

The literature in WTP and DCEs in health care is expanding rapidly and it is impossible to do it justice in a general textbook on economic evaluation. For those wishing to learn more, a good starting point would be the handbook by McIntosh et al. (2010). In addition, readers should consult the overviews by Donaldson et al. (2012) on contingent valuation and by Ryan et al. (2012) on DCEs. There are also several systematic reviews, by De Bekker-Grob et al. (2012) and Clark et al. (2014) on DCEs, Brennan and Dixon (2013) and Higgins et al. (2014) on estimating the value of process utility in assessing health care programmes, and Lin et al. (2013) on the value of

diagnostic technologies. Finally, given the rapidly changing nature of the literature, it is important to monitor journals that publish a substantial number of empirical studies, including *PharmacoEconomics* and *Value in Health*.

References

Arrow, K. (1963). Uncertainty and the welfare economics of medical care. *American Economic Review*, **53**, 941–73.

Bennett, J. and Blamey, R. (2001). *The choice modelling approach to environmental valuation*. Cheltenham: Edward Elgar.

Benning, T.M., Dellaert, B.G.C., Severens, J.L., and Dirksen, C.D. (2014). The effect of presenting information about invasive follow-up testing on individuals' noninvasive colorectal cancer screening participation decision: results from a discrete choice experiment. *Value in Health*, **17**, 578–87.

Bradley, R.A. and Terry, M.E. (1952). Rank analysis of incomplete block designs, I. the method of paired comparisons. *Biometrika*, **39**, 324–45.

Brandt, S., Lavin, F.V., and Hanemann, M. (2012). Contiguent valuation scenarios for chronic illnesses: the case of childhood asthma. *Value in Health*, **15**, 1077–83.

Brazier, J.E., Rowen, D., Mukuria, C., Whyte, S., Keetharuth, A., et al. (2013). Eliciting social preferences for burden of illness, therapeutic improvement and end-of-life for value-based pricing: a report of the main survey. *EEPRU Reseach Report 01/13*. Universities of Sheffield and York. Sheffield, October 2013.

Brennan, V.K. and Dixon, S. (2013). Incorporating process utility into quality adjusted life years: a systematic review of empirical studies. *PharmacoEconomics*, **31**, 677–91.

Bridges, J.F.P., Hauber, A.B., Marshal, D., et al. (2011). Conjoint analysis applications in health—a checklist: a report of the ISPOR Good Research Practices for Conjoint Analysis Task Force. *Value in Health*, **14**, 403–13.

Brisson, M. and Edmunds, W.J. (2006). Impact of model, methodological and parameter uncertainty in the economic analysis of vaccination programs. *Medical Decision Making*, **26**, 434–46.

Bryan, S., Gold, L., Sheldon, R., and Buxton, M. (2000). Preference measurement using conjoint methods: an empirical investigation of reliability. *Health Economics*, **9**, 385–95.

Carlsson, P. and Martinsson, P. (2003). Design techniques for stated preference methods in health economics. *Health Economics*, **12**, 281–94.

Caro, J.J., Ward, A.W., Deniz, H.B., O'Brien, J.A., and Ehreth, J. (2007). Cost-benefit analysis of preventing sudden cardiac death with an implantable cardioverter defibrillator versus amiodarone. *Value in Health*, **10**, 13–22.

Carson, R.T. (1997). Contingent valuation and tests of insensitivity to scope, in R.J. Kopp, W. Pommerhene, and N. Schwartz (ed.), *Determining the value of non-marketed goods: economic, psychological, and policy relevant aspects of contingent valuation methods*. Boston: Kluwer.

Carson, R.T. and Louviere, J.J. (2011). A common nomenclature for stated preference elicitation approaches. *Environmental and Resource Economics*, **49**, 539–559.

Clark, M.D., Determann, D., Petrou, S., Moro, D., and de Bekker-Grob, E.W. (2014). Discrete choice experiments in health economics: a review of the literature. *PharmacoEconomics*, **32**, 883–902.

Coast, J., Flynn, T.N., Natarajan, L., et al. (2008).Valuing the ICECAP capability index for older people. *Social Science and Medicine*, **67**, 874–82.

Deal, K., Marshall, D., Dabrowski, D., et al. (2013). Assessing the value of symptom relief for patients with gastroesophageal reflux disease treatment: willingness to pay using a discrete choice experiment. *Value in Health*, **16**, 588–98.

De Bekker-Grob, E.W., Ryan, M., and Gerard, K. (2012). Discrete choice experiments in health economics: a review of the literature. *Health Economics*, **21**, 145–72.

Devlin, N.J. and Krabbe, P.F.M. (2013). The development of new research methods for the valuation of EQ-5D-5L. *European Journal of Health Economics*, **14**(Suppl 1), S1–S3.

Diener, A., O'Brien, B., and Gafni, A. (1998). Health care contingent valuation studies: a review and classification of the literature. *Health Economics*, **7**, 313–26.

Donaldson, C. and Shackley, P. (1997). Does 'process utility' exist? A case study of willingness to pay for laparoscopic cholecystectomy. *Social Science and Medicine*, **44**, 699–707.

Donaldson, C., Mason, H., and Shackley, P. (2012). Contingent valuation in health care, in A. Jones (ed.), *The Elgar companion to health economics*, 2nd edition, pp. 425–34. Cheltenham: Edward Elgar.

Donaldson, C., Thomas, R., and Torgerson, D.J. (1997). Validity of open-ended and payment scale approaches to eliciting willingness to pay. *Applied Economics*, **29**, 79–84.

Finn, A. and Louviere, J.J. (1992). Determining the appropriate response to evidence of public concern: the case of food safety. *Journal of Public Policy and Marketing*, **11**, 12–25.

Fisher, A., Chestnut, L.G., and Violette, D.M. (1989). The value of reducing risks of death: A note on new evidence. *Journal of Policy and Management*, **8**, 88.

Flynn, T.N., Louviere, J.J., Peters, T.J., and Coast, J. (2007). Best–worst scaling: what it can do for health care research and how to do it. *Journal of Health Economics*, **26**, 171–89.

Frew, E.J., Whynes, D.K., and Wolstenholme, J.L. (2003). Eliciting willingness to pay: comparing closed-ended with open-ended and payment scale formats. *Medical Decision Making*, **23**, 150–9.

Frew, E.J., Wolstenholme, J.L., and Whynes, D.K. (2004). Comparing willingness-to-pay: bidding game format versus open-ended and payment scale formats. *Health Policy*, **68**, 289–98.

Gafni, A. (1991). Using willingness-to-pay as a measure of benefits: What is the relevant question to ask in the context of public decision-making? *Medical Care*, **29**, 1246–52.

Goosens, L.M.A., Utens, C.M.A., Smeenk, W.J.M., et al. (2014). Should I stay or should I go home? A latent class analysis of a discrete choice experiment on hospital-at-home. *Value in Health*, **17**, 588–96.

Haefeli, M., Elfering, A., McIntosh, E., Gray, A., Sukthankar, A., and Boos, N. (2008). A cost-benefit analysis using contingent valuation techniques: a feasibility study in spinal surgery. *Value in Health*, **11**, 575–88.

Hauber, A.B., et al. (2015, in press). Conjoint analysis statistical analysis: An ISPOR Conjoint Analysis Good Research Practices Task Force report. <http://www.ispor.org/TaskForces/Analyzing_Data_Conjoint_Analysis.asp> (Accessed 12 May 2015).

Heradia-Pi, I.B., Servan-Mori, E., Reynales-Shigermatsu, L.M., and Bautista-Arredondo, S. (2012). The maximum willingness to pay for smoking cessation method among adult smokers in Mexico. *Value in Health*, **15**, 750–8.

Higgins, A., Barnett, J., Meads, C., Singh, J., and Longworth, L. (2014). Does convenience matter in health care delivery? A systematic review of convenience-based aspects of process utility. *Value in Health*, **17**, 877–87.

Huber, J. and Zwerina, K. (1996). The importance of utility balance in efficient choice set designs. *Journal of Marketing Research*, **33**, 307–17.

ISPOR [International Society for Pharmacoeconomics and Outcomes Research] (2014). *Pharmacoeconomic guidelines around the world* <http://www.ispor.org/PEguidelines/index.asp> (Accessed 8 December 2014).

Johannesson, M. (1996a). *Theory and methods of economic evaluation of health care*. Dordrecht: Kluwer.

Johannesson, M. (1996b). *Ex ante* versus expected willingness-to-pay. *Social Science and Medicine*, **42**, 305–11.

Johannesson, M., Jönsson, B., and Karlsson, G. (1996). Outcome measurement in economic evaluation. *Health Economics*, **5**, 279–96.

Johansson, P.O. (1995). *Evaluating health risks*. Cambridge: Cambridge University Press.

Johnson, R.F. (2012). Why not real economics? *PharmacoEconomics*, **30**, 127–31.

Johnson, P., Bancroft, T., Barron, R., et al. (2014). Discrete choice experiment to estimate breast cancer patients' preferences and willingness to pay for prophylactic granulocyte colony-stimulating factors. *Value in Health*, **17**, 380–9.

Jones-Lee, M.W. (1976). *The value of a life: an economic analysis*. Chicago: University of Chicago Press.

Jones-Lee, M.W., Hammerton, M., and Phillips, P.R. (1985). The value of safety: results of a national sample survey. *Economic Journal*, **95**, 49–72.

Jutkowitz, E., Gitlin, L.N., and Pizzi, L.T. (2010). Evaluating willingness to pay thresholds for dementia caregiving interventions: application to the tailored activity program. *Value in Health*, **13**, 720–5.

Kartman, B., Andersson, F., and Johannesson, M. (1996). Willingness to pay for reductions in angina pectoris attacks. *Medical Decision Making*, **16**, 246–53.

Kilambi, V., Johnson, F.R., González, J.M., and Mohamed, A.F. (2014). Valuations of genetic test information for treatable conditions using discrete choice experiments: the case of colorectal cancer screening. *Value in Health*, **17**, 838–45.

Kløjgaard, M.E., Manniche, C., Pederson, L.B., Bech, M., and Søgaard, R. (2014). Patient preferences for treatment of low back pain—a discrete choice experiment. *Value in Health*, **17**, 390–6.

Lancsar, E. and Louviere, J. (2008). Conducting discrete choice experiments to inform health care decision making. *PharmacoEconomics*, **26**, 661–77.

Lancsar, E. and Savage, E. (2004). Deriving welfare measures from discrete choice experiments: inconsistency between current methods and random utility and welfare theory. *Health Economics Letters*, **13**, 901–7.

Li, C., Zeliadt, S.B., Hall, I.J., et al. (2012). Willingness to pay for prostate cancer treatment among patients and their family members at 1 year after diagnosis. *Value in Health*, **15**, 716–23.

Lin, P-J., Cangelosi, M.J., Lee, D.W., and Neumann, P.J. (2013). Willingness to pay for diagnostic technologies: a review of the contingent valuation literature. *Value in Health*, **16**, 797–805.

Linley, W.G. and Hughes, D.A. (2013). Societal views on NICE, cancer drugs fund and value-based pricing criteria for prioritising medicines: a cross-sectoral survey of 4118 adults in Great Britain. *Health Economics*, **22**, 948–64.

Louviere, J.J. and Fiebig, D.G. (2010). Benefit assessment for cost-benefit analysis studies in health care using discrete choice experiments: estimating welfare in a health care setting, in

E. McIntosh, P.M. Clarke, E.J. Frew, and J.J. Louviere (ed.), *Applied methods of cost-benefit analysis in health care*, pp. 211–29. Oxford: Oxford University Press.

Louviere, J.J. and Lancsar, E. (2009). Choice experiments in health: the good, the bad, the ugly and toward a brighter future. *Health Economics Policy and Law*, **4**, 527–46.

Louviere, J.J. and Woodworth, G.G. (1990). *Best–worst scaling: a model for largest difference judgements*. Working paper. Calgary: University of Alberta, Faculty of Business.

Louviere, J.J., Henscher, D.A., and Swait, J.D. (2000). *Stated choice methods. Analysis and application*. Cambridge: Cambridge University Press.

Luce, R.D. (1959). *Individual choice behaviours: a theoretical analysis*. New York: Wiley.

Marin, A. and Psacharopoulos, G. (1982). The reward for risk in the labour market: evidence from the United Kingdom and a reconciliation with other studies. *Journal of Political Economy*, **90**, 827–53.

Mason, H., Baker, R., and Donaldson, C. (2008). Willingness to pay for a QALY: past, present and future. *Expert Review of Pharmacoeconomics and Outcomes Research*, **8**, 575–82.

McIntosh, E., Clarke, P.M., Frew, E., and Louviere, J.J. (2010). *Applied methods of cost-benefit analysis in health care*. Oxford: Oxford University Press.

McIntosh, E. and Ryan, M. (2002). Using discrete choice experiments to derive welfare estimates for the provision of elective surgery: implications of discontinuous preferences. *Journal of Economic Psychology*, **23**, 367–82.

Mitchell, R.C. and Carson, R.T. (1989). *Using surveys to value public goods: the contingent valuation method*. Washington, DC: RFF Press/McGraw-Hill.

Mooney, G. (1977). *The valuation of human life*. London: Macmillan.

Mooney, G. (1992). *Economics, medicine and health care*. Hemel Hempstead: Wheatsheaf.

Morey, E.R., Rowe, R.D., and Watson, M. (1993). A repeated nested-logit model of Atlantic salmon fishing. *American Journal of Agricultural Economics*, **75**, 578–92.

Mushkin, S. (1978). Cost of disease and illness in the United States in the year 2000. *Public Health Reports*, **93**, 493.

NOAA [National Oceanic and Atmospheric Administration] (1993). Natural resource damage assessments under the oil pollution act of 1990. Notice of proposed rules. *Federal Register*, **58**, R4612.

Neumann, P. and Johannesson, M. (1994). The willingness-to-pay for *in vitro* fertilization: A pilot study using contingent valuation. *Medical Care*, **32**, 686–99.

O'Brien, B. and Gafni, A. (1996). When do the 'dollars' make sense? Toward a conceptual framework for contingent valuation studies in health care. *Medical Decision Making*, **16**, 288–99.

O'Brien, B. and Viramontes, J.L. (1994). Willingness-to-pay: A valid and reliable measure of health state preference? *Medical Decision Making*, **14**, 289–97.

O'Brien, B., Goeree, R., Gafni, A., et al. (1998). Assessing the value of a new pharmaceutical: a feasibility study of contingent valuation in managed care. *Medical Care*, **36**, 370—84.

Olsen, J.A. and Donaldson, C. (1998). Helicopters, hearts and hips: using willingness to pay to set principles for public sector health care programmes. *Social Science and Medicine*, **46**, 1–12.

Oppe, M., Devlin, N.J., van Hout, B., Krabbe, P.F.M., and de Charro, F. (2014). A programme of methodological research to arrive at the new international EQ-5D-5L valuation protocol. *Value in Health*, **17**, 445–53.

Pauly, M.V. (1995). Valuing health care benefits in money terms, in F.A. Sloan (ed.), *Valuing health care*, pp. 99–124. Cambridge: Cambridge University Press.

Potoglou, D., Burge, P., Flynn, T., et al. (2011). Best–worst scaling vs. discrete choice experiments: an empirical comparison using social care data. *Social Science and Medicine*, **72**, 1717–27.

Ratcliffe, J. (2000). The use of conjoint analysis to elicit willingness to pay. Proceed with caution? *International Journal of Technology Assessment in Health Care*, **16**, 270–90.

Ratcliffe, J., Buxton, M., McGarry, T., Sheldon, R., and Chancellor, J. (2004). Patients' preferences for characteristics associated with treatments for arthritis. *Rheumatology*, **43**, 337–45.

Reed Johnson, F., Fries, E.E., and Banzhaf, H.S. (1994). *Valuing morbidity: an integration of the willingness-to-pay and health status literatures*. Working Paper No. T-G401, Triangle Economic Research, North Carolina.

Ryan, M. and Farrar, S. (2000). Using conjoint analysis to elicit preferences for health care. *BMJ*, **320**, 1530–3.

Ryan, M. and Gerard, K. (2003). Using discrete choice experiment to value health care programmes: current practice and future research reflections. *Applied Health Economics and Health Policy*, **2**, 55–64.

Ryan, M., McIntosh, E., and Shackley, P. (1998). Methodological issues in the application of conjoint analysis in health care. *Health Economics*, **7**, 373–8.

Ryan, M., Scott, D.A., and Donaldson, C. (2004). Valuing health care using willingness to pay: a comparison of the payment card and dichotomous choice methods. *Journal of Health Economics*, **23**, 237–58.

Ryan, M., Gerard, K., and Currie, G. (2012). Using discrete choice experiments in health economics, in A. Jones (ed.), *The Elgar companion to health economics*, 2nd edition, pp. 437–46. Cheltenham: Edward Elgar.

Ryen, L. and Svensson, M. (2014). The willingness to pay for a quality adjusted life year: a review of the empirical literature. *Health Economics*, DOI: 10.1002/hec.3085.

Salomon, J.A., Vos, T., Hogan, D.R., et al. (2012). Common values in assessing health outcomes from disease and injury: disability weights measurement study for the Global Burden of Disease Study 2010. *Lancet*, **380**, 2129–43.

Scott, A., Watson, M.S., and Ross, S. (2003). Eliciting preferences of the community for out-of-hours care provided by general practitioners: a stated preference discrete choice experiment. *Social Science and Medicine*, **50**, 804–14.

Sculpher, M.J. and Claxton, K. (2012). Real economics needs to reflect real decisions: a response to Johnson. *PharmacoEconomics*, **30**, 133–6.

Sculpher, M.J., Bryan, S., Fry, P., de Winter, P., Payne, H., and Emberton, M. (2004). Patients' preferences for the management of non-metastatic prostate cancer: discrete choice experiment. *BMJ*, **328**, 382.

Shah, K.K., Lloyd, A., Oppe, M., and Devlin, N.J. (2013). One-to-one versus group setting for conducting computer-assisted TTO studies: findings from pilot studies in England and the Netherlands. *European Journal of Health Economics*, **14**(Suppl 1), S65–S73.

Skjoldborg, U.S. and Gyrd-Hansen, D. (2003). Conjoint analysis. The cost variable: an Achilles heel? *Health Economics*, **12**, 479–91.

Smith, R.D. (2003). Construction of the contingent valuation market in health care: a critical assessment. *Health Economics*, **12**, 609–28.

Stalhammer, N.O. (1996). An empirical note on willingness-to-pay and starting-point bias. *Medical Decision Making*, **16**, 2427.

Stalhammer, N.O. and Johannesson, M. (1996). Valuation of health changes with the contingent valuation method: a test of scope and question order effects. *Health Economics*, **5**, 531–41.

Stewart, J.M., O'Shea, E., Donaldson, C., and Shackley, P. (2002). Do ordering effects matter in willingness-to-pay studies of health care? *Journal of Health Economics*, **21**, 585–99.

Thompson, M.S. (1986). Willingness-to-pay and accepts risks to cure chronic disease. *American Journal of Public Health*, **76**, 392–6.

Viscusi, K.P. (1992). *Fatal trade-offs*. Oxford: Oxford University Press.

Weatherly, H.L.A., Cookson, R., and Drummond, M.F. (2014). Economic evaluation of public health interventions: methodological challenges. In A.J. Culyer (ed.), *Elsevier encyclopaedia of health economics*, Vol. 1, pp. 217–23. San Diego, CA: Elsevier.

Whynes, D.K., Frew, E.J., and Wolstenholme, J.L. (2003). A comparison of two methods for eliciting contingent valuations of colorectal cancer screening. *Journal of Health Economics*, **22**, 555–74.

Chapter 7

Cost analysis

7.1 Some basics

The analysis of the comparative costs of alternative treatments or health care programmes is common to all forms of economic evaluation and therefore most of the methodological issues discussed in this chapter are relevant to all analyses. Although many of the issues surrounding costing are context specific and the analyst's options are often limited by the availability of data, it is possible to give some general guidance. Three particularly thorny issues—the treatment of overhead costs (techniques for allocating shared overhead costs to individual projects), the allowance for differential timing of costs (the techniques of discounting and annuitization of capital expenditure), and the role and estimation of productivity costs—are discussed in some detail. However, the chapter begins by covering some of the basic questions that an evaluator might have when embarking on a costing study in the health field.

7.1.1 Which costs should be considered?

The main categories of costs of health care programmes or treatments were identified in Figure 3.1 of Chapter 3; these are the costs arising from the use of resources within the health sector, the resource use by patients and their families, the resource use in other sectors, and productivity changes. The particular range of costs included in a given study is likely to be decided upon as a result of considering the following four points.

7.1.1.1 What is the perspective for the analysis?

It is essential to specify the perspective because an item may be a cost from one point of view, but not a cost from another. For example, patients' travel costs are a cost from the patient's point of view and from society's point of view, but not a cost from the Ministry of Health's point of view. Workers' compensation payments are a cost to the paying government, a gain to the patient (recipient), and neither a cost nor a gain to society. (These money transfers, which do not reflect resource consumption, are called *transfer payments* by economists; costs are involved in their administration, but these are not measured by the amounts themselves.) Therefore, the study perspective is a key determinant of which costs are deemed relevant.

Possible perspectives include those of society, the Ministry of Health, other government ministries, the government in general, the patient, the employer, and the agency providing the programme. If the evaluation is being commissioned by a given body, this may give a clue to the relevant perspective(s). For example, most

sets of methods guidelines proposed by ministries of health and other official bodies specify the perspective to be adopted (ISPOR 2014). In instances where a particular perspective is not specified, one possibility would be to adopt a broad perspective, considering a wide range of costs. This would make it possible to re-analyse the data using different perspectives, depending on the eventual audience(s) for the study. However, remember from the discussion in Chapters 2 and 4, the choice of perspective is also linked to important value judgements underpinning the study.

The existence of different perspectives was highlighted by Byford et al. (2003) in their study of treatments for recurrent deliberate self-harm. Costs falling on the following sectors were considered: hospital services, social services, voluntary sector services, community accommodation, and the criminal justice system. Costs resulting from lost productivity, due to time off work, were also estimated. The relative costs of the two treatments depended on the perspective adopted. From a health care perspective, the relative annual costs per patient of the two programmes were fairly similar: £2395 for a new intervention, manual-assisted cognitive behaviour, and therapy, and £2502 for treatment as usual. However, when a broader perspective was adopted, including all costs, the annual cost difference per patient was £838 higher for treatment as usual (see Table 7.1).

7.1.1.2 Is the comparison restricted to the two or more programmes immediately under study?

If the comparison is restricted to the programmes or treatments immediately under study, costs common to both need not be considered as they will not affect the choice between the given programmes. (Elimination of such costs can save the evaluator a considerable amount of work.) However, if it is thought that at some later stage a broader comparison may be contemplated, including other alternatives not yet specified, it might be prudent to consider all the costs of the programmes.

7.1.1.3 Are some costs merely likely to confirm a result that would be obtained by consideration of a narrower range of costs?

Sometimes the consideration of patients' costs merely confirms a result that might be obtained from, say, consideration of only operating costs within the health sector. For example, treating a given condition by minimal-access surgery may be of lower cost to the patient, but also may be less costly to the health care system. Therefore, if consideration of patients' costs requires extra effort and the choice of programme is very unlikely to be changed, it may not be worthwhile to complicate the analysis unnecessarily. However, some justification for such an exclusion of a cost category should be given.

7.1.1.4 What is the relative order of magnitude of costs?

It is not worth investing a great deal of time and effort considering costs that, because they are small, are unlikely to make any difference to the study result (e.g. some laboratory tests). However, some justification should be given for the elimination of such costs, perhaps based on previous empirical work. It is still worthwhile identifying such cost categories in any event, although the estimation of them might not be pursued in any great detail.

Table 7.1 Twelve-month total cost per patient (£)

	MACT (*N* = 197)		TAU (*N* = 200)		Mean difference (MACT – TAU) (95% CI)	Adjusted	
	Mean (SD)	Total cost%	Mean (SD)	Total cost%		*P*	*P*[a]
Resource costs							
Hospital services	1 548 (3 326)	12	1796 (4754)	13	–248 (–1059 to 563)		
Community health services	678 (901)	5	566 (815)	4	112 (–58 to 281)		
Medication	169 (680)	1	140 (272)	1	29 (–73 to 131)		
Social services	252 (862)	2	470 (4384)	3	–218 (–844 to 408)		
Voluntary services	13 (51)	0	39 (245)	0	–26 (–61 to 9)		
Accommodation and living expenses	10 369 (2808)	77	10 570 (3138)	74	–200 (–788 to 388)		
Criminal justice services	126 (561)	1	355 (1746)	3	–229 (–485 to 27)		
Total resource costs	13 156 (5024)	98	13936 (7568)	98	–780 (–2050 to 489)	0.11	0.11
Productivity costs	294 (1019)	2	351 (1153)	2	–58 (–273 to 157)	0.60	0.48
Total costs	13450 (5313)	100	14288 (7669)	100	–838 (–2142 to 466)	0.21	0.09
Total costs per week	252 (100)		265 (144)		–14 (–38 to 11)	0.27	0.13

MACT, manual-assisted cognitive behaviour therapy; SD, standard deviation; TAU, treatment as usual.

[a] Adjusted for baseline characteristics: centre, gender, age, living situation (alone vs with others), parasuicide risk score, Beck hopelessness score, personality status (no disorder vs disorder), baseline costs.

Above all, the main point to remember when embarking on a costing study is that, to an economist, cost refers to the sacrifice (of benefits) made when a given resource is consumed in a programme or treatment. Therefore, it is important not to confine one's attention to expenditures, but to consider also other resources, the consumption of which is not adequately reflected in market prices: for example, volunteer time, patients' leisure time, and donated clinic space.

7.1.2 How should costs be estimated?

Once the relevant range of costs has been identified, the individual items must be measured and valued. That is, costing has two elements: measurement of the *quantities* of resource use (q) and the assignment of unit costs or *prices* (p). The measurement of resource quantities often depends on the context for the economic evaluation. For example, if an economic study is being conducted alongside a prospective clinical study, such as a clinical trial, data on the resource quantities may be collected on the case report forms and an analysis performed using the individual patient data. Such an analysis may also be possible using data extracted from patients' charts (case notes) or an administrative database. It may only be possible to estimate the quantities of some resources, such as domiciliary nursing visits, by asking patients, or by having them keep a diary during an ongoing clinical study.

However, many economic evaluations are conducted using summary data, such as costs in the literature from previously conducted clinical studies, or routinely available cost data. In such cases an analysis of individual patient data is not possible. Also, the extent to which resource quantities can be separated from prices will depend on how the data have been summarized.

Market prices will be available for many of the resource items. Although the theoretical proper price for a resource is its opportunity cost (i.e. the value of the forgone benefits because the resource is not available for its best alternative use), the pragmatic approach to costing is to take existing market prices unless there is some particular reason to do otherwise (e.g. the price of some resources may be subsidized by a third party such as a charitable institution). This is discussed further below.

Although the costing of most resource items is relatively unambiguous, the following issues commonly arise in costing studies.

7.1.2.1 How are values imputed for non-market items?

The major non-market resource inputs to health care programmes are volunteer time and patient/family leisure time. One approach to the valuation of these would be to use market wage rates (e.g. for volunteer time one might use unskilled wage rates). The market value of leisure time is harder to assess. One can argue for a value of lost leisure time of anything from zero, through average earnings, to average overtime earnings (time and a half or double time). Alternately, one might try to estimate the level of compensation an individual would require in order to give up some of their time. Strictly speaking, the value of time should depend on what is being sacrificed in terms of paid work, unpaid work, or leisure.

In the economic evaluation of health care treatments and programmes, an important component of unpaid time is that of caregivers. In a recent review of the literature in valuing informal care, Weatherly et al. (2014) acknowledge that informal care is rarely valued in

published economic evaluations, because the majority of these are conducted from the perspective of decision-makers allocating public funds. Nevertheless, they argue that valuing informal care might be important (1) for evaluating the cost-effectiveness of interventions to support carers (such as career breaks) (2) to assess interventions where there is an indirect impact on informal carers (such as those requiring the carer to administer medications), and (3) for use in designing and testing to evaluate the cost-effectiveness of different levels of access to more formal input (such as universal access versus means-tested access).

Koopmanschap et al. (2008) review several methods for valuing caregiver time. Van den Berg and Ferrer-i-Carbonell (2007) estimate the value by assessing the level of compensation caregivers would require to maintain the same level of well-being after providing informal care. Weatherly et al. (2014) discuss the key steps in the identification, measurement, and valuation of caregiving activities. Possible methods of valuation include identifying the opportunity cost (in forgone income), identifying the price of a close substitute for the activity concerned, the well-being approach (mentioned above), by identifying the maximum amount carers would be willing to pay for reducing caregiving activities (or the minimum amount they would be willing to accept to undertake more activities), or undertaking discrete choice experiments to estimate the relative value of attributes of a service which might include health, non-health, and process attributes. If a cost or price is added as an attribute, the monetary compensation for an hour of caregiving can be estimated (Mentzakis et al. 2011).

In situations where the valuation of time is thought to be controversial, a different approach would be merely to identify and measure units of (say) volunteer, family, or patient time input and to document these alongside the other costs when reporting results (see van den Berg et al. 2006 for a discussion of measurement issues). Presenting this information would enable the decision-maker to note those programmes relying heavily on volunteer or family support. It would then be up to the programme director (or advocate of the programme or therapy) to demonstrate that such an input could be obtained without an opportunity cost to other programmes arising from the diversion of volunteer or family time to the new programme.

A final approach would be to estimate the burden of informal care by estimating the impact of the carers' quality of life, using one of the approaches discussed in Chapter 5. Weatherly et al. (2014) discuss a number of instruments for measuring and valuing informal caregivers' quality of life. Van den Berg et al. (2014) estimate losses in well-being from caring, using a large longitudinal database from Australia and also provide some monetary estimates. Viewing the impact of informal caring as (say) losses in quality-adjusted life-years (QALYs) might be viewed as more relevant to some decision-makers than providing monetary estimates of the opportunity cost of informal carers' time. However, it is likely that the reductions in well-being experienced by informal carers are likely to be broader than the health-related effects typically measured by QALYs. As mentioned in Chapter 2, the relevance of non-health effects depends on whether the analyst is adopting a welfarist or non-welfarist approach.

7.1.2.2 When should existing market prices be adjusted?

It has long been recognized that, owing to the imperfections in health care markets, market prices may not reflect opportunity costs. For example, hospital charges may

deviate from costs if a hospital has a local monopoly or seeks to cross-subsidize one activity from another. Physician fees may not accurately reflect the relative skill level and time required for different procedures. Drug prices may be set in negotiations between a pharmaceutical company and the government, where the company's commitment to research and provision of employment might be taken into account, as well as the costs of discovery, production, and distribution of the drug in question.

Having said that, it is by no means clear when an analyst should attempt to adjust observed market prices to reflect true opportunity costs. Most economic evaluations are conducted from the perspectives of particular decision-makers, who would be most interested in the costs they incur on their budget, irrespective of whether these reflect true opportunity costs. However, if the intention is to conduct the analysis from a 'societal' perspective, then in principle all the costs should reflect social opportunity costs. However, as mentioned above, most studies use market prices unadjusted, even if attempting to reflect a broad perspective by considering a wide range of costs. For example, Garrison et al. (2010) note that the vast majority of published economic evaluations of drugs claiming to be from the 'societal' perspective use actual acquisition costs, rather the much lower social opportunity costs that would reflect only short-run manufacturing and distribution costs. Hay et al. (2010) point out that with drug costs in particular it is important to be clear on the perspective being adopted.

In order for analysts to attempt to adjust market prices, they should be convinced that:

- to leave prices unadjusted would introduce substantial biases into the study
- there is a clear and objective way of making the adjustments.

These issues have been explored most extensively in the context of hospital charges in the United States of America. An analysis was undertaken by Cohen et al. (1993) of in-hospital charges from the itemized hospital accounts of 3000 patients at Boston's Beth Israel Hospital (1990 and 1991). Costs were then derived by adjusting for department-specific cost/charge ratios by using data on actual resource consumption. Comparison of estimates led to the results shown in Table 7.2. It can be seen that while the ordering (in expense) of the procedures remains the same, the absolute differences change.

Table 7.2 Costs and charges for four procedures

	Standard hospital charges (SD)	**Costs (SD)**
PTCA	$8369 ($3885)	$5396($2829)
Atherectomy	$8391 ($2299)	$5726 ($2716)
Stent	$12 670 ($5247)	$7828 ($3270)
CABG	$27 739($7051)	$20 927 ($6048)

CABG, coronary artery bypass grafting; PTCA, percutaneous transluminal coronary angioplasty; SD, standard deviation.

Source: data from Cohen, D.J. et al., Economics of elective coronary revascularization: comparison of costs and charges for conventional angioplasty, directional atherectomy, stenting and bypass surgery, *Journal of the American College of Cardiology*, Volume 22, Issue 4, pp. 1052–9, Copyright © 2003.

In a more recent study Taira et al. (2003) compared four methods of estimating costs in three trials involving percutaneous coronary revascularization: (1) hospital charges; (2) hospital charges converted to costs by use of hospital-level cost-to-charge ratios; (3) hospital charges converted to costs by use of department-level cost-to-charge ratios; and (4) itemized laboratory costs with non-procedural hospital costs generated from department-level cost-to-charge ratios.

Their findings were similar to those of Cohen et al. (1993), in that, while there were big differences in the magnitude of the estimates obtained by the various methods, the method used to approximate costs did not affect the main results of the economic comparisons for any of the trials. They also concluded that conversion of hospital charges to costs on the basis of department-level cost-to-charge ratios appears to represent a reasonable compromise between accuracy and ease of implementation.

The methodology employed by many studies in the United States is to derive costs by adjusting for department-specific cost-to-charge ratios. (These are generally in the public domain.) This is probably an improvement on the uncritical use of charges, but it is still dependent on the quality of the accountancy study that generated the costs in the first place. Often this is difficult to assess. Nevertheless, adjustments by cost-to-charge ratios are becoming more commonplace in studies undertaken in the United States. For example, Nigrovic and Chiang (2000) calculated costs from charges 'using a standard cost-to-charge ratio of 0.65'. Zupancic et al. (2003) converted charges to costs 'using cost center-specific Medicare ratios of costs to charges for the Brigham and Women's Hospital for 1991'.

If the results of studies are relatively insensitive to the method used to approximate costs, should we be concerned about this issue? Only to the extent that, when costs or cost-effectiveness ratios for treatments are compared across studies, the differences observed may be partly dependent on the precise type of cost-to-charge adjustments. Also, although costs may differ from charges, it is worth noting that, if the study is being conducted from the perspective of a given payer, such as a government or health insurer, the prices actually paid for resource may be the most relevant cost estimates to use.

In general, there is probably no substitute for a well-conducted original costing study. In most countries, where hospital charges are not as detailed as in the United States, this is often the analyst's only alternative to using routinely available data, such as those on average hospital costs, or diagnosis-related group (DRG) rates. However, comparisons across studies could still be problematic because of the range of costing methods used.

Furthermore, in multi-country studies the availability of financial data and the variations in accounting practices can impact upon results, even if attempts are made to standardize costing methodology. Schulman et al. (1998) attempted to cost procedures used in the treatment of subarachnoid haemorrhage in seven countries. The results are shown in Table 7.3. It can be seen that there are considerable variations in estimates across countries, many of which do not appear to be systematic. Also, approximately 30% of the estimates had to be imputed because they were not available in the countries concerned.

The methodological and practical issues of costing health treatments and programmes were explored in the HealthBASKET project, funded by the European Union

Table 7.3 Reported procedure and *per diem* costs for study countries

	Costs (US$)						
	Germany	**Italy**	**France**	**Sweden**	**UK**	**Australia**	**Spain**
Procedure costs							
Burr holes	130	77	216	*372*	365	711	72
Chest tubes	87	210	*150*	175	*201*	120	93
CNS shunt	1148	1749	617	371	357	699	*526*
Craniofacial procedures	*350*	*471*	*628*	693	*843*	888	*673*
Cranioplasty	*590*	794	*1059*	*1557*	*1420*	*1197*	*1134*
Debridement of brain	824	357	740	1386	2247	717	552
Dialysis	*153*	206	*275*	*404*	*368*	*310*	*294*
Elevation of skull fracture	367	357	483	693	377	*505*	336
Evacuation of lesion	506	357	493	1386	476	*722*	705
Filtration for renal failure	*248*	*334*	*441*	*655*	*597*	759	234
Gastroscopy	*106*	245	63	347	*256*	156	*204*
Gastrostomy (procedure)	79	*148*	361	*290*	*264*	*223*	95
Humeral shaft fracture	*287*	*386*	106	*757*	1904	*582*	21
Intracranial drainage	273	432	*340*	175	365	389	259
Laparotomy (exploratory)	130	209	301	866	462	573	492
Lobectomy	*544*	830	*977*	1040	569	2251	705
Peritoneal lavage	*38*	117	*69*	*102*	*93*	23	34
Removal of bone flap	506	357	411	175	408	1650	332
Replacement of bone flap	809	604	524	*1203*	526	1308	616
Shunt placement	*642*	1749	*1152*	260	2087	*1302*	580
Spine operation	*1125*	1515	*2019*	*2970*	*2708*	*2283*	*2164*
Splenectomy	249	389	*483*	*711*	*648*	*547*	*518*
Swan–Ganz monitor	*207*	335	*371*	*546*	*498*	*420*	317
Superficial laceration	16	31	20	175	154	*68*	36
Tracheostomy	151	120	301	347	256	1105	132
Per diem costs							
Daily intensive care unit	445	601	774	1231	1159	945	876
Daily intermediate care unit	*169*	304	301	573	315	207	*324*
Daily routine care unit	134	187	350	267	173	159	236
Daily rehabilitation unit	140	324	210	336	*384*	186	464

Actual costs are in roman; market-basket imputed costs are in italics.

(Busse et al. 2008). It focused on the basket of services offered by nine EU member states, reviewed and developed methodologies to assess the costs and prices of individual services across those states, and developed and tested an innovative approach towards collecting and analysing cost variations at the micro-level for the purposes of international comparisons.

Among the difficulties addressed by the project were that the delivery and cost of a seemingly identical service might vary across countries due to variations in (1) the definition of the start and end of a service (e.g. whether rehabilitation following hip replacement is part of the hospital treatment or seen as a separate service with its own tariff); (2) the technology used (especially regarding the use of innovative and/or expensive technologies, e.g. cemented hip replacement versus costlier uncemented hip replacement); and (3) the accounting treatment of associated services (e.g. whether anaesthesia is included in the service 'surgical procedure' or counted and charged separately). Even for a comparable service, different factors might be included in the cost and/or price calculations (e.g. how overheads are treated; whether volume variable, fixed, amortization or investment costs are included; or whether any subsidies are made explicit). Therefore, it is clear that costs have to be generated within the jurisdiction of interest and not 'imported' from elsewhere.

7.1.2.3 For how long should costs be tracked?

It can be seen from Figure 7.1 that not only does the analyst have a choice about whose costs to consider but also a choice of time period. In assessing how long costs should be tracked, the main objective should be to avoid misleading the decision-maker or user. For example, an early comparison of the costs of coronary artery bypass grafting (CABG) versus percutaneous transluminal coronary angioplasty (PTCA) to hospital discharge has shown CABG to be substantially more expensive ($9138 versus $22 711) (Black et al. 1988). However, there is a possibility that patients receiving PTCA may require additional treatment subsequently, including CABG. In a costing study undertaken alongside a randomized controlled trial, Sculpher et al. (1993) showed that by

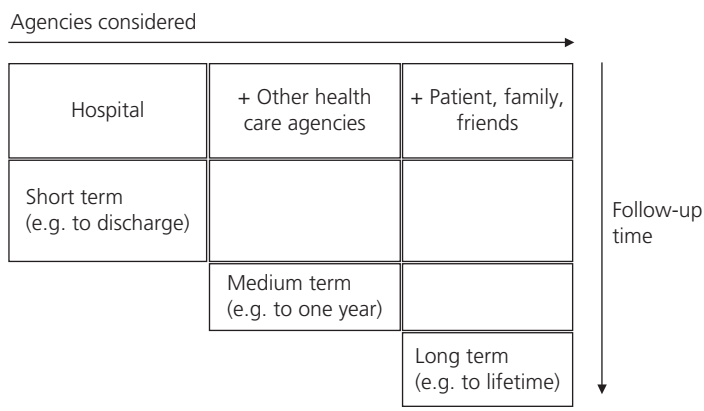


Fig. 7.1 Choices in the consideration of costs.

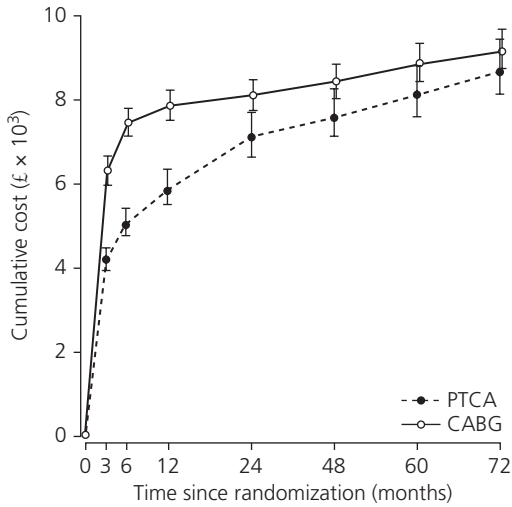


Fig. 7.2 Cumulative costs of percutaneous transluminal coronary angioplasty (PTCA) and coronary artery bypass grafting (CABG) over time (confidence intervals indicated by the bars).

Reprinted from *The Lancet*, Volume 352, Issue 9138, Henderson, R.A. et al., Long-term results of RITA-1 trial: clinical and cost comparisons of coronary angioplasty and coronary artery bypass grafting, pp.1419–1425, Copyright © 1998 Elsevier Ltd, with permission from Elsevier, <http://www.sciencedirect.com/science/journal/01406736>.

24 months after randomization, the cost difference between patients randomized to the alternative therapies had reduced substantially, mainly because some patients randomized to PCTA required a repeat procedure or a CABG. After 72 months the cumulative costs were virtually indistinguishable, with overlapping confidence intervals (Henderson et al. 1998) (see Figure 7.2).

In a more recent study, Faria et al. (2013) tracked costs over 5 years in the REFLUX clinical trial comparing laparoscopic fundoplication versus continued medical management for the treatment of gastro-oesophageal reflux disease. The patients randomized to surgery accrued a large proportion of total costs in the first year of follow-up, owing to the upfront cost of surgery, whereas costs in the continued medical management group were distributed across the 5 years. At the end of the first year, the incremental costs of the surgical group over the medical management group were 2363 euros (CI: 1951, 2775). By the end of 5 years, this had fallen to 1832 euros (CI: 1214, 2448). The reason for this was that, after the first year, patients on medical management accumulated costs (medications and visits to primary care) at a greater rate than those who underwent laparoscopic fundoplication, and the total difference in costs slowly converged over time.

There is fairly broad agreement among analysts that in the case of *therapy-specific* or *disease-specific* costs, the choice of follow-up period should not bias the analysis in favour of one intervention over another. In some cases this may involve tracking costs for lifetime, although the quantitative impact of costs (on the analysis) far into the future

will be reduced by discounting to present values (see Section 7.2). Nevertheless, most analysts feel that all *related* health care costs should be included.

However, this can sometimes lead to a decision-making dilemma. Manns et al. (2003) discuss the switch from cellulose to synthetic dialysers in the treatment of end-stage renal disease by haemodialysis. Synthetic dialysers were associated with only a small additional cost and led to a gain in QALYs, mainly through extending survival (ICER of $5036 per QALY gained). However, a direct consequence of improving survival was that haemodialysis costs themselves were higher. Consideration of these related costs increased the ICER to $83 501 per QALY gained, which might be considered by some to be on the borderline of being cost-effective. The main reason for this result is that haemodialysis itself might only be marginally cost-effective, but a decision to provide this treatment has already been taken. Therefore, unless one wanted to revisit the decision to provide haemodialysis to those people with end-stage renal disease, it makes sense to implement a cost-effective improvement to therapy.

Costs can also change over time, for example because of the existence of a 'learning curve' (Brouwer et al. 2001). That is, because health professionals learn how to become more efficient in the use of new health technologies, the costs in the early stages of use may not be a good predictor of costs in the long run. Examples include the dosage, administration, and wastage of drugs; the time taken to perform surgical procedures; and the monitoring of adverse events. Therefore, when costing new or emerging technologies it may be prudent to anticipate that learning effects may occur, although of course the timing of the economic evaluation will often be determined by the need to make a decision about the appropriate use of the new technology. In particular, in costing a new procedure over time it may be worthwhile checking whether the costs towards the end of the period are similar to those in the early months.

Another reason why costs, or at least prices, might change over time is because of changes in market conditions. For example, in 2003 the National Institute for Health and Care Excellence (NICE) in the United Kingdom appraised drug-eluting stents (DES) and found that they were more cost-effective than bare metal stents (BMS) for patients in whom the artery to be treated is less than 3 mm in diameter or the affected section of the artery is longer than 15 mm. However, when NICE undertook a reappraisal in 2008, it found that DES were no longer cost-effective from the perspective of the National Health Service because the market price of BMS had fallen substantially, causing the ICER (of DES over BMS) to be above NICE's threshold. Therefore, it revised its guidance to recommend the use of DES, only as long as their additional cost over BMS was no more than £300 (Drummond et al. 2009).

Another common price change relating to economic evaluations in the health care sector is the reduction in the price of drugs when they lose patent protection. Hoyle (2011) argues that this could be anticipated in economic evaluations involving the long-term use of drug therapy that goes beyond the period of patent protection. However, analysts should be cautious about anticipating future price changes. It would be more advisable to conduct the analysis using current prices and then to revise it if, and when, prices change.

7.1.2.4 Should health care costs unrelated to the programme or intervention under study be included?

The question of whether *unrelated* health care costs in the future should be included is much more open to debate. On the one hand, health care costs in later years of life are a clear consequence of keeping individuals alive. On the other hand, it does not seem totally fair to assign these costs to a prevention programme (e.g. hypertension screening), when they result from therapeutic decisions (e.g. to give cancer chemotherapy for advanced stages of disease) that should be considered on their own merits. Nevertheless, it is common in evaluations of prevention programmes to assign all the credit for life extension, or gains in QALYs, to the programme concerned. Therefore, it would make sense to assign all costs if a generic measure of outcome is being used.

In considering this issue it has to be remembered that all the forms of economic evaluation discussed in this book are what economists call partial equilibrium analyses. That is, while it is recognized that any change in economic activity (such as investment in health programmes) includes many ripples throughout the economy, it is argued that such investments can be assessed against a background of all else remaining constant. Therefore, a boundary is always being drawn around analyses (see the discussion in Chapter 4).

There is less agreement among economic analysts about whether *unrelated* health care costs in later years of life should be included (Gold et al. 1996). The main consideration here is the extent to which the provision of additional care in added years of life is a necessary consequence of the programme being evaluated. For example, if we were evaluating a new drug for treatment of septic shock in intensive care, it would be reasonable to assume that patients surviving an episode of septic shock were likely to have treatment for their underlying morbid condition. Therefore, these costs would be a direct consequence of giving the drug therapy (Schulman et al. 1991). The same would be true of the costs of diagnosing and treating cases of disease identified by a screening programme. These costs are very closely linked and it would make sense to evaluate the costs and consequences of screening, diagnosis, and treatment as a single package. Indeed, we might even consider these to be *related* health care costs.

On the other hand, if we were evaluating a new drug for treatment of hypercholesterolaemia, the added years of life, through reduction in the incidence of coronary heart disease, may be in the distant future. Treatment of unrelated disease (e.g. cancer) is not a necessary consequence of treatment of hypercholesterolaemia and may be determined by protocols that have not yet been defined. Few analysts attempt to track all these costs and consequences, although it is clear that additional costs will be incurred if individuals live longer. However, the fact that such costs and consequences are more distant is not the only consideration that leads to their frequent exclusion from economic evaluations. (Indeed, it could be argued that the costs of treating the coronary heart disease events are themselves distant, but most analysts would include these in an evaluation of drugs for hypercholesterolaemia.) In commenting on this debate, Weinstein and Manning (1997) argue that, in order to be consistent in the practice of including only 'related' costs, we would have to tease out which costs were truly 'related' and which were not.

Sometimes it is difficult to be more precise about the unrelated costs in added years of life than an average annual per capita health expenditure, perhaps age related. Therefore, one approach would be to include an estimate of age-related per capita health expenditure on the cost side of the equation for every year of life added by the intervention. This amount could either be included as a gross amount, or net of medical expenses that were already being included for treatment of the individual's main condition. Depending on the importance the analyst attaches to costs in added years of life, these could either be included in the primary analysis or a sensitivity analysis.

When estimates such as these have been included in economic evaluation of health care programmes they sometimes do not alter cost-effectiveness ratios by very much. For example, Drummond et al. (1993) found that adding an average expenditure figure for costs in added years of life only changed their estimate of the cost per life-year gained from treatment for hypercholesterolaemia by 2%. However, when Daly et al. (1992) added costs in extra years of life to their evaluation of hormone replacement therapy, this increased total programme costs, and the cost per life-year gained, by around 10%.

The small quantitative impact in the examples given is partly due to the fact that costs in added years of life are often heavily discounted and, in the words of one analyst, 'may amount to no more than a hill of beans' (Bush 1973). Therefore, in many instances it may be that unrelated health care costs in added years of life can be ignored without seriously biasing the analysis. However, the quantitative importance of costs in added years of life may vary from one evaluation to another and requires more empirical investigation. Nevertheless, if the health benefits in the study are projected over the individual's lifetime, all health care costs should be similarly projected.

Olchanski et al. (2013) reviewed the practice and implications for cost-effectiveness of including or excluding future costs in 44 economic evaluations of cancer therapies. Together, these studies generated a total of 59 incremental cost-effectiveness ratios. They found that, of the 59 ratios reviewed, all included direct medical costs related to the index therapy and 68% included direct medical costs related to both the index therapy and disease. None included unrelated medical costs, but 11% included nonmedical costs. Including only therapy costs made 26 additional ratios (68%) cost-saving and 4 more ratios (11%) cost-effective. Including all types of medical costs made 2 fewer ratios (5%) cost-saving and 3 fewer ratios (8%) cost-effective.

A much broader issue is that of whether related and unrelated *non-health care* costs should be included. Meltzer (1997) makes a strong case for considering *all* future costs in economic evaluations, including the impacts that treatments have on individuals' production and consumption. Studies have shown that this analytic judgement makes a difference to the results. Johannesson et al. (1997) found that, relative to other health care interventions, including unrelated non-health care costs improves the cost-effectiveness of life-saving programmes among younger individuals.

Weinstein and Manning (1997) argue that, from a 'welfarist' perspective, the inclusion of future non-health care costs is technically correct, 'but will give some practitioners pause to accepting the welfare-theoretical foundation of CEA'. This relates to the point made in Chapters 2 and 4, that economic evaluations can be conducted based on different sets of values. The 'welfarist' approach is just one of the possibilities and will

not be the most appropriate for many resource allocation decisions. Therefore analysts should indicate clearly the stance that they take on these issues and perhaps consider a sensitivity analysis of the inclusion and exclusion of costs in added years of life. We return to this issue when discussing the inclusion or exclusion of productivity changes in Section 7.3.

7.1.2.5 How should capital outlays be handled?

Capital costs are the costs to purchase the major capital assets required by the programme; generally equipment, buildings, and land. Capital costs differ from operating costs in a number of ways. First, they represent investments at a single point in time, often at the beginning of the programme, rather than annual sums like operating costs. Frequently, the capital costs are often not listed in the accounts or budgets of the organization because they have been funded in advance, perhaps by a one-time grant, while the budgets and accounts represent operating expenses only. Sometimes, the annual budgets and accounts contain an item called *depreciation*, which relates to capital costs, as explained below.

Capital costs represent an investment in an asset that is used over time. Most assets, such as equipment and buildings, wear out or depreciate with time. On the other hand, land is a non-depreciable asset because it maintains its value. There are two components of capital cost. One is the opportunity cost of the funds tied up in the capital asset. This is clearly seen in the case of land. Although an investment in non-depreciable land will return the original capital sum when sold, there is still a 'cost'. This cost is the lost opportunity to invest the sum in some other venture yielding positive benefits. It is usually valued by applying an interest rate (equal to the discount rate used in the study) to the amount of capital invested. (Discounting is discussed in Section 7.2.)

The second component of a capital cost represents the depreciation over time of the asset itself. Various accounting procedures (straight line, declining balance, double declining balance, and so on) are available for use in the accounts of the organization. Often, accounting practices relate more to the company tax laws governing the depreciation of assets than to the real change in the value of the asset.

There are several methods of measuring and valuing capital costs in an economic evaluation. The best method is to annuitize the initial capital outlay over the useful life of the asset; that is, to calculate the 'equivalent annual cost'. This method automatically incorporates both the depreciation aspect and the opportunity cost aspect of the capital cost. It is our preferred approach and is described in Section 7.2. An alternative but less exact method is to determine the depreciation cost each year using an accounting method and to determine the opportunity cost on the undepreciated balance for each year. Where market rates exist for the rental of buildings or lease of equipment, these may be used to estimate capital costs. This method also incorporates both the depreciation and the opportunity components of the cost.

If capital outlays relate to resources that are used by more than one programme they may require allocation in a similar fashion to 'overhead' costs. See the discussion of this point in Section 7.1.2.7.

7.1.2.6 What is the significance of the average cost/marginal cost distinction?

Economists tend to emphasize this point, but in fact, marginal cost and average cost are but two concepts relating costs to quantity (see Box 7.1 and Box 7.2).

The major significance of the average cost/marginal cost distinction to the analyst is as follows. First, when making a comparison of two or more programmes it is worth asking independently of each, 'What would be the costs (and consequences) of having a little more or a little less?' (e.g. suppose Neuhauser and Lewicki (1975) had been comparing the six-stool protocol for detecting colonic cancer with another diagnostic test. Perhaps the question of six- versus five-stool tests may never have been asked!). Second, when examining the effects (on cost) of small changes in output, it is likely that these will differ from average costs. For example, the extra cost of keeping patients in hospital for another day at the end of their treatment might be less than the average daily cost for the whole stay. (In fact, this issue usually arises in the opposite sense—the savings from a reduction of one day's stay are usually lower than the average daily cost; see Box 7.3.)

In practice, whereas it is important to acknowledge the difference between marginal and average costs (or savings), this issue can only really be explored in the context of specific locations or situations. For example, the extent to which costs can be saved when hospital stay is shortened depends on the flexibility available locally and the time period over which the change is made.

Therefore, in some studies analysts turn their attention to issues of marginal costs or savings in the discussion, after presenting average results as the primary analysis. For example, in a study investigating the costs and benefits of shortening time to discharge from a coronary intensive care unit by use of a more expensive sedative agent, Sherry

Box 7.1 Various definitions of cost

Total cost (TC)	Cost of producing a particular quantity of output
Fixed cost (FC)	Costs which do not vary with the quantity of output in the short run (about 1 year) and vary with time, rather than quantity: e.g. rent, equipment lease payments, some wages and salaries
Variable cost (VC)	Costs which vary with the level of output: e.g. supplies, food, fees for service
Cost function (TC)	$f(Q)$, total cost as a function of quantity
Average cost (AC)	TC/Q, the average cost per unit of output
Marginal cost (MC)	(TC of x + 1 units) − (TC of x units)
	= d(TC)/dQ evaluated at x
	= the extra cost of producing one extra unit of output

Box 7.2 Is it marginal or incremental?

The terms 'marginal' and 'incremental' are often used interchangeably in the literature. They both refer to a change in the scale of an activity. Strictly speaking, the *marginal cost* relates to the cost of producing *one extra* unit of output. However, it is often used to refer to the cost of producing the *next logical batch* of output, for example, in expanding a screening programme from high-risk people only to the whole population.

The term 'incremental' is sometimes also used to refer to such a change, but is more often used to refer to the difference, in cost or effect, between the two or more mutually exclusive programmes being compared in the evaluation.

In Figure 7.3, MC_A,Q_1 is the marginal cost of programme A evaluated at quantity (scale of activity) Q_1. MC_B,Q_1 is the equivalent estimate for programme B. The incremental cost, of programme A over programme B, evaluated at Q_1, is IC_{A-B},Q_1.

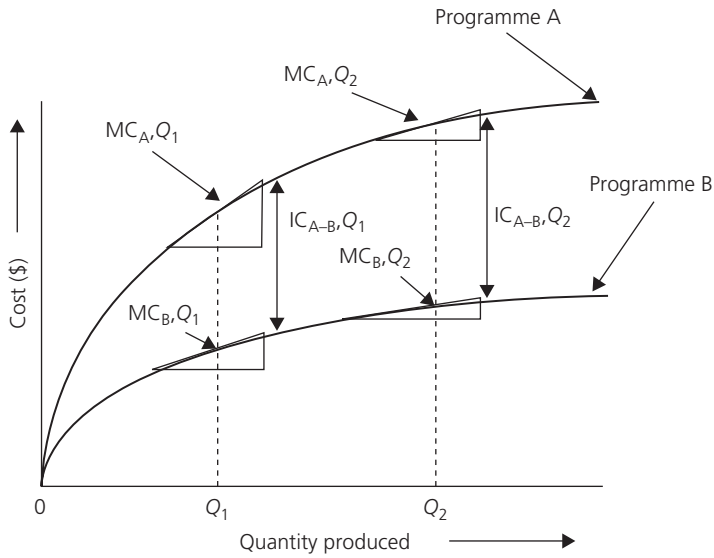


Fig. 7.3 Distinction between marginal and incremental cost.

et al. (1996) investigated the impact on nurse staffing requirements through fewer patients requiring intensive care during the night. It turned out that the hospital concerned had access to a bank of agency nurse staff that could be called in as required; so it was possible to realize potential savings from fewer patients requiring care. In another hospital, with different nurse staffing arrangements, the outcome could be quite different.

The study by Sherry et al. (1996) illustrates that costs, and cost savings, depend greatly on the local context. The importance of costing in context was also illustrated in a study by Chambers et al. (2010) of sugammadex, a newly developed agent for the

Box 7.3 Estimating the cost savings associated with reductions in hospital in-patient stay

Hospital cost can be considered to consist of two elements: the hotel cost, which is broadly constant over the length of stay, and the treatment cost, which may peak just after admission but then tail off in the later days of the stay (see Figure 7.4).

If the length of stay is reduced from d_1 to d_2, use of the average daily cost (c) would give an estimate of the saving of $c(d_1 - d_2)$. However, this would overestimate the actual saving, the shaded area on the diagram. Saving in this case means the value of the resources freed for alternative uses. Whether they *will* be usefully redeployed, or actual expenditure saved, also needs to be investigated.

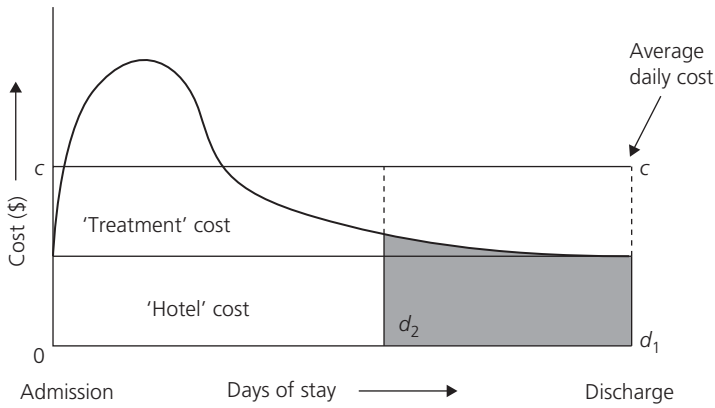


Fig. 7.4 Typical hospital cost profile by length of stay.

reversal of neuromuscular blockade induced by rocuronium or vecuronium. Sugammadex can reverse profound blockade and can be given for immediate reversal. Its use would also avoid the potentially serious adverse effects of the currently used agent, succinylcholine. In addition, sugammadex can reverse neuromuscular blockade more quickly and predictably than existing agents.

The authors explored two possible valuations of the benefits of quicker reversal: in the first, the value of each minute of recovery time saved was estimated as being the pro rata cost of employing the operating room staff (on the basis that all time savings would be achieved in the operating room); in the second, the value of each minute saved was estimated as the pro rata cost of employing a single nurse in the recovery room (on the basis that all time savings would be achieved in the recovery room). Of course, the major uncertainty is the extent to which any time saved in recovery could be put to alternative productive use, for example in caring for another patient or some other activity. This was unknown, as no evidence was identified in the literature. It is also possible that extra operations could be scheduled as a result of any reduced recovery

time but again there was a lack of suitable evidence on the associated impact on costs and health effects.

Since the clinical strategies were assumed to have identical health outcomes but generally different costs, the analysis effectively simplifies to a cost minimization. Given the fact that particular variables are unknown, a threshold analysis was undertaken. The critical variables in this analysis were the reduction in recovery time by using sugammadex and the value of each minute of recovery time saved.

Under the base-case assumptions and the estimates made of the recovery time saved with sugammadex, sugammadex is cost-effective in patients with moderate (profound) blockade where the value of each minute of recovery time saved with sugammadex is approximately £2.40 or greater. It was estimated that time saved in the operating room had a value of £4.44 per minute, whereas time saved in the recovery room had a value of £0.33 per minute. Sugammadex therefore appeared to be cost-effective for the routine reversal of rocuronium-induced moderate (profound) blockade at the current list price, if all reductions in recovery time that are associated with sugammadex are achieved in the operating room, but does not appear cost-effective if all of the reductions in recovery time are achieved in the recovery room. Where savings in recovery time are achieved in both the operating room and the recovery room, or where there is additional value in reducing recovery times (e.g. in preventing operations from being delayed or forgone), sugammadex could also be cost-effective. Thus, the cost-effectiveness of sugammadex was highly dependent on the setting in which it was administered and the benefits that could be obtained in the particular context.

Very rarely do analysts undertake a 'costing in context'. It was mentioned in Chapter 3 that economic evaluations tacitly assume that freed resources will be redeployed to other cost-effective activities. Clearly this is not always the case and it is the responsibility of analysts to at least point this out, even if they do not explore the implications in great detail. The report of the United States Public Health Service Panel on *Cost-effectiveness in health and medicine* (Gold et al. 1996), recommended that when information on capacity utilization in hospitals or other health care facilities is not available, analysts should use the benchmark assumption that capacity is utilized at the rate of 80%, under a long-run perspective. However, the prime motivation for this was to encourage some consistency in study reporting and the 80% figure is not etched in stone. It is very unlikely to apply in all locations or all health care systems.

7.1.2.7 How should shared (or overhead) costs be handled?

The term '*overhead costs*' is an accounting term for those resources that serve many different departments and programmes, for example, general hospital administration, central laundry, medical records, cleaning, porters, power, and so on. If individual programmes are to be costed, these shared costs may need to be attributed to programmes.

The main point to note at the outset is that there is no unambiguously *right* way to apportion such costs. The approach that is favoured by economists is to employ marginal analysis. That is, to see which (if any) of such costs would change if a given programme were added to, or subtracted from, the overall activity. This is fine up to a point, but

the most common situation is that the choice is not such an addition or subtraction, but one between two programmes, each of which would consume the given central services (perhaps because they are competitors for the same space in the hospital). For example, suppose the question concerned space in the hospital that could be used either for anticoagulant therapy for pulmonary embolism, or for renal dialysis. If the economic evaluation concerned a choice between these two programmes then there would be no methodological problem; the costs associated with use of the space would be common to both and could be excluded from the analysis. However, typically the comparison might be between the anticoagulant therapy and another programme in the same field. This could be a programme of more definitive diagnosis of pulmonary embolism, which would avert some hospitalization. In such an instance it would be relevant to obtain an estimate of the value of the freed resources (e.g. hospital floor space) that could be diverted to other uses.

A number of methods can be used to determine a more accurate cost of a programme in a hospital or other setting where shared (or overhead) costs are involved. The methods are illustrated below in terms of a hospital setting. The basic idea is to determine the quantities of service consumed by the patient (days of stay in ward A, B, or C, number of laboratory tests of each type, number of radiological procedures, number of operations, and so on), to determine a full cost (including the proper share of overhead, capital, and so on) for a unit of each type of service, and to multiply these together and sum up the results. The allocation methods described below are different ways to determine the cost per unit for each type of service. In these methods the overhead costs (e.g. housekeeping) are allocated to other departments (e.g. radiology) on the basis of some measure, called an *allocation basis*, judged to be related to usage of the overhead item (e.g. square metres of floor space in the radiology department might be used to allocate housekeeping costs to radiology).

In deciding which of the following approaches to use, the comments made in Section 7.1.1 should be borne in mind. That is, the more important the cost item is for the analysis, the greater the effort that should be made to estimate it accurately. There may conceivably be evaluations for which simple per diem or average daily costs will suffice, because the result is unlikely to change irrespective of the figure assumed for the cost of hospital care. However, we suspect that such situations are in the minority, given the relative order of magnitude of hospital costs compared with other elements of health care expenditures.

Alternatively, an intermediate approach may suffice. Here the per diem cost is purged of any items relating to medical care costs, leaving just the 'hotel' component of hospital expenditure. It is then assumed that all patients are 'average' in respect of their hotel costs and that this expenditure can therefore be apportioned on the basis of patient-days. Thus, the hotel cost can be calculated for the patients in the programme of interest and combined with the medical care costs attributable to those patients to give the total costs of the programme. (The medical care costs would be estimated separately, using data specifically relating to the patients in the programme.)

If a more detailed consideration of costs is required, various methods for allocating shared (or overhead) costs are available:

1 *Direct allocation (ignores interaction of overhead departments)*. Each overhead cost (e.g. central administration or housekeeping) is allocated directly to final cost centres (e.g. programmes such as day surgery, or departments such as wards or radiology). Therefore, a given ward's share of central administration would be equal to the total cost of central administration, multiplied by the ward's share (or proportion) of the allocation basis (say, paid hours for staff). Note that the ward's share is its paid hours divided by total paid hours of all final cost centres, not total paid hours for the whole organization. The latter method would underestimate the costs in all final cost centres.

2 *Step-down allocation (partial adjustments for interaction of overhead departments)*. The overhead departments are allocated in a stepwise fashion to all of the remaining overhead departments and to the final cost centres. Typically, the process starts with those departments that service the broadest number of other departments, such as the hospital administrative office and the power plant.

3 *Step-down allocation with iterations (full adjustment for interaction of overhead departments)*. The overhead departments are allocated in a stepwise fashion to all of the other overhead departments and to the final cost centres. The procedure is repeated a number of times (about three) to eliminate residual unallocated amounts.

4 *Simultaneous allocation (full adjustment for interaction of overhead departments)*. This method uses the same data as (2) or (3) but it solves a set of simultaneous linear equations to give the allocations. It gives the same answer as method (3) but involves less work. (The method is shown diagrammatically in Figure 7.5.)

The effort that one would put into overhead cost allocation would depend on the likely importance of overhead costs (in quantitative terms) for the whole analysis. A much simpler, but cruder, approach is to do the following:

1 Identify those hospital costs unambiguously attributable to the treatment or programme in question (e.g. physicians' fees, laboratory tests, and drugs); these are known as the *directly allocatable costs*. Allocate these directly and immediately to the programme.

2 Deduct, from total hospital operating expenses, the cost of departments already allocated above and departments known not to service the programme being costed.

3 Allocate the remainder of hospital operating expenses on the basis of number of patient-days, for example:

$$\text{Hospital cost of the programme} = \text{Directly allocatable costs} + \frac{\text{Net hospital expenditure}}{\text{Total number of hospital patient-days}} \times \text{Hospital patient-days attributable to the programme}$$

4 Finally, undertake a sensitivity analysis.

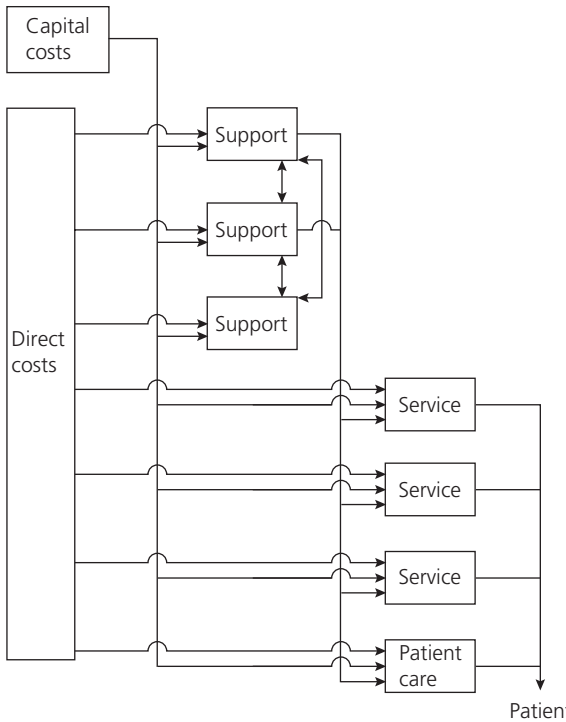


Fig. 7.5 Schematic illustration of cost allocations.

Reproduced from Boyle, M.H., et al., *A cost analysis of providing neonatal intensive care to 500–1499-gram birth-weight infants*, Research Report No. 51, Programme for Quantitative Studies in Economics and Population McMaster University, Hamilton, Ontario, Canada, Copyright © 1982, by permission of the author.

Although there is nothing to suppose that this method is anything but crude, if the choice between programmes is fairly insensitive to the value derived it may suffice.

Typically the choice of allocation methodology is driven by the accounting conventions followed by a particular organization. Within the context of a given economic evaluation, the analyst will not be directly involved in the accounting methods for allocating overheads. Rather, he or she will be using cost data that already embody particular accounting conventions. Therefore, it is important to understand what these conventions imply for the inclusion of categories of cost and allocation of overheads.

7.1.3 Overall, how accurate does costing have to be?

Costing can take considerable time and effort and it is not possible to do a perfect job every time. However, it is important not to make the perfect the enemy of the merely good. Therefore, analysts need to form a judgement on how accurate (or precise) cost estimates need to be within a given study.

Box 7.4 Levels of precision in hospital costing

Most precise

Least precise

Micro-costing
Each component of resource use (e.g. laboratory tests, days of stay by ward, drugs) is estimated and a unit cost derived for each

Case-mix group
Gives the cost for each category of case or hospital patient. Takes account of length of stay. Precision depends on the level of detail in specifying the types of cases

Disease-specific per diem (or daily cost)
Gives the average daily cost for treatments in each disease category. These may still be quite broad (e.g. orthopaedic surgery)

Average per diem (or daily cost)
Averages the per diem over all categories of patient. Available in most health care systems

Box 7.4 indicates the different levels of precision in costing for hospital costs. The least precise estimates are likely to be based on average per diems (or daily costs); the most precise estimates are likely to be based on micro-costing.

The guidance for deciding on the accuracy of costing is similar to that for deciding on the inclusion or exclusion of costs discussed earlier. Clearly a major factor is the likely quantitative importance of each cost category in the evaluation. For example, in an evaluation comparing two drug therapies it is likely that the study result will be sensitive to the costs of the drugs themselves. Therefore, it will be important to record dosages and routes of administration carefully, to facilitate micro-costing. On the other hand, if the drugs concerned have side effects that may infrequently cause hospitalizations, it may suffice to use a per diem or case-mix group cost for these, if one is available.

Similarly, even if it has been decided to follow a micro-costing approach, different levels of accuracy can be applied to different cost items. For example, it is well known that many laboratory tests cost only a few cents/pence each. Therefore, it does not make sense to invest considerable effort in costing these accurately: an average laboratory charge may suffice. On the other hand, nursing costs are often a major component of overall hospital costs. Therefore, it may be important to record the numbers and grades of nursing staff in the ward where the patients of interest are being cared for.

In using routinely available cost data, such as hospital per diems, or case-mix group costs (e.g. DRGs), it is important to pose the following questions. First, how was the cost estimate derived? Namely, what categories of costs are included? Second, how up to date is the cost estimate? Simple adjustments for inflation will not suffice if recent technological advances have dramatically changed the costs of the treatment concerned (e.g. the introduction of DES in coronary care).

The choice between using routinely available data and undertaking micro-costing often depends on the impact that the treatment of interest has on resource use. Suppose one were evaluating a new drug in the field of cardiology. If the main effect of the drug is to reduce the frequency of cardiac events (e.g. myocardial infarctions) requiring hospitalization, it may be sufficient to use the relevant case-mix costs (e.g. the DRG cost for the event concerned). However, if the main effect of the drug is to reduce the severity of the events, or the level of intensity of treatment, a micro-costing approach may be required.

In addition, it is easier to undertake micro-costing if the economic evaluation is being undertaken alongside a prospective clinical study (see Chapter 8), because it is possible to have access to individual patient data. If, on the other hand, a decision-analytic modelling study is being undertaken (see Chapter 9), it is more likely that routinely available data would have to be used.

Finally, it is worth bearing in mind that the calculation of total cost requires the quantities of resources to be multiplied by the prices (unit costs) of those resources. Therefore, when deciding on the level of precision in the estimation of resource quantities, it is worthwhile considering what degree of detail will be available on the costs, or vice versa. For example, it may not be worthwhile collecting considerable detail on the resource quantities if, for example, only average per diem costs are available in a given setting.

7.2 Allowance for differential timing of costs (discounting and the annuitization of capital expenditures)

7.2.1 Time preference

As mentioned in Chapter 3 (Section 3.1.7) and Chapter 4 (Section 4.5.2), some allowance needs to be made for the differential timing of costs and consequences. That is, even in a world with zero inflation, it would be an advantage to receive a benefit earlier or to incur a cost later—it gives you more options. Economists call this the notion of *time preference*.

There are a number of reasons why individuals may have a *positive rate of time preference*; that is, a preference for benefits today rather than in the future. First, they may have a short-term view of life; living for today rather than thinking about the future. Second, the future is uncertain, so, as the saying goes, 'a bird in the hand is worth two in the bush'. Third, with positive economic growth and the long-term trend since the Second World War, individuals might expect to be more wealthy in the future. Therefore, a dollar today would be of higher value than one in the future when you are richer. Finally, since most individuals appear to have a positive rate of time preference, one can usually obtain a positive return when making a riskless investment.

7.2.2 Comparing programmes or interventions with different time profiles

The notion of preferring benefits today, or wanting to postpone costs, extends beyond money transactions and could extend to goods and services that could not easily be traded. It is of most significance for those economic evaluations that compare programmes or interventions with different time profiles. For example, if two options for

Table 7.4 Yearly costs for two health care programmes

Year	Cost of Programme A ($000s)	Cost of Programme B ($000s)
1	5	15
2	10	10
3	15	4

dealing with heart disease were (1) expanding funding for CABG, and (2) a health education campaign to influence diet and lifestyle, we might expect option (1) to deliver benefits earlier. Therefore, if a positive rate of time preference were acknowledged, it would look more attractive, compared with the preventive option, than would otherwise be the case.

Typically, economic evaluation texts discuss the situation where the costs of the alternative programmes A and B can be identified by the year in which they occur (see Table 7.4). In this example, B might be a preventive programme that requires more outlay in Year 1 with the promise of lower cost in Year 3. The crude addition of the two cost streams shows B to be of lower cost (29 000 versus 30 000), but the outlays under A occur more in the later years.

A comparison of A and B (adjusted for the differential timing of resource outlays) would be made by discounting future costs to present values. The calculation is as follows. If P = present value, F_n = future cost at year n, and r = annual interest (discount) rate (e.g. 0.05 or 5%), then

$$P = \sum_{n=1}^{3} F_n (1+r)^{-n} = \frac{F_1}{(1+r)} + \frac{F_2}{(1+r)^2} + \frac{F_3}{(1+r)^3}$$

$$= \frac{F_1}{(1.05)} + \frac{F_2}{(1.05)^2} + \frac{F_3}{(1.05)^3}$$

In our example this gives the present value of cost of A = 26.79; present value of cost of B = 26.81. This assumes that the costs all occur at the end of each year. An alternative assumption that is commonly used is to assume that the costs all occur at the beginning of each year. Then, Year 1 costs need not be discounted, Year 2 costs should be discounted by 1 year, and so on. Calculated in this way, the previous example is

$$P = \sum_{n=0}^{2} F_n (1+r)^{-n} = F_0 + \frac{F_1}{(1+r)} + \frac{F_2}{(1+r)^2}$$

The present value of A = 28.13 and the present value of B = 28.15.

The factor $(1 + r)^{-n}$ is known as the *discount factor* and can be obtained for a given n and r from Table A7.2.1 in Annex 7.2. For example, the discount factor for three periods (years) at a discount rate of 5% is 0.8638.

7.2.3 Annuitization and equivalent annual cost

The approach described in Section 7.2.2 is the most convenient for many programme comparisons, but a more common situation is that where most of the costs are easily

expressed on an annual recurring basis and it is only capital costs which differ from year to year (typically these will be at the beginning of the programme, or Year 0). Here it might be more convenient to express all the costs on an annual basis, obtaining an *equivalent annual cost (E)* for the capital outlay by an amortization or annuitization procedure. This works as follows.

If the capital outlay is K, we need to find the annual sum E which over a period of n years (the life of the facility), at an interest rate of r, will be equivalent to K. This is expressed by:

$$K = \frac{E_1}{(1+r)} + \frac{E}{(1+r)^2} + \cdots + \frac{E}{(1+r)^n}$$

$$K = E\frac{1-(1+r)^{-n}}{r}$$

$$K = E\left[\text{Annuity factor, } n \text{ period, interest } r\right]$$

As before, the annuity factor is easily obtainable from Table A7.2.2 in Annex 7.2. For example, in the cost analysis of providing long-term oxygen therapy, Lowson et al. (1981) found the total capital (set up) costs (K) to be £2153. Therefore, applying the formula given above,

$$2153 = \frac{E}{(1+r)} + \frac{E}{(1+r)^2} + \frac{E}{(1+r)^3} + \frac{E}{(1+r)^4} + \frac{E}{(1+r)^5}$$

$$2153 = E(\text{annuity factor, 5 years, interest rate 7\%})$$

$$2153 = E(4.1002) \text{ (from Table A7.2.2 in Annex 7.2)}$$

$$E = \text{£}525 \text{ (as shown in Table III of Lowson } et\ al. \text{ (1981).}$$

Note that Lowson et al. (1981) assumed that the annuity was in arrears, that is, due at the end of the year. It might be argued that a more realistic assumption is for it to be payable in advance. This is equivalent to the formula

$$2153 = E + \frac{E}{(1+r)} + \frac{E}{(1+r)^2} + \frac{E}{(1+r)^3} + \frac{E}{(1+r)^4}$$

The value for E can still be obtained from Table A7.2.2 by taking one less period and adding 1.000. This gives a lower value for E = £491. This is logical because the repayments are being made earlier (at the beginning of each year) rather than in arrears.

7.2.4 Useful life and resale value

This approach can be generalized to handle the situation where the equipment or buildings have a resale value at the end of the programme. If S is the resale value, n is the useful life of the equipment, r is the discount (interest rate), $A(n, r)$ is the annuity

factor (n years at interest rate r), K is the purchase price/initial outlay, and E is the equivalent annual cost, then

$$E = \frac{K - (S/(1+r)^n)}{A(n,r)}$$

The method described above is unambiguous for new equipment. For old equipment, there are two choices:

1 Use the replacement cost of the equipment or the original cost indexed to current dollars and a full life.

2 Use the current market value of the old machine and its remaining useful life.

Choice 1 is usually better as the results are more generalizable—less situational. Note that using the undepreciated balance from the accounts of the organization is never a method of choice.

It can be seen that the equivalent annual cost of buildings or equipment to a given programme depends on the values of n, r, and S, all of which must be assumed at the time of evaluation.

It is important to make a distinction between the physical life of a piece of equipment and its useful clinical life. The latter is highly dependent on technological change. Obviously one can undertake a sensitivity analysis using different values for n, but in general it is best to be conservative and assume a short life (say, around 5 years) for clinical equipment.

7.2.5 Choice of discount rate

Traditionally there have been *two competing theories* regarding the proper measure for the discount rate (r) for public projects (the social discount rate):

- r = the real rate of return (to society) forgone in the private sector (known as the social opportunity cost approach)—this can be estimated empirically, although not without controversy
- r = the social rate of time preference.

The theoretical basis for discounting and the determination of the discount rate was discussed at length in Chapter 4. In the context of a particular study, the relevant discount rate may be given in official guidelines for the conduct of economic evaluations in the jurisdiction concerned. (See ISPOR 2014 for details of the current methods guidance in various jurisdictions.) If an official rate is not given in the jurisdiction in which the study is being conducted, the best approach would be to conduct the analysis using rates existing in the literature, typically 3–5% per annum, since this would facilitate comparisons with other studies. The most common approach in the literature is to discount costs and benefits by the same rate.

7.2.6 How to handle inflation

If it is assumed that all the items of cost in the programme will inflate at the same rate and that this will be the same rate as inflation in general, there are two equivalent choices:

- Inflate all future costs by this predicted inflation rate and then use a larger discount rate that allows for the effect of general inflation (the inflation-adjusted

discount rate). For example, if the real discount rate is 5% and general inflation is 8%, then the inflation-adjusted $r = 1.05 \times 1.08 = 1.134$ or 13.4%

- Do not inflate any future costs (i.e. use constant dollars) and use a smaller discount rate that does not allow for inflation (the real discount rate).

The second method is the simpler and preferred approach. All the announced rates, and the rates recommended by analysts, are real rates.

If it is assumed that different items of cost in the programme will inflate at different rates, there are also two equivalent choices:

- Inflate all future costs by their particular predicted inflation rates and then use a larger discount rate that allows for the effect of general inflation (the inflation-adjusted discount rate).
- Do not inflate any future costs (i.e. use constant dollars) and use a smaller discount rate that does not allow for inflation (the real discount rate), but adjust the discount rate for each item to account for the differential inflation rate between this item and the 'general' rate of inflation, for example, if general inflation is 8%, this item is expected to inflate by 10%, and the real r is equal to 4%, then r adjusted for this item is

$$r = 1.04 \times \frac{1.08}{1.10} = 1.021, \text{ i.e. } 2.1\%$$

The second method is again the preferred approach. In general, however, most studies perform the whole analysis in constant price terms and use a single discount rate (see Annex 7.1 for a tutorial on methods of measuring and valuing capital costs).

7.3 Productivity changes

7.3.1 Role in policy

The extent to which productivity effects should be incorporated into economic evaluation and, if so, the method by which this is done, is an area of controversy in the field, not least in accounting for the productivity effects of the health care programme(s) likely to be displaced by adoption of the new treatment under consideration (see Chapter 4, Section 4.3.2). The relevant productivity changes that may be considered relevant are those arising from the patient or family member being unable to participate in work activities as a result of ill-health. The major debate about the role and estimation of productivity changes relates to their consideration as a major consequence of health care programmes. That is, as a result of treatment, the patient may be able to return to work or be more productive at work. Also, in the case of life-saving therapy, extension of life may also imply extension of the patient's working life. It is important to understand that there are different components to productivity effects. First, if individuals are in paid work then a proportion of their earned income will be used to finance a range of consumption activities for themselves and their families which may have some impact on their perceptions of the quality of life they associate with different levels of health. Secondly, a proportion of the individuals' remuneration will, mainly through taxation, benefit other individuals. These different elements have implications for the practicalities of how productivity effects are measured.

As was indicated earlier, the relevance of productivity changes depends on the perspective for the analysis. As discussed in Chapters 2 and 4, establishing the 'correct' perspective is not straightforward. One approach is to take the costs that are relevant to the decision-maker for whom the analysis is being undertaken, and this will typically be the costs falling on the budget for which they are responsible. This is why most guidelines for economic evaluation to support decision-making bodies specify the cost perspective of the health care systems and do not include productivity effects which impact more widely on the economy. However, defining a 'correct' perspective is fundamentally a normative question about defining what should count as a source of social value in decisions in public policy. As discussed in Chapter 2, this is a contested area: although mainstream economics offers the tools of welfarism to guide social decisions, these are driven by individual preferences and imply some strong value judgements which are not universally accepted, particularly in the field of health care where extra-welfarist approaches to evaluation predominate. Despite this, productivity effects are recommended for incorporation into economic evaluations in some countries' guidelines, where preferred methods otherwise focus on health effects using cost-effectiveness methods—for example, in Sweden and the Netherlands.

The inclusion of productivity changes, either as costs or consequences, is, therefore, contentious. The issue might arise as follows. Suppose one were evaluating two programmes in the field of mental health. Existing practice requires institutionalization of the patient for a given period; a new alternative is a community-based programme using community psychiatric nurses in association with outpatient hospital visits, resulting in patients remaining in their own homes. (For simplicity, let us assume that the programmes turn out to be equivalent in their impact on health, as assessed by some agreed measure of clinical symptomatology.)

Suppose it turns out that the community care programme has higher costs to the health care system, but that the number of work days lost by the cohort of patients on the community regimen is lower, as many more of them can remain at work. Would it be right to deduct these production gains from the higher health care costs of the community care programme? If so, how would the production gains be valued?

One might take the view that the production gains should be included in the analysis, since in principle there is no difference between these resource savings and any of the other labour inputs included in the health care cost estimates. Although the approach followed above is quite defensible, it gives rise to a number of wider considerations that should be noted. First, the approach assumes that the community loses production if the institutional-based programme removes patients from employment. However, it may be that, given a pool of unemployed labour, the jobs vacated by patients admitted to institutional care would be filled by other members of the community. If this were the case there may be few overall production gains from adopting the community care programme. Second, the community care programme imposes additional costs on the health care budget which, assuming it is fixed, will displace other activities. This will be expected to have negative health consequences in other types of patients and can be reflected in the cost-effectiveness threshold (see Chapter 4). Furthermore, these opportunity costs may include productivity effects which also need to be factored in. In other words, if productivity effects are important with respect to a new programme, they are also important in terms of opportunity costs.

It may be that, at some later stage, the cost-effectiveness estimates obtained in this study are compared with those obtained in other fields of health care, say a community care programme for people with learning difficulties or for elderly people. Because the patients benefiting from these programmes are unlikely to be in employment, there is less potential for production gains. This would make the community care programme for mental illness patients seem relatively inexpensive in terms of net cost, particularly if it were for workers earning high incomes, such as business executives, psychiatrists, or, dare we say, economists! Thus, in making a choice on the basis of net cost-effectiveness estimates, decision-makers may be tacitly accepting priorities different from their stated ones—if these are for the care of elderly people or those with learning difficulties.

7.3.2 Estimating productivity costs

There are at least four concerns about the inclusion of productivity changes in evaluations undertaken from the broader 'societal' perspective. The first concern is related to the *estimation* of changes in productivity. As mentioned above, these are typically estimates using the gross earnings (including employment overheads and benefits) of those in employment. (In the literature, this is known as the *human capital* approach.) Also, some studies impute an equivalent value for those not in paid employment (e.g. homemakers) by one of a number of methods. These include the use of average wages, the cost of replacing the role fulfilled by the individual, or the opportunity cost of the production they could have contributed were they not at home.

However, it is frequently argued that these valuations overestimate the true cost to society if individuals were to be taken out of the workforce, either through illness or to receive health care. For example, for short-term absences, losses in production could be compensated for by the worker on their return to work, or by colleagues. Also, for many categories of worker the value of the productivity lost at the margin is likely to be lower than the average wage, on the grounds that all jobs contain tasks that are more or less important, and it is the less important ones that are usually forgone as a result of a short period of absence. Finally, for long-term absences the employer is likely to hire a replacement worker. Therefore, the amount of productivity lost depends on the time and cost of organizing the replacement, and the resulting adjustments in the economy more generally. That is, if the President gets sick, sooner or later one person will be removed from the ranks of the unemployed!

We should note that many of these points arose in the context of the valuation of health care costs above. Namely, it was argued that costs or savings at the margin may not be reflected by average costs, and that there are frequently costs or inefficiencies associated with changes in resource allocation. For example, the closure of a large mental illness institution cannot take place overnight and there may be times during the closure process when wards are underoccupied.

In the context of productivity losses, Koopmanschap et al. (1995) have proposed that these should be estimated by the *friction cost method*. The basic idea is that the amount of production lost due to disease depends on the time span organizations need to restore the initial production level. This friction period is likely to differ by location, industry, firm, and category of worker. For example, it may only take half a day to train

Table 7.5 Human capital versus friction costs: Netherlands (1988, billions of guilders)

Cost category	Human capital	Friction costs
Absence from work	23.8	9.2
Disability	49.1	0.15
Mortality	8.0	0.15
Total	89.9	9.5
Percentage of net national income	18%	2.1%

a replacement hamburger server for a fast-food chain, but at least two days to train a replacement health economist!

The challenge is therefore to estimate the relevant friction periods and some calculations have been made for the Netherlands (Koopmanschap et al. 1995; Koopmanschap and Rutten 1996). These give estimates of lost production much lower than those obtained from traditional methods, such as the human capital approach (see Table 7.5). Also, Goeree et al. (1999) compared the human capital and friction cost approaches in estimating the productivity costs due to premature mortality from schizophrenia in Canada. The estimates using the human capital approach were 69 times higher than those obtained using the friction cost approach. In a sample of 40 studies, Pritchard and Sculpher (2000) found that the vast majority (26 out of 40) had estimated productivity costs using the human capital method, whereas only 7 had used the friction cost method. However, irrespective of the chosen approach, questionnaires have been developed to estimate productivity changes more precisely (e.g. Reilly et al. 1993; van Roijen et al. 1996).

A further measurement complication is that productivity may be lost even though the worker remains at work. This is often called 'presenteeism' and has been argued to be a major proportion of the productivity lost through mood disorders. Several questionnaires have been developed to estimate the productivity losses associated with presenteeism, but this issue remains controversial (Despiegel et al. 2012).

The second concern relates to *double counting*, especially in relation to productivity gains. If the value of improved health estimated in a given study already includes the value of the increased productivity that would result, then it would not be appropriate to include an additional estimate of the value of this item. This is most likely to be a problem in the case of the two forms of evaluation yet to be discussed, cost–utility and cost–benefit analysis. Here health state scenarios are presented to individuals for valuation, either in utility or monetary terms. Unless specifically told to ignore the impact that return to work would have on their income, respondents may factor this into their response.

This was a concern of the Public Health Service Panel on Cost-effectiveness in Health and Medicine (Gold et al. 1996). The panel's recommendation that the impact

of productivity gains was best captured in the denominator of the cost-effectiveness ratio gave rise to a lively debate (Brouwer et al. 1997a; Weinstein et al. 1997). In the final exchange in the debate, Brouwer et al. (1997b) argued that, even if individuals did consider income when assessing the value to them of improved health, individual income may only have a weak link with production change, particularly in settings where individuals have protection against loss of income (e.g. through sickness benefit payments) or where they experience reduced productivity while remaining at work.

Whether or not respondents take account of productivity effects in their valuations of health states is still a matter for debate. Some instruments for valuing health states explicitly ask respondents not to take into account the impact on ability to work, by asking them to imagine that their income would be protected by unemployment insurance. However, many instruments are silent on this matter. Tilling et al. (2010) argue that avoiding mentioning income effects in health state valuations may induce a minority of respondents to include them, but the impacts on the QALY estimates are minor. In a further study, where respondents used the time trade-off method to value EQ-5D states, explicit inclusion or exclusion of income effects had little impact on the health state valuations (Tilling et al. 2012).

Therefore the best approach would be to estimate the value of improved health while asking individuals to ignore income effects and then to estimate productivity changes separately, for inclusion in the numerator of the cost-effectiveness ratio. See Currie et al. (2002) and Sculpher (2001) for more discussion of the double counting issue and the estimation of productivity changes.

The third concern relates to the issue of *objectives and perspective* in the use of economic evaluation. For example, an analyst following the extra-welfarist approach (see Chapters 2 and 4) would argue that when the measure of benefit in an economic evaluation is health specific (e.g. life-years gained, or QALYs gained), the opportunity cost of scarce health care resources is defined in terms only of health forgone. It then follows that the opportunity cost of interest in the context of cost-effectiveness analysis or cost–utility analysis is determined by the best alternative use of small increases to total health care budgets and not opportunity costs elsewhere in the economy. It would thus not be relevant to include productivity changes, or indeed other non-health care costs such as patients' time, volunteer time, and costs falling on other agencies.

The fourth concern is the one hinted at in the discussion of the mental health programmes above; namely that the inclusion of productivity changes in an evaluation raises *equity* considerations. This takes us back to the different perspectives on the role of economic evaluation in health care, as discussed in Chapter 2. As mentioned in the discussion of non-health care costs in future years (above), the extent to which productivity changes are included in the analysis may depend on the view one takes about equity. Olsen and Richardson (1999) argue that the value of productivity effects may be included to the extent that it results in increased resources being made available for health care.

Other ways the problem might be alleviated are:

- expressing productivity changes as the number of days of work or normal activity lost or gained, rather than the monetary amount

- using a general wage rate to value productivity changes, rather than the actual wages of individuals affected by the health programme being evaluated.

Given the controversy surrounding the inclusion and estimation of productivity changes, we would suggest the following approach.

1 Report productivity changes separately so that the decision-maker can make a decision on whether or not to include them.
2 Report the quantities (in days of work, or normal activity lost or gained) separately from the prices (e.g. earnings) used to value the quantities. (This mirrors the recommendation made earlier for costing.)
3 Consider whether earnings adequately reflect the value of lost production at the margin and whether an approach based on the adjustments necessary to restore productivity (e.g. the friction approach) would be more valid.
4 Pay attention to the equity implications of the inclusion of productivity changes, and, where equity concerns are important, continue to conduct the base-case analysis using the actual estimates of the impact of the programme; but also consider a sensitivity analysis to explore the impact of using more equitable estimates, for example, a general wage rate rather than age-, gender-, or disease-specific rates.
5 Consider whether the inclusion of productivity changes represents double counting. (As indicated above, this is particularly pertinent when undertaking a cost–utility or cost–benefit analysis, but less likely when the effectiveness measure does not incorporate any *valuation* of the health consequences.)
6 Take account of any official guidelines for conducting economic evaluation existing in the jurisdiction concerned.

7.4 Exercise: costing alternative radiotherapy treatments

7.4.1 The task

A clinical trial is being carried out comparing two forms of radiotherapy for patients with head and neck cancer and carcinoma of the bronchus. Patients receiving *conventional therapy* are treated once per day, 5 days per week, for about 6 weeks. They normally travel to a hospital-based radiotherapy centre daily to receive care. Patients receiving *continuous hyperfractionated accelerated radiotherapy* (CHART) are treated three times on each of 12 consecutive days, including the weekend. Because of the intensity and frequency of treatment, patients normally stay in hospital during therapy, either in a regular hospital ward or in a hostel owned by the hospital.

The different treatment regimens obviously give rise to different costs. However, in addition, there may be differences in the period following treatment for the following reasons:

- The higher intensity of the CHART regimen might give rise to more side effects, and hence a greater need for community care after hospital discharge.
- The CHART regimen might give better tumour control, thereby slowing down the progression of the disease.

- CHART might reduce the extent of late radiation changes, and a lower incidence of necrosis may also reduce the need for salvage surgery.

The clinical trial will provide an opportunity to gather data on the use of resources by patients in the two treatment groups. You are asked to:

- *identify* which categories of resource you feel it would be important to assess
- indicate how you might *measure* the use of these resources in physical units
- say how you might *value* the resource consumption in money terms.

7.4.2 Identification of resource categories

Resource use can be considered under the broad headings outlined in Chapters 2 and 3. (See Table 7.6.)

7.4.3 Measurement of resource use

The fact that a clinical trial is taking place greatly increases the opportunity for accurate data collection, as case report forms are completed for patients enrolled in the trial. Normally these record data on clinical events, but they can be modified to include resource use, such as number and type of investigations, and date of hospital admission and discharge. Also, the fact that patients are enrolled in a trial provides the opportunity to interview them about resource use in community care, time taken to travel to hospital, and personal expenditure. They can also be given diary cards to record expenditure or time spent by relatives in home nursing.

In the absence of a trial the two major sources of data on resource use are routine statistics kept at the hospital or by other agencies, and patients' case notes (charts). The quality of these records varies by agency, and data are usually more comprehensive at the main place (clinic) where the patient is being treated. In addition, there are no routine records for patient and family resource use.

Turning to the specific resource items identified above, we might expect to record quantities used as outlined in Table 7.7.

Table 7.6 Categories of resource use

Health care resource use	Categories
Hospital resources	Radiotherapy, bed days, outpatient attendances, overheads
Community care resources	GP visits, nurse visits (types of nurse will vary by country or setting), ambulance or hospital car
Patient and family resource use	Patients' time, time of relatives, out-of-pocket expenses for transport (public or private)
Resource use in other sectors	Social worker visits, home help (homemaker) visits

Table 7.7 Measurements of resource quantities

Item	Possible measurements
(a) Hospital care	
Radiotherapy	The number of treatment sessions could be recorded, possibly differentiating by length of session and time of day (e.g. normal working hours, after hours, weekends)
Bed days	The number of bed days could be recorded, differentiating by type of hospital ward
Outpatient attendances	The number of attendances could be recorded
Overheads	These would probably be related to the number of bed days or other suitable resource item
(b) Community care	
GP visits	The number could be ascertained, either by asking patients, or by consulting the GP. It may make sense to differentiate between home visits and visits to the GP's office
Nurse visits	The number could be recorded as for GP visits above. The purpose of the nurse visit and type of nurse (e.g. general nurse, specialist cancer nurse) would be recorded
Ambulance and hospital car	The number and length of trips could be recorded. Length of trip could be ascertained from the patient's place of residence
(c) Patient and family resources	
Patients' time	The time taken in seeking and receiving care could be estimated by asking the patient. Time off work could be estimated separately
Relatives' time	Relatives could spend time in home nursing and in accompanying patients to hospital. It could be estimated as for patients' time above
Out-of-pocket expenses	Some may be estimated directly in money terms (e.g. bus fares). Others may be estimated by asking patients (e.g. distance travelled in private car)
(d) Resources in other sectors	
Social worker and home help visits	These would be estimated in a similar way to nurse visits above

GP, general practitioner (family doctor).

7.4.4 Valuation of resource items

It is extremely difficult to give general advice on this because it is so dependent on the availability of local financial data. In some settings, such as the United States, there may be data on hospital billings or charges. In other settings, detailed costing studies would be necessary. As mentioned elsewhere in this chapter, when using charge data it is important to:

- investigate the relationship between charges and costs
- record physical quantities as well as charges, so as to facilitate generalization of study results to other settings.

The general strategies for costing, ranging from the use of average costs (or per diems) to micro-costing, are outlined in Box 7.4. The skill in costing is to match the level of precision (and effort) to the importance (in quantitative terms) of the cost item. Turning to the specific resource items measured above, we might expect to value them as follows.

7.4.4.1 Hospital care

Radiotherapy treatment sessions In some settings there may be charge data, or average cost figures, for radiotherapy sessions. However, even if these exist, which is unlikely in many locations, they may not differentiate by type of session (e.g. normal hours, out-of-hours, or weekend). This distinction is critical to understanding the relative costs of conventional radiotherapy and CHART. Therefore, it is likely that micro-costing would be required.

In micro-costing the approach would be to derive the cost of a treatment session from its component parts, namely consultant (medical) time, radiographer time, medical physics time, consumables, equipment, buildings, and departmental overheads. Some survey work may be required, plus data from the hospital finance department on staff salaries, overtime allowances, and equipment prices. Costing of equipment and buildings will require assumptions to be made about useful life and resale value. It would be necessary to express these costs first as *equivalent annual costs* (see the methods outlined in Section 7.2) and then to apportion them to individual treatment sessions. Judgements would also need to be made about which components of hospital overheads (e.g. cleaning, building maintenance, or administration) are most appropriately allocated to departments and the allocation basis (e.g. square metres, cubic metres, number of staff, and so on). Some elements of overhead may be better allocated on the basis of in-patient days or number of patients.

Bed days It may be possible to use the average daily costs (or per diems) for different types of wards, including hostel wards. However, these may be considered too imprecise, in which case micro-costing might be undertaken. This would derive a daily cost for a particular category of ward by considering nurse staffing levels, medical (consultant) input, and overheads.

Because hostel wards may not feature in the standard hospital accounts, micro-costing may be required for these: for example, they may be slightly off site or rely

partly on staffing by volunteers. An opportunity cost for volunteer time may have to be imputed. In costing hospital beds it may be decided to make an allowance for the fact that there is usually less than 100% occupancy.

Outpatient attendances There may be an average cost or charge available for an outpatient visit, although this may not differentiate between oncology and other clinical specialties. Depending on the quantitative importance of this item, micro-costing may be undertaken.

Overheads As mentioned above, these could be allocated to the radiotherapy treatments, to outpatient attendances, or to hospital bed days, depending on the overhead item.

7.4.4.2 Community care

General practitioner visits Data may be available on physician fees for various types of visit (e.g. general assessment, home visit, etc.). Alternatively, there may be nationally available data on the average costs of various general practitioner services. Failing this, micro-costing may be required. This would calculate the cost of practitioners' time (per minute or per hour) and add the cost of travel for home visits. Drug costs would also need to be considered.

Nurse visits The agencies providing the nurses may have data on the average cost of a visit. This may even distinguish between various types of visit. Failing this, micro-costing would have to be employed, taking into account nursing salaries, length of visits, travel time, and nurses' time spent in general administration. There may also be some consumables to be accounted for in the cost of nurse visits.

Ambulance and hospital car Estimates may be available for the average cost per mile travelled. This could be combined with data on the distances involved to generate total costs.

7.4.4.3 Patient and family resources

Patients' time If the time was taken from work time, the gross salary (including employment benefits) could be used. Different assumptions could be made about the opportunity cost of leisure time.

Relatives' time In general the valuation of this raises the same issues as the valuation of patients' time. The valuation of time spent in informal nursing care is complicated because the relative may also be able to carry out other tasks at the same time.

Out-of-pocket expenses In general, the financial expenditures made (e.g. public transport, hospital car parking charges) would suffice. However, for some items, such as use of one's private car, the expenditures (say) on fuel would underestimate the true cost. Motoring organizations can often provide data on the cost (per mile or kilometre) of running a car.

Finally, a few rare events, such as hospital admission for particular types of surgery, may be handled separately. Depending on how quantitatively important they seem,

case-mix group costs or disease-specific per diems may suffice. Alternatively, micro-costing may be undertaken.

7.5 Concluding remarks

Cost analysis is a central feature of all economic evaluations, but it has received relatively little attention from analysts to date. Two reviews of the literature (Graves et al. 2002; Halliday and Darba 2003) have commented on the inadequacies of current practice. Deficiencies exist in the specification of the study perspective, the estimation of both quantities and prices, and even the identification of the year(s) to which costs apply. Therefore, users of economic evaluations would be wise to subject the costing methods to a fair degree of scrutiny. In particular they should be suspicious of any study that does not clearly state separately the sources and methods of estimating the quantities and prices used in the calculation of costs.

References

Black, A.J., Roubin, G.S., Sutor, C., et al. (1988). Comparative costs of percutaneous transluminal coronary angioplasty and coronary bypass grafting in multivessel coronary artery disease. *American Journal of Cardiology*, **62**, 809–11.

Brouwer, W., Koopmanschap, M., and Rutten, F. (1997a). Productivity costs measurement through quality of life? A response to the recommendation of the Washington Panel. *Health Economics*, **6**, 253–9.

Brouwer, W., Koopmanschap, M., and Rutten, F. (1997b). Productivity costs in cost-effectiveness analysis: numerator or denominator: a further discussion. *Health Economics*, **6**, 511–4.

Brouwer, W., Rutten, F., and Koopmanschap, M. (2001). Costing in economic evaluations, in M.F. Drummond and A. McGuire (ed.), *Economic evaluation in health care: merging theory and practice*, pp. 68–93. Oxford: Oxford University Press.

Bush, J.W. (1973). Discussion, in R.L. Berg, (ed.), *Health status indexes*, p. 203. Chicago, IL: Hospital Research and Educational Trust.

Busse, R., Schreyögg, J., and Smith, P. (2008). Analysing the variation of health care treatment costs in Europe. *Health Economics*, **17**(S1), S1–S104.

Byford, S., Knapp, M., Greenshields, J., et al. (2003). Cost-effectiveness of brief cognitive behaviour therapy versus treatment as usual in recurrent deliberate self-harm: a decision-making approach. *Psychological Medicine*, **33**, 977–86.

Chambers, D., Paulden, M., Paton, F., et al. (2010). Sugammadex for the reversal of muscle relaxation in general anaesthesia: a systematic review and economic assessment. *Health Technology Assessment*, **14**(39), 1–211.

Cohen, D.J., Breall, J.A., Kalon, K.L.H., et al. (1993). Economics of elective coronary revascularization: comparison of costs and charges for conventional angioplasty, directional atherectomy, stenting and bypass surgery. *Journal of the American College of Cardiology*, **22**, 1052–9.

Currie, G.G., Donaldson, C., O'Brien, B.J., Stoddart, G.L., Torrance, G.W., and Drummond, M.F. (2002). Willingness-to-pay for what? A note on alternative definitions of health care program benefits for contingent valuation studies. *Medical Decision-Making*, **22**, 493–7.

Daly, E., Roche, M., Barlow, D., Gray, A., McPherson, K., and Vessey, M. (1992). HRT: an analysis of benefits, risks and costs. *British Medical Bulletin*, **48**, 368–400.

Despiegel, N., Danchenko, N., Francois, C., Lensberg, B. and Drummond, M.F. (2012). The use and performance of productivity scales to evaluate presenteeism in mood disorders. *Value in Health*, **15**, 1148–61.

Drummond, M.F., McGuire, A.L., and Fletcher, A. (1993). *Economic evaluation of drug therapy for hypercholesterolaemia in the United Kingdom*, Discussion Paper 104. York: Centre for Health Economics, University of York.

Drummond, M.F., Griffin, A., and Tarricone, R. (2009). Economic evaluation for devices and drugs: same or different? *Value in Health*, **12**, 402–4.

Faria, R., Bojke, L., Epstein, D., Corbacho B., and M. Sculpher; REFLUX trial group (2013). Cost-effectiveness of laparoscopic fundoplication *versus* continued medical management for the treatment of gastro-oesophageal reflux disease based on long-term follow-up of the REFLUX trial. *British Journal of Surgery*, **100**, 1205–13.

Garrison, L., Mansley, E.C., Abbott, T.A., Bresnahan, B.W., Hay, J.W., and Smeeding, J. (2010). Good research practices for measuring drug costs in cost-effectiveness analyses: a societal perspective: the ISPOR Drug Cost Task Force Report—Part II. *Value in Health*, **12**, 8–13.

Goeree, R., O'Brien, B.J., Blackhouse, G., Agro, K., and Goering, P. (1999). The valuation of productivity costs due to premature mortality: a comparison of the human-capital and friction-cost methods for schizophrenia. *Canadian Journal of Psychiatry*, **44**, 455–63.

Gold, M.R., Siegel, J.E., Russell, L.B., and Weinstein, M.C. (ed.) (1996). *Cost-effectiveness in health and medicine*. New York: Oxford University Press.

Graves, N., Walker, D., Raine, R., Hutchings, A., and Roberts, J.A. (2002). Cost data for individual patients included in clinical studies: no amount of statistical analysis can compensate for inadequate costing methods. *Health Economics*, **11**, 735–9.

Halliday, R.G. and Darba, J. (2003). Cost data assessment in multinational economic evaluations: some theory and review of published studies. *Applied Health Economics and Health Policy*, **2**, 149–55.

Hay, J.W., Smeeding, J., Carroll, N.V., et al. (2010). Good research practices for measuring drug costs in cost-effectiveness analyses: issues and recommendations: the ISPOR Drug Cost Task Force Report—Part 1. *Value in Health*, **12**, 3–7.

Henderson, R.A., Pocock, S.J., Sharp, S.J., et al. (1998). Long-term results of RITA-1 trial: clinical and cost comparisons of coronary angioplasty and coronary-artery bypass grafting. *Lancet*, **352**, 1419–25.

Hoyle, M. (2011). Accounting for the drug life cycle and future drug prices in cost-effectiveness analysis. *PharmacoEconomics*, **29**, 1–15.

ISPOR [International Society for Pharmacoeconomics and Outcomes Research] (2014). *Pharmacoeconomic guidelines around the world*. <http://www.ispor.org/PEguidelines/index.asp> (Accessed 8 December 2014).

Johannesson, M., Meltzer, D., and O'Conor, R.M. (1997). Incorporating future costs in medical cost-effectiveness analysis: implications for the cost-effectiveness of the treatment of hypertension. *Medical Decision Making*, **17**, 382–9.

Koopmanschap, M.A. and Rutten, F.F.H. (1996). Indirect costs: the consequence of production loss or increased costs of production. *Medical Care*, **34** (suppl.), DS59–68.

Koopmanschap, M.A., Rutten F.F.H., van Ineveld, B.M., and van Roijen, L. (1995). The friction cost method for measuring indirect costs of disease. *Journal of Health Economics*, **14**, 171–89.

Koopmanschap, M.A., van Exel, J.N.A., van den Berg, B., and Brouwer, W.B.F. (2008). An overview of methods and applications to value informal care in economic evaluations of healthcare. *PharmacoEconomics*, **26**, 269–80.

Lowson, K.V., Drummond, M.F., and Bishop, J.M. (1981). Costing new services: long-term domiciliary oxygen therapy. *Lancet*, **i**, 1146–9.

Manns, B., Melzer, D., Taub, K., and Donaldson, C. (2003). Illustrating the impact of including future costs in economic evaluations: an application to end-stage renal disease care. *Health Economics*, **12**, 949–58.

Meltzer, D. (1997). Accounting for future costs in medical cost-effectiveness analysis. *Journal of Health Economics*, **16**, 33–64.

Mentzakis, E., Ryan, M., McNamee, P. (2011). Using discrete choice experiments to value informal care tasks: exploring preference heterogeneity. *Health Economics*, **20**, 930–44.

Neuhauser, D. and Lewicki, A.M. (1975). What do we gain from the sixth stool guaiac? *New England Journal of Medicine*, **293**, 226–8.

Nigrovic, L.E. and Chiang, V.W. (2000). Cost analysis of enteroviral polymerase chain reaction in infants with fever and cerebrospinal fluid pleocytosis. *Archives of Pediatrics and Adolescent Medicine*, **154**, 817–21.

Olchanski, N., Zhong, V., Saret, C., Cohen, Y.T., Bala, M., and Neumann, P.J. (2013). Economics of costs in added years of life: a review of methodologic practices and consequences for cost-effectiveness. *Value in Health*, **16**, A19.

Olsen, J.A. and Richardson, J. (1999). Production gains for health care: what should be included in cost-effectiveness analysis? *Social Science and Medicine*, **49**, 17–26.

Pritchard, C. and Sculpher, M. (2000). *Productivity costs: principles and practice in economic evaluation*. London: Office of Health Economics.

Reilly, M.C., Zbrozek, A.S., and Dukes, E.M. (1993). The validity and reproducibility of a work productivity and activity impairment instrument. *PharmacoEconomics*, **4**, 353–65.

Schulman, K., Burke, J., Drummond, M.F., et al. (1998). Resource costing for multinational neurologic trials. *Health Economics*, **7**, 629–38.

Schulman, K.A., Glick, H.A., Rubin, H., and Eisenberg, J.M. (1991). Cost-effectiveness of HA-lA monoclonal antibody for gram-negative sepsis. *JAMA*, **266**, 3466–71.

Sculpher, M.J. (2001). The role and estimation of productivity costs in economic evaluation, in M.F. Drummond and A. McGuire (ed.), *Economic evaluation in health care: merging theory with practice*, pp. 94–112. Oxford: Oxford University Press.

Sculpher, M.J., Seed, P., Henderson, R.A., Buxton, M.J., Pocock, S.J., and Parker, J. (1993). Health service costs of coronary angioplasty and coronary artery bypass surgery: the randomized intervention treatment of angina (RITA) trial. *Lancet*, **244**, 927–30.

Sherry, K.M., McNamara, J., Brown, J.S., and Drummond, M.F. (1996). An economic evaluation of propofol/fentanyl compared with midazolam/fentanyl on recovery of the ICU following cardiac surgery. *Anaesthesia*, **51**, 312–7.

Taira, D.A., Seto, T.B., Siegrist, R., Cosprove, R., Berezin, R., and Cohen, D.J. (2003). Comparison of analytic approaches for the economic evaluation of new technologies alongside multicenter clinical trials. *American Heart Journal*, **145**, 452–8.

Tilling, C., Krol, M., Tsuchiya, A., Brazier, J., and Brouwer, W. (2010). In or out? Income losses in health state valuations: a review. *Value in Health*, **13**, 298–305.

Tilling, C., Krol, M., Tsuchiya, A., Brazier, J., van Exel, J., and Brouwer, W. (2012). Does the EQ-5D reflect lost earnings? *PharmacoEconomics*, **30**, 47–61.

van den Berg, B. and Spauwen, P. (2006). Measurement of informal care: an empirical study into the valid measurement of time spent on informal caregiving. *Health Economics*, **15**, 447–60.

van den Berg, B. and Ferrer-i-Carbonell, A. (2007). Monetary valuation of informal care: the well-being valuation method. *Health Economics*, **16**, 1227–1244.

van den Berg, B., Fiebig, D.G., and Hall, J. (2014). Well-being losses due to care-giving. *Journal of Health Economics*, **35**, 123–31.

van Roijen, L., Essink-Bot, M.L., Koopmanschap, M.A., Bonsel, G., and Rutten, F.F.H. (1996). Labor and health status in economic evaluation of health care. *International Journal of Technology Assessment in Health Care*, **12**, 405–15.

Weatherly, H., Faria, R., and van den Berg, B. (2014). Valuing informal care for caregiving, in A.J. Culyer (ed.), *Elsevier encyclopaedia of health economics*. San Diego, CA: Elsevier.

Weinstein, M.C. and Manning, W.G. (1997). Theoretical issues in cost-effectiveness analysis (editorial). *Journal of Health Economics*, **16**, 121–8.

Weinstein, M.C., Seigel, J.E., Garber, A.M., et al. (1997). Productivity costs, time costs and health-related quality of life: a response to the Erasmus Group. *Health Economics*, **6**, 505–10.

Zupancic, J.A.F., Richardson, D.K., O'Brien, B.J., Eichenwald, E.C., and Weinstein, M.C. (2003). Cost-effectiveness analysis of predischarge monitoring for apnea of prematurity. *Pediatrics*, **111**, 146–52.

Annex 7.1 Tutorial on methods of measuring and valuing capital costs

The examples given here should help to clarify the treatment of capital costs.

As a first note, we need to distinguish two classes of 'capital'—land and equipment. This is an important consideration, because in costing exercises we assume land does not depreciate, while of course capital equipment does. We can think of there being a continuum along which materials and supplies depreciate or are used up instantaneously and so are costed fully in the year of use; capital equipment depreciates more slowly and may be handled in a variety of ways; land does not depreciate at all.

As a second note, recall that capital equipment costs have three components—depreciation, opportunity cost, and actual operating costs. We will ignore the last of these here.

First, consider equipment, and let us use an example of a machine costing $200 000 that, at the end of 5 years, has a resale value of $20 000. Assume straight-line depreciation and a discount rate of 4%. There are, then, four approaches to costing:

1 We can assume all costs accrue at time 0. This amounts to treating the equipment as one would less durable materials and supplies (Table A7.1.1). Alternatively, but equivalently, one can treat the machine as instantaneously depreciating, except for the $20 000 resale value, which then is maintained through the 5 years (Table A7.1.2).

2 We can compute depreciation and opportunity costs separately. They are related in that the opportunity cost of equipment refers to the use of the resources embodied in the equipment, in their next best use—this is 'approximated' by calculating the return on the funds implicit in the undepreciated value of the equipment at each point in time. Hence, the higher the rate of depreciation, the lower the opportunity cost, all else equal. Again, one has the choice of building the $20 000 resale in at the end, or just depreciating less of the machine. It works out the same (Tables A7.1.3 and A7.1.4).

3 We can compute an equivalent annual cost (E). This may be useful in a situation where other operating costs are the same each year, making necessary the

Table A.7.1.1 Costs all assumed to occur at time zero

Time	0	1	2	3	4	5
Depreciation	200 000	0	0	0	0	(20 000)
Undepreciated balance at beginning of period	0	0	0	0	0	
Opportunity cost	0	0	0	0	0	
Depreciation + opportunity cost	200 000	0	0	0	0	(20 000)
Present value (PV)	200 000	0	0	0	0	(16 439)
Net present value (NPV) of equipment cost = $183 561						

Table A7.1.2 Machine instantaneously depreciates

Time	0	1	2	3	4	5
Depreciation	180 000	—	—	—	—	—
Undepreciated balance at beginning of period		20 000	20 000	20 000	20 000	20 000
Opportunity cost		800	800	800	800	800
Depreciation + opportunity cost		800	800	800	800	800
PV	180 000	769	740	711	684	658
NPV of equipment cost = $183 562						

Table A7.1.3 Computing depreciation and opportunity costs separately

Time	1	2	3	4	5
Depreciation	36 000	36 000	36 000	36 000	36 000
Undepreciated balance at beginning of period	200 000	164 000	128 000	92 000	56 000
Opportunity cost	8000	6 560	5 120	3 680	2240
Depreciation + opportunity cost	44 000	42 560	41 120	39 680	38 240
PV	42 308	39 349	36 556	33 919	31 430
NPV of equipment cost = $183 562					

comparison of only a single year of cost data for each alternative in the economic evaluation:

$$\text{NPV} = E \times \text{AF}_{5.4\%}$$

where $\text{AF}_{5.4\%}$ is the annuity factor for 5 years at an interest rate of 4% (see Table A7.2.2 in Annex 7.2);

$$\$183\,562 = E \times 4.4518 \rightarrow E = \$41\,233.$$

In other words, an *equal* stream of costs amounting to $41 233 in *each* of the 5 years of the programme has a present value equivalent to any of the *unequal* cost streams in (1) or (2) above. Note, therefore, that the equivalent annual cost embodies both depreciation and opportunity cost.

Table A7.1.4 Computing depreciation and opportunity costs separately (resale value at end)

Time	1	2	3	4	5
Depreciation	40 000	40 000	40 000	40 000	20 000
Undepreciated balance at beginning of period	200 000	160 000	120 000	80 000	40 000
Opportunity cost	8000	6 400	4 800	3 200	1 600
Depreciation + opportunity cost	48 000	46 400	44 800	43 200	21 600
PV	46 154	42 899	39 827	36 928	17 754
NPV of equipment cost = $183 562					

Table A7.1.5 Cost stream for a land purchase of $200 000 at time zero

Time	1	2	3	4	5
Depreciation	—	—	—	—	—
Undepreciated balance at beginning of period	200 000	200 000	200 000	200 000	200 000
Opportunity cost	8 000	8 000	8 000	8 000	8 000
Depreciation + opportunity cost	8 000	8 000	8 000	8 000	8 000
PV	7 692	7 396	7 112	6 838	6 575
NPV = $35 613					

4 We can use equivalent or actual rental costs, if available or estimable. Note that because the renter will need to recover not only depreciation of the rental equipment but also a rate of return at least as good as that from the next best use of the resource, one can take rental cost to embody both depreciation and opportunity cost.

Second, the treatment of *land* is quite different because of the lack of depreciation. A land purchase of $200 000 at time 0 would generate the cost time stream shown in Table A7.1.5. Converted to an equivalent annual cost:

$$\text{NPV} = E \times \text{AF}_{5.4\%}$$

$$\$35\,613 = E \times 4.4518$$

It comes as no particular surprise that E = $8000!

Acknowledgement

Examples in Annex 7.1 reproduced courtesy of Morris Barer, University of British Columbia, Canada. We are indebted to Morris Barer for producing these examples, which should clarify the treatment of capital costs.

Annex 7.2 Discount tables

This annex contains discount tables for the present value of $1 (Table A7.2.1) and the present value of annuity of $1 in arrears (Table A7.2.2).

Table A7.2.1 Present value of $1

N	1%	2%	3%	4%	5%	6%	7%	8%	9%	10%	11%	12%	13%	14%	15%
1	0.9901	0.9804	0.9709	0.9615	0.9524	0.9434	0.9346	0.9259	0.9174	0.9091	0.9009	0.8929	0.8850	0.8772	0.8696
2	0.9803	0.9612	0.9426	0.9246	0.9070	0.8900	0.8734	0.8573	0.8417	0.8264	0.8116	0.7972	0.7831	0.7695	0.7561
3	0.9706	0.9423	0.9151	0.8890	0.8638	0.8396	0.8163	0.7938	0.7722	0.7513	0.7312	0.7118	0.6931	0.6750	0.6575
4	0.9610	0.9238	0.8885	0.8548	0.8227	0.7921	0.7629	0.7350	0.7084	0.6830	0.6587	0.6355	0.6133	0.5921	0.5718
5	0.9515	0.9057	0.8626	0.8219	0.7835	0.7473	0.7130	0.6806	0.6499	0.6209	0.5935	0.5674	0.5428	0.5194	0.4972
6	0.9420	0.8880	0.8375	0.7903	0.7462	0.7050	0.6663	0.6302	0.5963	0.5645	0.5346	0.5066	0.4803	0.4556	0.4323
7	0.9327	0.8706	0.8131	0.7599	0.7107	0.6651	0.6227	0.5835	0.5470	0.5132	0.4817	0.4523	0.4251	0.3996	0.3759
8	0.9235	0.8535	0.7894	0.7307	0.6768	0.6274	0.5820	0.5403	0.5019	0.4665	0.4339	0.4039	0.3762	0.3506	0.3269
9	0.9143	0.8368	0.7664	0.7026	0.6446	0.5919	0.5439	0.5002	0.4604	0.4241	0.3909	0.3606	0.3329	0.3075	0.2843
10	0.9053	0.8203	0.7441	0.6756	0.6139	0.5584	0.5083	0.4632	0.4224	0.3855	0.3522	0.3220	0.2946	0.2697	0.2472
11	0.8963	0.8043	0.7224	0.6496	0.5847	0.5268	0.4751	0.4289	0.3875	0.3505	0.3173	0.2875	0.2607	0.2366	0.2149
12	0.8874	0.7885	0.7014	0.6246	0.5568	0.4970	0.4440	0.3971	0.3555	0.3186	0.2858	0.2567	0.2307	0.2076	0.1869
13	0.8787	0.7730	0.6810	0.6006	0.5303	0.4688	0.4150	0.3677	0.3262	0.2897	0.2575	0.2292	0.2042	0.1821	0.1625
14	0.8700	0.7579	0.6611	0.5775	0.5051	0.4423	0.3878	0.3405	0.2992	0.2633	0.2320	0.2046	0.1807	0.1597	0.1413
15	0.8613	0.7430	0.6419	0.5553	0.4810	0.4173	0.3624	0.3152	0.2745	0.2394	0.2090	0.1827	0.1599	0.1401	0.1229
16	0.8528	0.7284	0.6232	0.5339	0.4581	0.3936	0.3387	0.2919	0.2519	0.2176	0.1883	0.1631	0.1415	0.1229	0.1069
17	0.8444	0.7142	0.6050	0.5134	0.4363	0.3714	0.3166	0.2703	0.2311	0.1978	0.1696	0.1456	0.1252	0.1078	0.0929

Table A7.2.1 (continued) Present value of $1

N	1%	2%	3%	4%	5%	6%	7%	8%	9%	10%	11%	12%	13%	14%	15%
18	0.8360	0.7002	0.5874	0.4936	0.4155	0.3503	0.2959	0.2502	0.2120	0.1799	0.1528	0.1300	0.1108	0.0946	0.0808
19	0.8277	0.6864	0.5703	0.4746	0.3957	0.3305	0.2765	0.2317	0.1945	0.1635	0.1377	0.1161	0.0981	0.0829	0.0703
20	0.8195	0.6730	0.5537	0.4564	0.3769	0.3118	0.2584	0.2145	0.1784	0.1486	0.1240	0.1037	0.0868	0.0728	0.0611
21	0.8114	0.6598	0.5375	0.4388	0.3589	0.2942	0.2415	0.1987	0.1637	0.1351	0.1117	0.0926	0.0768	0.0638	0.0531
22	0.8034	0.6468	0.5219	0.4220	0.3418	0.2775	0.2257	0.1839	0.1502	0.1228	0.1007	0.0826	0.0680	0.0560	0.0462
23	0.7954	0.6342	0.5067	0.4057	0.3256	0.2618	0.2109	0.1703	0.1378	0.1117	0.0907	0.0738	0.0601	0.0491	0.0402
24	0.7876	0.6217	0.4919	0.3901	0.3101	0.2470	0.1971	0.1577	0.1264	0.1015	0.0817	0.0659	0.0532	0.0431	0.0349
25	0.7798	0.6095	0.4776	0.3751	0.2953	0.2330	0.1842	0.1460	0.1160	0.0923	0.0736	0.0588	0.0471	0.0378	0.0304
26	0.7720	0.5976	0.4637	0.3607	0.2812	0.2198	0.1722	0.1352	0.1064	0.0839	0.0663	0.0525	0.0417	0.0331	0.0264
27	0.7644	0.5859	0.4502	0.3468	0.2678	0.2074	0.1609	0.1252	0.0976	0.0763	0.0597	0.0469	0.0369	0.0291	0.0230
28	0.7568	0.5744	0.4371	0.3335	0.2551	0.1956	0.1504	0.1159	0.0895	0.0693	0.0538	0.0419	0.0326	0.0255	0.0200
29	0.7493	0.5631	0.4243	0.3207	0.2429	0.1846	0.1406	0.1073	0.0822	0.0630	0.0485	0.0374	0.0289	0.0224	0.0174
30	0.7419	0.5521	0.4120	0.3083	0.2314	0.1741	0.1314	0.0994	0.0754	0.0573	0.0437	0.0334	0.0256	0.0196	0.0151
35	0.7059	0.5000	0.3554	0.2534	0.1813	0.1301	0.0937	0.0676	0.0490	0.0356	0.0259	0.0189	0.0139	0.0102	0.0075
40	0.6717	0.4529	0.3066	0.2083	0.1420	0.0972	0.0668	0.0460	0.0318	0.0221	0.0154	0.0107	0.0075	0.0053	0.0037
45	0.6391	0.4102	0.2644	0.1712	0.1113	0.0727	0.0476	0.0313	0.0207	0.0137	0.0091	0.0061	0.0041	0.0027	0.0019
50	0.6080	0.3715	0.2281	0.1407	0.0872	0.0543	0.0339	0.0213	0.0134	0.0085	0.0054	0.0035	0.0022	0.0014	0.0009

Table A7.2.2 Present value of annuity of $1 in arrears

N	1%	2%	3%	4%	5%	6%	7%	8%	9%	10%	11%	12%	13%	14%	15%
1	0.9901	0.9804	0.9709	0.9615	0.9524	0.9434	0.9346	0.9259	0.9174	0.9091	0.9009	0.8929	0.8850	0.8772	0.8696
2	1.9704	1.9416	1.9135	1.8861	1.8594	1.8334	1.8080	1.7833	1.7591	1.7335	1.7125	1.6901	1.6681	1.6467	1.6257
3	2.9410	2.8839	2.8286	2.7751	2.7232	2.6730	2.6243	2.5771	2.5313	2.4869	2.4437	2.4018	2.3612	2.3216	2.2832
4	3.9020	3.8077	3.7171	3.6299	3.5460	3.4651	3.3872	3.3121	3.2397	3.1699	3.1024	3.0373	2.9745	2.9137	2.8550
5	4.8534	4.7135	4.5797	4.4518	4.3295	4.2124	4.1002	3.9927	3.8897	3.7908	3.6959	3.6048	3.5172	3.4331	3.3522
6	5.7955	5.6014	5.4172	5.2421	5.0757	4.9173	4.7665	4.6229	4.4859	4.3553	4.2305	4.1114	3.9975	3.8887	3.7845
7	6.7282	6.4720	6.2303	6.0021	5.7864	5.5824	5.3893	5.2064	5.0330	4.8684	4.7122	4.5638	4.4226	4.2883	4.1604
8	7.6517	7.3255	7.0197	6.7327	6.4632	6.2098	5.9713	5.7466	5.5348	5.3349	5.1461	4.9676	4.7988	4.6389	4.4873
9	8.5660	8.1622	7.7861	7.4353	7.1078	6.8017	6.5152	6.2469	5.9952	5.7590	5.5370	5.3282	5.1317	4.9464	4.7716
10	9.4713	8.9826	8.5302	8.1109	7.7217	7.3601	7.0236	6.7101	6.4177	6.1446	5.8892	5.6502	5.4262	5.2161	5.0188
11	10.3676	9.7868	9.2526	8.7605	8.3064	7.8869	7.4987	7.1390	6.8052	6.4951	6.2065	5.9377	5.6869	5.4527	5.2337
12	11.2551	10.5753	9.9540	9.3851	8.8633	8.3838	7.9427	7.5361	7.1607	6.8137	6.4924	6.1944	5.9176	5.6603	5.4206
13	12.1337	11.3484	10.6350	9.9856	9.3936	8.8527	8.3577	7.9038	7.4869	7.1034	6.7499	6.4235	6.1218	5.8424	5.5831
14	13.0037	12.1062	11.2961	10.5631	9.8986	9.2950	8.7455	8.2442	7.7862	7.3667	6.9819	6.6282	6.3025	6.0021	5.7245
15	13.8651	12.8493	11.9379	11.1184	10.3797	9.7122	9.1079	8.5595	8.0607	7.6061	7.1909	6.8109	6.4624	6.1422	5.8474
16	14.7179	13.5777	12.5611	11.6523	10.8378	10.1059	9.4466	8.8514	8.3126	7.8237	7.3792	6.9740	6.6039	6.2651	5.9542
17	15.5623	14.2919	13.1661	12.1657	11.2741	10.4773	9.7632	9.1216	8.5436	8.0216	7.5488	7.1196	6.7291	6.3729	6.0472
18	16.3983	14.9920	13.7535	12.6593	11.6896	10.8276	10.0591	9.3719	8.7556	8.2014	7.7016	7.2497	6.8399	6.4674	6.1280
19	17.2260	15.6785	14.3238	13.1339	12.0853	11.1581	10.3356	9.6036	8.9501	8.3649	7.8393	7.3658	6.9380	6.5504	6.1982

Table A7.2.2 (continued) Present value of annuity of $1 in arrears

N	1%	2%	3%	4%	5%	6%	7%	8%	9%	10%	11%	12%	13%	14%	15%
20	18.0456	16.3514	14.8775	13.5903	12.4622	11.4699	10.5940	9.8181	9.1285	8.5135	7.9633	7.4694	7.0248	6.6231	6.2593
21	18.8570	17.0112	15.4150	14.0292	12.8212	11.7641	10.8355	10.0168	9.2922	8.6487	8.0751	7.5620	7.1016	6.6870	6.3125
22	19.6604	17.6580	15.9369	14.4511	13.1630	12.0416	11.0612	10.2007	9.4424	8.7715	8.1757	7.6446	7.1695	6.7429	6.3587
23	20.4558	18.2922	16.4436	14.8565	13.4886	12.3034	11.2722	10.3711	9.5802	8.8832	8.2664	7.7184	7.2297	6.7921	6.3988
24	21.2434	18.9139	16.9355	15.2470	13.7986	12.5504	11.4693	10.5288	9.7066	8.9847	8.3481	7.7843	7.2829	6.8351	6.4338
25	22.0232	19.5235	17.4131	15.6221	14.0939	12.7834	11.6536	10.6748	9.8226	9.0770	8.4217	7.8431	7.3300	6.8729	6.4641
26	22.7952	20.1210	17.8768	15.9828	14.3752	13.0032	11.8258	10.8100	9.9290	9.1609	8.4881	7.8957	7.3717	6.9061	6.4906
27	23.5596	20.7069	18.3270	16.3296	14.6430	13.2105	11.9867	10.9352	10.0266	9.2372	8.5478	7.9426	7.4086	6.9352	6.5135
28	24.3164	21.2813	18.7641	16.6631	14.8981	13.4062	12.1371	11.0511	10.1161	9.3066	8.6016	7.9844	7.4412	6.9607	6.5335
29	25.0658	21.8444	19.1885	16.9837	15.1411	13.5907	12.2777	11.1584	10.1983	9.3696	8.6501	8.0218	7.4701	6.9830	6.5509
30	25.8077	22.3965	19.6004	17.2920	15.3725	13.7648	12.4090	11.2578	10.2737	9.4269	8.6938	8.0552	7.4957	7.0027	6.5660
35	29.4086	24.9986	21.4872	18.6646	16.3742	14.4982	12.9477	11.6546	10.5668	9.6442	8.8552	8.1755	7.5856	7.0700	6.6166
40	32.8347	27.3555	23.1148	19.7928	17.1591	15.0463	13.3317	11.9246	10.7574	9.7791	8.9511	8.2438	7.6344	7.1050	6.6418
45	36.0945	29.4902	24.5187	20.7200	17.7741	15.4558	13.6055	12.1084	10.8812	9.8628	9.0079	8.2825	7.6690	7.1232	6.6543
50	39.1961	31.4236	25.7298	21.4822	18.2559	15.7619	13.8007	12.2335	10.9617	9.9148	9.0417	8.3045	7.6752	7.1327	6.6605

Chapter 8

Using clinical studies as vehicles for economic evaluation

8.1 Introduction to vehicles for economic evaluation

There is a long history of using a single clinical study as the basis or *vehicle* for undertaking an economic evaluation. That is, the study provides all sources of data and a framework for the overall evaluation. This approach to economic evaluation is in contrast to the use of decision-analytic modelling which is covered in detail in Chapter 9. Increasingly, however, data from effectiveness studies (at the level of the individual patient) are used to estimate inputs for (to parametrize) decision models, a topic discussed in Chapter 10.

The main focus of this chapter is the use of the single clinical (or effectiveness) study as the main means of delivering an economic evaluation. Particular issues relate to the design of such studies, and to the statistical analysis of clinical and economic data which are collected. The chapter also discusses some general issues with clinical (or effectiveness) studies which are relevant to all types of economic evaluation.

8.2 Alternative vehicles for economic evaluation

8.2.1 The randomized controlled trial

8.2.1.1 Internal validity versus generalizability

The randomized controlled trial (RCT) is a widely used study design to measure the effectiveness of health care interventions. Its value is seen as coming primarily as a source of 'internal validity'. This concept is worth looking at in more detail. Figure 8.1 shows three panels characterizing issues relating to the internal validity and generalizability of RCTs. Panel A shows the sample of patients that is being randomized (the node marked 'R') to two interventions (A and B). By measuring outcomes in the two groups, an estimate of treatment effectiveness is generated which, because patients are randomly allocated to those groups, is considered to have high 'internal validity'. This is achieved by randomization because patients will be similar in terms of both observed characteristics which are considered prognostic (i.e. affecting outcomes) and unobserved characteristics (i.e. those that are prognostic but are unknown to the analyst).

The concept of internal validity, however, relates to the specific sample of patients that is randomized. The purpose of RCTs is to estimate treatment effectiveness for a target population rather than for a specific sample of patients at a given point in time. This broader target population is illustrated in Panel B of Figure 8.1, where all these 'real

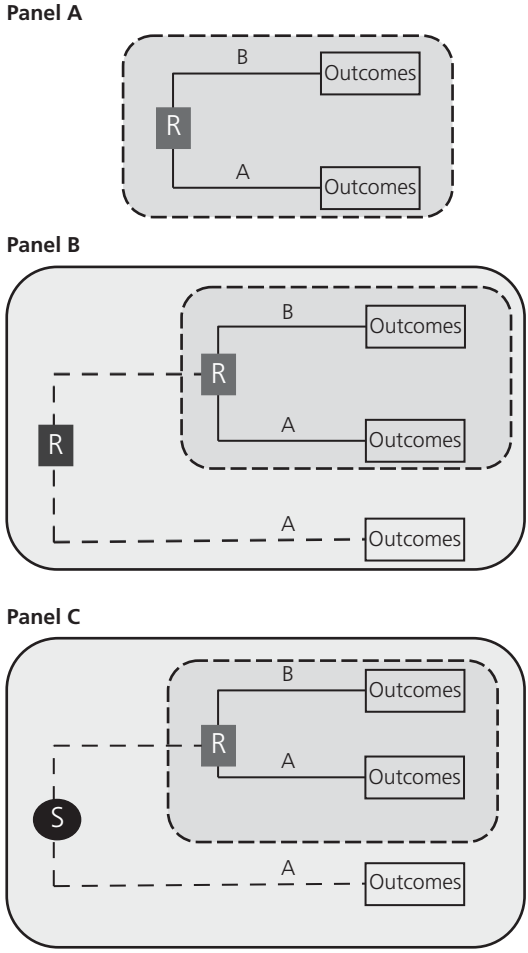


Fig. 8.1 Issues relating to the internal validity and generalizability of randomized controlled trials (RCTs). Panel A illustrates an RCT where patients are allocated to Intervention A or Intervention B. Panel B shows how patients entering the trial are drawn from a wider target population on a random basis. Panel C shows the trial sample being drawn from the wider target population using some form of selection.

world' patients are assumed to receive intervention A. The sample for the RCT is drawn from this population and the node marked 'R' indicates that the trial sample is randomly drawn from the target population. The combination of a random sample from the target population and the within-sample randomization results in both high internal validity and high generalizability. Not only will the randomized groups be similar in terms of observed and unobserved characteristics which affect outcomes, but both of these groups will be similar to the target population in both sets of characteristics.

Most RCTs do not, however, identify a sample randomly from a target population. In reality patients are *selected* on the basis of a set of criteria which might be justified in

order to be able to show a treatment effect more quickly or with fewer patients (e.g. more severe patients) or to reduce the incidence of side effects (e.g. non-pregnant women). This is shown in Panel C of Figure 8.1. In some situations the process of selection into a RCT sample is quite removed from some parts of the target population. For example, the trial might be undertaken in eastern Europe but the results are intended, in part, to be used to guide practice and policy in western Europe and North America. In all RCTs, patients have to agree to be randomized and, consequently, these individuals could be quite different to those more generally in the target population.

When patients are selected into trials, there is no threat to the internal validity of the RCT as long as they are reliably randomized. However, the generalizability of the RCT may be limited if two situations apply. The first is that potentially prognostic characteristics of trial patients are not be similar to those in the target population because of the selection process. The second is that the treatment effect (i.e. the relative effectiveness of the intervention being compared) is systematically impacted by one or more of those characteristics. Furthermore, the difference between observed and unobserved characteristics is important in this respect. Even when the sample for the RCT is selected from the target population, it is possible that it will be similar in terms of observed characteristics (i.e. representative). Indeed, even if a trial is not representative to start with, it may be possible to use statistical methods to adjust estimates of effectiveness for observed characteristics and, therefore, better reflect implications for the target population (Hartman et al. 2015; Sculpher et al. 2004b). This is not sufficient, however, for it to be generalizable if the selected sample differs from the target population in terms of unobserved prognostic characteristics which impact on the treatment effect. By virtue of these characteristics being unobserved, it is much more difficult to adjust for them statistically. We will return to the challenge of analysing effectiveness studies in the face of selection when we consider observational studies as a vehicle for economic evaluation.

Although there may be challenges for its generalizability, the scientific strength of the RCT is based largely on its internal validity. As a result, it is widely used in medical research and health evaluation. For health care interventions, such as new pharmaceuticals, most countries have formal requirements for the provision of safety and efficacy evidence before products can be licensed. The accepted standard for the collection of such data is the RCT. Hence these trials are typically a necessary condition for the successful licensing of a pharmaceutical. Furthermore, many countries invest public resources in RCTs to provide evidence to support policy and decisions in their health care systems. For example, the National Institute for Health Research (NIHR) in the United Kingdom funds a large number of RCTs relating to programmes and interventions in areas such as public health, medical devices, screening, and service delivery (<http://www.nihr.ac.uk>). In the United States, comparative effectiveness research (CER) has been defined as 'the generation and synthesis of evidence that compares the benefits and harms of alternative methods to prevent, diagnose, treat and monitor a clinical condition, or to improve the delivery of care' (Sox and Greenfield 2009). Although a range of methods is used in CER, including analysis of routine datasets, the RCT is seen as central to evidence generation for CER, despite the fact that the design of such studies poses challenges as discussed below.

8.2.1.2 Trial-based economic evaluation

Given their high internal validity and wide use as a source of effectiveness evidence, a reasonable question is whether RCTs should also be used to generate economic evidence. Since the 1980s, researchers have answered this positively, although with caveats given the challenge of achieving generalizability in estimated effects. As well as a *source* of economic data, such studies have been used as the basis (or vehicle) for economic analysis. This has become known as 'trial-based economic evaluation' (Ramsey et al. 2015). Such a study would be characterized as follows:

- The options being compared in the economic analysis are determined by the interventions to which participants are randomly allocated in the trial.
- Resource use and economically relevant outcome data are collected within the trial on all or a subset of trial participants.
- Data from individual patients in the trial are aggregated to generate estimates of mean costs and mean health outcomes (e.g. life-years or quality-adjusted life-years, QALYs) for each option being compared and, therefore, as the basis of incremental analysis (see Chapters 3 and 4).
- The time horizon for the economic evaluation is determined by the follow-up period of the trial.

There are advantages of having economic evaluation data collected prospectively as part of the trial. First, consistent with what is described above, having patient-specific data on both costs and outcomes is potentially attractive for analysis and internal validity. Secondly, given the (typically) large fixed costs incurred in collecting clinical data, the marginal cost of collecting economic data may be modest. Thirdly, and relatedly, collecting economic data in a trial might provide the most rapid source of relevant evidence for economic evaluation if that study is underway or has recently been completed.

There are, however, numerous issues and problems that researchers face when conducting economic evaluation as part of a trial, and these are discussed in the following sections.

Choice of comparison therapy As described in Section 8.2.1.1, RCTs may lack generalizability to the target population of interest. In practical terms this can take numerous forms. One problem is when the comparison therapy is not the most relevant for the policy question being addressed. In many countries, a placebo comparison plays an important role in regulatory approval of new medicines. For trial-based economic evaluation, the relevance of a placebo-controlled study depends upon whether the new drug is intended as *adjunctive* therapy or as a *substitute* for an existing therapy that is the current standard of care. For example, the efficacy of new biological drugs for progressive psoriatic arthritis, which were primarily designed to substitute for older disease-modifying anti-rheumatic drugs, was mainly assessed using placebo-controlled RCTs which were undertaken for regulatory purposes and would not have facilitated trial-based economic evaluation (Rodgers et al. 2011). This was one reason why Bojke and colleagues used indirect comparison and modelling as the basis of their cost-effectiveness analysis (Bojke et al. 2011). Chapter 9 provides further discussion of

indirect comparison as a means of addressing the problem of limited or inappropriate comparators in economic evaluation.

In some circumstances, placebo comparison data may be appropriate for trial-based economic studies; this is typically the case where the new drug will not be a substitute for another but will be a new *adjunctive* therapy; that is, all randomized groups receive standard therapies (which might be best supportive care), but participants in the experimental arm(s) also receive(s) the new intervention(s) and those in the control group have the addition of a placebo. An example would be the placebo-controlled trials of misoprostol for prophylaxis against gastrointestinal complications in persons taking anti-inflammatory drugs; a number of economic studies were undertaken using these trial data (Drummond et al. 1992). Here the placebo was considered a close approximation to no additional therapy.

Measurement in trials versus routine practice Explanatory RCTs often employ measurements for outcomes that are more detailed, invasive, or frequent than is customary in usual care. For example, many health care systems have assessed the cost-effectiveness of drug-eluting stents for patients with coronary artery disease. One of the key outcomes from an economic perspective is the number of times patients require repeat revascularization using percutaneous stenting, as this has impacts on costs and outcomes (risk of mortality and negative impact on health-related quality of life). However, many of the RCTs of drug-eluting compared to standard ('bare metal') stents assessed this rate using angiographic follow-up, which is not used routinely in clinical practice (at least in many countries). This has tended to exaggerate the number of repeat revascularizations in the bare metal stent control groups and the absolute reduction in these procedures offered by drug-eluting stents, thus potentially overstating the latter's cost-effectiveness (Bagust et al. 2006). One option to overcome this is to use evidence on baseline (i.e. bare metal stent) repeat revascularization rates from routine (non-RCT) sources and to assume that the RCT-derived estimate of the proportionate reduction in these procedures with drug-eluting stents is reliable. This would need to be implemented using decision-analytic modelling rather than trial-based methods, however, as shown by Bagust et al. (2006). The limitations for economic evaluation of some of the trials of drug-eluting stents has resulted in some health systems investing in cost-effectiveness studies using routine data for costs and effectiveness; for example, in Ontario, Canada (Goeree et al. 2009).

Intermediate versus final health outcomes Some interventions are expected ultimately to have a beneficial impact on health outcomes, but this may not be anticipated for some considerable period after the end of the trial, or only when the intervention in used in a sufficient number of patients (trials often tend to be too small to measure these health outcomes with anything other than low precision). Pharmaceuticals with such features have often been assessed in RCTs that have been designed to detect differences in one or more intermediate biomedical markers. If there is considered to be reasonable evidence that the intermediate marker is predictive of the ultimate measure of health gain such as reduced long-term mortality, then showing differences in the intermediate end point may be sufficient for the product to be licensed. The early trials of cholesterol-lowering drugs are a good example, where the outcome was the

measured change in total blood cholesterol or some subfraction. Although many of these products have now been evaluated in large RCTs where 'hard end points' such as mortality have been assessed, they obtained their initial licences on the basis of intermediate end points. Other examples abound, including trials of new HIV interventions which are designed to measure changes in viral load or CD4 count; studies of new cancer drugs which focus on progression-free survival or recurrence rates rather than overall survival; and trials of new coronary interventions which estimate differences in major adverse cardiac events.

For decisions about resource allocation, however, knowing that an intervention has a positive impact on an intermediate measure of effect is generally not sufficient to show cost-effectiveness, and the impact on *final* health outcomes such as mortality and morbidity will usually have to be indirectly quantified. As discussed more fully in Chapter 9, this quantitative link is often made using decision modelling informed by clinical and epidemiological evidence from outside the trial. The uncertainty associated with the cost-effectiveness of an intervention may be strongly related to the extent and quality of this 'external' evidence. In the case of cholesterol-lowering drugs, for example, data from cohort studies such as the Framingham study were originally used with models to predict changes in final outcomes (e.g. deaths and myocardial infarctions) from changes in risk factors such as blood serum cholesterol (Morris et al. 1997). Over the longer term, trials may emerge that seek to confirm the link between the intermediate and final outcomes by directly measuring the impact of an intervention on changes in health. In the case of cholesterol-lowering therapies, for example, this was the case with the Scandinavian Simvastatin Survival Study (4S) and the Heart Protection Study (HPS) (Heart Protection Study Collaborative Group 2002), both of which had a follow-up period of several years. These 'outcome trials' have supported trial-based economic evaluations to inform future resource use decisions including those based on 4S (Johannesson et al. 1997) and HPS (Mihaylova et al. 2005), but this sort of trial is typically unavailable when new therapies enter the health care system.

A related issue is the potential for inappropriate sample sizes in RCTs. The need to link intermediate markers with ultimate measures of health outcomes often arises because the trial was too small to generate a sufficiently precise estimate of the ultimate measures. This may be a problem for many health systems hoping to understand something about the cost-effectiveness of new interventions, but pharmaceutical regulators and some clinical audiences will be willing to use precise estimates of an intermediate clinical effect to inform their decisions. The general issue of the determination of trial sample sizes is, in fact, an important economic decision as it balances costs (e.g. recruiting and running a trial) and benefits (e.g. the value of the information generated for future patients). For this reason, the traditional rules of inferential statistics whereby trial sample size is determined to get a 'balance' between a type 1 error (accepting an intervention as (cost)-effective when it is not) and type 2 error (rejecting an intervention when it is (cost)-effective) has a limited role in economic analysis. The use of decision theory in general, and value of information methods in particular, to determine trial sample size offers a way of linking trial design to the resource constraints to undertake research and the types of decisions these studies are set up to inform. Methods of statistical analysis for trial-based economic evaluation are considered further

in Section 8.3.2. Chapter 11 deals with the economics of research prioritization and design.

Inadequate patient follow-up A feature of many RCTs is that patient follow-up and data collection often terminate abruptly when the patient experiences one of the clinical outcome 'events' of interest. From the perspective of the economic analyst this can be frustrating as the cost-effectiveness of an intervention may be dependent on the patient's prognosis subsequent to the event. Many examples can be found relating to regulatory trials for new cancer therapies where efficacy is assessed in terms of the new product's ability to delay the time until disease progression, but the trial evidence on overall survival may be limited. In principle this problem can be addressed by longer-term data collection, but this imposes additional costs on trials which are often set up to estimate clinical effects rather than costs and cost-effectiveness.

Protocol-driven costs and outcomes A problem with basing cost estimates on data gathered as part of an RCT is the extent to which the resource use being captured is associated with the effects of the trial *per se* (i.e. including the resource implication of doing the research) rather than the resource effects of providing the therapy. These so-called *protocol-driven costs* can arise in a number of different ways. For example, to preserve blinding in their comparison of oral gold (auranofin) versus placebo, Thompson and colleagues required regular blood tests for patients randomized to placebo (Thompson et al. 1989); in the analysis they excluded these costs from the placebo control group. However, excluding these protocol-driven costs may not be the only factor requiring adjustment because there may also be some form of ascertainment bias in that patients in both groups were seeing a physician more regularly for tests than in routine practice and therapy may have been modified based on observations that would not occur outside the trial. This is a similar issue to the one relating to the evaluation of drug-eluting stents discussed above. The point to stress is that, at the outset of any clinical trial where an economic question is also being addressed, it is important to establish the extent to which patient management and resource use reflects regular practice.

As experience with RCT-based economic evaluation has increased, there have been more subtle protocol biases to consider. For example, the requirement in many trials that the physician be blind to the treatment assigned to a patient may have a bearing on the way that patient is managed in the trial. In routine practice, *knowing* that the patient is receiving a given treatment may make the physician less cautious in terms of frequency of observation or test-ordering; this therefore poses a threat to the generalizability of the cost data collected within the trial (Freemantle and Drummond 1997).

Another central feature of many RCTs is the emphasis on conforming to the rules mandated by the protocol, the principle of compliance by physicians and patients. Great efforts are typically made in the conduct such studies to ensure that patients consume their prescribed medications and that physicians prescribe therapies according to protocol. Outside the trial, when the drug is used in routine practice, there are no such guarantees. Hughes et al. (2001) provide a good review of the issues related to patient compliance in economic evaluation. Methods work has also looked at analytic approaches to adjust for non-compliance (Drummond et al. 1992).

8.2.2 Explanatory versus pragmatic RCTs

The nature and extent of the challenges of using RCTs as a vehicle for economic evaluation partly depends on the primary purpose and design of the RCT. Studies that have been designed primarily for clinical purposes (e.g. to support licensing applications for new pharmaceuticals) are often described as 'explanatory' in nature in that they seek to estimate the *efficacy* of interventions in ideal or experimental settings. The strength of such 'efficacy' studies is that they potentially have very high levels of internal validity. They also usually have a clear definition of the interventions being compared and the relevant patient population, which can be an advantage when different studies are being combined using meta-analysis (see Chapter 10).

Rather than attempting to piggyback an economic evaluation on to an existing explanatory clinical trial primarily designed to address safety and efficacy questions, an alternative option is to design trials specifically as a vehicle for economic evaluation. The alternative to the 'explanatory' orientation (i.e. *can* the intervention work?) is the 'pragmatic' orientation (i.e. *does* the intervention work?), a distinction introduced by Schwartz and Lellouch (1967), and the general aim of a RCT undertaken to support economic evaluation is to be more pragmatic. The intention of such a study is to offer some compromise between the goals of internal validity and generalizability. The pragmatic trial retains the concept of subjects being randomly allocated to treatments, but has fewer restrictions regarding how patients are recruited and followed up after randomization, thus seeking to increase generalizability. In other words, these studies aim to evaluate the effectiveness or cost-effectiveness of an intervention under something closer to the 'real world' conditions that would prevail once the intervention is in routine use.

Rather than there being a strict dichotomy between a pragmatic and an explanatory trial, the relevant characteristics can be seen as being on a pragmatic–explanatory continuum. The pragmatic–explanatory continuum indicator summary (PRECIS) was developed to assess and display the position of any given trial within this continuum (Thorpe et al. 2009). The aim of PRECIS was to help trialists assess the degree to which design decisions align with the trial's stated purpose of either supporting decision-making (pragmatic) or providing explanation. Table 8.1 shows the ten domains of the continuum defined by the PRECIS tool. These relate to:

1 The criteria used to establish the eligibility of trial participants
2 How prescriptively the protocol defines the use of the experimental intervention
3 The level of clinical expertise in applying and monitoring the 'new' intervention
4 How prescriptively the protocol defines the use of the control intervention
5 The level of clinical expertise in applying and monitoring the control intervention
6 How intensively trial participants are followed up
7 The type of primary outcome measure used
8 How much attention is given to measuring compliance with therapies and whether attempts are made to improve compliance
9 How much attention is given to measuring practitioners' adherence to the trial's protocol and whether attempts are made to improve adherence
10 The analysis of the primary outcome.

Table 8.1 PRECIS domains illustrating the extremes of explanatory and pragmatic approaches to each domain

Domain	Pragmatic trial	Explanatory trial
Participants		
Participant eligibility criteria	All participants who have the condition of interest are enrolled, regardless of their anticipated risk, responsiveness, comorbidities, or past compliance	Stepwise selection criteria are applied that (a) restrict study individuals to those previously shown to be at highest risk of unfavourable outcomes, (b) further restrict these high-risk individuals to those who are thought likely to be highly responsive to the experimental intervention, and (c) include just those high-risk, highly responsive study individuals who demonstrate high compliance with pretrial appointment-keeping and mock intervention
Interventions and expertise		
Experimental intervention—flexibility	Instructions on how to apply the experimental intervention are highly flexible, offering practitioners considerable leeway in deciding how to formulate and apply it	Inflexible experimental intervention, with strict instructions for every element
Experimental intervention—practitioner expertise	The experimental intervention typically is applied by the full range of practitioners and in the full range of clinical settings, regardless of their expertise, with only ordinary attention to dose setting and side effects	The experimental intervention is applied only by seasoned practitioners previously documented to have applied that intervention with high rates of success and low rates of complications, and in practice settings where the care delivery system and providers are highly experienced in managing the types of patients enrolled in the trial. The intervention often is closely monitored so that its 'dose' can be optimized and its side effects treated; co-interventions against other disorders often are applied

(*continued*)

Table 8.1 (continued) PRECIS domains illustrating the extremes of explanatory and pragmatic approaches to each domain

Domain	Pragmatic trial	Explanatory trial
Comparison intervention—flexibility	'Usual practice' or the best alternative management strategy available, offering practitioners considerable leeway in deciding how to apply it	Restricted flexibility of the comparison intervention; may use a placebo rather than the best alternative management strategy as the comparator
Comparison intervention—practitioner expertise	The comparison intervention typically is applied by the full range of practitioners and in the full range of clinical settings, regardless of their expertise, with only ordinary attention to their training, experience and performance	Practitioner expertise in applying the comparison intervention(s) is standardized to maximize the chances of detecting whatever comparative benefits the experimental intervention might have
Follow-up and outcomes		
Follow-up intensity	No formal follow-up visits of study individuals. Instead, administrative databases (e.g. mortality registries) are searched for the detection of outcomes	Study individuals are followed with many more frequent visits and more extensive data collection than would occur in routine practice, regardless of whether patients experienced any events
Primary trial outcome	The primary outcome is an objectively measured clinically meaningful outcome to the study participants. The outcome does not rely on central adjudication and is one that can be assessed under usual conditions (e.g. special tests or training are not required)	The outcome is known to be a direct and immediate consequence of the intervention. The outcome is often clinically meaningful but may sometimes (e.g. early dose-finding trials) be a surrogate marker of another downstream outcome of interest. It may also require specialized training or testing not normally used to determine outcome status or central adjudication
Compliance/adherence		
Participant compliance with 'prescribed intervention'	There is unobtrusive (or no) measurement of participant compliance. No special strategies to maintain or improve compliance are used	Study participants' compliance with the intervention is monitored closely and may be a prerequisite for study entry. Both prophylactic strategies (to maintain) and 'rescue' strategies (to regain) high compliance are used

Table 8.1 (continued) PRECIS domains illustrating the extremes of explanatory and pragmatic approaches to each domain

Domain	Pragmatic trial	Explanatory trial
Practitioner adherence to study protocol	There is unobtrusive (or no) measurement of practitioner adherence. No special strategies to maintain or improve adherence are used	There is close monitoring of how well the participating clinicians and centres are adhering to even the minute details in the trial protocol and 'manual of procedures'
Analysis		
Analysis of primary outcome	The analysis includes all patients regardless of compliance, eligibility, and others (intention-to-treat analysis). In other words, the analysis attempts to see if the treatment works under the usual conditions, with all the noise inherent therein	An intention-to-treat analysis is usually performed. However, this may be supplemented by a per-protocol analysis or an analysis restricted to 'compliers' or other subgroups in order to estimate maximum achievable treatment effect. Analyses are conducted that attempt to answer the narrowest, 'mechanistic' question (whether biological, educational or organizational)

An example of a trial that may be considered more at the pragmatic end of the PRECIS continuum is the REFLUX trial, which compared laparoscopic surgery with medical management in individuals with chronic gastro-oesophageal reflux disease (Grant et al. 2013). The trial is characterized using the 10 PRECIS domains in Table 8.2. In general, this suggests the trial was of a more pragmatic nature, particular its primary outcome, analysis, and the intensity of participant follow-up. The eligibility criteria, flexibility of interventions, and practitioner expertise may be considered less pragmatic.

The REFLUX trial was used as the source of evidence for an initial economic evaluation which, largely because of the relatively short initial follow-up period in the trial (1 year), used a modelling framework to explore alternative scenarios relating to longer-term cost-effectiveness (Epstein et al. 2009). Indeed, despite examples of successful pragmatic randomized trials for economic evaluation, they retain many of the general problems associated with using trials as a vehicle for economic analysis. As well as the relatively short follow-up, these include the difficulty of comparing more than two or three options and the fact that other trial evidence may exist and need to be considered in the economic analysis (Sculpher et al. 2006). The limitations of trial-based economic evaluation and the role of decision-analytic modelling are discussed more fully in Chapter 9.

Furthermore, the pragmatic trial may not overcome the challenges of achieving generalizability as described in Section 8.2.1 and illustrated in Figure 8.1. Such studies may provide a way of ensuring that the RCT sample is representative of the wider population from which it is drawn in terms of observed characteristics. However, if that trial is being used to inform decisions relating to a somewhat different population, a pragmatic study can be more difficult to interpret than a more explanatory trial. For example, an explanatory trial may be designed to consider the efficacy of a new intervention compared to a placebo control. This will provide no direct evidence on the new intervention against other available active treatments; however, the use of indirect comparison and network meta-analysis would provide a framework for synthesizing different explanatory trials to offer estimates of treatment efficacy across a range of active therapies (see Chapter 10). In contrast, a pragmatic trial may randomize patients to the new intervention or to a 'usual care' control group where the clinician chooses the therapy based on judgement and custom. This may enhance generalizability and usefulness in supporting decision-making as long as the control group treatments are representative of those used in the target population. If this is not the case—for example, other new treatments are being used which were not in the control arm of the pragmatic trial—the generalizability of the study is limited and it is much more difficult to address the problem using network meta-analysis. More generally, as discussed in Section 8.2.1, the pragmatic trial may not be similar to the target population in terms of unobserved characteristics which are harder to adjust for.

8.2.3 Observational studies

In observational studies, patients receive treatments on the basis of routine decisions resulting from the interaction of patients, their clinicians, and the health system more generally. This contrasts with the randomized allocation of patients to treatments in

Table 8.2 Use of the PRECIS domains to describe the REFLUX trial (Grant et al. 2008a, 2008b) on the explanatory–pragmatic continuum

Participant eligibility criteria	For inclusion, patients required more than 12 months' symptoms requiring maintenance medical treatment for reasonable control; evidence of GORD on the basis of endoscopic and/or 24 hour pH monitoring; suitability for either policy; and the recruiting doctor was uncertain which management policy to follow. Exclusion criteria were morbid obesity; Barrett's oesophagus of >3 cm or with evidence of dysplasia; para-oesophageal hernia; and oesophageal stricture. These criteria were perhaps not the most pragmatic feature of the trial but it could be argued that, for patients to be considered for invasive surgery, some key criteria needed to be met
Experimental intervention—flexibility	Although the trial specifically focused on a laparoscopic form of fundoplication, the type of procedure was left to the surgeon, thus largely pragmatic
Experimental intervention—practitioner expertise	Surgeons were required to have undertaken at least 50 laparoscopic fundoplication procedures to be included in the trial. This may be considered more explanatory in nature
Comparison intervention(s)—flexibility	Patients in the medical management arm had their treatment reviewed by a local gastroenterologist to be 'best medical management,' based on clinical guidelines. The option of surgery (i.e. 'crossover') if a clear indication developed after randomization. This is also at the more explanatory end of the continuum
Comparison intervention(s)—practitioner expertise	Care was provided by a gastroenterologist with no further requirements specified, more pragmatic
Follow-up intensity	No non-routine hospital visits were scheduled, with outcomes assessed largely through routine hospital information and questionnaires sent to patients by post. This would be considered more towards the pragmatic end of the continuum
Primary trial outcome	The primary outcome was a disease-specific patient reported outcome measure incorporating assessment of reflux and other gastrointestinal symptoms and the side effects and complications of both treatments. As a patient-important outcome, this is highly pragmatic
Participant compliance with 'prescribed' intervention	The uptake of surgery (and repeat surgery) in both trial arms, and the use of antireflux medications was measured as an outcome and to facilitate cost analysis, but no attempt was made to change this uptake. This can again be considered pragmatic but not at the extreme end of the continuum
Practitioner adherence to study protocol	Surgeons were required to record the type of surgery they used, but gastroenterologists were not monitored for the medical therapies they prescribed. Again, at the pragmatic end of the continuum
Analysis of primary outcome	The primary outcome was analysed on an intention-to-treat basis, thus pragmatic

GORD, gastro-oesophageal reflux disease.

Source: data from Grant, A.M. et al, The effectiveness and cost-effectiveness of minimal access surgery amongst people with gastro-oesophageal reflux disease—a UK collaborative study: The REFLUX trial, *Health Technology Assessment*, Volume 12, Issue 31, pp. 1–204, Copyright © 2008; and Grant, A.M. et al., Minimal access surgery compared with medical management for chronic gastro-oesophageal reflux disease: UK collaborative randomised trial, *BMJ*, Volume 337, a2664, Copyright ©2008.

a RCT sample. These studies can be considered to be at the extreme pragmatic end of the explanatory–pragmatic spectrum discussed in Section 8.2.2 in that they generally impose little or no 'protocol restrictions' on routine practice. For this reason, observational designs have been labelled 'real world' studies. The disadvantage of such studies, however, is that their lack of randomization, whether these studies are used to estimate effectiveness or cost-effectiveness (Jones and Rice 2011). These issues with observational studies are illustrated in Figure 8.2 (contrasting with Figure 8.1 for RCTs).

Panel A of Figure 8.2 shows the comparison of two interventions where treatment allocation is determined through some selection process based on clinician and/or

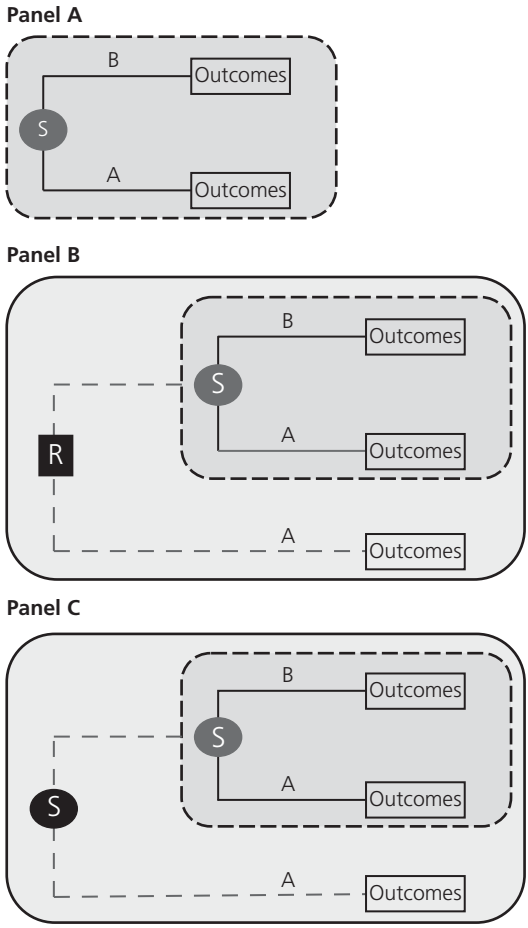


Fig. 8.2 Issues with observational designs. Panel A illustrates an observational study where patients are selected for Intervention A or Intervention B rather than randomized. Panel B shows how patients entering the observational study are drawn from a wider target population on a random basis. Panel C shows the sample for the study being drawn from the wider target population using some form of selection.

patient preference or some other mechanism. Unlike for RCTs, the groups are unlikely to be similar in terms of patients' observed or unobserved characteristics because they may be related to the mechanism of selection (e.g. more severe patients are allocated to intervention A). To the extent that these characteristics are also prognostic (they predict patients' outcomes), any estimate of treatment effectiveness in such a study will be *confounded* as a result of this imbalance. Unlike RCTs, statistical modelling is needed to adjust for imbalance in both observed and unobserved characteristics related to treatment selection and potentially predictive of outcomes. Furthermore, generating a reliable estimate of treatment effects requires that there is enough variation between patients with similar characteristics in the interventions they receive. For example, if disease severity is considered prognostic and all patients with severe disease receive intervention A and all patients with mild disease undergo intervention B, there is no basis to estimate the comparative effectiveness of the two interventions.

This can be challenging enough, but the task can become more difficult when the observational study is conducted in a sample of patients from the broader population. Panel B shows a situation where that sample is randomly selected, in which case it should be similar to the population in terms of observed and unobserved characteristics. Examples of such studies may be the use of routine datasets used for administrative purposes, such as claims data from Medicare in the United States. This could include all patients in that system with a particular condition, and a random sample may be used for analysis to estimate treatment effectiveness to guide policy for Medicare patients. In this case, the challenge of statistical analysis relates to the within-sample observational study alone.

Panel C shows a more complex situation where some selection process is used to identify patients for the sample. As with the RCT, this could be selection based on clinical criteria, geography, or patient preference, but it will result in the study sample probably differing from the population in terms of observed and unobserved characteristics related to the selection process and potentially predictive of outcome. Analysis of such a study would, therefore, need to tackle two types of selection: that determining who goes into the comparative observational study and that relating to treatment selection. The statistical techniques used to try to address the risk of confounding and selection bias are considered further in Section 8.3.4.

In principle, observational studies can provide valuable evidence on the (cost-) effectiveness of interventions. The value of such studies is particularly high when RCTs are simply impractical. For example, policy initiatives are rarely introduced in such a way that experimental designs are feasible. In such situations the use of routine observational data can provide a vehicle for evaluation. An example is the analysis by Sutton and colleagues of the impact of a 'pay for performance' initiative which provided additional funding to hospitals for achieving improvements in 28 quality measures covering 5 clinical areas (Sutton et al. 2012). Using 30-day in-hospital mortality data, the study found a clinically significant reduction in mortality in patients admitted for pneumonia, heart failure, or acute myocardial infarction.

However, the selection processes in observational studies provide an analytical challenge, the extent of which depends on the design of the study. For example, the design in Panel B of Figure 8.2, where patients are randomly allocated to the study sample,

would generally be preferred to that in Panel C, where patients are selected using a range of criteria. Studies can either be undertaken prospectively as *de novo* studies, or be based on retrospective analysis of existing datasets which are often developed for administrative purposes. There are several advantages of *de novo* studies. One is that the full range of patients' baseline characteristics (i.e. socio-demographic and clinical details at the point that interventions are selected) can be identified and measured in the study. There may also be scope to establish (and collect data relating to) factors that may explain why patients are given particular types of treatment, some of which may be unrelated to their expected outcome. Planning and then collecting these types of data maximizes the opportunity to use the statistical techniques discussed in Section 8.3.4 to adjust for how patients are selected for specific treatments.

In principle, a second advantage of prospective observational studies is the possibility of including the range of interventions that a cost-effectiveness analysis would seek to compare. An example of a prospective observational study providing the vehicle for an economic evaluation is a comparison of alternative therapies for stress urinary incontinence in women (Mihaylova et al. 2010). This study included 3739 women from 10 European countries who had been diagnosed with stress urinary incontinence. Data were collected on resource use and health-related quality of life (HRQoL) over a 12 month follow-up period, facilitating estimates of costs and QALYs for a range of alternative treatments.

Many non-randomized prospective studies take the form of treatment registers: a collection of baseline and follow-up data on patients receiving particular treatments. Sometimes these relate to one treatment only and such studies offer little basis to estimate relative (cost-) effectiveness. Some registers relate to several treatments, often new therapies—for example, the use of novel biological therapies for rheumatoid arthritis (McErlane et al. 2013). Consequently they lack information on the characteristics and outcomes of patients on older therapies or without active treatment, potentially limiting their value as a vehicle for cost-effectiveness analysis and, indeed, for estimating treatment effects of the full range of interventions.

Retrospective observational data may lack some of the baseline characteristics of patients and treatment decisions and may not include the full range of comparators. It may also be more difficult to model the selection process into the study than with a prospective design. However, they have the potential advantage of being relatively low-cost sources of data. There has been much interest in recent years in the use for research purposes of information collected routinely by health systems. In some health care systems, data have been collected for many years for administrative purposes including determining payments for health care providers. A number of health systems are now trying to link together a range of administrative datasets, sometimes augmented by new data collection facilitated by computerization in a range of provider settings, to provide datasets spanning large numbers of patients across a wide range of clinical areas. For example, a cost-effectiveness analysis was undertaken comparing a number of widely used hip prostheses using three linked NHS data sources from England: the National Patient Reported Outcome Measures programme, the National Joint Registry for England and Wales, and the Hospital Episode Statistics (Pennington et al. 2013).

8.2.4 The practicalities of collecting individual patient data for economic evaluation

Increasingly, economists have an opportunity to work closely with those designing clinical studies to determine the type of data to be collected to facilitate a 'trial-based' economic evaluation. There are a number of practical issues which need to be considered.

8.2.4.1 Resource use data

As discussed in Section 8.2.1.2, it is important to identify and minimize resource consequences measured in trials that are due to the research protocol and do not characterize the delivery of care in the normal setting. But how are resource quantities associated with interventions actually collected as part of the evaluation? The first point is that, to the extent possible, it makes sense to build upon the research infrastructure that will already be in place for collecting the clinical data. For example, many clinical trials of hospital-based acute therapies collect data using case report forms (CRFs) designed for completion by, for example, a study nurse. To facilitate collection of resource quantities associated with interventions, pages can be added to the same CRF and extend the responsibility of data collection by the study nurse to include key items of resource use. Studies will obviously vary in the amount of detail and precision needed in the collection of resource use quantities. At a minimum, for a hospital-based study, it would probably be desirable to know the total length of stay in hospital, the length of stay in high-cost areas such as intensive care, major diagnostic or therapeutic procedures such as MRI scans or surgery, and use of expensive pharmaceuticals. As discussed in Chapters 3 and 7, depending on the intervention being evaluated, the precision of costing may need to be much greater. For example, if two alternative treatments in the intensive care unit are being compared, much more detailed information on the resources consumed in the intensive care unit would probably be appropriate.

In addition to the resource quantities associated with the initiation of therapy, it is usually necessary to capture downstream resource consequences of the treatment or the disease. Typically these downstream costs could be an exacerbation of the problem warranting readmission to hospital, or a mild complication resulting in consultation with a family physician or attendance at an emergency room. Capturing these data is more problematic for a number of reasons. First, if a person is rehospitalized at a hospital that is not part of the clinical trial, then knowledge of this rehospitalization and access to information on resource consumption from that hospital may be limited. To facilitate information retrieval it may be necessary to ensure that patients have given approval for such data gathering as part of the informed consent documentation. The monitoring of such events can be achieved either by patient recall or, depending on local circumstances, it may be possible to use routine data sources such as those relating to physician reimbursement claims.

In many trials, ambulatory physician visits are often recorded using patient recall. For example, patients could be asked whether they have seen a family doctor in the past 3 months for a reason associated with their hypertension or its treatment. As with any survey technique, the reliability of patient recall comes into question, particularly when one is studying population groups where recall may be a problem (e.g. elderly people).

The use of diaries to be completed regularly by patients may be a way around the challenge of recall. The method of follow-up contact could be by mail survey, SMS texting, or telephone follow-up, and depending on the patient group being studied, there are advantages and disadvantages of these alternative methods. Readers who are interested in more detailed discussion of principles of data collection for economic data as part of trials and design of CRFs should consult a range of other papers (Glick et al. 2001; Mauskopf et al. 1996; Ramsey et al. 2015).

A systematic review was undertaken of resource use data collection methods in 100 trial-based economic evaluations undertaken as part of the United Kingdom's NIHR Health Technology Programme (Ridyard and Hughes 2010). The authors concluded that there was some convergence of methods, in particular the consistent use of routine data sources and of patient recall. However, they also noted marked variation in the methods used including the chosen cost perspective and the time period used for patient recall. The study found little evidence that resource use data capture was being shaped by systematic approaches to reviewing other studies in the field or by piloting of data collection instruments, and validation of instruments was rare. Following on from this work the authors have led a consortium to develop the Database of Instruments for Resource Use Measurement (DIRUM), which is available to review at <www.dirum.org> (Ridyard et al. 2012). Its purpose is to provide a tool to those planning trial-based economic studies to help in the choice of instruments or parts of instruments which are relevant to researchers' particular interests (i.e. disease, population, settings of care).

8.2.4.2 Health-related quality of life (HRQoL) data

Chapters 5 and 6 have covered some of the key approaches to measuring and valuing HRQoL for economic evaluation. For economic evaluations alongside clinical studies which aim to estimate the health impact of treatments in terms of QALYs, there is a need to collect relevant data from study patients.

A number of issues need to be considered in planning for this. The first is what approach to measurement is appropriate. There are some examples of clinical studies which include direct elicitation of patients' preferences using the methods described in Chapter 5. For example, an RCT assessing the effectiveness and cost-effectiveness of drug interruption and intensification in advanced HIV measured the health preferences of patients using both direct standard gamble and time trade-off in a subset of patients (Joyce et al. 2009).

It is also possible to measure patients' HRQoL using a disease-specific instrument for which preference weights are available using some form of statistical model which 'maps' between the disease-specific measure and a generic one (see Chapter 5). Generally, however, primary studies such as RCTs use established preference-based measures of HRQoL with associated preference weights suitable for estimating QALYs at the level of the individual patient. These have a practical advantage in that they simply involve asking patients to complete a questionnaire to 'describe' their HRQoL, with the link to preference values having been established on the basis of separate valuation studies. They also provide a standardized approach to measuring and valuing HRQoL which provides consistency between evaluations over time and between clinical areas. These generally take the form of generic measures such as those described in Chapter 5,

including the EuroQol (EQ-5D) and the Health Utilities Index. Patients typically complete the descriptive part of these instruments at baseline and various points of follow-up. Depending on their responses, each patient allocates themselves to a health state which is associated with a unique weight derived from public preferences. In the case of the EQ-5D, for example, there are 245 health states defined by responses to the three-level version of the instrument. With 1 for good health and 0 for dead, other weights range between −0.429 and 0.848 (based on a UK valuation study; Dolan et al. 1996).

A QALY is calculated for each patient based on their weights and survival status over time. This involves working out a 'QALY profile' by interpolating between the points and working out the area under the profile (Matthews et al. 1990). For example, in the REFLUX trial referred to in Section 8.2.2 evaluating laparoscopic surgery compared to continued drug therapy in patients with oesophageal reflux disease, EQ-5D data were collected on all patients at baseline, 3 months after randomization, and annually until 5 years after randomization (Grant et al. 2013). Table 8.3 shows the estimated mean EQ-5D weights in each arm at each point of measurement, together with the mean differences with 95% confidence intervals (CI). The results suggest that patients' mean HRQoL in both groups is improving over time, that surgery seems to be associated with higher HRQoL than medical management, and that this difference is at its maximum at 3 years after which it begins to fall. When translated into QALYs, and in the absence of any difference between the groups in terms of mortality, surgery generates an estimated mean gain in QALYs over 5 weeks of 0.1976 (95% CI −0.0857, 0.4810).

A second issue which needs to be considered, therefore, is the frequency and interval of HRQoL data collection. In principle, measurements should be taken at key points when the patient's health is expected to change. For example, in a trial-based cost-effectiveness analysis of laparoscopic-assisted hysterectomy and standard hysterectomy, EQ-5D data were collected at baseline and then 6 weeks, 2 months, and 12 months after randomization (Sculpher et al. 2004a). The main difference between the arms of the trial was expected to be in the duration of convalescence and its impact on HRQoL. At baseline, the difference in mean EQ-5D values (based on UK preferences) between women randomized to laparoscopic and abdominal hysterectomy was 0.026. By 6 weeks follow-up, this had narrowed to 0.001. It is possible that this interval missed a key difference in HRQoL between the groups. Any such difference in HRQoL is, however, used to quality-adjust a relatively short period in terms of QALYs. Therefore, achieving finer granularity in the difference between trial arms in HRQoL measurement may ultimately have little impact on health outcomes in terms of QALYs or cost-effectiveness. A further consideration is that additional measurements require patients to devote more time to data collection which may result in a lower questionnaire completion rate and more missing data. In addition, more frequent measurement represents an additional cost to the study and needs to be balanced against the possibility of a more precise estimate of outcome.

Generally HRQoL data are only collected in clinical studies when there is expected to be some difference between the studied groups in that type of outcome, either because one treatment will increase HRQoL or another will reduce it. Therefore, clinicians designing a study to compare two alternative interventions which are expected to impact

Table 8.3 Estimates of mean EQ-5D at baseline and over follow-up from the REFLUX study (Grant et al. 2013)

Completed questionnaires returned at each time point		Follow-up	Mean (SD) EQ-5D		Difference in EQ-5D (surgery – medical management) (95% CI)[b,c]
Surgery (n = 178[a])	Medical management (n = 179[a])		Surgery	Medical management	
171	173	Baseline	0.7107 (0.2581)	0.7201 (0.2545)	–0.0094 (–0.0638 to 0.0445)
149	153	3 months	0.7881 (0.2328)	0.6894 (0.3012)	0.0987 (0.0376 to 0.1597)
152	164	Year 1	0.7537 (0.2468)	0.7097 (0.2715)	0.0440 (–0.0136 to 0.1016)
122	138	Year 2	0.7619 (0.2718)	0.7172 (0.3127)	0.0447 (–0.0273 to 0.1167)
129	132	Year 3	0.8034 (0.2312)	0.7474 (0.2621)	0.0560 (–0.0043 to 0.1163)
125	127	Year 4	0.7713 (0.2438)	0.7544 (0.2719)	0.0169 (–0.0472 to 0.0810)
124	117	Year 5	0.7743 (0.2590)	0.7612 (0.2815)	0.0131 (–0.0555 to 0.0817)

[a] n refers to the number of patients originally randomized to each trial arm.

[b] Confidence intervals estimated using OLS regression.

[c] Unadjusted for baseline EQ-5D.

only on mortality may decide not to measure HRQoL at all. However, for an economic evaluation using QALYs as the measure of outcome, it is necessary to estimate HRQoL during the additional survival duration generated by the more effective treatment. Where no appropriate HRQoL data are collected in a clinical study, it will probably be necessary to move to a decision modelling framework to bring in other sources of such data (see Chapters 9 and 10).

8.2.4.3 Clinical data

For a clinical study being used as a vehicle for economic evaluation it might be expected that the focus would be on analysing resource use and HRQoL data. However, clinical data may be used in a range of ways in an economic analysis. Often data are collected for clinical purposes that can feed directly into cost estimation. An example is data collected on concomitant medicines used by patients in clinical studies (i.e. drugs used in addition to any being directly evaluated in the study). For clinical reasons, this information is often useful—for example, to judge whether patients are receiving appropriate care. They are also a direct source of resource use information. By applying unit costs to these data, this can be the basis of a category of differential cost between the interventions subject to evaluation. As for resource use data more generally, the analyst would need to consider whether this route to cost analysis is complete and appropriate.

Some cost-effectiveness analyses present differential effects in terms of clinical endpoints (or in a way derived from clinical endpoints). For instance, Sculpher and Buxton assessed the cost-effectiveness of two medicines for asthma in terms of incremental cost per day free of asthma episodes (Sculpher and Buxton 1993). However, as discussed in Chapters 4 and 5, the role of these types of studies to support decision-making is unclear, with an increasing use of generic measures of outcome, typically QALYs.

The QALY, however, also depends on clinical data of course, as it is necessary to know something about the mortality risks of patients in the study over time. Some interventions are assumed to have no mortality impact and QALYs are estimated solely in terms of differences in HRQoL. In other studies, however, differential mortality will be a key endpoint and this will be measured carefully for clinical purposes. Although these data can feed directly into economic analysis, there is often a mismatch between the study follow-up and the time horizon of an economic evaluation. In most clinical studies, a proportion of patients remains alive at the end of the study; that is, they are censored because the study follow-up is completed before they die. However, the appropriate time horizon for an economic analysis is the period over which costs and benefits potentially differ and, where there are mortality differences and the outcome measure includes gains in survival duration, the time horizon should be a patient's expected lifetime. Chapters 9 and 10 discuss issues with survival data in economic evaluation and the frequent need to extrapolate beyond what is observed in clinical studies.

Clinical data can also be used in economic evaluations as a means of assessing the characteristics of patients receiving one intervention and those undergoing another (i.e. patients' baseline characteristics). As discussed in Section 8.2.1, in an RCT it is usually possible to ensure similar types to patients in the alternative treatment groups within the sample entering the study. This is generally not the case with non-randomized studies, so clinical data collected at baseline are necessary as a basis for some form of

statistical analysis to identify comparable groups (in terms of observed and unobserved characteristics) receiving alternative interventions (see Section 8.3.4.). Indeed, even in randomized studies, baseline clinical data are essential to economic evaluations: to assess how generalizable the study is to routine practice, possibly in different jurisdictions; as a basis of statistical techniques to make the estimates of costs and effectiveness more precise; and as a basis for subgroup estimates of cost and effectiveness (see Section 8.3.3).

8.3 Analytical issues with individual patient data

Undertaking economic evaluation based on individual patient data from clinical trials requires a range of analytical issues to be addressed. In principle, the availability of individual patient data provides a number of major advantages. There is an opportunity to move from *deterministic* to *stochastic* analysis for trial-based economic evaluation. In deterministic economic analysis, cost and effect variables are analysed as point estimates. In the context of trials, this could be because there has been a focus solely on mean estimates, with the uncertainty in those estimates ignored. If we consider a treatment that is both more costly and more effective than control, the economic comparison is illustrated in the top right quadrant of the cost-effectiveness plane (see Box 3.2). The slope of the line extending from the origin (the control) through our study estimate, point A, represents the incremental cost-effectiveness ratio (ICER) of the treatment relative to control. A *stochastic cost-effectiveness analysis* is where both costs and effects are determined from data sampled from the same patients in a study. If cost and effect data are sampled and variances are available, then formal statistical methods can be used on observed differences in costs (treatment–control) or effects.

This section examines some of the statistical issues that arise in the conduct of stochastic economic evaluation based on individual patient data. This is an area of methods that has developed rapidly over the last 15 years, and it is outside the scope of this book to provide a comprehensive guide to all the issues. Rather, the purpose of this section is to summarize the main developments. For more detailed overviews, see the following references (Glick et al. 2014; Willan and Briggs 2006).

8.3.1 The nature of economic data

One of the reasons why there has been so much interest in statistical methods for economic evaluation alongside trials is that economic data are not synonymous with clinical data and have required the development of new methods, or the adaptation of existing techniques, to analyse them appropriately. Some of the characteristics of economic data are described in Sections 8.3.3.1–8.3.1.4.

8.3.1.1 Skewed cost data

Cost data (and the resource use data which underlie them) often exhibit some challenging characteristics for statistical analysis. A particular problem is that cost data are usually right-skewed because costs are naturally bounded by zero (they cannot be negative), but they have no logical upper bound. In the context of a clinical study, it is quite common to have a small proportion of patients with very high costs, perhaps

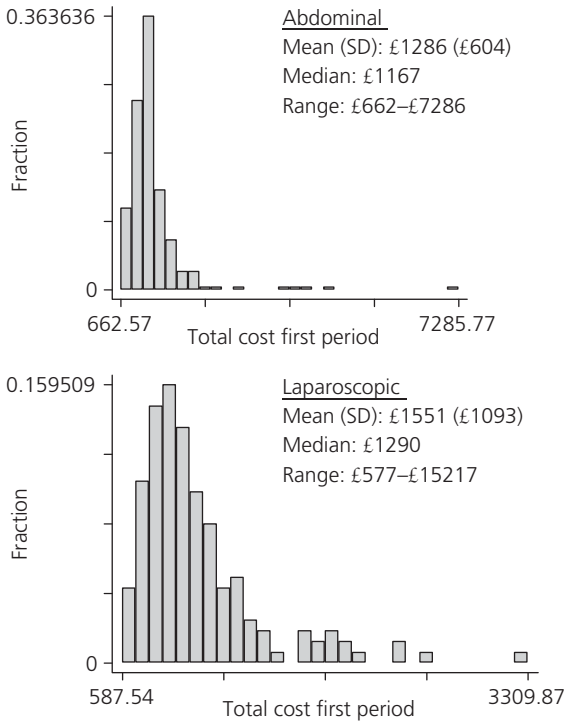


Fig. 8.3 Right-skewed cost data from a randomized trial comparing laparoscopic and abdominal hysterectomy (Garry et al. 2004; Sculpher et al. 2004a). Costs relate to an initial period of 6 weeks following surgery.

Source: data from Sculpher, M.J. et al., Cost-effectiveness analysis of laparoscopic-assisted hysterectomy in comparison with standard hysterectomy: results from a randomised trial, *BMJ*, Volume 328, pp. 134–40, Copyright © 2004; and Garry, R. et al., EVALUATE hysterectomy trial: a multicentre randomised trial comparing abdominal, vaginal and laparoscopic methods of hysterectomy, *Health Technology Assessment,* Volume 8, Number 26, Copyright © NETSCC 2004.

reflecting serious adverse effects with low incidence. The cost of this small number of patients has a much bigger effect on mean cost than the median, giving the distribution its characteristic right-skewed shape. This feature of cost data is illustrated in Figure 8.3 using results from a trial-based economic evaluation of laparoscopic compared to open abdominal hysterectomy (Garry et al. 2004; Sculpher et al. 2004a). In addition to right skewness, cost data often tend to have 'heavy tails'; that is, there is a relatively large proportion of patients with relatively large values. Furthermore, the distribution of costs by individual patients can be multimodal, often because a large proportion of patients have zero costs because they consumed no relevant resources. Individually and together, these features pose problems for the statistical analysis undertaken as part of an economic evaluation alongside a clinical study.

Faced with these characteristics, a standard approach in clinical evaluation would be to use non-parametric methods. This usually results in a focus on a summary measure

of the distribution in the form of the median. In the context of costs, however, this is inappropriate given that the decision-maker needs to be able to link the summary measure of cost per patient to the overall budget impact, and this can only be achieved by estimating the mean cost per patient.

Mihaylova and colleagues undertook a detailed literature review to describe the various methods which have been suggested—and sometimes used—to address these problems (Mihaylova et al. 2011). Alternative approaches included, first, methods based on the normal distribution such as the use of ordinary least squares (OLS) regression. These methods have the advantage of ease of implementation, but estimates of mean outcomes can be sensitive to extreme values for individual costs. Mihaylova and colleagues recommend the use of these methods when sample sizes are sufficiently large to ensure approximate normality of sample means. This is on the basis of the central limit theorem (CLT) which states that, if the size of a sample from a population is large enough, the mean of all samples from the same population will be approximately equal to the mean of the population and the samples will follow an approximate normal distribution (i.e. there is an assumption of asymptotic normality).

A second approach is to transform cost or resource use data onto a scale which ameliorates the skewness and facilitates a more reliable comparison of means or use of OLS, usually with transformation to the log scale (although this cannot be used when there are zeros in the data). Although this can be a useful tool in some circumstances, such as dealing with heavy-tailed distributions, care has to be taken regarding back-transformation and an appropriate transformation has to be used. A third approach is non-parametric methods and, in particular, commonly uses non-parametric bootstrapping. This uses resampling from the data with replacement to generate an empirical estimate of the sampling distribution of means costs (Section 8.3.2 provides more detail on bootstrapping methods). Nixon and colleagues compared bootstrapping and methods relying on normal distributions and concluded that, even with small samples from skewed data, both provide accurate estimates of the mean with the latter generating at least as accurate estimates of the uncertainty in mean values (Nixon et al. 2010). They conclude that both methods are potentially appropriate but methods invoking the CLT are generally easier to implement.

Mihaylova and colleagues also describe more complex methods which, in certain circumstances, may have advantages (Mihaylova et al. 2011). General linear models are increasingly used to analyse economic data. These models directly use a family of statistical distributions—for example, the gamma distribution to model costs and the Poisson to analyse resource use—and allow the use of covariables. However, the choice of distribution needs to be justified and sensitivity analysis undertaken to assess the robustness of estimates to this choice. Another analytic approach is two-part models which are useful for multimodal data, most typically when resource use or cost data are characterized by a proportion of zero counts. The first part of the model estimates the probability of a non-zero cost or resource use, and the second part estimates the mean cost/resource use conditional on a positive value.

8.3.1.2 Missing data

In all economic studies involving the collection of patient-level data, it is likely that there will be some missing data. This can happen for a variety of reasons including patients failing to respond to questionnaires or being lost to follow-up, or items in CRFs not being completed by clinical or research staff. Missing data on economic measurements can be a particular problem if these items are accorded less priority by clinical and research staff. The problem of missing data also exists in clinical evaluation, but it has some particularly important implications for economic analysis. The estimation of mean cost will typically be based on a number of items of resource use data, each of which is multiplied by a relevant unit cost, and total cost is the aggregation across these items (see Chapter 7). If any one of these resource items is missing for a particular patient, then it is not clear how to estimate that patient's total cost. A similar situation exists with the analysis of individual patient data on HRQoL for the purpose of QALY estimation. If an instrument like the EQ-5D is used to measure HRQoL, patients will typically be asked to indicate their level of health at baseline and various points of follow-up. It is possible to have missing data for a given dimension of health at a specific point in time (e.g. no box is ticked for ability to undertake 'usual activities' at a 1 month follow-up), a patient may fail to complete the entire instrument at a specific time point but then complete all others subsequently, or the patient may be lost to follow-up and all HRQoL data beyond a particular time point are missing. As the calculation of a QALY at the level of the individual patient relies on all measures of HRQoL, methods are necessary to deal with these different types of missingness. Table 8.4 shows the numbers of questionnaires (relating to EQ-5D and resource use) returned and completed in the REFLUX trial described in Section 8.2.2 (Grant et al. 2013). The complete case analysis (patients who completed all five questionnaires) was only 49% in the surgery arm and 47% in the medical management group.

In identifying the best way of dealing with missing data, it is necessary to assess the likely mechanisms of missingness (Little and Rubin 1987). Data are *missing completely at random* (MCAR) when the likelihood of their being missing is not a function of known or unknown measurements (covariables). An example of this might be when an outcome measurement is not possible in a given patient because of equipment failure. A second mechanism is *missing at random* (MAR) where the likelihood of data being missing does not depend on unobserved measurements but can be determined on the basis of other measurements. An example of data being MAR could be where a patient is withdrawn from a study because their condition becomes too severe on the basis of a predefined scale. A third mechanism is *missing not at random* (MNAR) when neither MCAR nor MAR applies; that is, when the likelihood of missingness cannot be explained by observed data alone and depends also on unobserved measurements. However, it is not easy to determine which of the mechanisms applies for particular missing observations unless the reasons for missing data are well understood (e.g. faulty equipment, as mentioned earlier). Therefore, there is always a need to use sensitivity analysis to assess how different methods of handling missing data affect ultimate estimates of interest in a study.

Table 8.4 Questionnaires returned and completed in the REFLUX trial comparing laparoscopic fundoplication with medical management in patients with gastro oesophagel reflux disease

Year	Questionnaires returned, *n* (%)		Completed questionnaires,[a] *n* (%)	
	Surgery	Medical management	Surgery	Medical management
1	154 (87)	164 (92)	134 (75)	147 (82)
2	128 (72)	142 (79)	121 (68)	134 (75)
3	132 (74)	134 (75)	112 (63)	119 (66)
4	126 (71)	129 (72)	114 (64)	118 (66)
5	127 (71)	119 (66)	115 (65)	113 (63)
Number of patients in complete case analysis			88 (49)	84 (47)

[a]Completed questionnaires means that all of the questions on health care resource use and EQ-5D were filled in.

Reproduced from Grant, A.M., et al., Clinical and economic evaluation of laparoscopic surgery compared with medical management for gastro-oesophageal reflux disease: 5-year follow-up of multicentre randomised trial (the REFLUX trial), *Health Technology Assessment*, Volume 17, Number 22, Copyright © Queen's Printer and Controller of HMSO 2013.

The various available methods to address missing data, depending on the mechanism of missingness, have been considered in several papers (Briggs et al. 2003; Gomes et al. 2013; Manca and Palmer 2001; Marshall et al. 2009). None of these methods is specific to economic data but, as outlined above, missing data can be a particular problem in economic evaluation. A simple method is to analyse the costs of those patients for whom complete data are available; this is referred to as *complete case analysis* (CCA). However, in studies where there are a lot of resource use data on each patient, a large proportion of patients may have at least one item of missing data. This will mean that CCA will relate to relatively few patients and relevant data will effectively be 'thrown away', resulting in mean costs being estimated with less precision than necessary (i.e. the approach is *inefficient)*. Furthermore, when data are not MCAR, the use of CCA can result in biased estimates of mean values. Another simple approach is mean imputation where, if a measurement is missing for a particular patient, the mean value of that measurement in other patients is used. This will also be biased if the missingness mechanism is not MCAR, and will always lead to an under-representation of uncertainty in the relevant measurement (and those derived from it).

A generally preferred approach to dealing with missing data is *multiple imputation* (MI). This is based on a mechanism of MAR and, by generating a number of alternative sets of imputed data, seeks to reflect the uncertainty associated with the imputation process within the ultimate results of the study. There are various versions of MI and different software with which to apply the method (Carpenter and Kenwood 2013; Yu et al. 2007). The validity of MI depends on the credibility of the assumption of MAR, and this can be enhanced by using a full range of covariables and outcomes as the basis

of imputation—e.g. patients' baseline characteristics, observed costs, clinical measures during follow-up, and any reasons for missingness (Marshall et al. 2009).

Noble and colleagues reviewed 88 trial-based cost-effectiveness studies to establish what methods they had used to address missing data (Noble et al. 2010). They found that these approaches were generally poorly reported, and that few studies used sensitivity analysis to assess the implications of alternative methods of handling missing data on study results. Overall, CCA was the most widely adopted method and its use had increased over time. Despite the potential for bias with this method, less than half the studies considered the possibility of bias. MI was used in 6/34 studies between 2003 and 2005 and 10/54 studies between 2006 and 2009; however, Noble and colleagues argue that there should be more transparency in the use of MI including the procedure, software and approach to reflecting the implications of the multiple imputation datasets in the measure of uncertainty in cost-effectiveness. They urge future studies to follow recent guidelines for MI (Sterne et al. 2009).

8.3.1.3 Censored cost data

A specific type of missing data relates to situations where patients have been followed up for differential time periods, which leads to *censored* data. One example of this is administrative censoring, when patients are recruited into a trial over a period of time (e.g. 2 years), but analysis is undertaken at a specific time based on all available data at that point (e.g. 1 year after the last patient is recruited). This means that the extent of follow-up data will vary between patients—in this example, the minimum follow-up period will be 1 year and the maximum 3 years.

A key assumption when analysing censored data is that censoring is 'non-informative'; that is, patients who are fully followed up (uncensored) are representative of those who are censored. Standard survival analysis methods exist for the analysis of censored time-to-event data (e.g. time until death) (Machin et al. 2006). However, these methods are not appropriate to analyse censored cost data and will lead to biased estimates (Etzioni et al. 1999). This is because different patients can accumulate costs at different rates over time, so two patients who are censored at the same time, but with different accumulated cost, would ultimately be expected to have different total costs if they had been fully followed up. Hence the assumption of non-informative censoring is no longer tenable.

Various methods have been developed in recent years to analyse censored costs appropriately. It is not within the scope of this book to describe these methods in detail, but good reviews are available (Glick et al. 2014; O'Hagan and Stevens 2004; Wijeysundera et al. 2012), as well as comparisons of the available methods (Raikou and McGuire 2012).

8.3.1.4 Difficulties with incremental cost-effectiveness ratios (ICERs)

As described in Chapters 3, and 4, the ICER is the traditional summary result from a cost-effectiveness analysis. A statistical issue that has been of some interest in the literature since the mid-1990s relates to how to present the statistical uncertainty around the ICER when it is derived from sampled patient-level data on costs and effects. Chapter 11 covers the analysis and policy uses of uncertainty in economic evaluation in

general, whether these are based on model-based studies or clinical studies with individual patient data. Many of the issues discussed here are also relevant to that chapter.

The development of appropriate and easily implementable ways of expressing the implications of sampling variation in individual patient data for uncertainty in the ICER has been hampered by some particular features of the ICER (Briggs 2001). One problem relates to the issue of negative ICERs. The consideration of the decision rules of cost-effectiveness analysis in Chapter 4 make clear that, if an intervention is less costly and more effective than a comparator (the bottom right quadrant of the cost-effectiveness plane in Figure 3.2), it is dominant and unequivocally cost-effective. Conversely, if an intervention is more costly and less effective than a comparator (the top left quadrant in Figure 3.2), it is dominated and cannot be considered cost-effective. When working with estimates of mean costs and effects, it is necessary to report the fact that an intervention is dominated or dominant because simply reporting an ICER, which in both these situations would be negative, would not indicate whether or not the intervention should be regarded as cost-effective.

This process is more complex, however, when presenting sampling uncertainty around the ICER. This can be described with reference to Figure 8.4 which reproduces a cost-effectiveness plane and shows two different comparisons of a new intervention relative to a standard intervention. In the top left quadrant, comparison A shows a situation where the standard intervention dominates the new intervention, with lower costs and higher effects. In the bottom right quadrant, the new intervention is dominant (comparison B). Despite these quite different results, the ICER is identical for the two comparisons (−10). This feature of the ICER is not necessarily a problem for a deterministic analysis based on a point estimate of differential cost and effect because it is clear in which quadrant the comparison is located. However, when the uncertainty around the point estimates is allowed for, the ICER and its uncertainty could span more than one quadrant and the feature of the ICER shown in Figure 8.4 can make this uncertainty difficult to present.

A second, and related, problem occurs when the comparison is located in the top left or bottom right quadrants of the cost-effectiveness plane where one option is dominant and the ICER is negative; the magnitude of the ICER has no meaning. Consider three different comparisons in the bottom right quadrant:

- X results in 1 QALY gained, a saving of £2000, and hence an ICER of −£2000
- Y results in 2 QALYs gained, a saving of £2000, and hence an ICER of −£1000
- Z results in 2 QALYs gained, a saving of £1000, and hence an ICER of −£500.

In terms of ICERs, Z would be preferred to Y, which would be preferred to X. However, in terms of changes in costs and QALYs, Y would clearly be preferred to X and Z as it has the highest combination of QALY gain and cost saving.

A third problem is, if the denominator of the ICER (the difference in effects) is zero, the ratio itself is infinite. Again, this is not a major problem with a deterministic analysis as one treatment is likely to dominate by virtue of having lower mean costs. With a formal analysis of uncertainty, however, this feature of the ICER will present problems when the uncertainty in the effect difference is allowed for if there is a non-negligible probability of that difference being zero. Furthermore, because the difference in effects could be negative, this can cause a discontinuity in the ICER,

Treatment	Costs	Effects	ICER
Comparison A			
New Intervention	80	5	
Standard Intervention	60	7	
Difference	+20	–2	–10
Comparison B			
New Intervention	65	12	
Standard Intervention	85	10	
Difference	–20	+2	–10

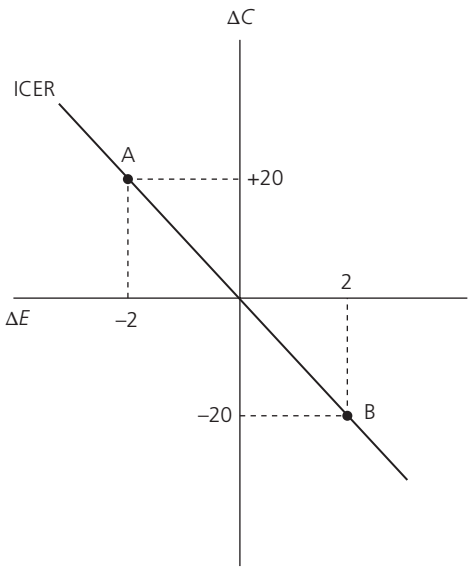


Fig. 8.4 The problem of negative incremental cost-effectiveness ratios (ICERs).

which means there is no mathematically tractable expression for the variance of the ratio (Briggs et al. 2002).

A fourth problem with the ICER is that, as a ratio statistic, it is not easy to use as a dependent variable within a regression analysis. However, such analysis may be necessary as it may be important to adjust the cost-effectiveness estimates generated in a trial for differences in patient case mix between the intervention groups at baseline. Indeed, in the context of an observational study, this process of adjusting for known (and where possible unknown) confounding variables is essential (see Section 8.3.4). It may also be important to assess the extent to which the cost-effectiveness of a given intervention varies between different subgroups of patients. The process of adjusting for differences in baseline case mix and of undertaking subgroup analysis would ideally be undertaken within a regression framework. Section 8.3.2 describes methods that have been developed to overcome some of the problems with ICERs.

8.3.2 Quantifying uncertainty in cost-effectiveness using patient-level data

It is helpful to distinguish two general approaches to representing uncertainty in cost-effectiveness results: (1) hypothesis testing and (2) estimation, which are dealt with in Sections 8.3.2.1 and 8.3.2.2 respectively. In Chapter 11 the link between these approaches to uncertainty and decision-making is considered.

8.3.2.1 Hypothesis testing

When patient-level sample data are available on costs and effects, it is possible to use formal hypothesis testing as a way of reflecting the uncertainty in cost, effects, and cost-effectiveness. In the analysis of effect data with associated sampling variation, the null hypothesis is usually that there is no difference in outcome between experimental and control therapy, and this is tested against either a one-tailed alternative (usually that the experimental treatment is more effective) or a two-tailed alternative (that the experimental treatment is more or less effective than the control).

The purpose of economic evaluation is, however, to inform decisions about resource allocation rather than to make inferences about particular phenomena. Although hypothesis testing has been explored in economic evaluation (O'Brien and Drummond 1994; O'Brien et al. 1994), its role has been limited. One problem with hypothesis testing as a way of reflecting uncertainty is that an overemphasis tends to be placed on the statistical significance of results (the size of the *P*-value) in isolation from the magnitude of the effect size. This can be illustrated with reference to Box 1.3. If a formal hypothesis test suggests that there is no statistically significant difference in effects between treatment and control, then it may seem logical to adopt the approach of cost-minimization analysis whereby the less costly treatment is the most cost-effective. However, unless the measure of effect in the cost-effectiveness study is the primary outcome measure, it is very unlikely that the trial would have been 'powered' to find a statistically significant difference in effects. This is likely to be the case, for example, when QALYs are the measure of effect in the study, but the trial was not powered to show a statistical difference in mortality or HRQoL. Therefore, there is likely to be an important risk of rejecting the alternative hypothesis (that treatment is more effective than control) when in fact that hypothesis is correct (i.e. a high type II error), and this error will be reflected in the economic evaluation (Briggs and O'Brien 2001).

Even if the trial had been formally powered to test a hypothesis around the effect measure used in the cost-effectiveness study, it would still be inappropriate to conclude that the lack of a statistically significant difference in effectiveness between a new treatment and control is synonymous with a difference of zero. The difference between the two sample means remains the best estimate of effect difference rather than zero. The importance of this point can be illustrated using the example of the cost-effectiveness analysis of laparoscopic versus open abdominal hysterectomy introduced above and in Figure 8.3 (Garry et al. 2004; Sculpher et al. 2004a). Over the full follow-up period of 1 year (as opposed to the short-term follow-up in Figure 8.3), laparoscopic hysterectomy had an additional mean cost of £186 (95% CI –£26 to £375) and additional mean QALYs of 0.007 (95% CI –0.008 to 0.023). In neither case, therefore, were these

differences statistically significant as the 95% CIs around them both crossed zero. How should a decision-maker react to these results? One option would be to interpret the data as saying that the lack of a statistically significant difference in costs and effects is synonymous with a zero difference in these end points, and hence the options are identical in terms of cost-effectiveness. This is inappropriate as the mean differences remain the best estimates of differential costs and effects.

A second option is to adopt conventional rules of statistical inference and conclude that the data provide no basis, using a threshold *P*-value of 0.05, to reject the null hypotheses that open abdominal hysterectomy is less costly and more effective than the new laparoscopic technique, and that abdominal method should remain the preferred option. However, this assumes that the error probability inherent in the 0.05 *P*-value is appropriate for the decision. A third option is to reject conventional methods of statistical inference and focus on mean differences in costs and effects, in which case laparoscopic hysterectomy would be preferred if the ICER (£186/0.007 = £26 571) is below the cost-effectiveness threshold. However, basing the decision on mean costs and effect differences alone would ignore the uncertainty associated with those mean values.

8.3.2.2 Estimation

To what extent are these difficulties overcome by using estimation for cost-effectiveness rather than hypothesis testing? This is consistent with clinical evaluation for which statistical guidelines in a number of medical journals have generally recommended that, in preference to reporting only *P*-values, analyses should report the observed effect size with an associated CI (Gardner and Altman 1986). The advantage of the CI is that it yields information on the *magnitude* of the observed difference (quantitative significance or importance), and this may be useful in presenting the results of economic evaluation.

There have been various ways in which CIs for ICERs have been addressed. Of course, cost-effectiveness incorporates both cost and effect differences. The combination of these simple 95% CIs for cost and effect differences can be portrayed as two-dimensional confidence regions for cost-effectiveness as in Figure 8.5 (O'Brien et al. 1994; O'Brien and Drummond 1994). The simplest definition of the confidence region is the 'confidence box' bounded by *abcd*. Rays from the origin passing through points *a* and *d* define a slice of pie based on the upper limits of each CI. The box approach assumes that the difference in costs is independent of (uncorrelated with) the difference in effects which would be unlikely in most situations. It is also the case that, when costs and effects are independent, the confidence box will represent 90% CIs although it is based on 95% intervals on individual costs and effects (Briggs 2001).

In order to reflect the covariance in cost and effect differences rather than assume independence in these two parts of the ICER, the concept of the confidence ellipse was suggested (O'Brien et al. 1994; O'Brien and Drummond 1994; Van Hout et al. 1994). The precise shape of the ellipsoid confidence region will depend upon covariation between costs and effects. This is illustrated in Figures 8.6 and 8.7, which show situations where there is, respectively, positive and negative correlation between the numerator and denominator of the ICER. The ellipses are derived assuming a joint normal distribution in costs and effects. The rays drawn from the origin as tangents to the ellipses

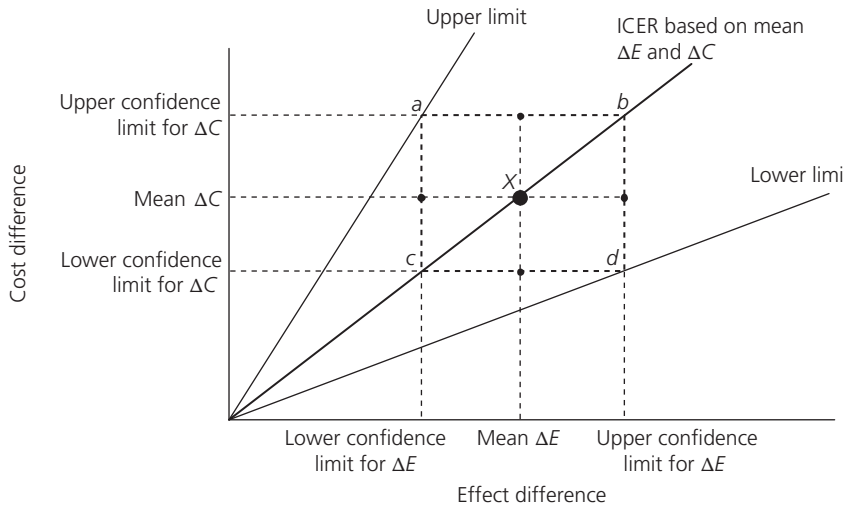


Fig. 8.5 The top right ('north-east') quadrant of the cost-effectiveness plane. This shows mean cost and effect differences (of treatment compared to control) at point X and the associated incremental cost-effectiveness ratio (ICER) represented by the bold line from the origin through that point. The 'confidence box' *abcd* is the combination of confidence intervals in cost and effect differences.

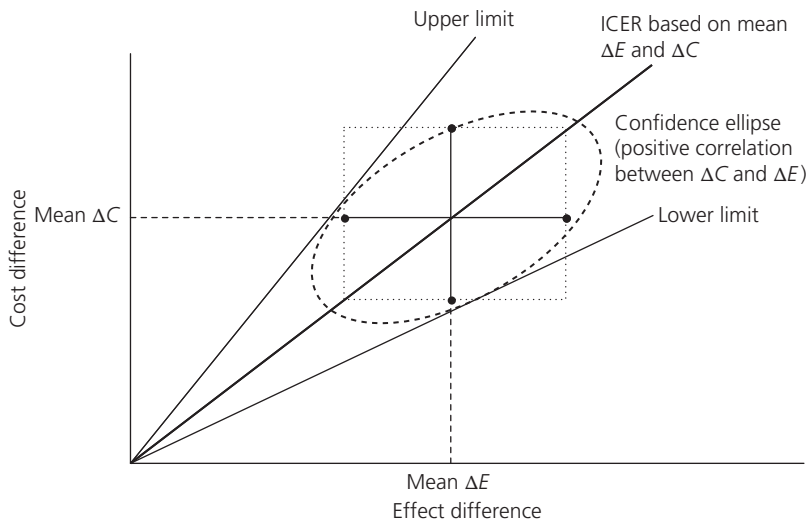


Fig. 8.6 The top right ('north-east') quadrant of the cost-effectiveness plane showing a confidence ellipse when there is positive correlation between costs and effects. ICER, incremental cost-effectiveness ratio.

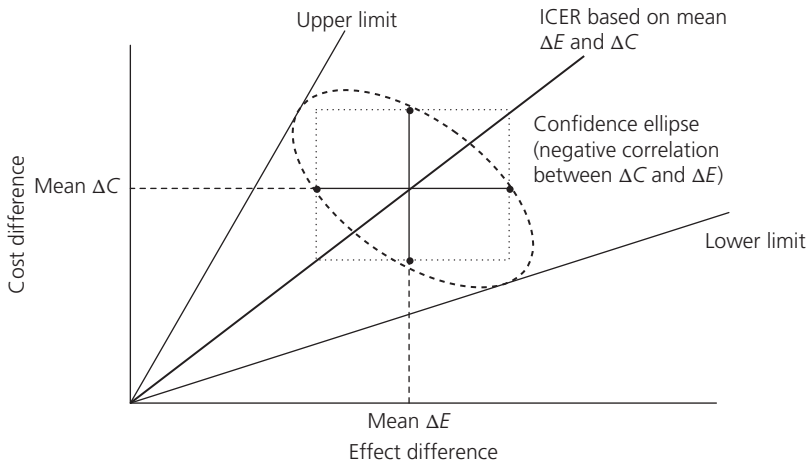


Fig. 8.7 The top right ('north-east') quadrant of the cost-effectiveness plane showing a confidence ellipse when there is negative correlation between costs and effects. ICER, incremental cost-effectiveness ratio.

represent approximations to the 95% CI of the ICER. It is clear from Figures 8.6 and 8.7, however, that the width of the CIs defined by this method is highly dependent on the correlation between costs and effects and would only be the same as that defined by the confidence box by coincidence.

The statistical characteristics of the ICER, which were discussed earlier, mean that it has not been straightforward to define methods to calculate CIs for this measure of cost-effectiveness. Several methods have been suggested to overcome these difficulties (see Briggs 2001), but two methods have been used in the applied literature. The first is Fieller's method, which is based on work on ratio statistics undertaken in the 1930s. This was developed as a means of calculating an exact CI for the ICER allowing for the possible skewness in the sampling distribution of the ratio (Willan and O'Brien 1996). Like the ellipse method, however, the key assumption of Fieller's theorem is that the numerator and denominator of the ICER follow a joint normal distribution. This assumption may be overly strong given the skewness of sampled cost data.

The second method for deriving CIs for the ICER is non-parametric bootstrapping (Efron and Tibshirani 1993). It was discussed more generally in Section 8.3.1.1 and, unlike Fieller's method, it has been widely used in applied cost-effectiveness studies. Rather than making assumptions about the underlying distributions in the ICER, this method resamples from the original data to build an empirical estimate of the sampling distribution of the ICER. Box 8.1 summarizes the bootstrap method as applied to derivation of CIs for the ICER.

8.3.2.3 Moving to net benefits

It is clear that there are various statistical issues with the ICER that have complicated efforts to quantify its sampling uncertainty in a trial-based economic evaluation. In the

Box 8.1 Summary of the stages of non-parametric bootstrap methods

1 Draw a sample from (and of equal size to) the observations of the treatment group by simple random sampling with replacement. Compute $\overline{C}_T^*$ and $\overline{E}_T^*$, the bootstrap replicates of $\overline{C}_T$ and $\overline{E}_T$.

2 Draw a sample from (and of equal size to) the observations of the control group by simple random sampling with replacement. Compute $\overline{C}_C^*$ and $\overline{E}_C^*$, the bootstrap replicates of $\overline{C}_C$ and $\overline{E}_C$.

3 Compute the bootstrap replicate $\hat{R}_b^*$:

$$\hat{R}_b^* = \frac{\overline{C}_T^* - \overline{C}_C^*}{\overline{E}_T^* - \overline{E}_C^*}$$

4 Repeat steps 1–3 a large number of times (say B) and obtain the independent bootstrap replications $\hat{R}_1^*, \hat{R}_2^*, \ldots \hat{R}_B^*$. This is the empirical estimate of the sampling distribution of the ICER.

5 Several methods are then available to calculate CIs from this empirical estimate. For example, a simple approach would be to base a 95% CI on the 2½ and 97½ centiles from the empirical sampling distribution.

face of these statistical difficulties, the use of net benefit (NB) is increasing. The concept of NB was introduced in Chapter 4 as a way of moving away from a ratio and placing both costs and effects on a single scale, either net monetary benefit (NMB) or net health benefit (NHB). In NMB, the difference in effects between two options being evaluated is rescaled into monetary value using the cost-effectiveness threshold as a value for each unit of effect, and the difference in costs between the options is subtracted from this value. NHB is a rescaling of the measure of cost-effectiveness into health by dividing the difference in costs by the cost-effectiveness threshold and then subtracting this from the difference in health effects. Of course, whether NMB, NHB, or the ICER is used, it will provide exactly the same answer in terms of whether a given intervention is cost-effective compared with alternatives for a given cost-effectiveness threshold. Note that some authors define the difference in costs and effect, when rescaled into NB, as *incremental* net benefit.

For statistical analysis, compared to the ICER, either form of NB has the advantage of being a linear expression, which means it is more tractable and has a sampling distribution that is easier to work with. The variance and confidence intervals for NB can be easily defined using parametric methods or bootstrapping (Briggs et al. 2002; Glick et al. 2014), although the issues covered earlier in this section and reviewed by Mihaylova and colleagues are also relevant to NBs (Mihaylova et al. 2011). Given that the value of the cost-effectiveness threshold may not be known with certainty by the analyst (see Chapter 4 for a discussion of this threshold), it is possible to present either

form of NB, and its sampling uncertainty, diagrammatically as a function of the threshold. An example of this, in terms of (incremental) NMB, is shown in Figure 8.8. The example is taken from the laparoscopic versus abdominal hysterectomy study (Sculpher et al. 2004a) using 6-week follow-up data. The bold line on the figure shows the estimated mean NMB (derived as explained in the previous paragraph). Where this line intersects the *x*-axis, NMB is zero and the threshold value on the *x*-axis is equal to the ICER (£203 846 as shown). This line meets the *y*-axis at a value of NMB that is equal to the negative value of the difference in mean cost between the two options (£265 as shown). The slope of the NMB line is equal to the difference in effects. The upper and lower 95% CIs in NMB are also shown in the figure as the lighter curves below and above the mean NMB line. Hence, for a given cost-effectiveness threshold, it is possible to read off the mean NMB as well as the upper and lower 95% CIs. When these CI lines cross the *x*-axis, it is possible to define the 95% CIs for the ICER. In the example shown, the lower 95% CI is defined (£35 000) but this is not the case with the upper 95% CI.

As a result of the statistical flexibility of NB, they can used be used with conventional hypothesis testing. For a given cost-effectiveness threshold, it would be possible to undertake a one- or two-tailed hypothesis test and report the *P*-value. As with estimation, the general lack of clarity about the cost-effectiveness threshold (at least in some jurisdictions) may

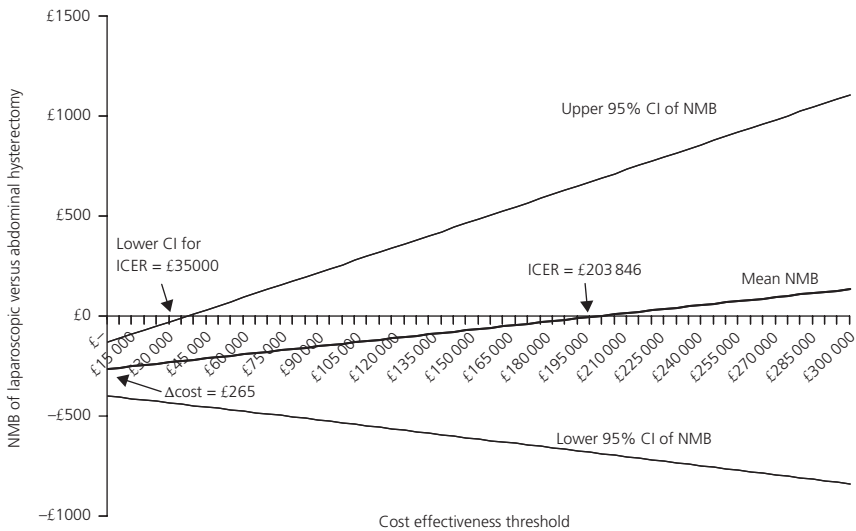


Fig. 8.8 Example of the presentation of incremental net monetary benefit (NMB), together with its sampling uncertainty, as a function of the cost-effectiveness threshold. The example is based on the comparison of laparoscopic and abdominal hysterectomy based on 6-week follow-up (Sculpher et al. 2004a). CI, confidence interval; ICER, incremental cost-effectiveness ratio; QALY, quality-adjusted life-year.

Source: data from Sculpher, M.J. et al., Cost-effectiveness analysis of laparoscopic-assisted hysterectomy in comparison with standard hysterectomy: results from a randomised trial, *BMJ*, Volume 328, pp. 134–40, Copyright © 2004

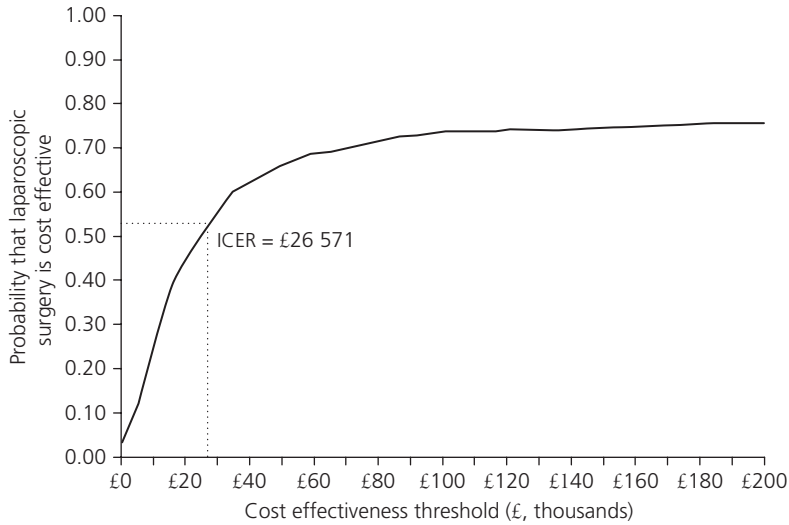


Fig. 8.9 A cost-effectiveness acceptability curve for laparoscopic hysterectomy based on the uncertainty in cost and effect differences. These results are based on 1-year follow-up. ICER, incremental cost-effectiveness ratio.

lead the analyst to repeat hypothesis testing at different thresholds. Figure 8.9 shows these results in the form of a *cost-effectiveness acceptability curve* (CEAC) (Van Hout et al. 1994). Again using the example of laparoscopic versus abdominal hysterectomy, the figure shows, on the *y*-axis, 1 – *P*-value on a one-sided hypothesis of a difference between the treatments in terms of mean NB, as a function of the cost-effectiveness threshold. The CEAC also embodies other information about the comparison. Where it cuts the vertical axis is the *P*-value for a (one-sided) hypothesis test for the difference in mean costs. At the other end, the curve tends towards 1 – *P*-value for a (one-sided) hypothesis test for the difference in mean QALYs. The ICER is also shown on the CEAC. As the ICER is based on mean cost and effect differences, and the 50% point on the CEAC effectively corresponds to the *median* in those differences, the ICER will not necessarily fall at the 50% point (Fenwick et al. 2001). The shape of the CEAC will vary depending on the joint uncertainty in cost and effect differences on the cost-effectiveness plane (Fenwick et al. 2004).

In effect, the *y*-axis shows 'error probabilities' that are typically interpreted as the probability that a treatment is cost-effective on the basis of the available data. Chapter 11 deals in more detail with these error probabilities, their different interpretations, and their use in decision-making.

8.3.3 Explaining variability in cost-effectiveness analysis—the use of regression analysis

As mentioned in Section 8.3.2, in analysing patient-level data on costs and effects, explaining variability in cost-effectiveness can be very important. If the focus of cost-effectiveness analysis is only on deterministic measures of the ICER, formal regression

analysis to explain variability is not possible. One of the implications of quantifying variability in costs and effects using patient-level economic data is that regression methods are made feasible. As discussed in Section 8.3.2, there is a range of regression methods which has been used on cost data which have been reviewed by Mihaylova et al. (2011). The 'statistically challenging' nature of costs described in that section is also true of HRQoL data which, among other things, are often skewed and have spikes at 0 and 1; these may also require quite sophisticated regression methods (Basu and Manca 2012).

Very often regression methods are used to estimate parameters for decision models—for example, a mean cost or a mean HRQoL weight associated with a clinical event. This use of regression analysis is considered in more detail in Chapter 10. In trial-based economic evaluation, where cost-effectiveness is estimated directly from the data, regression methods are important for two reasons. First, regression is a way of adjusting estimated cost-effectiveness for any differences between the types of patients in the treatment groups being compared (i.e. differences in patients' demographic and clinical characteristics). This is essential in non-randomized studies where unbiased mean estimates of cost-effectiveness cannot be assured through random allocation and will always need some form of statistical analysis (see Section 8.3.4). Even in randomized studies, however, statistical adjustment using regression analysis is a means of increasing the precision in mean estimates. This is because even randomized studies will have slight imbalance between treatment groups in terms of baseline characteristics, which occurs simply by chance. This type of adjustment is particularly important where QALYs are being estimated, for each individual patient, based on HRQoL measures at baseline and follow-up. Here the use of regression methods to estimate differences in HRQoL at follow-up between intervention groups needs to adjust for potential differences between the groups in baseline HRQoL (Manca et al. 2005a).

The second reason for considering regression analysis is to assess how patients' characteristics impact on the cost-effectiveness of interventions, and to use this approach as a means of estimating cost-effectiveness for different subgroups of patients. For example, if it were shown that variation in cost-effectiveness could be explained, to some degree, by the gender of the patient, it might be appropriate to present estimates of the cost-effectiveness of a treatment separately for men and women. As discussed further in Chapters 9 and 10, this consideration of heterogeneity in cost-effectiveness as a key source of information for decision-making. There is a range of important caveats in how subgroups are identified and presented, including the need for their biological plausibility, and many would argue that this can only be assured if the subgroup effects to be estimated are defined before the data are available.

Hoch and colleagues made an important initial contribution to the development of regression methods for cost-effectiveness by using the concept of NMB (see Section 8.3.2) as the dependent variable, calculated for each individual (i) (Hoch et al. 2002). Using a model regressing this patient-level NMB_i against the treatment arm dummy variable (t_i), Hoch and colleagues demonstrated the equivalence of a regression-based approach to CEA with a 'standard' cost-effectiveness analysis. Their regression framework is illustrated in the equation below:

$$\text{NMB}_i = \alpha + \beta t_i + e_i$$

In this formulation, the NMB for the *i*th patient in the trial is the patient-level NB defined in terms of their absolute costs and effects, and t_i represents a treatment dummy taking the value 0 for 'standard' or comparator therapy and 1 for the new intervention. In the context of a trial, this dummy variable would be defined in terms of the group to which the individual patient is randomized. The coefficients α and β are, respectively, the intercept and the slope term obtained from a standard OLS regression. The term e_i is an error term with constant variance usually assumed to be normally distributed. In terms of the interpretation of the results from the OLS regression, the estimated coefficient α represents the mean NMB in the group receiving the 'standard treatment' in the trial, the sum of the two estimated coefficients, $\alpha + \beta$, is the mean NMB in the new intervention arm, and β is the incremental NMB between the two arms of the trial.

A potential limitation of NB regression is that the use of a single dependent variable, NB, risks neglecting important differences between the costs and effects from which it is derived. For example, the covariables which provide important explanatory variables for costs may not do so for effects, and vice versa. An alternative approach to regression analysis for cost-effectiveness is to estimate two separate statistical models, one for costs and one for effects. As noted in Section 8.3.2, however, costs and effects are correlated, so it would be inappropriate to assume that these two regression models are independent. Willan and colleagues used the method of 'seemingly unrelated regression' where different covariables and functional forms for the two equations can be used for the cost and effect regressions, but the error terms are correlated (Willan et al. 2004). The use of statistical modelling using regression methods for economic data is discussed more fully in Chapter 10.

A third contribution offered by regression analysis is to assess whether cost-effectiveness varies according to the location in which the patient was treated (e.g. the centre or country in a trial setting). Cook and colleagues used a range of statistical analyses to assess whether there is heterogeneity in cost-effectiveness between alternative countries in a randomized trial (Cook et al. 2003). The use of formal regression methods for this purpose was explored by others. Multilevel regression modelling is an approach which has been used to assess the extent to which costs, effects, and cost-effectiveness vary systematically across locations in RCTs and other studies (Grieve et al. 2005, 2007; Manca et al. 2005b; Willan et al. 2005). This approach can also be used to generate location-specific estimates of costs, outcomes, or cost-effectiveness which takes into consideration the variation in data between and within locations. The use of regression methods to explore issues relating to the transferability of economic data from setting to setting is discussed further in Chapter 9.

8.3.4 Analysis of observational studies

The statistical analysis of observational studies presents particular challenges because the absence of randomized treatment allocation risks biased estimates of means costs, effects and cost-effectiveness (see Section 8.2.3). Various statistical methods are available to try to minimize this risk of bias: in general, they can be grouped into three types, which follow the general principle of allowing for observed differences in baseline characteristics between the treatment and control groups. The first uses regression

methods simply to adjust estimated treatment effects using patients' clinical and other characteristics at study outset. The second approach is matching, where an attempt is made to estimate treatment effects while achieving a balance in observed characteristics between the comparison groups in the matched samples (Sekhon and Grieve 2011). A popular approach to matching is to do so on the basis of propensity scores which are the predicted probability of a given patient being assigned to a particular intervention on the basis of a set of observed covariables (Mitra and Indurkhya 2005). In such studies it is necessary to establish that, after matching on the propensity score, the key baseline prognostic characteristics are balanced between the groups being compared (Kreif et al. 2012).

The problem with regression methods and matching is that they assume that all potentially confounding variables relating to an estimated treatment effect have been observed (i.e. data have been collected on them within the study). A method which, in principle, also allows for unobserved confounding variables is the use of *instrumental variables* (IVs) (Grootendorst 2007). If appropriate IVs can be identified, they address the challenge of unobserved confounding variables and the risk of selection bias. In a statistical model, IVs need to be able to predict patients' treatment allocation conditional on other observed covariables in the model; but they should be statistically unrelated to the outcome of interest. An example of the use of IVs is an evaluation of invasive treatment of acute myocardial infarction in elderly patients based on observational data from Medicare in the United States (McClellan et al. 1994). The selected IV was the distance between patients' homes and the hospitals providing different types of care. This variable was found to predict accurately the type of treatment patients received but was not correlated with health outcomes. The analysis used differential distances effectively to randomize patients to different likelihoods of receiving intensive treatments.

The nature of the assumptions that need to be made in order to estimate treatment effects (on costs or outcomes) from observational studies is important to understand. This is true for those designing and analysing observational studies and for those using them in decision-making. A recent study developed a checklist for assessing the quality of such analyses (Kreif et al. 2013). This includes a series of questions about methods including whether or not the study assessed the validity of the assumption of no unobserved confounding, whether an assessment was made regarding the extent to which the distributions of the baseline covariables overlapped between the treatment groups, and whether covariate balance was assessed after applying a matching method. The checklist was used to review 81 systematically identified economic evaluations based on observational data. Of these, 51% used regression methods to address potential selection bias, whereas 47% using matching methods (based on propensity scores or covariables) and only 2% used IVs. The authors concluded that economic evaluations based on observational studies rarely assess the main assumptions used in undertaking statistical analyses to address selection bias.

8.4 **Conclusions**

The use of clinical studies as vehicles for economic evaluation is widespread in the published literature. There has also been considerable interest from researchers in

developing statistical methods to analyse such studies. These include the use of individual patient data on costs and effects to improve the way uncertainty is characterized, and the use of regression methods to reflect the heterogeneity in cost-effectiveness between patients and locations. There are, however, some major concerns with the use of single data sources as a vehicle for economic evaluation to support decision-making including the inappropriate selection of comparators and the failure to incorporate all relevant evidence. These issues are discussed more fully in Chapter 9.

It is often the case, however, that many interventions, programmes and policy changes are not easy to evaluate using experimental (randomized) designs. Furthermore, many health care systems are investing in the routine collection of clinical and other data on individual patients, either for administrative or research purposes. These routine observational datasets provide a valuable opportunity to assess the effectiveness and cost-effectiveness of alternative ways of allocating resources (Heckman 2007). However, the statistical methods required to analyse these data in a way that reflects known and unknown sources of confounding are challenging and developing rapidly.

8.5 **Exercise**

Read the following journal article:

> Mihaylova, B., R. Pitman, D. Tincello, H. van der Vaart, R. Tunn, L. Timlin, D. Quail, A. Johns and M. Sculpher (2010). 'Cost-effectiveness of duloxetine: the Stress Urinary Incontinence Treatment (SUIT) study.' *Value in Health*, **13**, 565–72.

Address the following questions:

- To what extent would this study be susceptible to selection bias? Provide some examples of potential sources of bias.
- How do the authors seek to address the risk of bias?
- What are the strengths and limitations of the study to inform decision-making about the value of duloxetine in this indication?

References

Bagust, A., Grayson, A.D., Palmer, N.D., et al. (2006). Cost-effectiveness of drug-eluting coronary artery stenting in a UK setting: cost-utility study. *Heart*, **92**, 68–74.

Basu, A. and Manca, A. (2012). Regression estimators for generic health-related quality of life and quality-adjusted life years. *Medical Decision Making*, **32**, 56–69.

Bojke, L., Epstein, D., Craig, D., et al. (2011). Modelling the cost-effectiveness of biologic treatments for psoriatic arthritis. *Rheumatology*, **50**(suppl 4), iv39–iv47.

Briggs, A. (2001). Handling uncertainty in economic evaluation and presenting the results, in M. Drummond and A.J. McGuire (ed.), *Economic evaluation in health care: merging theory with practice*, pp. 172–214. Oxford: Oxford University Press.

Briggs, A.H. and O'Brien, B.J. (2001). The death of cost-minimisation analysis? *Health Economics*, **10**, 179–84.

Briggs, A.H., Clark, T., Wolstenholme, J., et al. (2003). Missing . . . presumed at random: cost-analysis of incomplete data. *Health Economics*, **12**, 377–92.

Briggs, A.H., O'Brien, B.J., and Blackhouse, G. (2002). Thinking outside the box: Recent advances in the analysis and presentation of uncertainty in cost-effectiveness studies. *Annual Review of Public Health*, **23**, 377–401.

Carpenter, J.R. and Kenwood, M.G. (2013). *Multiple imputation and its application*. Chichester: Wiley.

Cook, J.R., Drummond, M., Glick, H., et al. (2003). Assessing the appropriateness of combining economic data from multinational clinical trials. *Statistics in Medicine*, **22**, 1955–76.

Dolan, P., Gudex, C., Kind, P., et al. (1996). Valuing health states: a comparison of methods. *Journal of Health Economics*, **15**, 209–31.

Drummond, M.F., Bloom, B.S., Carrin, G., et al. (1992). Issues in the cross-national assessment of health technology. *International Journal of Technology Assessment in Health Care*, **8**, 671–82.

Efron, B. and Tibshirani, R. (1993). *An introduction to the bootstrap*. New York: Chapman & Hall.

Epstein, D., Bojke, L., Sculpher, M.J., et al. (2009). Laparoscopic fundoplication compared with medical management for gastro-oesophageal reflux disease: cost effectiveness study. *BMJ*, **339**, b2576.

Etzioni, R.D., Feuer, E.J., Sullivan, S.D., et al. (1999). On the use of survival analysis techniques to estimate medical care costs. *Journal of Health Economics*, **18**, 365–80.

Fenwick, E., Claxton, K., and Sculpher, M. (2001). Representing uncertainty: the role of cost-effectiveness acceptability curves. *Health Economics*, **10**, 779–89.

Fenwick, E., O'Brien, B.J., and Briggs, A. (2004). Cost-effectiveness acceptability curves—facts, fallacies and frequently asked questions. *Health Economics*, **13**, 405–15.

Freemantle, N. and Drummond, M. (1997). Should clinical trials with concurrent economic analyses be blinded? *JAMA*, **277**, 63–4.

Gardner, M.J. and Altman, D.G. (1986). Confidence intervals rather than P values: estimation rather than hypothesis testing. *BMJ*, **292**, 746–50.

Garry, R., Fountain, J., Brown, J., et al. (2004). EVALUATE hysterectomy trial: a multicentre randomised trial comparing abdominal, vaginal and laparoscopic methods of hysterectomy. *Health Technology Assessment*, **8**(26), 1–154.

Glick, H., Polsky, D., and Schulman, K. (2001). Trial-based economic evaluations: an overview of design and analysis, in M.F. Drummond and A. McGuire (ed.), *Economic evaluation in health care: merging theory with practice*, pp. 113–40. Oxford: Oxford University Press.

Glick, H.A., Doshi, J.A., Sonnad, S.S., et al. (2014). *Economic evaluation in clinical trials*, 2nd edition. Oxford: Oxford University Press.

Goeree, R., Bowen, J.M., Blackhouse, G., et al. (2009). Economic evaluation of drug-eluting stents compared to bare metal stents using a large prospective study in Ontario. *International Journal of Technology Assessment in Health Care*, **25**, 196–207.

Gomes, M., Díaz-Ordaz, K., Grieve, R., et al. (2013). Multiple imputation methods for handling missing data in cost-effectiveness analyses that use data from hierarchical studies: an application to cluster randomized trials. *Medical Decision Making*, **33**, 1051–63.

Grant, A.M., Boachie, C., Cotton, S.C., et al. (2013). Clinical and economic evaluation of laparoscopic surgery compared with medical management for gastro-oesophageal reflux disease: 5-year follow-up of multicentre randomised trial (the REFLUX trial). *Health Technology Assessment*, **17**(22), 1–167.

Grant, A., Wileman, S., Ramsay, C., et al. (2008a). The effectiveness and cost-effectiveness of minimal access surgery amongst people with gastro-oesophageal reflux disease—a UK collaborative study. The REFLUX trial. *Health Technology Assessment*, **12**(31), 1–204.

Grant, A.M., Wileman, S.M., Ramsay, C.R., et al. (2008b). Minimal access surgery compared with medical management for chronic gastro-oesophageal reflux disease: UK collaborative randomised trial. *BMJ*, **337**, a2664.

Grieve, R., Nixon, R., Thompson, S., et al. (2005). Using multilevel models for assessing the variability of multinational resource use and cost data. *Health Economics*, **14**, 185–96.

Grieve, R., Thompson, S., Nixon, R., et al. (2007). Multilevel models for estimating incremental net benefits in multinational studies. *Health Economics*, **16**, 815–26.

Grootendorst, P. (2007). A review of instrumental variables estimation of treatment effects in the applied health sciences. *Health Services and Outcomes Research Methodology*, **7**, 159–79.

Hartman, E., Grieve, R., Ramsahai, R., and Sekhon, J.S. (2015). From sample average treatment effect to population average treatment effect on the treated: Combining experimental with observational studies to estimate population treatment effects. *Journal of the Royal Statistical Society Series A*, **178**, 757–78.

Heart Protection Study Collaborative Group (2002). MRC/BHF Heart Protection Study of cholesterol lowering with simvastatin in 20536 high-risk individuals: a randomised placebo-controlled trial. *Lancet*, **360**, 7–22.

Heckman, J.J. (2007). Econometric evaluation of social programs, Part I: causal models, structural models and econometric policy evaluation, in J.J. Heckman and E.E. Leamer (ed.), *Handbook of econometrics*, Vol. 6, Part B, pp. 4875–5143. Amsterdam: Elsevier.

Hoch, J.S., Briggs, A.H., and Willan, A. (2002). Something old, something new, something borrowed, something BLUE: a framework for the marriage of health econometrics and cost-effectiveness analysis. *Health Economics*, **11**, 415–30.

Hughes, D.A., Bagust, A., Haycox, A., et al. (2001). The impact of non-compliance on the cost-effectiveness of pharmaceuticals: a review of the literature. *Health Economics*, **10**, 601–15.

Johannesson, M., Jönsson, B., Kjekshus, J., et al. (1997). Cost effectiveness of simvastatin treatment to lower cholesterol levels in patients with coronary heart disease. *New England Journal of Medicine*, **336**, 332–6.

Jones, A.M. and Rice, N. (2011). Econometric evaluation of health policies, in S. Glied and P.C. Smith (ed.), *The Oxford handbook of health economics*. Oxford: Oxford University Press.

Joyce, V.R., Barnett, P.G., Chow, A., et al. (2009). Health-related quality of life in a randomized trial of antiretroviral therapy for advanced HIV disease. *Journal of Acquired Immunodeficiency Syndromes*, **50**, 27–36.

Kreif, N., Grieve, R., Radice, R., et al. (2012). Methods for estimating subgroup effects in cost-effectiveness analyses that use observational data. *Medical Decision Making*, **32**, 750–63.

Kreif, N., Grieve, R., and Sadique, M.Z. (2013). Statistical methods for cost-effectiveness analyses that use observational data: a critical appraisal tool and review of current practice. *Health Economics*, **22**, 486–500.

Little, R.J.A. and Rubin, D.B. (1987). *Statistical analysis with missing data*. New York: Wiley.

Machin, D., Cheung, Y.B., and Parmar, M.K.B. (2006). *Survival analysis: a practical approach*, 2nd edition. Chichester: Wiley.

Manca, A. and Palmer, S. (2001). Handling missing data in patient-level cost-effectiveness analysis alongside randomised controlled trials. *Applied Health Economics and Health Policy*, **4**, 65–75.

Manca, A., Hawkins, N., and Sculpher, M. (2005a). Estimating mean QALYs in trial-based cost-effectiveness analysis: the importance of controlling for baseline utility. *Health Economics*, **14**, 487–96.

Manca, A., Rice, N., Sculpher, M.J., et al. (2005b). Assessing generalisability by location in trial-based cost-effectiveness analysis: the use of multilevel models. *Health Economics*, **14**, 471–85.

Marshall, A., Billingham, L.J., and Bryan, S. (2009). Can we afford to ignore missing data in cost-effectiveness analyses? *European Journal of Health Economics*, **10**, 1–3.

Matthews, J.N.S., Altman, D., and Campbell, M.J. (1990). Analysis of serial measurements in medical research. *BMJ*, **300**, 230–5.

Mauskopf, J., Schulman, K., Bell, L., et al. (1996). A strategy for collecting pharmacoeconomic data during phase II/III clinical trials. *PharmacoEconomics*, **9**, 264–77.

McClellan, M., McNeil, B.J., and Newhouse, J.P. (1994). Does more intensive treatment of acute myocardial infarction in the elderly reduce mortality? Analysis using instrumental variables.[see comment]. *JAMA*, **272**, 859–66.

McErlane, F., Foster, H.E., Davies, R., et al. (2013). Biologic treatment response among adults with juvenile idiopathic arthritis: results from the British Society for Rheumatology Biologics Register. *Rheumatology (Oxford)*, **5**, 1905–13.

Mihaylova, B., Briggs, A., Armitage, J., et al. (2005). Cost-effectiveness of simvastatin in people at different levels of vascular disease risk: economic analysis of a randomised trial in 20 536 individuals. *Lancet*, **365**, 1779–85.

Mihaylova, B., Pitman, R., Tincello, D., et al. (2010). Cost-effectiveness of duloxetine: the Stress Urinary Incontinence Treatment (SUIT) study. *Value in Health*, **13**, 565–72.

Mihaylova, B., Briggs, A., O'Hagan, A., et al. (2011). Review of statistical methods for analysing healthcare resources and costs. *Health Economics*, **20**, 897–916.

Mitra, N. and Indurkhya, A. (2005). A propensity score approach to estimating the cost effectiveness of medical therapies from observational data. *Health Economics*, **14**, 805–15.

Morris, S., McGuire, A., Caro, J., et al. (1997). Strategies for the management of hypercholesterolaemia: a systematic review of the cost-effectiveness literature. *Journal of Health Services Research and Policy*, **2**, 231–50.

Nixon, R.M., Wonderling, D., and Grieve, R.D. (2010). Non-parametric methods for cost-effectiveness analysis: central limit theorem and the bootstrap compared. *Health Economics*, **19**, 316–333.

Noble, S.M., Hollingworth, W., and Tilling, K. (2010). Missing data in trial-based cost-effectiveness analysis: the current state of play. *Health Economics*, **21**, 187–200.

O'Brien, B.J. and Drummond, M.F. (1994). Statistical versus quantitative significance in the socioeconomic evaluation of medicines. *PharmacoEconomics*, **5**, 389–98.

O'Brien, B.J., Drummond, M.F., Labelle, R.J., et al. (1994). In search of power and significance: issues in the design and analysis of stochastic cost-effectiveness studies in health care. *Medical Care*, **32**, 150–63.

O'Hagan, A. and Stevens, J.W. (2004). On estimators of medical costs with censored data. *Journal of Health Economics*, **23**, 615–25.

Pennington, M., Grieve, R., Sekhon, J.S., et al. (2013). Cemented, cementless, and hybrid prostheses for total hip replacement: cost effectiveness analysis. *BMJ*, **346**, f1026.

Raikou, M. and McGuire, A. (2012). Estimating costs for economic evaluation, in A.M. Jones (ed.), *The Elgar companion to health economics*. Cheltenham: Edward Elgar.

Ramsey, S.D., Willke, R.J., Glick, H., et al. (2015). Cost-effectiveness analysis alongside clinical trials II—An ISPOR Good Research Practices Task Force Report. *Value in Health*, **18**, 161–72.

Ridyard, C.H. and Hughes, D.A. (2010). Methods for the collection of resource use data within clinical trials: a systematic review of studies funded by the UK Health Technology Assessment Program. *Value in Health*, **13**, 867–72.

Ridyard, C.H., Hughes, D.A.; DIRUM Team (2012). Development of a database of instruments for resource-use measurement: purpose, feasibility, and design. *Value in Health*, **15**, 650–5.

Rodgers, M., Epstein, D., and Bojke, L., et al. (2011). Etanercept, infliximab and adalimumab for the treatment of psoriatic arthritis: a systematic review and economic evaluation. *Health Technology Assessment*, **15**(10), 1–329.

Schwartz, D. and Lellouch, J. (1967). Explanatory and pragmatic attitudes in therapeutic trials. *Journal of Chronic Diseases*, **20**, 637–48.

Sculpher, M.J. and Buxton, M.J. (1993). The episode-free day as a composite measure of effectiveness. *PharmacoEconomics*, **4**, 345–52.

Sculpher, M.J., Manca, A., Abbott, J., et al. (2004a). Cost-effectiveness analysis of laparoscopic-assisted hysterectomy in comparison with standard hysterectomy: results from a randomised trial. *BMJ*, **328**, 134–40.

Sculpher, M.J., Pang, F.S., Manca, A., et al. (2004b). Generalisability in economic evaluation studies in health care: a review and case studies. *Health Technology Assessment*, **8**(49), 1–213.

Sculpher, M.J., Claxton, K.P., Drummond, M.F., et al. (2006). Whither trial-based economic evaluation for health care decision making? *Health Economics*, **15**, 677–87.

Sekhon, J. and Grieve, R. (2011). A matching method for improving covariate balance in cost-effectiveness analyses. *Health Economics*, **21**, 695–714.

Sox, H.C. and Greenfield, S. (2009). Comparative effectiveness research: a report from the Institute of Medicine. *Annals of Internal Medicine*, **151**, 203–5.

Sterne, J.A.C., White, I.R., Carlin, J.R., et al. (2009). Multiple imputation for missing data in epidemiological and clinical research: potential and pitfalls. *BMJ*, **338**, b2393.

Sutton, M., Nikolova, S., Boaden, R., et al. (2012). Reduced mortality with hospital pay for performance in England. *New England Journal of Medicine*, **367**, 1821–8.

Thompson, M.S., Read, J.L., Hutchings, H.C., et al. (1989). The cost-effectiveness of auranofin: results from a randomised controlled trial. *Journal of Rheumatology*, **15**, 35–42.

Thorpe, K.E., Zwarenstein, M., Oxman, A.D., et al. (2009). A pragmatic–explanatory continuum indicator summary (PRECIS): a tool to help trial designers. *Journal of Clinical Epidemiology*, **62**, 464–75.

Van Hout, B.A., Al, M.J., Gordon, G.S., et al. (1994). Costs, effects and c/e-ratios alongside a clinical trial. *Health Economics*, **3**, 309–19.

Wijeysundera, H.C., Wang, X., Tomlinson, G., et al. (2012). Techniques for estimating health care costs with censored data: an overview for health services researchers. *ClinicoEconomics and Outcomes Research*, **4**, 145–55.

Willan, A.R. and Briggs, A.H. (2006). *Statistical analysis of cost-effectiveness data*. Chichester: Wiley.

Willan, A and O'Brien, B. (1996). Confidence intervals for cost-effectiveness ratios: an application of Fieller's theorem. *Health Economics*, **5**, 297–305.

Willan, A.R., Briggs, A.H., and Hoch, J.S. (2004). Regression methods for covariate adjustment and subgroup analysis for non-censored cost-effectiveness data. *Health Economics*, **13**, 461–75.

Willan, A.R., Pinto, E.M., O'Brien, B.J., et al. (2005). Country specific cost comparisons from multinational clinical trials using empirical Bayesian shrinkage estimation: the Canadian ASSENT-3 economic analysis. *Health Economics*, **14**, 327–38.

Yu, L.M., Burton, A., and Rivero-Aris, O. (2007). Evaluation of software for multiple imputation of semi-continuous data. *Statistical Methods in Medical Research*, **16**, 243–58.

Chapter 9

Economic evaluation using decision-analytic modelling

9.1 Some basics

Chapter 8 considered economic evaluation based on a single study such as a randomized controlled trial (RCT), where data collected from a sample of patients or other study participants on resource use and outcomes facilitate estimates of differential mean costs and effects relating to different interventions. Although such studies are still undertaken and published, there is a growing use of decision-analytic modelling as an alternative vehicle for economic evaluation.

The ultimate purpose of economic evaluation is to inform different types of decision-maker about the efficient allocation of health care resources. In recent years, economic evaluation has been increasingly undertaken for specific decision-makers, who formally require economic evidence. For example, a number of health care systems now use economic evaluation to help them to decide whether new health technologies (particularly pharmaceuticals) represent sufficient value for money to be funded (see <http://www.ispor.org/HTARoadMaps/Default.asp> for details). The role of economic evaluation in decision-making is discussed in more detail in Chapter 2, but the greater use of these methods to inform particular decisions in specific jurisdictions has had implications for economic evaluation (Claxton et al. 2002). In particular, it has indicated that relying on a randomized trial as a single vehicle for economic evaluation has a number of limitations (Sculpher et al. 2006). As a result, economic evaluation for decision-making will usually need to draw on evidence from a range of sources. These could include clinical, resource use, and outcome data collected alongside randomized trials, but are also likely to include evidence from other types of study such as cohort studies and surveys. A decision-analytic model provides a means of bringing together this full range of evidence and directing it at a specific decision problem being addressed by a health system at a given point in time and in a particular jurisdiction.

Decision-analytic modelling has its theoretical foundations in statistical decision theory (Raiffa 1968), and shares common theoretical origins with expected utility theory (discussed in Chapter 5). It also has a close association with Bayesian statistics where statistical analysis is closely related to decision-making (Spiegelhalter et al. 2004). Decision analysis has been widely used outside health care, such as in business and engineering. It has an established basis as a framework for *clinical* decision-making—that is, decision-making relating to individual patients when costs are not necessarily a primary consideration—and several good texts exist introducing decision analysis in health care generally (Hunink et al. 2014; Weinstein and Fineberg 1980).

This chapter introduces the use of these methods in economic evaluation in particular. More advanced material on decision analysis for economic evaluation is available (Briggs et al. 2006).

Decision-analytic modelling provides a framework for decision-making under conditions of uncertainty. More specifically, a decision-analytic model defines a set of mathematical relationships between entities (usually health states or pathways) characterizing the range of possible disease prognoses and the impacts of alternative interventions. These entities themselves predict the quantities we are interested in for economic evaluation: costs and health effects. As a set of methods, decision analysis can satisfy five important objectives for any economic evaluation as discussed briefly here and more fully in subsequent sections.

- *Structure*. It can provide a structure that appropriately reflects the possible prognoses that individuals of interest may experience, and how the interventions being evaluated may impact on these prognoses. The structure will have to reflect the variability between apparently similar individuals in their prognoses and in the effect of interventions. The individuals will often be patients with a particular condition, but may be healthy or asymptomatic people in the context, for example, of a screening or public health programme.
- *Evidence*. It offers an analytic framework within which the full range of evidence relevant to the study question can be brought to bear. This is achieved partly through the structure of the model, but also in the estimates of the input parameters of the model.
- *Evaluation*. It provides a means of translating the full extent of relevant evidence into estimates of the cost and effects of the alternative options being compared. Using appropriate decision rules (see Chapter 4), the option that the analysis identifies as the best can be identified based on existing evidence.
- *Uncertainty, variability, and heterogeneity*. It facilitates an assessment of the various types of uncertainty relating to the evaluation. As described in detail in Chapter 11, this includes uncertainty relating to model structure and input parameters. Models also provide flexibility to characterize heterogeneity across different subgroups of individuals.
- *Future research*. Through the assessment of uncertainty, it can estimate the value of future research, and identify likely priorities for it. This can provide a more nuanced set of decision options relating to research development to be considered alongside a more standard 'adopt' or 'reject' (see Chapter 11).

9.2 The role of decision-analytic models for economic evaluation

In considering the role of decision models in economic evaluation, it is useful to contrast two different activities in health care evaluation—measurement and decision analysis—which are summarized in Box 9.1. In part, economic evaluation is concerned with the process of measurement through the collection of data relating to effectiveness, resource

Box 9.1 Contrasting activities in economic evaluation: measurement versus decision analysis

Health care evaluation in general, and economic evaluation in particular, involves two important but separate activities: measurement and decision analysis. Both are important in establishing the most economically appropriate form of management or intervention, but they have distinct roles. The process of measurement has the following features:

- A focus on estimating and testing hypotheses relating to particular parameters and relationships between parameters (e.g. the rate of clinical events, relative treatment effects, resource use, and HRQoL effects).
- A concentration on relatively few parameters.
- A primary interest in randomized trials as a vehicle for measurement, particularly of relative treatment effects.
- The focus of uncertainty analysis is on parameters, usually represented in terms of hypothesis tests.

The activity of decision analysis can be characterized as:

- The primacy of identifying an appropriate course of action from amongst a full range of alternatives for a specific recipient group in a particular location/jurisdiction.
- The process of informing decisions based on all relevant and currently available evidence.
- Identification of a preferred option based on the expected values of the alternatives (e.g. expected cost-effectiveness) rather than on individual parameters.
- An explicit acceptance that decisions will always be taken under conditions of uncertainty.

use, unit costs, and health-related quality of life (HRQoL) weights. Ultimately, though, economic evaluation is concerned with informing appropriate decisions in health care about resource allocation in specific jurisdictions under conditions of uncertainty. This will require the appropriate synthesis of all relevant evidence (see Chapter 10). Being clear about these different roles for economic evaluation emphasizes that decision models and randomized trials (and other clinical studies) are not competing alternatives. Rather, the latter are focused on measuring different effects of interventions on relevant costs and outcomes. Decision models, on the other hand, are concerned with informing specific decisions (Sculpher et al. 2006); they draw upon the measurements undertaken in clinical studies but also provide an explicit framework for the inevitable assumptions and judgements needed in decision-making.

This section considers the features of economic evaluation for decision-making that frequently necessitate the use of decision-analytic models rather than reliance on a

single data source such as an RCT. There are several different requirements for an economic evaluation that are relevant, and these are detailed in Sections 9.2.1–9.2.6.

9.2.1 The need to compare all options

As described in Chapters 2, 3, and 4, economic evaluation is about comparing the value for money of alternative courses of action (or options) for particular recipient groups. It is possible that decision-makers will be misled by a study that fails to compare all the relevant options, which might reflect the fact that more than one option is currently used in practice as well as the new option(s) available. Indeed, defining all relevant options will often involve specifying strategies rather than specific treatments. In the evaluation of pharmaceuticals this can take the form of the comparison of sequences of interventions. For example, Woolacott and colleagues compared the cost-effectiveness of alternative sequences of treatments for psoriasis where a treatment is typically used until it ceases to be effective or exhibits side effects (Woolacott et al. 2006). The alternative sequences consisted of several new biological treatments as well as older therapies and best supportive care. In the economic evaluation of diagnostics, strategies are specified in terms of sequences of tests and alternative decision rules about therapeutic choices conditional on test results. As a result, the number of options to compare can be very large. A particularly notable example is the 12^{14} (i.e. 1 283 918 464 548 860) options initially compared for prenatal screening and treatment to prevent group B streptococcal and other bacterial infections in early infancy (Colbourn et al. 2007). Clearly, in such circumstances, approaches need to be identified to make the number of options practical. In this study, uncertainty analysis was used to remove options from consideration that had a very limited probability of being cost-effective.

However, as discussed in Chapter 8, in randomized trials it is rarely the practice to compare all relevant options and a subset is typically studied. In some studies, moreover, the comparator is not an active intervention at all, but a placebo. To compare all relevant options, it is likely that effectiveness data will have to be taken from several trials. For example, in a cost-effectiveness analysis (CEA) of primary angioplasty compared to medical therapy in patients with myocardial infarction, 22 randomized trials were used as the basis of estimating the relative effectiveness for the analysis (Bravo Vergel et al. 2007). Meta-analysis and other forms of evidence synthesis will be necessary to synthesize this type of evidence (see Chapter 10), but the decision model will provide the framework to combine it with other types of evidence such as the underlying risk of clinical events (sometimes called natural history), resource use, and HRQoL weights. Models also provide a means of structuring the relationships between clinical variables and how their magnitudes change over time.

9.2.2 The need to reflect all relevant evidence

To offer a decision-maker guidance on the best course of action from an economic perspective, for a given patient group, it is important that all relevant evidence is brought to bear on the decision problem. This is consistent with the axioms of evidence-based medicine, where appropriate evidence is systematically and comprehensibly used to

make clinical decisions (Sackett et al. 1996). For economic evaluation, however, it is not just effectiveness evidence that is required. In addition, evidence relating to resource use, unit costs, and HRQoL is necessary, as well as the relationship between different parameters and how they change with time. This array of evidence is often not collected in trials and, if it is, it is unlikely to be the only source. Again, the decision model is used to combine all sources of evidence.

9.2.3 The need to link intermediate to final end points

As discussed in Section 8.2.1.2, many clinical studies measure effects in terms of end points which are clinically meaningful but which are only indirectly related to the ultimate measures of health that are central to most economic evaluations (e.g. changes in life expectancy and/or patients' HRQoL). Examples are the use of CD4 count and viral load in studies of the efficacy of therapies for HIV, time until progression in cancer, and cases detected in screening studies. Intermediate end points are often a challenge when assessing the cost-effectiveness of diagnostic strategies as clinical studies often focus on the accuracy of tests using, for example, estimates of sensitivity and specificity. A focus on these *intermediate* end points will limit the value of clinical studies as a vehicle for economic evaluation as they are unlikely to provide a reliable estimate of ultimate health outcomes. Decision analysis provides a means of linking intermediate and final outcomes. Evidence is still required on the relationships between the different types of outcomes, and these may be taken from observational studies. The clinical plausibility of these links and the associated uncertainty needs to be fully considered in developing the model. Box 9.2 contains some examples of studies which used models to link intermediate to final outcomes for economic evaluation.

9.2.4 The need to extrapolate over the appropriate time horizon of the evaluation

The issue of the appropriate time horizon for an economic evaluation was discussed in Chapter 3. In principle, the time horizon should be the period over which the costs and/or effects of the alternative options being compared might be expected to differ. Often the appropriate time horizon will need to be the patient's lifetime to capture these differences fully. For example, in the case of the treatment of chronic disease, the initiation of an intervention in a middle-aged patient may have cost and effect implications for the remainder of their life. Except in rare cases where palliative treatments are being compared (e.g. for advanced cancer), clinical studies will not follow all patients up until they die. An important role of decision models, then, is to bridge the gap between what has been observed in trials and what would be expected to happen, in terms of costs and effects, over a long-term time horizon.

It is frequently necessary to extrapolate when the options being compared differ in terms of mortality and this difference is to be expressed in terms of life-years or quality-adjusted life-years (QALYs) gained. This is illustrated in Figure 9.1 in the form of the survival curves of two interventions (treatment and control) being compared in a randomized trial. These show the proportion of patients surviving until particular

Box 9.2 Examples of studies where decision models have been used to link intermediate end points to measures of ultimate health outcome

Study	Intermediate endpoints	Method of extrapolation
A model to assess the cost-effectiveness of donepezil in mild to moderate Alzheimer's disease (AD) (Neumann et al. 1999)	Treatment effect from trial between baseline and 24 weeks in terms of transition between Clinical Dementia Rating (CDR) scale 2 (moderate) and CDR 0.5 or 1 (mild)	Markov model with states defined in terms of mild, moderate, and severe disease (in terms of CDR); also a dead state. Baseline disease progression between states was taken from a longitudinal cohort study. HRQoL weights were estimated for each Markov state using the Health Utilities Index (HUI)-2 (see Chapter 5) in a cross-sectional survey of caregivers of AD patients. The treatment effect was applied to the baseline transitions, with several alternative durations. After coming off the drug, patients return to baseline progression rate
A model to assess the cost-effectiveness of infliximab in rheumatoid arthritis (Kobelt et al. 2003)	Treatment effect in terms of change in Health Assessment Questionnaire (HAQ)- score, which is focused on functional disability—over a period of 54 weeks	A Markov model was developed with seven states—six relating to HAQ levels (functional disability) and one for death. HRQoL weights for each HAQ state were estimated using the EQ5D. The trial data were used to show short-term movements between the HAQ states on infliximab and for its comparator. Data from cohort studies were used to estimate longer-term transitions between states
A model to assess the cost-effectiveness of cardiovascular magnetic resonance in the diagnosis of coronary heart disease (Walker et al. 2013)	The sensitivity and specificity of a range of testing strategies was estimated in a non-randomized study with a reference standard of coronary angioplasty	A combination of a decision tree, to capture alternative diagnostic pathways, and a Markov model, to reflect treatments and their effect on outcomes, was used. The prevalence of disease requiring revascularization, together with the sensitivity and specificity of each test, were used to estimate the probability of a given patient being in one of four groups: true positive, false negative, true negative, and false positive. Based on group allocation, treatments received or delayed were modelled. Evidence from other studies linked treatments to outcomes in terms of rates of non-fatal cardiovascular events and mortality. The HRQoL effects, costs, and prognostic implications (in terms of mortality risk) of non-fatal events were included

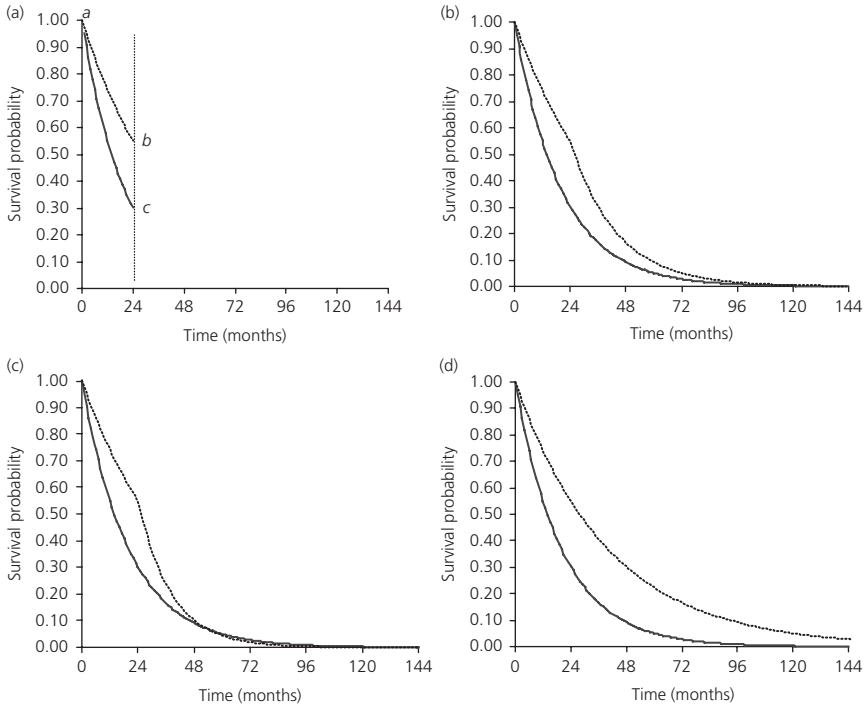


Fig. 9.1 Alternative extrapolation assumptions relating to survival data observed in a trial over a 24-month follow-up. Solid line, control group; dotted line, treatment group. (a) The curves as observed in the trial, with a 50% reduction in mortality rate with treatment compared with control. (b) The survival curves extrapolated over 144 months with the assumption of a 'one-time' benefit to patients. (c) Extrapolation with a rebound effect. (d) Assumption of a continuous treatment effect.

points of follow-up (measured in months). Figure 9.1a relates to the maximum follow-up during the trial of 24 months. It indicates that, in terms of mortality, treatment is more effective than control as it reduces the mortality rate by 50%. If the time horizon of the study is taken to be the same as the maximum follow-up in the trial, the measure of life-years gained from surgery is equivalent to the area between the two curves over 2 years (area abc in Figure 9.1a). This estimate of the *restricted* mean life-years over a 2 year time horizon assumes that patients who remain alive at the end of trial follow-up receive a maximum benefit of 2 years additional life expectancy—this effectively assumes that they die at the end of the trial! In reality, the patients living at the end of the trial will continue living afterwards, including the additional patients alive having received the treatment, so this 'within-trial' measure of life-years gained will inevitably be an underestimate.

The use of modelling to *extrapolate* beyond the follow-up period in the clinical study involves predicting what the survival curves will look like beyond what has been observed. A key question with extrapolation relates to appropriate assumptions about the shape of the survival curves after follow-up. For interventions that take place only

during the trial period, one assumption is that the more effective treatment during the trial confers a 'one-time' benefit to patients. As illustrated in Figure 9.1b, this means that, beyond the period of the trial, the rate of death per period of time, conditional on surviving until the end of the trial, is the same for patients originally allocated to treatment and to control. In other words, beyond the 24 month period, the 50% treatment effect (the reduction in the rate of mortality in the trial) is assumed to end and both groups become identical in terms of the mortality rate. The area between the two survival curves represents the gain in mean survival duration with treatment. This assumption of a one-time benefit has been used in several studies, particularly in the cardiovascular field. For example, in the base-case of their CEA of alternative thrombolytic therapies for acute myocardial infarction, Mark and colleagues used a model to extrapolate beyond the 1 year period of trial follow-up (Mark et al. 1995). To do this they used a separate source of data (a register of patients who had experienced acute myocardial infarction and survived the first year) to provide 15 year estimates of risk of mortality; beyond 15 years these risks were based on general population data.

In some contexts, it might be more appropriate to assume that the survival curves converge more rapidly after the trial follow-up period. That is, the conditional rate of death beyond trial follow-up becomes higher with treatment compared to control. This is illustrated in Figure 9.1c, where it is assumed that, beyond the trial follow-up period, the mortality rate increases by 40% in the treatment arm compared to control. This could happen, for example, if the more effective intervention delays the death of a high-risk subgroup of patients who, once treatment is ended (at the end of trial follow-up), die at a faster rate than those patients surviving in the other arm. This scenario is sometimes known as a 'rebound effect'. It can be seen that the area between the survival curves (the gain in mean survival duration with treatment) is less than when a 'one-time' benefit is assumed. The scenario was one considered in a CEA of endovascular aneurysm repair which extrapolated from a randomized trial (Epstein et al. 2008).

At the other extreme, it may be reasonable to assume that the treatment confers a continuous benefit beyond trial follow-up, and this is shown in Figure 9.1d. That is, the curves continue to diverge in the longer term, the 50% reduction in mortality of treatment continues, and patients randomized to that option continue to die at a slower rate. It can be seen that the area between the curves, assuming a continuous benefit, is larger than in Figures 9.1b and 9.1c. This may be a more appropriate assumption when treatment is still ongoing at the end of trial follow-up when, of course, costs are extrapolated as well as benefits. This was the base-case assumption of a CEA of ivabradine in heart failure, although alternative scenarios were used in sensitivity analysis (Griffiths et al. 2014).

The issue of the most reasonable assumption can be informed by the shape of the survival curves within the trial. For example, if they are ceasing to diverge in the latter period of follow-up, an assumption of continued divergence in the extrapolation is likely to be unwarranted. External non-trial data may also hold some clues. It is also important, however, to identify an appropriate assumption on the basis of what is known about the biology of the intervention—for example, in the case of a pharmaceutical, the length of time it remains active in the patient's body. Chapter 10 considers the use of evidence to model long-term extrapolation in more detail.

The choice of assumption made regarding extrapolation may have major implications on study results. For example, in an early cost-effectiveness study of therapy for patients with HIV, Schulman et al. (1991) estimated the incremental cost per life-year gained under two alternative assumptions about the effect of zidovudine on the development of AIDS and hence on mortality: a one-time effect and a continuous effect. The incremental cost per life-year gained of therapy ranged widely, from $6553 to $70 526 under those two scenarios. This shows that it is important to run alternative scenarios regarding plausible extrapolation assumptions. The judgements about plausibility are usually based on our knowledge about the epidemiology of the disease and the effects of other treatments that have been evaluated in the past.

It should be noted that extrapolating survival curves relates not just to mortality; rather, this can relate to any evidence on the time until a particular event and can include, for example, time until cancer progression and time until non-fatal cardiovascular event. Extrapolation can also relate to other assumptions regarding the effectiveness of treatments, for example what happens to outcomes once treatment is discontinued. An example is a CEA of biological therapies for psoriatic arthritis (Bojke et al. 2011). Evidence on efficacy was available for three biological therapies based on trial data with 3 month follow-up, and clinical treatment effects were mapped to HRQoL weights to estimate QALYs. It was assumed that HRQoL improvement continued while the patient was on therapy. When biological treatment was withdrawn, the effect on patients' HRQoL was modelled in terms of two alternative profiles. Firstly, that their HRQoL returned to what it was prior to treatment; or, alternatively, it rebounded to what it would have been had they never received treatment. The first of these was used in the base case, informed by a formal elicitation of opinion from clinical experts (see Chapter 10). Sensitivity analysis showed cost-effectiveness to be highly sensitive to this assumption.

9.2.5 The need to make results applicable to the decision-making context

Another situation where there may be a gap between the available evidence, particularly from randomized trials, and the requirements for a decision, relates to situations where the decision problem being addressed is inconsistent with the nature of the available clinical evidence. An example of the use of decision models to relate available evidence to a particular decision context is when some types of evidence are available from outside the jurisdiction where the decision is being taken and may not generalize to that location. In such a context, a model can be used to combine this outside evidence with other information and explicit assumptions to inform the relevant decision. An example of a study where the generalizability of clinical evidence to the decision-making jurisdiction, the United Kingdom, was subject to doubt was a CEA of glycoprotein IIb/IIIa antagonists for acute coronary syndrome where effectiveness was assessed in terms of the rate of fatal and non-fatal cardiovascular events (Palmer et al. 2005). As most of the trials of these therapies had randomized patients from outside the United Kingdom and, at the time, some aspects of cardiac care in the United Kingdom were considered to be at variance with those in the trials, the relevance of this evidence to the decision needed to be carefully assessed. The approach adopted was to distinguish the underlying (or baseline) rates of cardiovascular events under

usual practice (i.e. without the new drug therapy) from the *relative* effectiveness of the glycoprotein IIb/IIIa antagonists on that baseline risk. The former type of evidence was taken from a longitudinal observational study undertaken in the United Kingdom rather than from the randomized trials. Relative effectiveness, however, was based on a meta-analysis of the trials. Brought together within the model, these two types of evidence generated estimates of the absolute reduction in the risk of cardiovascular events generated by the new therapy. The explicit assumption with this approach was that the *relative* risk of events with glycoprotein IIb/IIIa antagonists (compared with usual care) from the trials generalizes to routine care in the United Kingdom; however, the absolute baseline risk does not generalize from the trials and needed to be estimated from UK sources. This example highlights a more general point with RCTs: that they are mainly designed to estimate relative effectiveness, and that the baseline rates of event may vary significantly between jurisdictions and indeed between patient subgroups. A key role for decision models, then, is often to bring together jurisdiction- or subgroup-specific estimates of baseline risk, often derived from non-randomized studies, with relative effectiveness estimates from a single RCT or meta-analysis.

Issues of generalizability in economic evaluation in general, and the use of models to facilitate the generalizability of evidence, have been considered (Sculpher et al. 2004). More recently, an International Society of Pharmacoeconomics and Outcomes Research (ISPOR) Task Force considered good practice in assessing and implementing transferability of economic evaluations across jurisdictions and made recommendations in areas including the interpretation of studies, economic evaluations based on individual patient studies (see Chapter 8), modelling studies, and further research activities (Drummond et al. 2009).

9.2.6 Using models to assess heterogeneity

Decision models can also be used to identify subgroups of a wider population of patients in whom an intervention is cost-effective, based on patients' clinical and socio-demographic characteristics. Understanding heterogeneity in cost-effectiveness between different types of patients is important in decision-making because, in principle, different decisions regarding the funding of interventions can be made for different subgroups (Sculpher 2008). Indeed, failing to reflect heterogeneity in economic evaluation (and hence in decisions) can impose costs on the health system in terms of opportunities for health gain forgone or resources wasted. This is because some types of patients will receive treatments which are not the most cost-effective given their characteristics (Coyle et al. 2003). The term *expected value of individualized care* has been developed (Basu and Meltzer 2007) to represent the added value of reflecting heterogeneity analytically and in decisions.

Chapter 8 considered how regression methods can be used to assess heterogeneity based on patient-specific estimates of total costs and benefits based on single studies such as trials. An advantage of the use of decision models in this context is that heterogeneity in different types of evidence can be more explicitly considered. Interest in subgroups for CEA extends beyond the clinical trialist's usual concerns regarding whether the relative effectiveness of an intervention is consistent across different types of patients (formally, whether any patient characteristics at baseline are treatment effect

modifiers). Economic evaluation includes consideration of possible heterogeneity in baseline risks, costs, and HRQoL. Heterogeneity in patients' preferences for different states of health has also been considered using modelling (Basu and Meltzer 2007; Owens and Nease 1997; Sculpher and Gafni 2001).

As discussed in Section 9.2.5, the methods used to estimate cost-effectiveness for specific subgroups often divides the absolute clinical benefit of a treatment upon which cost-effectiveness is based (e.g. the absolute reduction in the risk of an event such as the rate of cardiovascular events in the glycoprotein IIb/IIIa antagonists example in Section 9.2.5) into two elements: baseline risks and relative treatment effects. Baseline risks are the measure of events under (one of) the comparator intervention(s). The relative treatment effect is often a ratio (e.g. an odds ratio, relative risk, or hazard ratio) representing the effectiveness of the newer therapy *relative* to the comparator intervention, which is typically the main focus when the clinical results of a randomized trial are reported. Often the clinical report of a trial will indicate that there is no evidence of differences between subgroups in terms of *relative treatment effect*. However, cost-effectiveness is driven by *absolute* benefit, and there may still be important heterogeneity between subgroups in baseline event rates. Indeed, an assumption of constant relative effects being applied to subgroup-specific baseline event rates is common in cost-effectiveness models.

An example of a modelling study using this assumption compared lifetime costs and QALYs of two alternative hip prostheses used in primary hip replacement (Briggs et al. 2004). The effectiveness of the two prostheses, in terms of their failure rate over time, was taken from a large register developed in Sweden. Reflecting the register data, the model allowed the baseline failure rate (that for the 'usual care' prosthesis) to vary by the patient's age and sex. The relative reduction in failure rate with the newer prosthesis was, however, assumed constant over those subgroups. It is also important to note that the 'background' mortality rate (i.e. the population rate from all causes) is known to vary by age and sex. As death from reasons unrelated to hip replacement was also included in the model (as a 'competing risk'), this increases the differences between age and sex subgroups in terms of the cost-effectiveness of the newer prosthesis. Using the top-right and bottom-right quadrants of the cost-effectiveness plane (see Figure 3.2), Figure 9.2 shows how the cost-effectiveness of the newer prosthesis varied by age and sex. The top line relates to incremental costs and effects of the newer prosthesis relative to the older one for females, with various points on that line representing subgroups based on patients' age. The lower line shows the similar relationship for males. It can be seen that, for both males and females, the newer prosthesis is more cost-effective for the younger age groups: the newer device is dominant for men aged 70 or younger and for women aged 60 or younger. This is because the lifetime risk of failure with the older prosthesis is higher in younger patients because they impose greater wear and tear on their hips due to their greater activity, and because they would be expected to live longer. Therefore, a constant relative reduction in the failure rate with the newer prosthesis will confer a greater absolute benefit in these younger age groups. In addition to heterogeneity in the cost-effectiveness of the newer prosthesis by age, it is also less cost-effective in women (i.e. the line in Figure 9.2 is higher). Again, this is because men have a higher baseline event rate with the existing prosthesis as their activity levels place greater stress on their hips.

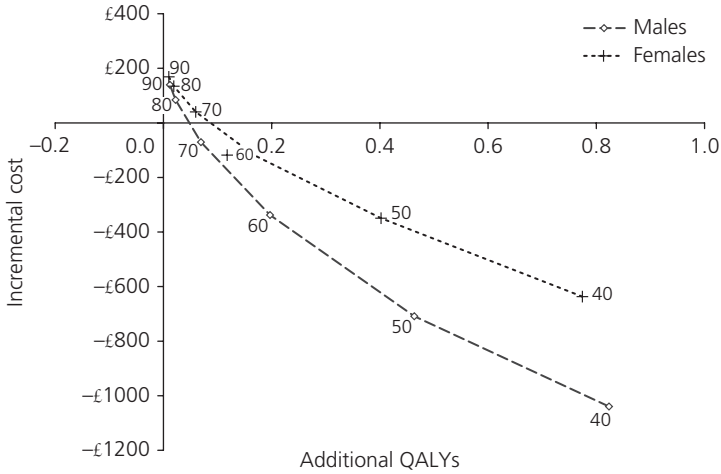


Fig. 9.2 Results of a cost-effectiveness study by Briggs et al. (2004) showing heterogeneity in the cost-effectiveness of a newer prosthesis by patient subgroup defined in terms of age and sex. QALY, quality-adjusted life-year.

In some clinical settings the cost-effectiveness of an intervention is very sensitive to a range of parameter values which are heterogeneous across a number of different patient-specific characteristics. For example, Henriksson and colleagues considered the cost-effectiveness of early intervention with angiography and possible revascularization, compared to pharmaceutical therapy alone, in patients with non-ST-elevation myocardial infarction (Henriksson et al. 2008). Based on evidence from an RCT, the study quantified the probability of non-fatal myocardial infarction or cardiovascular death as a function of the intervention received as well as a series of baseline patient characteristics including age, presence of diabetes, whether or not there had been a prior myocardial infarction, and smoking status. Mean costs during the initial hospitalization and 1 year afterwards were also estimated as a function of treatment received as well as covariables including age, sex, and severity of angina. Finally, patients' HRQoL at baseline and the change over follow-up was also estimated as a function of a similar set of covariables. Ultimately, the cost-effectiveness of early intervention was shown to be heterogeneous with respect to a patient's underlying risk of a further clinical event. Incremental cost-effectiveness ratios ranged from £12 750 per additional QALY in the most high-risk patients to £53 760 per additional QALY in the most low-risk patients. When heterogeneity in the relative effectiveness of early intervention—as well as the underlying risk of events—was allowed for, early intervention became dominated by pharmaceutical therapy in the most low-risk patients with an incremental cost-effectiveness ratio (ICER) of £10 476 in the most high-risk patients.

9.3 Key elements of decision-analytic modelling

There are some key elements to decision analysis that are common to all models. These are the use of probabilities to reflect the likelihood of events or changes in health, and the expected values to inform decisions. Different model types such as decision trees and Markov models could also be added to this list, but these are discussed more fully in Section 9.4.

9.3.1 Probabilities

Probabilities are used widely in quantitative methods in many fields, and have an important role in clinical decision-making (Weinstein and Fineberg 1980). A common way of thinking about probability is as the measured frequency of an event in a given sample or population. For example, if a sample of 200 patients is treated with a particular medicine over 1 year and 10 patients have an adverse event, the proportion of 0.05 can be taken as an estimate of the 1 year probability of a patient experiencing an adverse event with that therapy. Assuming this estimate generalizes to other patients, a future patient will either experience the adverse event or not; when the decision is being taken regarding whether to administer the therapy, the uncertain outcome for the individual can be expressed in terms of a probability estimated from the experiences of similar patients.

This concept of probability as a number indicating our 'state of knowledge' regarding whether an event will or will not take place is a feature of Bayesian statistics, and is not shared with the classical or 'frequentist' statistical methods that are widely used in the analysis of randomized trials (Spiegelhalter et al. 2004). This emphasizes the common origins of decision analysis and Bayesian statistics. This concept of probability can be generalized to represent strength of belief which, for a given individual, is based on their previous knowledge and experience. This view of probability is important in decision analysis as, in many analyses, especially when the probability of particular events may not have been informed by formal studies such as trials, estimates from relevant experts may need to be elicited. Given that decisions about the use of finite resources have to be taken regardless of the strength of the evidence available, on the basis of assumptions and judgements, decision analysis provides an analytical framework within which this can be done explicitly. Chapter 10 considers the use of elicitation of expert beliefs as an input into decision models in more detail. Bayesian statistical methods are also valuable in thinking about uncertainty in the context of decision-making (see Chapter 11).

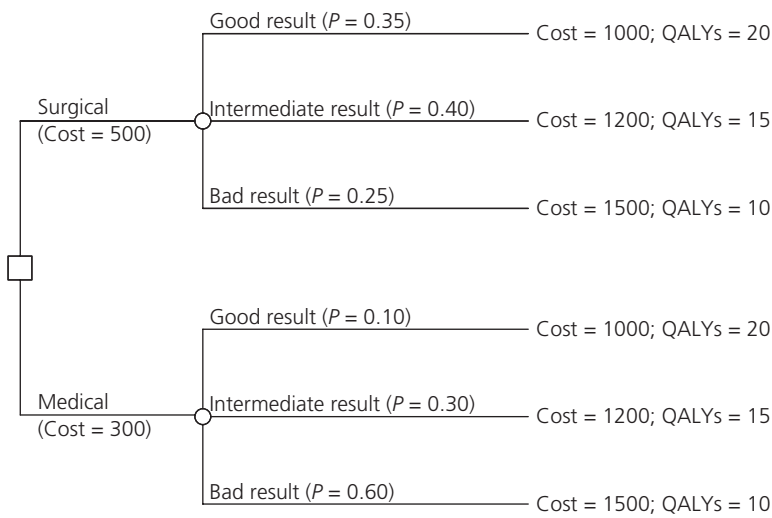
Some specific probability principles are also important in decision analysis, and these are summarized in Box 9.3.

9.3.2 Expected values

A key concept in decision analysis is the expected value of the costs or outcomes or a measure of cost-effectiveness of an option. This is illustrated in Figure 9.3, which compares two alternative interventions, medical and surgical. For each intervention, a given patient can follow one of three possible pathways which result, respectively, in a bad, intermediate, or good outcome. Before treatment, it is unknown which pathway a specific patient will follow, but probabilities are used to express the likelihood of each

Box 9.3 Probability concepts

Joint probability	The probability of two events occurring concomitantly. In terms of notation, the joint probability of events A and B: P[A and B]. When the events are independent, P[A and B] = P[A] × P[B]
Conditional probability	The probability of an event A given that an event B is known to have occurred. The notation is P[A|B]. Joint and conditional probabilities are related in the following equation: P[A and B] = P[A|B] × P[B]
Independence	Events A and B are independent if P[A] is the same as P[A|B]



Expected cost of surgery: 500 + (0.35 × 1000) + (0.40 × 1200) + (0.25 × 1500) = 1705
Expected QALYs of surgery: (0.35 × 20) + (0.40 × 15) + (0.25 × 10) = 15.5
Expected cost of medicine: 300 + (0.10 × 1000) + (0.30 × 1200) + (0.60×1500) = 1660
Expected QALYs of medicine: (0.10 × 20) + (0.30 × 15) + (0.60 × 10) = 12.5

Incremental cost per QALY gained of surgery: (1705 – 1660)/(15.5 – 12.5) = 15

Fig. 9.3 Simple decision tree showing example of the calculation of expected values. QALY, quality-adjusted life-year.

occurring. These are likely to differ by therapy. For the alternative therapies, each pathway has a cost and an outcome expressed in terms of QALYs; there is also a cost of the intervention itself which is incurred whatever pathway the patient follows. For each of the therapies, an expected cost and expected outcome can be calculated. The expected cost is the cost of the intervention plus the therapy-specific sum of the costs of the three pathways, weighted by the probability of a patient following each pathway with that

treatment. The same idea is applied to calculate expected outcome. On that basis, it is clear that surgery has both a higher expected cost and higher expected QALYs. Using the methods of incremental analysis introduced in Chapter 3, the incremental cost per additional QALY generated by surgery can be calculated.

The concept of expected value is clearly analogous to the mean value of an end point when sample data are available. In a trial-based CEA, for example, the mean costs and QALYs across patients in each of the randomized groups are used as the basis of the incremental analysis (see Chapter 8). As for mean values in studies based on a single trial, the expected value from a decision model represents the best estimate of the end-points of interest for decision-making. As decision analysis shares common theoretical origins with expected utility theory described in Chapter 5, the expected values calculated in decision models were originally seen as being strictly von Neumann–Morgenstern utilities. Given that expected utility theory is a normative framework for decision-making under conditions of uncertainty, the expected utilities from decision models would provide a clear indication of the preferred option from those being compared. However, decision analysis is widely used for situations when outcomes other than von Neumann–Morgenstern utilities are used. The expected value can still provide the key input to guide decision-making as long as the outcomes have been chosen appropriately.

9.4 Stages in the development of a decision-analytic model

The development of a decision-analytic model for economic evaluation involves a number of stages. This section considers each of the stages to provide a fuller understanding of the role of decision modelling in this field.

9.4.1 Defining the decision problem

One of the key stages in the development of the model is the specification of the question being addressed, sometimes called the decision problem. This process closely mirrors the specification of the study question for economic evaluation in general as discussed in Chapters 2 and 3. In particular, there is a need to define the recipient group (patients or others) and the relevant options being compared. It is important to emphasize that, in defining these options, this may include more than specific interventions. It may include, for instance, starting and stopping rules for treatments—for example, when to start and stop medical therapy for a particular chronic condition. As discussed in Section 9.2.1, for some evaluations the options will represent *clinical treatment strategies* or *pathways*, such as the sequence of therapies that might be used for the treatment of a condition characterized by treatment failure with some therapies. An example of such a study is a decision model looking at the cost-effectiveness of alternative therapies for epilepsy where assumptions were made about which therapies patients were placed on if they failed on initial treatment (Wilby et al. 2005).

9.4.2 Defining the boundaries of the model

All models are simplifications of reality so, in developing an analysis, decisions have to be taken about what to include. In part, this relates to general issues in economic

evaluation such as the choice of perspective, the appropriate measure of effect/benefit, and the time horizon (see Chapter 2). However, it is important to consider how far a model should go to cover all the possible implications of an intervention or programme. A simple example would be whether or not to include rare side effects in a model. A more complex example is that, in considering the cost-effectiveness of antibiotic treatment for a given condition, the issue of the cost and health effects of antibiotic resistance could be an important consideration for decision-making, but relatively few cost-effectiveness models consider this aspect—that is, they have drawn the boundaries of the model to exclude it from consideration. For decision-making, it would be important to consider whether this exclusion limits the value of the model. The CEA of alternative preventive strategies for group B streptococcal infection referred to in Section 9.2.1 assessed the extent to which resistance would have to impact negatively on population health to change the cost-effectiveness ordering of the options evaluated (Colbourn et al. 2007).

Another example is a decision model that was developed to assess the cost-effectiveness of routine antenatal HIV testing (Ades et al. 1999). These authors assessed the impact of different screening strategies on the extent to which a woman's HIV status was known during pregnancy. Through the use of interventions, this knowledge could have three beneficial health effects on: (1) the woman through earlier use of antiretroviral therapy; (2) the child through the use of interventions to reduce the mother-to-child (vertical) transmission rate; and (3) the child through the earlier use of antiretroviral therapy and prophylaxis if the child is born with HIV. The broader health benefit relating to reductions in infections to others through changes in sexual behaviour (horizontal transmission) was not, however, considered in the model, thus defining the model boundary.

Decisions about the boundaries in decision models will partly be based on the availability of data and complexity of the modelling task, but they should mainly be driven by the extent to which extending the boundaries (adding complexity) is considered likely to impact on the cost-effectiveness of the options being compared and hence the most appropriate decision.

9.4.3 Conceptualizing a decision model

A key stage in the development of a decision model is the process of deciding on a structure. Formally, this involves a series of decisions concerning how the input parameters in the model are to be related and, in particular, choices about how to characterize the clinical events and health states of interest (e.g. episodes of a disease, disease progression, case identification). Each economic evaluation brings with it different structural issues, but a few common ones are given below.

- Do the events of interest occur just once (e.g. death) or could they happen several times over the relevant time horizon (e.g. a non-fatal myocardial infarction)?
- Are patients at risk of several events over time (i.e. *competing* risks); for example, the risk of a heart attack but also of stroke?
- As discussed in Section 9.2.4, when extrapolating events over time, what is the durability of the effectiveness of an intervention relative to comparators?

- Do the probabilities of events change as time elapses or are they constant with respect to time?
- Are all important events included and has double counting of events been avoided?
- Does a patient's prognosis partly depend on the events they have already experienced in the model?
- For the management of a chronic disease, does the structure of the model allow for the costs and effects of subsequent therapies to be included?
- Is the clinical prognosis of a given patient partly dependent on the clinical status of other patients as might be the case, for example, with infectious diseases?

The way in which these issues are handled can essentially be defined as a series of mathematical relationships between parameters. Indeed, some studies present the structure of their decision models in terms of a series of equations (e.g. see Spiegelhalter and Best 2003). Most decision models used in economic evaluation, however, present the structure of their model schematically. The collaboration between the ISPOR and the Society of Medical Decision Making (SMDM) developed a number of useful guidance documents, one of which focused on model conceptualization (Roberts et al. 2012).

9.4.4 Implementing a model—the decision tree

The judgements that are taken by an analyst to characterize the nature of a disease and the impacts of alternative interventions are implemented in a specific decision model. Various types of model are used for economic evaluation, but one of the most widely used is the *decision tree*.

The decision tree represents individuals' possible prognoses, following some sort of intervention, by a series of pathways. Figure 9.3 used a simple decision tree to explain the concept of expected value. A clinical example is used to illustrate the use of decision models in more detail—the comparison of two antiemetic prophylactic therapies for patients undergoing chemotherapy for cancer. The example is based on a published study comparing ondansetron and metoclopramide which was undertaken before a price had been determined for ondansetron (Buxton and O'Brien 1992). The study considered the acquisition cost of the therapies, as well as the cost of the adverse events and of their treatment, and the cost of treatment failure (i.e. an episode of emesis). Effects were expressed in terms of the probability of a patient being successfully treated, which was defined as the absence of emesis and adverse events. The decision tree is shown in Figure 9.4, and this can be used to describe a series of general features with this sort of model structure.

9.4.4.1 Nodes

Decision nodes The square box at the start of the tree is a decision node and represents the decision being addressed in the model: here, which of ondansetron and metoclopramide is the more cost-effective in preventing episodes of emesis without adverse events.

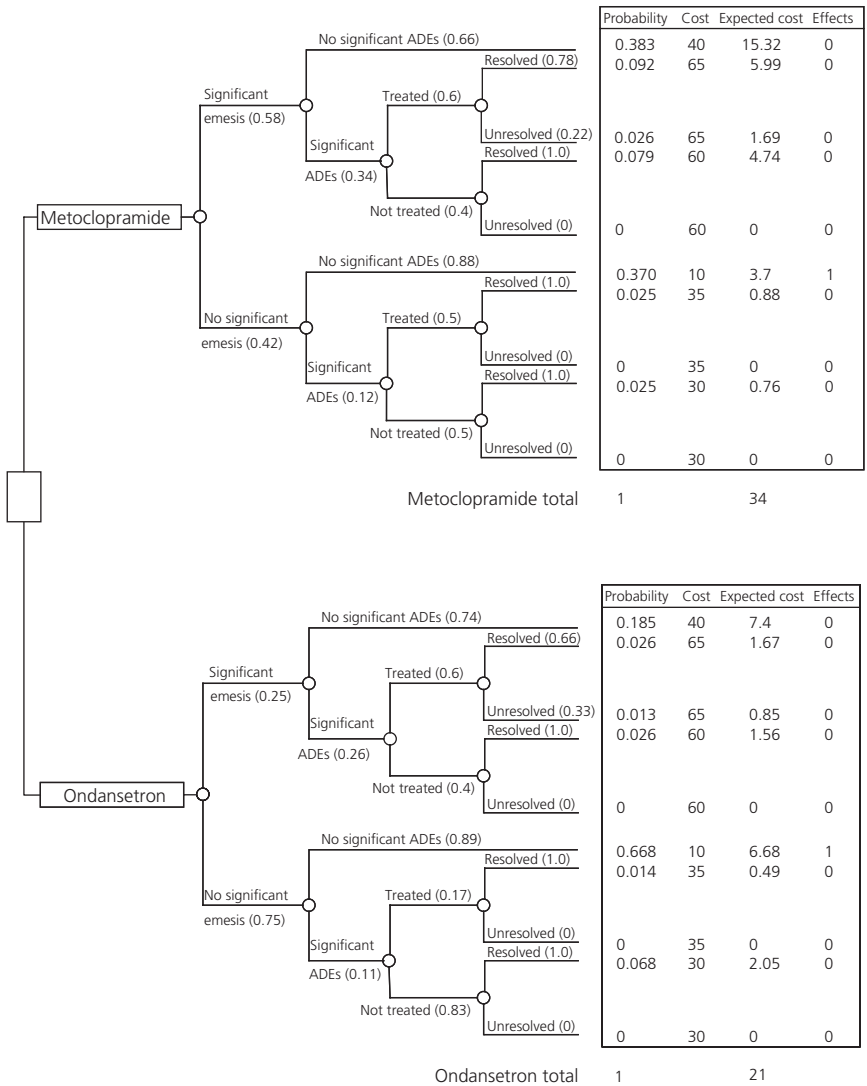


Fig. 9.4 Example of a decision tree taken from Buxton and O'Brien (1992). Other inputs into the model: price of both treatments, £10; cost of an episode of emesis, £30; cost of side effects, £20; cost of treating side effects, £5. ADE, adverse drug events.

Chance nodes Coming out of the decision node is the range of possible pathways that characterize the effects of the alternative therapies. The pathways are built up through a series of branches representing particular events. Here the events are significant emesis, significant adverse events, the treatment of adverse events and the resolution of adverse

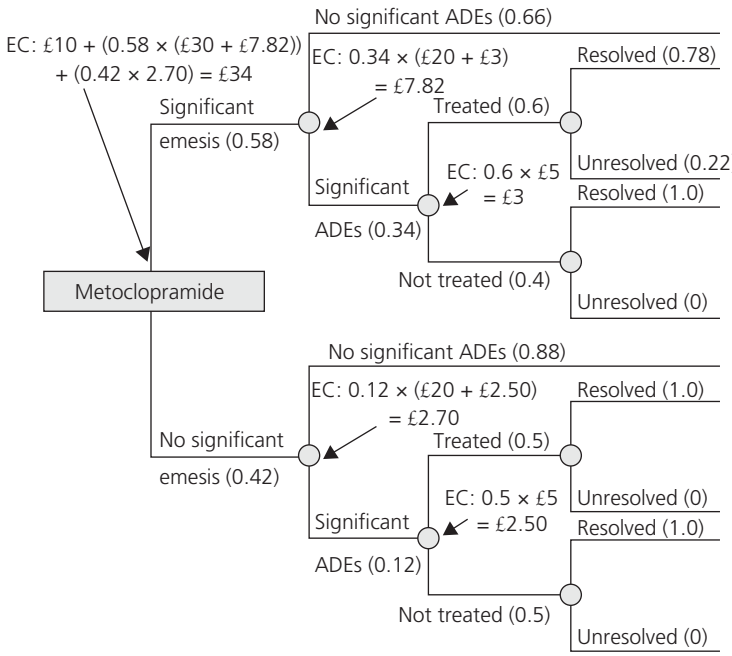


Fig. 9.5 The process of 'rolling back' a decision tree to calculate expected values. The example shows the calculation of expected cost for metoclopramide in the decision tree shown in Figure 9.4. ADE, adverse drug events; EC, expected costs.

Source: data from Buxton, M.J. and O'Brien, B.J., Economic evaluation of ondansetron: preliminary analysis using clinical trial data prior to price setting, *British Journal of Cancer*, Volume 66, Supplement XIX, pp. S64–S67, Copyright © 1992.

events. Given that, *ex ante*, it is not known whether a particular patient will experience a given event and follow a particular branch, the circular nodes (chance nodes) define points of uncertainty for an individual in the tree.

9.4.4.2 Probabilities

Branch probabilities The branches issuing from a chance node represent the possible events patients may experience at that point in the tree. The likelihood of the event is represented in terms of branch probabilities. For both treatments, the first chance node relates to whether or not a patient experiences an episode of emesis, and the probability of emesis and its complement (i.e. 1 – the probability of that event) are shown on the respective branches.

Conditional probabilities Moving from left to right, chance nodes show subsequent uncertain events. The probabilities of these events are *conditional probabilities* (see Box 9.3) because they can differ according to whether or not patients have experienced particular previous events. For example, for those patients who have experienced an episode of emesis on metoclopramide, the conditional probability of significant adverse events is 0.34; and the probability of treatment conditional on having experienced such an event is 0.6. It can be seen that, although the possible events are the same for the

two therapies, the probabilities in the two parts of the tree are not the same. Specifically, the efficacy of ondansetron was considered higher than that of metoclopramide, so there is a lower probability of emesis. Ondansetron was also considered less toxic, as the probabilities of adverse events (whether or not the drug had been efficacious) are lower than those for metoclopramide. There are also differences between the two therapies in the probabilities of an adverse event being treated and of it resolving.

9.4.4.3 Pathways

The combination of the different branches in the tree determines a series of pathways along which patients can pass in the tree. For both treatments, there are ten possible pathways. The top pathway for each treatment incorporates significant emesis and no adverse events; the second pathway contains significant emesis and a significant adverse event which is treated and resolves, and so on. The final pathway for each treatment relates to no significant emesis and an adverse event that is not treated and that does not resolve. These pathways are mutually exclusive (a given patient can only follow one of the pathways) and exhaustive (a given patient must follow one of the pathways).

Pathway probabilities To the right of the decision tree in Figure 9.4 is a series of columns of numbers. The first contains the probability of a given patient passing along each of the pathways. These probabilities are calculated by multiplying the initial branch probability by subsequent conditional probabilities. So the probability of the first pathway, with ondansetron, is the product of the probability of significant emesis (0.25) and the probability of no significant adverse events conditional on significant emesis (0.74), which equals 0.185. As the pathways are mutually exclusive, the probabilities for a given treatment must sum to 1.

Pathway costs Each pathway in the tree also has costs associated with it. These represent the sum of the costs of each of the events patients experience in that pathway. For the first pathway, for example, the relevant costs are the cost of the drug itself (£10) and the cost of significant emesis (£30), totalling £40. The second pathway cost is the sum of the drug cost (£10), significant emesis (£30), and significant adverse events (£20) which are treated (£5), equalling £65. The same principle is applied to the other pathways in the tree. It can be seen that the pathway costs are the same for metoclopramide and ondansetron as it is assumed that the two products have the same acquisition price and event costs.

9.4.4.4 Expected values

The expected cost for the two therapies can be calculated by weighting each pathway cost by its respective probability, and then summing across all the pathways. This can be seen in the expected cost column of Figure 9.4, and adding down this column generates expected costs for metoclopramide and ondansetron of £34 and £21, respectively.

This decision model used the probability of successful treatment as the relevant measure of effect in the CEA (i.e. the probability of no emesis and no adverse events). In terms of expected values, this is equivalent to giving the pathway of no emesis and no adverse events the value 1 and all other pathways the value 0. Assuming equal prices for the products, this version of the model indicates that ondansetron is dominant, as it has

a lower expected cost than metoclopramide and a higher expected effect. The original paper considered a number of sensitivity analyses involving alternative assumptions about the prices for the products (Buxton and O'Brien 1992). Another way of working out the expected costs and effectiveness for a given option in a decision tree is by 'rolling back' the tree. It will give exactly the same answer as the approach outlined above, but involves working from the right-hand side of the tree towards the left, calculating expected values at each chance node. Figure 9.5 shows how the decision tree for the alternative antiemetics is rolled back to estimate expected costs for metoclopramide. It could be argued that the use of the probability of successful treatment as the effectiveness measure is a weakness of this analysis as it assumes that all other pathways have the same (zero) value. Weighting each pathway using HRQoL and calculating expected QALYs for each treatment would probably have been a more informative analysis.

9.4.4.5 Limitations of the decision tree

The decision tree is widely used in economic evaluation, but has important limitations. The first is that events are implicitly taken as occurring over an instantaneous discrete period. In the antiemetic case study discussed above, for example, costs and effects over an undefined treatment period were considered. In other words, time is not explicitly defined in a decision tree unless the analyst does so in characterizing the different branches. Therefore, those elements of an economic evaluation that are time dependent can be difficult to implement. This is true of discounting, where the time at which costs and outcomes are accrued is very important. It also applies to the process of adjusting survival duration for HRQoL in calculating QALYs where it is necessary to know when a change in health status occurs (e.g. to reflect the impact of age on HRQoL).

The second, and related, limitation of decision trees is that they can become very complex when they are used to model complicated long-term prognoses, particularly related to chronic diseases. For example, to model the future prognosis of a woman with early-stage breast cancer, a decision tree would have to characterize a whole series of competing risks that a patient would have to face including adverse treatment effects, cancer recurrence (of various types), remission from cancer, and death (from various causes). Once an event is experienced in one time period (e.g. cancer recurrence), a series of new risks may present themselves for future time periods. In principle, these recurring event risks could be structured using a decision tree where a set of chance nodes and branches could be used to characterize events in a particular time period, and the same or similar ones could be used for subsequent time periods. However, for a long-term chronic disease, where a patient is at risk of events for many years, the tree could become very 'bushy', with many mutually exclusive pathways. A model of this type would probably be very time consuming to programme and analyse.

9.4.5 Implementing a model—the Markov model

The limitations of the decision tree are the main reason why another model structure—the Markov model—is also widely used in economic evaluation to handle particular decision problems (Briggs and Sculpher 1998; Sonnenberg and Beck 1993). Whereas decision trees characterize possible prognoses in terms of alternative branches, Markov models are based on a series of 'states' that a patient can occupy at a given point

in time. Time elapses explicitly with a Markov model, with the probability of a patient occupying a given state assessed over a series of discrete time periods, called cycles. The length of these cycles will depend on the disease and interventions being evaluated, but might be a month or a year. In judging the appropriate cycle length, a key consideration is to limit the probability that a given patient could experience more than one event the period of the cycle. Each state in the model generally has a cost associated with it and, when QALYs are used as the ultimate outcome measure, a HRQoL weight. The time duration during which the average patient occupies the various states in the model will, when weighted by the relevant cost or HRQoL weight, be used to calculate expected costs and outcomes. The speed with which patients move between the states in the model is determined by a set of transition probabilities.

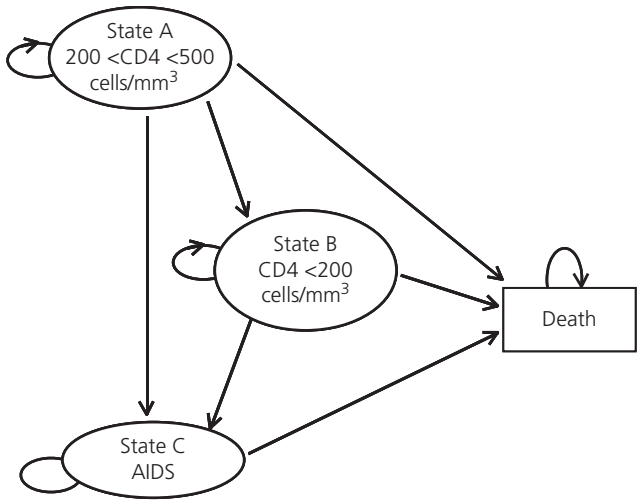
These concepts are described in more detail using an example from the published literature, which evaluated the cost-effectiveness of two antiretroviral therapies (zidovudine monotherapy versus zidovudine plus lamivudine combination therapy) for patients with HIV infection (Chancellor et al. 1997). Although more recent models for HIV are more complex, this study provides a good example to understand the key features of Markov models.

9.4.5.1 Markov states

Figure 9.6 shows a schematic of the Markov model used in the HIV example. The model is structured in terms of four Markov states. Two of these are related to a patient's CD4 count, which indicates the strength of their immune system. State A represents the healthiest patients with relatively high CD4 counts, and State B includes patients with lower CD4 counts. State C includes patients who have progressed to AIDS, and the patient moves to State D when they die. The arrows in the model show how patients can progress through the model over the cycles, which were taken to be 1 year. If a patient starts in State A in the first cycle, various transitions are possible in the second cycle: the patient can (1) remain in State A; (2) move to State B as their CD4 count drops; (3) move to State C if they suffer an AIDS-defining illness; or (4) move to State D if they die. Once a patient has moved to State B, in the next cycle they can remain in this state, or progress to State C or to State D. In this particular model, it is not possible for a patient's health to improve, so they cannot, for example, move from State B to State A. Once in State C, in the next cycle they can remain in that state or die, but not move back to States A or B. State D (death) is an absorbing state from which, sadly, there is no escape!

9.4.5.2 Transition probabilities

Figure 9.6 shows the transition probabilities that define the speed at which patients move between the Markov states under monotherapy, and the cycle length is 1 year. The matrix shows the state in which the patient starts the cycle, and the probabilities associated with the various transitions during one cycle conditional on a starting state (so these are conditional probabilities). For example, if a patient is in State A on monotherapy, there is a probability of 0.721 that they will remain in that state in the next cycle, of 0.202 that they will progress to State B, of 0.067 that they will progress to State C, and of 0.01 that they will die. The zeros in the matrix represent situations where backwards transitions are not considered feasible in this particular model. It can be seen that, because a patient always has to be in one of the states, the sum of the



Transition probabilities—monotherapy

	Transition to			
Transition from	State A	State B	State C	State D
State A	0.721	0.202	0.067	0.01
State B	0	0.581	0.407	0.012
State C	0	0	0.75	0.25
State D	0	0	0	1

Fig. 9.6 Markov diagram for a cost-effectiveness model in HIV taken from Chancellor et al. (1997). Below the diagram are the transition probabilities used for the monotherapy treatment.

probabilities across the lines must always equal 1. These 'baseline' probabilities, relating to what was then current practice, were taken from a longitudinal cohort study, which is a common type of evidence source for this type of parameter. Appropriate analytical methods are required to translate longitudinal data into transition probabilities over discrete cycles—see Chapter 10, also Briggs et al. (2006). A second set of transition probabilities was calculated for combination therapy in an attempt to reflect its effectiveness, compared to monotherapy, based on a meta-analysis of RCTs.

In this model the transition probabilities are the same for every cycle in the model. This implies, for example, that a patient with AIDS is at the same risk of death over the next year regardless of factors such as their age or the duration of time they have had AIDS. Markov models with fixed transition probabilities with respect to time are known as Markov chains. Some transition probabilities can also be allowed to vary over time depending on the structure of the model (Briggs et al. 2006).

9.4.5.3 Costs and outcomes

In the Markov model, costs are typically implemented each cycle according to the state a patient occupies (although they can also be applied to the proportion making a transition, e.g. to a 'dead' state). For the two therapies being evaluated in the HIV example, the cost of being in a given state is the same, and the only difference is in the acquisition price of the therapies. Hence, as for the decision tree example above, the only elements of the Markov model that differ between the two therapies are the acquisition cost of the therapies themselves and the probabilities that determine how a patient moves through the model. On the outcomes side, the HIV model used expected survival duration (life-years) as the measure of effectiveness, and this was evaluated over a lifetime time horizon (i.e. until the probability of being alive is very small).

9.4.5.4 Expected values

The process of calculating expected costs and effectiveness with a Markov model is very similar to that in a decision tree. This because both models are examples of *cohort models* which are set up to calculate the outcomes of interest in the *average patient*, and there is no consideration to how *individual patients* vary between each other. Instead of summing pathway costs and effects and weighting by their probabilities as with the decision tree (or rolling back the tree), the costs and values of each Markov state are weighted by the time a patient spends in that state. This is made up of two stages. In the first, the probability of a patient being in a given state for each cycle is calculated (this can also be understood as the proportion of the patient cohort in that state at a point in time). This is usually done in a spreadsheet or similar software using a method known as the cohort simulation method, which produces a 'Markov trace' showing the proportion of the cohort in each state over time.

This is illustrated in Figure 9.7 with respect to the monotherapy intervention in the HIV example, using a time horizon of 20 annual cycles. It is assumed that 1000 patients begin in the cohort, but the number is irrelevant in a cohort model as the focus is the average patient and only the proportions of the cohort in particular states at a given time point matter. One patient or one million patients could be used as the starting cohort, and the answer will be the same. For each cycle, the proportion of the cohort in each state is calculated on the basis of the proportions in the various states in the last cycle and the transition probabilities. Figure 9.7 shows the calculations for the first cycle. In a spreadsheet or other software, once the equations have been determined for the first cycle, it is normally a case of simply copying down the formulas for subsequent cycles. As more and more cycles are added, the proportion of the cohort in the absorbing state (here death) increases, and all but a very small proportion should have died once a cycle number that is consistent with the relevant life expectancy has been reached.

Once the proportion of patients (or, in other words, the probability of a given patient being) in each state for each cycle has been calculated, the second stage involves working out expected costs and effects. On the cost side, this involves calculating an expected cost per cycle by adding the cost of each state weighted by the proportion of the cohort in each state. The overall expected cost simply involves summing the expected cost of all cycles. Implementing discounting is straightforward, with the standard formula (see Chapter 4) used to adjust the expected cost of every individual cycle. Expected outcomes are calculated on a similar basis. In the case of survival duration, this simply involves

Cycle	State A	State B	State C	State D	Total
0	1000	0	0	0	1000
	1000 × 0.721	1000 × 0.202	1000 × 0.067	1000 × 0.01	
	↓	↓	↓	↓	
1	721	202	67	10	1000
2	520	263	181	36	1000
3	375	258	277	90	1000
4	270	226	338	166	1000
5	195	186	363	256	1000
6	140	147	361	351	1000
7	101	114	340	445	1000
8	73	87	308	532	1000
9	53	65	271	611	1000
10	38	48	234	680	1000
11	27	36	197	739	1000
12	20	26	164	789	1000
13	14	19	135	831	1000
14	10	14	110	865	1000
15	7	10	89	893	1000
16	5	7	72	916	1000
17	4	5	57	934	1000
18	3	4	45	948	1000
19	2	3	36	959	1000
20	1	2	28	968	1000

Fig. 9.7 The results of the Markov trace for the monotherapy group in the HIV example shown in Figure 9.6. The trace assumes a starting cohort of 1000 beginning in State A.

weighting the proportion of patients in each state per cycle by 1 if they are alive, and by 0 if they are dead. Adding up across the cycles (with discounting as necessary) will provide the expected number of life-years experienced by the cohort. In the case of QALYs, this is slightly different because the proportion of the cohort in each state is weighted by the HRQoL value associated with that state, and then summed across the cycles.

A cohort simulation is undertaken for each option being evaluated. In the case of the HIV model, cohort simulations were undertaken separately for monotherapy and combination therapy. On this basis, combination therapy was found to be more costly and more effective, with an incremental cost per life-year gained of £6276.

9.4.6 Implementing a model—other model types

This chapter has considered two popular model types used in economic evaluation. The Markov model is used in situations when the decision tree would become too unwieldy, typically when events can recur over a long time horizon. The HIV example in Section 9.4.5 used a Markov model in a conventional manner, where a cohort simulation is undertaken for each of the respective options being compared. The two interventions differed only in terms of the initial (acquisition) costs of the interventions and the

transition probabilities. In some situations, a decision analysis may involve the combination of both a decision tree and a Markov model. This was the case in the evaluation of glycoprotein IIb/IIIa antagonists example discussed in Section 9.2.5, for example (Palmer et al. 2005). A short-term decision tree was used to establish the proportion of patients with each therapy experiencing non-fatal myocardial infarction or death over an initial 6 month period reflecting trial results; a Markov model was used to calculate long-term expected costs and quality-adjusted survival duration conditional on which events had been experienced in the short-term model.

9.4.6.1 Time dependency and the memoryless property

A simple Markov model may be unsuitable to capture the key aspects of some prognoses. The HIV example described above, for example, assumed that transition probabilities did not vary over time. However, in some situations, this assumption may be difficult to sustain because of evidence suggesting that the probability increases or decreases with time. Some forms of 'time dependency' in transition probabilities can be handled quite easily in Markov models. This is the case, for instance, when the transition probability changes as the age of the patient increases. This can be implemented simply by having a different transition probability for each cycle in the cohort simulation. Equivalently, it is not difficult to implement transition probabilities that change as a function of the time a patient has been in a state, as long as all patients start in that state, and none returns to it once they have left it. The use of time-dependent transition probabilities in Markov models is dealt with in more detail in Briggs et al. (2006).

In other situations, it is less straightforward to implement time dependency in transition probabilities because of the key assumption underlying Markov models. This 'Markov assumption' is often described as the 'memoryless' feature of these models. It holds that the probability of a given transition in the model is independent of the nature or timing of earlier transitions. This can be illustrated using the HIV model shown in Figure 9.6. In this model, patients can enter the AIDS state (State C) from either State A or State B. However, once a patient has entered the AIDS state, the model cannot 'remember' where the patient came from; that is, it cannot distinguish the origin of the patients in the state at a given time point and treats them as homogenous. So, if the prognosis (cost and/or health impact) of being in the AIDS state was considered likely to differ according to the CD4 count a patient had at the point of entering that state, this particular Markov model would not be able to reflect this.

This assumption might be difficult to justify if evidence suggests, for example, that mortality risk is higher in patients who have experienced AIDS-defining events having previously had lower CD4 counts. If the memoryless feature represents an oversimplification of the epidemiological evidence, then one approach is to add additional states to the Markov model. In the example above, for example, two AIDS states could be used: one containing patients who moved there from State A and the other including patients who arrived from State B. These two AIDS states could then differ with respect to the risk of mortality and, if necessary, in terms of the cost per cycle of occupying that state.

Giving a Markov model additional 'memory' by adding states can, however, become unwieldy if numerous additional states have to be added. In this situation, one option is to move to a different modelling approach. The decision tree and Markov model are examples of cohort models. As shown above, this involves calculating the proportions of

a homogeneous cohort that would move along particular pathways or occupy specific Markov states. Using these proportions to weight the costs and outcomes associated with pathways or Markov states, this provides the route to calculating overall expected costs and outcomes for each of the options being compared.

9.4.6.2 Individual sampling models

The difficulty in incorporating 'memory' and time dependency are two of the limitations of cohort models. An alternative approach to decision modelling is to move away from the cohort model towards modelling individual patients moving through models. These *individual sampling models* (ISM) calculate the costs and effects of a large number of simulated patients and average across these patients to estimate expected values for the alternatives under evaluation. When ISM is used in the context of state transition models with discrete cycles, this is referred to as micro-simulation or first-order Monte-Carlo simulation (Siebert et al. 2012). This type of modelling literally tracks the process of individual simulated patients through particular states, and allows them to accumulate costs and benefit over time. They have the potential to offer greater flexibility than cohort models as the future prognosis of a given patient can vary according to their 'history'. In the case of the HIV example, the use of micro-simulation would mean that a patient who experiences an AIDS event with a CD4 count of 400 could have a different risk of future events than a patient who experienced such an event with a CD4 count of 200. Given the focus of economic evaluation on expected values, such a model has to simulate the costs and outcomes of a large number of patients and estimate the average over those simulations.

An alternative form of an ISM is *discrete event simulation* (DES) (Karnon et al. 2012). Whereas micro-simulation retains the concept of the state and discrete cycles as with a Markov model, DES simulates the time until the next event for a given simulated patient. The way time is handled in DES means that these models can advance to the next time a given simulated patient has an event, thus avoiding modelling time and effort in unnecessary interim computations. More detail about the use of ISM models in economic evaluation can be found elsewhere (Barton et al. 2004; Caro et al. 2010; Davies 1985; Standfield et al. 2014).

ISMs have some important limitations, however. The opportunity to incorporate patient history with such models may allow greater structural flexibility, but it will typically require additional evidence to populate such models. This is because parameters representing possible future prognoses for a given patient need to be conditional on history, thus increasing the number of parameters to be estimated. A second limitation is that the simulation requirements of these models can be time consuming, even with modern computers. This is particularly the case when probabilistic sensitivity analysis (PSA) is undertaken to quantify parameter uncertainty (see Chapter 11).

9.4.6.3 Dynamic transmission models

All of the types of models considered so far in this chapter assume that the individuals being modelled are independent from each other with respect to their health. That is, the health of one individual does not impact on the health of one or more others. This independence assumption may be untenable in the context of infectious disease where the incidence of new infections depends on the existing number of individuals who are infected. During an epidemic this number changes dynamically. Therefore, decision-analytic

models relating to infectious diseases may need to consider explicitly this dynamic feature of such diseases. Allowing for the interaction between individuals in the context of infectious disease may be necessary for interventions including vaccination, screening, and treatments where the transmissibility of the individual is affected. Many economic evaluations of these types of intervention for infectious disease often retain the standard 'static' decision model which does not allow for the non-linear interactions between individuals. For example, it has been found that most models used for the economic evaluation of screening for *Chlamydia trachomatis* used static methods and, therefore, were unlikely to have correctly estimated its cost-effectiveness (Roberts et al. 2006).

The importance of using dynamic transmission models for economic evaluation is now more fully understood. Useful introductions to the specifics of these models are available (Brisson and Edmunds 2003; Jit and Brisson 2011; Pitman 2014). There has also been a recent report from a task force on infectious disease modelling established by SMDM and ISPOR which, as well as providing another accessible introduction to the area, also offers a useful guide to good practice (Pitman et al. 2012). Further details of different model types can be found elsewhere (Barton et al. 2004; Brennan et al. 2006).

9.4.7 Selecting a model

Identifying an appropriate structure, and the type of model with which to implement it, is an extremely important stage of the decision modelling process. It is not possible to provide definitive guidelines for the selection of a particular model structure, as these have to depend on the overall objective of the economic evaluation as well as the nature of the disease process and impacts of the interventions. It has to be emphasized that all models are simplifications of reality, and the ultimate objective in selecting an appropriate structure for a decision model is to make the model no more complex than it has to be to address the policy questions appropriately. The value of a model is limited if it has been made highly complex in order to provide a more 'accurate' estimate of expected cost-effectiveness but its complexity precludes full uncertainty analysis.

9.5 Critical appraisal of decision-analytic models

Although decision-analytic modelling can provide a valuable framework for economic evaluation, its results are always conditional on the evidence used and on structural assumptions. In other words, there are good and bad decision models. As for any other form of evaluation, it is crucial that decision models are subject to careful critical review, and their results should not be used blindly in decision-making. There are a number of examples in the literature of papers that have discussed the characteristics of a 'good model' (Eddy et al. 2012; McCabe and Dixon 2000; Sculpher et al. 2000; Shemilt et al. 2013).

Methods guidelines in decision modelling were reviewed to compare and contrast their recommendations for good practice by Philips et al. (2004). There was a fair amount of consistency between the papers in their guidelines, but some areas of conflict. One example of this was the extent to which the availability of data should constrain the structure of a model as opposed to structure being determined based on the understanding of a condition and the effect of a treatment. The authors argued that, in principle, structure should not be influenced by the extent or quality of the data available to populate a model

but, in practice, this will not always be possible and more detailed guidance would be of value for analysts. Philips and colleagues went on to synthesize the available guidelines and, based on this, came up with a checklist to apply to specific decision models used in economic evaluation. It should be emphasized that this checklist relates to models and is not a substitute for those that relate to economic evaluation in general, including the one introduced in Chapter 3. This modelling checklist is reproduced in Annex 9.1.

9.6 Conclusions

Decision modelling is an important vehicle for economic evaluation, particularly where there is a specific resource allocation decision to be taken. The value of a formal analytic framework for decision-making is that it offers a means of synthesizing available evidence from a range of sources rather than relying on a *single study*, provides a way of relating the available evidence to the specific decision problem being posed, provides a framework within which the limitations of a single clinical study *as a vehicle for economic evaluation* can be addressed, helps decision-makers identify optimal interventions under conditions of uncertainty, and can contribute to the process of setting research priorities.

This chapter has provided an introduction to these methods, and further reading is available (Briggs et al. 2006). There remain important methods questions to address in decision modelling. These include how to develop efficient methods to identify evidence relating to all parameters in decision models and not just those relating to treatment effects, how to synthesize all available evidence in models and reflect the uncertainty and correlation in these data, and how to deal with uncertainty in the structure of decision models and reflect this in the value of information analysis. These issues are dealt with in Chapters 10 and 11. Although decision modelling has the potential to provide a powerful input for decision-making, there are good and bad applications of these methods, and critical appraisal is essential.

9.7 Exercise: developing a decision-analytic model

9.7.1 Background

Imagine that you have been asked to advise local decision-makers on the cost-effectiveness of antenatal HIV testing (i.e. testing pregnant women for HIV infection). You undertake a literature search to identify published economic evaluations, but find nothing to help you in your analysis. You quickly realize that you will have to undertake a decision analysis of your own using evidence from available sources.

9.7.2 The evidence

From a literature search you identify publications that provide you with the following information:

- If a woman has HIV and her infection is not known during pregnancy, the probability that she will transmit the infection to her child is 26%.
- If a woman's infection is known during pregnancy, however, it is possible to use risk-reduction interventions such as caesarean section, antiretroviral therapy, and bottle-feeding. These interventions cost £800 more than a normal delivery and

reduce the probability of vertical transmission to 7%, but only 95% of infected women accept them.

- Discussion with midwifery staff indicates that offering the test to women could be achieved at negligible additional cost, but your pathology laboratories suggest that each blood test will cost £10; they also indicate that the tests are 100% accurate (i.e. there are no false negatives or false positives).
- A published paper suggests that the prevalence of previously undetected HIV in the antenatal population in your area is 5%.

9.7.3 Assumptions

Discussions with professional staff indicate that the following assumptions can be justified:

- No woman will select to terminate on discovering she has HIV infection.
- All women who are tested positive will be offered risk-reduction interventions.

9.7.4 The task

1 Your task is the following:
   a To structure a decision tree characterizing the decision regarding whether or not to offer antenatal HIV testing.
   b To calculate the expected cost per true positive case detected.
2 What are likely to be the key sensitivity analyses to undertake?
3 What are the weaknesses of the analysis?

9.7.5 Solutions

1a The decision tree is shown in Figure 9.8.

1b Expected cost of testing

$= (810 \times 0.0033) + (810 \times 0.0441) + (10 \times 0.0007) + (10 \times 0.0019) + (10 \times 0.95)$

$=$ £47.92.

Probability of vertical transmission with testing $= 0.0033 + 0.0007 = 0.004$.

Expected cost of no testing $= 0$.

Probability of vertical transmission $= 0.013$.

Additional expected cost per HIV-infected child avoided:

- additional cost $=$ £47.92
- reduced vertical transmission $= 0.013 - 0.004 = 0.009$
- additional cost per HIV-infected birth avoided $= 47.92/0.009 =$ £5324.

2 The following sensitivity analyses would be warranted:
   - Due to parameter uncertainty:
     - probabilities of transmission
     - probability of acceptance of interventions
     - prevalence.

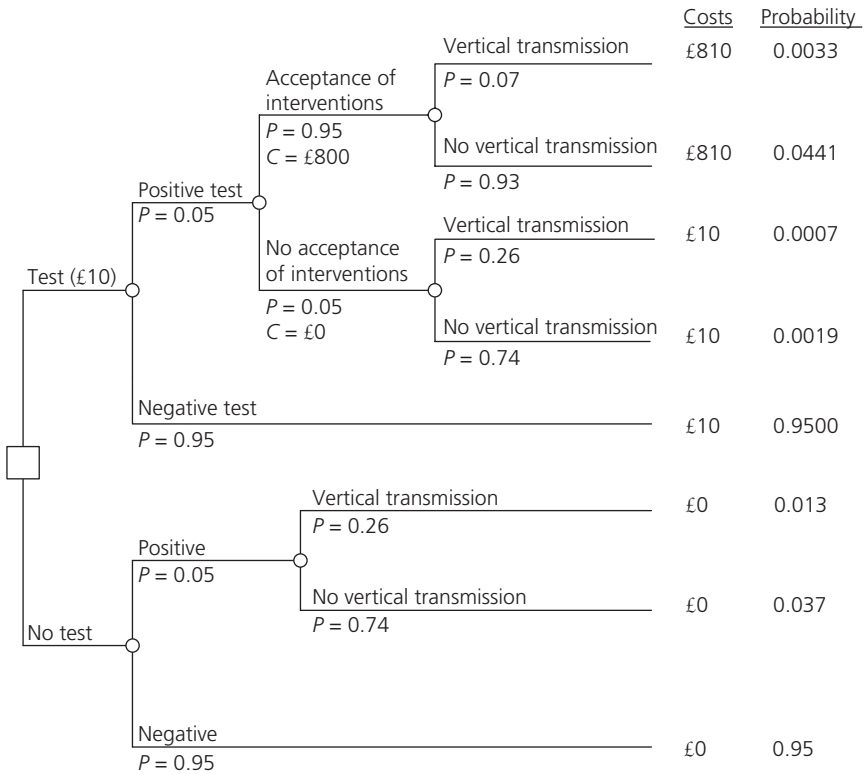


Fig. 9.8 Decision tree for antenatal HIV testing.

- Due to heterogeneity:
  - costs
  - probability of acceptance of interventions
  - prevalence.
- Due to structural uncertainty:
  - accuracy of test (need to restructure tree)
  - termination rate: in this model, termination may increase or decrease costs; there would be a difficulty in how this is dealt with on the outcomes side
  - uptake of test will not affect cost-effectiveness.

3 Weaknesses of the analysis:

- In reality, the relevant options to compare would be more likely to be universal testing versus high-risk group testing versus on-demand testing, rather than testing versus not.
- A full PSA would ideally be undertaken for to assess parameter uncertainty.
- The scope of analysis is limited:
  - It should have a longer-term model to include an assessment of the costs and (quality-adjusted) life-years conditional on HIV transmission.

- There should be a consideration of lifetime costs and outcomes.
- The model should consider the effect of testing on the women themselves in terms of the costs and outcomes of earlier treatment than would be expected without testing.
- Possible effects on horizontal transmission might be considered in a wider scope.

References

Ades, A.E., Sculpher, M.J., Gibb, D.M., et al. (1999). Cost effectiveness analysis of antenatal HIV screening in United Kingdom. *BMJ*, **319**, 1230–4.

Barton, P., Bryan, S., and Robinson, S. (2004). Modelling in the economic evaluation of health care: selecting the appropriate approach. *Journal of Health Services Research and Policy*, **9**, 110–18.

Basu, A. and Meltzer, D. (2007). Value of information on preference heterogeneity and individualized care. *Medical Decision Making*, **27**, 112–27.

Bojke, L., Epstein, D., Craig, D., et al. (2011). Modelling the cost-effectiveness of biologic treatments for psoriatic arthritis. *Rheumatology*, **50**(suppl 4), iv39–iv47.

Bravo Vergel, Y., Palmer, S., Asseburg, C., et al. (2007). Results of a comprehensive decision analysis. Is primary angioplasty cost effective in the UK? *Heart*, **93**, 1238–43.

Brennan, A., Chick, S.E., and Davies, R. (2006). A taxonomy of model structures for economic evaluation of health technologies. *Health Economics*, **15**, 1295–310.

Briggs, A. and Sculpher, M.J. (1998). An introduction to Markov modelling for economic evaluation. *PharmacoEconomics*, **13**, 397–409.

Briggs, A., Sculpher, M., Dawson, J., et al. (2004). The use of probabilistic decision models in technology assessment: the case of hip replacement. *Applied Health Economics and Policy*, **3**, 79–89.

Briggs, A., Claxton, K., and Sculpher, M. (2006). *Decision modelling for health economic evaluation*. Oxford: Oxford University Press.

Brisson, M. and Edmunds, W.J. (2003). Economic evaluation of vaccination programs: the impact of herd-immunity. *Medical Decision Making*, **23**, 76–82.

Buxton, M.J. and O'Brien, B.J. (1992). Economic evaluation of ondansetron: preliminary analysis using clinical trial data prior to price setting. *British Journal of Cancer*, **66**, S64–S67.

Caro, J.J., Moller, J., and Getsios, D. (2010). Discrete event simulation: the preferred technique for health economic evaluations? *Value in Health*, **13**, 1056–60.

Chancellor, J.V., Hill, A.M., Sabin, C.A., et al. (1997). Modelling the cost effectiveness of lamivudine/zidovudine combination therapy in HIV infection. *PharmacoEconomics*, **12**, 1–13.

Claxton, K., Sculpher, M.J., and Drummond, M.F. (2002). A rational framework for decision making by the National Institute for Clinical Excellence. *Lancet*, **360**, 711–15.

Colbourn, T., Asseburg, C., Bojke, L., et al. (2007). Prenatal screening and treatment strategies to prevent group B streptococcal and other bacterial infections in early infancy: cost-effectiveness and expected value of information analyses. *Health Technology Assessment*, **11**(29), 1–240.

Coyle, D., Buxton, M.J., and O'Brien, B.J. (2003). Stratified cost-effectiveness analysis: a framework for establishing efficient limited use criteria. *Health Economics*, **12**, 421–7.

Davies, R. (1985). An assessment of models of a health system. *Journal of the Operational Research Society*, **36**, 679–87.

Drummond, M.F., Barbieri, M., Cook, J., et al. (2009). Transferability of economic evaluations across jurisdictions: ISPOR Good Research Practices Task Force Report. *Value in Health*, **12**, 409–18.

Eddy, D.M., Hollingworth, W., Caro, J.J., et al. (2012). Model transparency and validation: a report of the ISPOR-SMDM Modeling Good Research Practices Task Force-7. *Value in Health*, **15**, 843–50.

Epstein, D.M., Sculpher, M.J., Manca, A., et al. (2008). Modelling the long-term cost-effectiveness of endovascular or open repair for abdominal aortic aneurysm. *British Journal of Surgery*, **95**, 183–90.

Griffiths, A., Paracha, N., Davies, A., et al. (2014). The cost effectiveness of ivabradine in the treatment of chronic heart failure from the UK National Health Service perspective. *Heart*, **100**, 1031–6.

Henriksson, M., Epstein, D.M., Palmer, S.J., et al. (2008). The cost-effectiveness of an early interventional strategy in non-ST-elevation acute coronary syndrome based on the RITA 3 trial. *Heart*, **94**, 717–23.

Hunink, M., Weinstein, M.C., Wittenberg, E., et al. (2014). *Decision making in health and medicine: integrating evidence and values*, 2nd edition. Cambridge: Cambridge University Press.

Jit, M. and Brisson, M. (2011). Modelling the epidemiology of infectious diseases for decision analysis. A primer. *PharmacoEconomics*, **29**, 371–86.

Karnon, J., Stahl, J., Brennan, A., et al. (2012). Modeling using discrete event simulation: a report of the ISPOR-SMDM Modeling Good Research Practices Task Force-4. *Value in Health*, **15**, 821–7.

Kobelt, G., Jonsson, B., Young, A., et al. (2003). The cost-effectiveness of infliximab (Remicade) in the treatment of rheumatoid arthritis in Sweden and the United Kingdom based on the ATTRACT study. *Rheumatology*, **42**, 326–35.

Mark, D.B., Hlatky, M.A., Califf, R.M., et al. (1995). Cost effectiveness of thrombolytic therapy with tissue plasminogen activator as compared with streptokinase for acute myocardial infarction. *New England Journal of Medicine*, **332**, 1418–24.

McCabe, C. and Dixon, S. (2000). Testing the validity of cost-effectiveness models. *PharmacoEconomics*, **17**, 501–13.

Neumann, P.J., Hermann, R.C., and Kuntz, K.M. (1999). Cost-effectiveness of donepezil in the treatment of mild or moderate Alzheimer's disease. *Neurology*, **52**, 1138–45.

Owens, D.K. and Nease, R.F. (1997). A normative analytic framework for development of practice guidelines for specific clinical populations. *Medical Decision Making*, **17**, 409–26.

Palmer, S., Sculpher, M., Philips, Z., et al. (2005). Management of non-ST-elevation acute coronary syndromes: how cost-effective are glycoprotein IIb/IIIa antagonists in the UK National Health Service? *International Journal of Cardiology*, **100**, 229–40.

Philips, Z., Ginnelly, L., Sculpher, M., et al. (2004). A review of guidelines for good practice in decision-analytic modelling in health technology assessment. *Health Technology Assessment*, **8**(36), 1–158.

Pitman, R. (2014). Infectious disease modelling, in A. J. Culyer (ed.), *Encyclopedia of health economics*. Amsterdam: Elsevier.

Pitman, R., Fisman, D., Zaric, G.S., et al. (2012). Dynamic transmission modeling: A report of the ISPOR-SMDM Modeling Good Research Practices Task Force Working Group-5. *Medical Decision Making*, **32**, 712–21.

Raiffa, H. (1968). *Decision analysis: introductory lectures on choices under uncertainty*. Reading, MA: Addison-Wesley.

Roberts, M., Russell, L.B., Paltiel, A.D., et al. (2012). Conceptualizing a model: A report of the ISPOR-SMDM Modeling Good Research Practices Task Force-2. *Value in Health*, **15**, 804–11.

Roberts, T.E., Robinson, S., Barton, P., et al. (2006). Screening for Chlamydia trachomatis: a systematic review of the economic evaluations and modelling. *Sexually Transmitted Infections*, **82**, 193–200.

Sackett, D.L., Rosenberg, W.M.C., Gray, J.A.M., Haynes, R.B., and Richardson, W.S. (1996). Evidence-based medicine: what it is and what it isn't. *BMJ*, **312**, 71–2.

Schulman, K.A., Glick, H.A., Rubin, H., and Eisenberg, J.M. (1991). Cost-effectiveness of HA-lA monoclonal antibody for gram-negative sepsis. *JAMA*, **266**, 3466–71.

Sculpher, M.J. (2008). Subgroups and heterogeneity in cost-effectiveness analysis. *PharmacoEconomics*, **26**, 799–806.

Sculpher, M.J. and Gafni, A. (2001). Recognizing diversity in public preferences: the use of preference sub-groups in cost-effectiveness analysis. *Health Economics*, **10**, 317–24.

Sculpher, M., Fenwick, E., and Claxton, K. (2000). Assessing quality in decision-analytic cost-effectiveness models. A suggested framework and example of application. *PharmacoEconomics*, **17**, 461–77.

Sculpher, M.J., Pang, F.S., Manca, A., et al. (2004). Generalisability in economic evaluation studies in health care: a review and case studies. *Health Technology Assessment*, **8**(49), 1–213.

Sculpher, M.J., Claxton, K.P., Drummond, M.F., et al. (2006). Whither trial-based economic evaluation for health care decision making? *Health Economics*, **15**, 677–87.

Shemilt, I., Wilson, E., and Vale, L. (2013). Quality assessment in modeling in decision-analytic models for economic evaluation, in A.J. Culyer (ed.), *Encyclopedia of health economics*. Amsterdam: Elsevier.

Siebert, U., Alagoz, O., Bayoumi, A.M., et al. (2012). State-transition modeling: a report of the ISPOR-SMDM Modeling Good Research Practices Task Force-3. *Value in Health*, **15**, 812–20.

Sonnenberg, F.A. and Beck, J.R. (1993). Markov models in medical decision making. *Medical Decision Making*, **13**, 322–38.

Spiegelhalter, D.J., Abrams, K.R., and Myles, J.P. (2004). *Bayesian approaches to clinical trials and health-care evaluation*. Chichester: Wiley.

Spiegelhalter, D.J. and Best, N.G. (2003). Bayesian approaches to multiple sources of evidence and uncertainty in complex cost-effectiveness modelling. *Statistics in Medicine*, **22**, 3687–709.

Standfield, L., Comans, T., and Scuffham, P. (2014). Markov modelling and discrete event simulation in health care: a systematic comparison. *International Journal of Technology Assessment in Health Care*, **30**(2), 1–8.

Walker, S., Girardin, F., McKenna, C., et al. (2013). Cost-effectiveness of cardiovascular magnetic resonance in the diagnosis of coronary heart disease: an economic evaluation using data from the CE-MARC study. *Heart*, **99**, 873–81.

Weinstein, M.C. and Fineberg, H.V. (1980). *Clinical decision analysis*. Philadelphia: W.B. Saunders.

Wilby, J., Kainth, K., Hawkins, N., et al. (2005). A rapid and systematic review of the clinical effectiveness, tolerability and cost effectiveness of newer drugs for epilepsy in adults. *Health Technology Assessment*, **9**(15), 1–157.

Woolacott, N., Hawkins, N., Mason, A., et al. (2006). Efalizumab and etanercept for the treatment of psoriasis. A systematic review. *Health Technology Assessment*, **10**(46), 1–258.

Annex 9.1 Checklist for assessing quality in decision-analytic models

Table A9.1, from Philips et al. (2004), provides a suggested checklist for assessing quality in decision-analytic models.

Table A9.1 A suggested checklist for assessing quality in decision-analytic models

Dimension of quality		Attributes of good practice	Questions for critical appraisal
Structure			
S1	Statement of decision problem/objective	There should be a clear statement of the decision problem prompting the analysis. The objective of the evaluation and of the model should be defined. The primary decision-maker should be stated clearly.	Is there a clear statement of the decision problem? Is the objective of the evaluation and model specified and consistent with the stated decision problem? Is the primary decision-maker specified?
S2	Statement of scope / perspective	The perspective of the model (relevant costs and consequences) should be stated clearly, and the model inputs should be consistent with the stated perspective and overall objective of the model. The scope of the decision model should be specified and justified. The outcomes of the model should reflect the perspective and scope of the model and should be consistent with the objective of the evaluation.	Is the perspective of the model stated clearly? Are the model inputs consistent with the stated perspective? Has the scope of the model been stated and justified? Are the outcomes of the model consistent with the perspective, scope, and overall objective of the model?

(*continued*)

Table A9.1 (continued) A suggested checklist for assessing quality in decision-analytic models

Dimension of quality		Attributes of good practice	Questions for critical appraisal
S3	Rationale for structure	The structure of the model should be consistent with a coherent theory of the health condition under evaluation and the treatment pathways (disease states or branches) should be chosen to reflect the underlying biological process of the disease in question and the impact of the intervention. The structure should not be dictated by current patterns of service provision. All sources of evidence used to develop and inform the structure of the model (i.e. the theory of disease) should be described. The structure should be consistent with this evidence.	Is the structure of the model consistent with a coherent theory of the health condition under evaluation? Are the sources of data used to develop the structure of the model specified? Are the causal relationships described by the model structure justified appropriately?
S4	Structural assumptions	All structural assumptions should be transparent and justified. They should be reasonable in the light of the needs and purposes of the decision-maker.	Are the structural assumptions transparent and justified? Are the structural assumptions reasonable given the overall objective, perspective, and scope of the model?
S5	Strategies/comparators	There should be a clear definition of the options under evaluation. All feasible and practical options relating to the stated decision problem should be evaluated. Options should not be constrained by the immediate concerns of the decision-maker, or data availability, nor limited to current clinical practice.	Is there a clear definition of the options under evaluation? Have all feasible and practical options been evaluated? Is there justification for the exclusion of feasible options?

Table A9.1 (continued) A suggested checklist for assessing quality in decision-analytic models

Dimension of quality		Attributes of good practice	Questions for critical appraisal
S6	Model type	The appropriate model type will be dictated by the stated decision problem and the choices made regarding the causal relationships within the model.	Is the chosen model type appropriate given the decision problem and specified causal relationships within the model?
S7	Time horizon	A model's time horizon should extend far enough into the future in order for it to reflect important differences between options. It is important to distinguish between the time horizon of the model, the duration of treatment and the duration of treatment effect.	Is the time horizon of the model sufficient to reflect all important differences between options? Are the time horizon of the model, the duration of treatment, and the duration of treatment effect described and justified?
S8	Disease states/pathways	Disease states/pathways should reflect the underlying biological process of the disease in question and the impact of interventions.	Do the disease states (state transition model) or the pathways (decision tree model) reflect the underlying biological process of the disease in question and the impact of interventions?
S9	Cycle length	For discrete time models, the cycle length should be dictated by the natural history of disease. It should be the minimum interval over which the pathology or symptoms are expected to alter.	Is the cycle length defined and justified in terms of the natural history of disease?

(*continued*)

Table A9.1 (continued) A suggested checklist for assessing quality in decision-analytic models

Dimension of quality		Attributes of good practice	Questions for critical appraisal
Data			
D1	Data identification	Methods for identifying data should be transparent and it should be clear that the data identified are appropriate given the objectives of the model. There should be justification of any choices that have been made about which specific data inputs are included in a model. It should be clear that particular attention has been paid to identifying data for those parameters to which the results of the model are particularly sensitive. Where expert opinion has been used to estimate particular parameters, sources and methods of elicitation should be described.	Are the data identification methods transparent and appropriate given the objectives of the model? Where choices have been made between data sources, are these justified appropriately? Has particular attention been paid to identifying data for the important parameters in the model? Has the quality of the data been assessed appropriately? Where expert opinion has been used, are the methods described and justified?
D2	Data modelling	All data modelling methodology should be described and based on justifiable statistical and epidemiological methods. Specific issues to consider include those below.	Is the data modelling methodology based on justifiable statistical and epidemiological techniques?
D2a	Baseline data	Baseline probabilities may be based on natural history data derived from epidemiological/observational studies or relate to the control group of an experimental study. Rates and interval probabilities should be transformed into transition probabilities appropriately. If there is evidence that time is an important factor in the calculation of transition probabilities in state transition models, this should be incorporated. If a half-cycle correction has not been used on all transitions in state transition model (costs and outcomes), this should be justified.	Is the choice of baseline data described and justified? Are transition probabilities calculated appropriately? Has a half-cycle correction been applied to both cost and outcome? If not, has this omission been justified?

Table A9.1 (continued) A suggested checklist for assessing quality in decision-analytic models

Dimension of quality		Attributes of good practice	Questions for critical appraisal
D2b	Treatment effects	Relative treatment effects derived from trial data should be synthesized using recognized meta-analytic techniques. The methods and assumptions that are used to extrapolate short-term results to final outcomes should be documented and justified. This should include justification of the choice of survival function (e.g. exponential or Weibull forms). Alternative assumptions should be explored through sensitivity analysis. Assumptions regarding the continuing effect of treatment once treatment is complete should be documented and justified. If evidence regarding the long-term effect of treatment is lacking, alternative assumptions should be explored through sensitivity analysis.	If relative treatment effects have been derived from trial data, have they been synthesized using appropriate techniques? Have the methods and assumptions used to extrapolate short-term results to final outcomes been documented and justified? Have alternative assumptions been explored through sensitivity analysis? Have assumptions regarding the continuing effect of treatment once treatment is complete been documented and justified? Have alternative assumptions been explored through sensitivity analysis?
D2c	Costs	Costing and discounting methods should accord with standard guidelines for economic evaluation.	Are the costs incorporated into the model justified? Has the source for all costs been described? Have discount rates been described andjustified given the target decision-maker?
D2d	Quality of life weights (utilities)	Utilities incorporated into the model should be appropriate for the specified decision problem.	Are the utilities incorporated into the model appropriate? Is the source for the utility weights referenced? Are the methods of derivation for the utility weights justified?

(*continued*)

Table A9.1 (continued) A suggested checklist for assessing quality in decision-analytic models

Dimension of quality		Attributes of good practice	Questions for critical appraisal
D3	Data incorporation	All data incorporated into the model should be described and the sources of all data should be given and reported in sufficient detail to allow the reader to be aware of the type of data that have been incorporated. Where data are not mutually consistent in the model, the choices and assumptions that have been made should be explicit and justified. The process of data incorporation should be transparent. It should be clear whether data are incorporated as a point estimate or as a distribution. If data have been incorporated as distributions as part of probabilistic analysis, the choice of distribution and its parameters should be described and justified.	Have all data incorporated into the model been described and referenced in sufficient detail? Has the use of mutually inconsistent data been justified (i.e. are assumptions and choices appropriate)? Is the process of data incorporation transparent? If data have been incorporated as distributions, has the choice of distribution for each parameter been described and justified? If data have been incorporated as distributions, is it clear that second-order uncertainty is reflected?
D4	Assessment of uncertainty	In assessing uncertainty, modellers should distinguish between the four principal types of uncertainty.	Have the four principal types of uncertainty been addressed? If not, has the omission of particular forms of uncertainty been justified?
D4a	Methodological	Methodological uncertainty relates to whether particular analytic steps taken in the analysis are the most appropriate.	Have methodological uncertainties been addressed by running alternative versions of the model with different methodological assumptions?
D4b	Structural	There should be evidence that structural uncertainties have been evaluated using sensitivity analysis.	Is there evidence that structural uncertainties have been addressed via sensitivity analysis?
D4c	Heterogeneity	It is important to distinguish between uncertainty resulting from the process of sampling from a population and variability due to heterogeneity (i.e. systematic differences between patient subgroups).	Has heterogeneity been dealt with by running the model separately for different subgroups?

Table A9.1 (continued) A suggested checklist for assessing quality in decision-analytic models

Dimension of quality		Attributes of good practice	Questions for critical appraisal
D4d	Parameter	Where data have been incorporated into the model as point estimates, the ranges used for sensitivity analysis should be stated and justified. Probabilistic analysis is the most appropriate method of handling parameter uncertainty because it facilitates assessment of the joint effect of uncertainty over all parameters (see data incorporation).	Are the methods of assessment of parameter uncertainty appropriate? If data are incorporated as point estimates, are the ranges used for sensitivity analysis stated clearly and justified?
Consistency			
C1	Internal consistency	There should be evidence that the internal consistency of the model has been evaluated in terms of its mathematical logic.	Is there evidence that the mathematical logic of the model has been tested thoroughly before use?
C2	External consistency	The results of a model should be explicable. Results should either make intuitive sense or counter-intuitive results should be fully explained. All relevant available data should be incorporated into a model. Data should not be withheld for purposes of assessing external consistency. The results of a model should be compared with those of previous models and any differences should be explained.	Are any counter-intuitive results from the model explained and justified? If the model has been calibrated against independent data, have any differences been explained and justified? Have the results of the model been compared with those of previous models and any differences in results explained?

Chapter 10

Identifying, synthesizing, and analysing evidence for economic evaluation

10.1 Introduction to evidence in economic evaluation

Chapter 8 considered the methods and practice of economic evaluation undertaken using a vehicle of a single trial or observational study. A more general approach to economic evaluation is where evidence is drawn together from a range of sources within a decision-analytic model. The rationale for modelling studies and their key characteristics were described in Chapter 9. An important feature of any model-based economic evaluation is the evidence to include in the analysis. This evidence generally relates to what can be described as 'clinical' parameters such as the relative effectiveness of a treatment, the underlying or baseline risks of particular clinical events, the prevalence or incidence of conditions, or the accuracy of particular tests. These types of parameters are also relevant in non-clinical settings such as evaluations of public health or prevention interventions. In addition, and depending on the type of economic analysis, evidence on the impact of a disease or intervention on health-related quality of life (HRQoL) may be needed, as is evidence for parameters relating to resource use and the cost of those resources. In addition to the need to acquire appropriate evidence on input parameters in models, the process of designing and structuring a model will also require various types of evidence. These include evidence relating to disease natural history, the effects of interventions, and the configuration of existing services.

The process of evidence-gathering for economic evaluation brings economics together with other disciplines such as clinical epidemiology, statistics, and information specialists. This chapter considers how different types of evidence are identified, synthesized, and appropriately analysed to feed into models. The chapter starts by considering how to define 'relevant evidence' for economic evaluation. It then considers methods for identifying and synthesizing evidence. The chapter also deals with some issues regarding the estimation of some types of parameters.

10.2 Defining relevant evidence

It is inevitable that any economic evaluation will need to make judgements regarding the evidence to include, and a key consideration is the *'relevance'* of the evidence to the decision problem being informed by the economic evaluation. This can be characterized on several dimensions. The first is the nature of the populations and subpopulations

being considered. For an evaluation of alternative therapeutic interventions, this could cover a patient group with a given diagnosis and subgroups with, for example, different levels of severity of the diagnosed condition. In a study of diagnostics, the population may be a group of individuals with signs and symptoms suggestive of one or more possible conditions. For an analysis of preventive interventions, the population could be individuals at risk of developing a given disease. Suitable evidence for an analysis will relate to the specifics of these defined populations.

The second aspect of a decision problem is the options being compared. In some studies it is possible to distinguish interventions of interest and comparators; this is true, for example, of economic evaluations of newly licensed pharmaceutical interventions compared to the range of existing therapies for the population(s) (comparators). Other economic evaluations simply compare the range of options available without making a distinction between interventions and comparators. However defined, the evidence gathered for the study needs to relate to the full range of options being compared.

A third element of the decision problem that helps to define *relevant* evidence is the jurisdiction, health care system, or subsystem where the resource allocation decision is to be made. For example, the problem may be whether to introduce a screening programme for diabetic retinopathy in a primary care setting in the United Kingdom. In judging the relevance of available evidence for use in an economic evaluation, the analyst needs to assess whether it was generated in this setting and jurisdiction. If not, (s)he has to assess whether the evidence is sufficiently generalizable to be reasonably comparable to the evidence that could have been collected in that setting. Issues of generalizability—in particular in contrast to internal validity—are discussed in Chapter 8. The degree of generalizability of evidence across settings requires an assessment of the extent of differences in factors such as the recipient populations, types of medical practice, and the nature of the health care system (e.g. its funding arrangements) (Sculpher et al. 2004). Some types of evidence may be more generalizable than others (e.g. treatment effectiveness may be expected to generalize across settings more than estimates of resource use). Furthermore, issues of generalizability can be relevant across time as well as between settings. Even if a particular type of evidence has been collected in the system for which the economic evaluation is being undertaken, if the study from which that evidence is taken was completed 20 years ago, it may be of limited relevance today as a result, for example, of changes in clinical practice.

10.3 Identifying and reviewing evidence

10.3.1 Reviewing economic evaluations

A key principle of evidence-based medicine and decision-making is that evidence should not be selected to suit a particular point of view. Rather, it should be the entirety of the evidence available for a particular parameter or, at least, it should be representative of the entire evidence base. A starting point in any economic evaluation should be to consider whether such a study has been undertaken before and, if so, whether further work is necessary. Therefore, reviewing previous economic evaluations is an important precursor for this type of research. There are resources available to support

this including databases, search terms for electronic databases like Medline, and checklists (Anderson and Shemilt 2010). Even if it is clear that an additional study is needed, there is value in a reviewing earlier research. For example, such studies can help to establish the best approach to developing a decision model for a specific decision problem, sources of evidence, and appropriate comparators. In some instances it may be possible to take an earlier economic evaluation and to update it using recent evidence, or using data which are more appropriate for a different jurisdiction. An example of this is a cost-effectiveness analysis (CEA) of a new pharmaceutical product for psoriatic arthritis which was based on an earlier model developed for similar products (Cummins et al. 2011). Although there are some examples in the literature (Anderson and Shemilt 2010), it is rare for studies to synthesize the various cost-effectiveness results from a systematic review. The reasons for this relate to the requirements of economic evaluation to support decision-making considered in Chapter 9. It is unlikely that a systematic review would identify a number of studies that have used sufficiently similar methods and all appropriate comparators suitable for the jurisdiction in which the decision is to be taken. Even if that were the case, a formal synthesis would only make sense if, together, they included all relevant evidence but there was no overlap in the evidence they used. In reality, the studies identified are likely to be heterogeneous in their methods and data, and to have been focused on different decision problems in different locations. In those circumstances formal synthesis would make little sense.

10.3.2 Identifying evidence for decision models

Many economic evaluations will, therefore, be based on newly developed decision models (albeit perhaps influenced by earlier versions in the literature). As discussed in Chapter 8, decision models provide a framework to direct all relevant evidence at the specific decision problem which is defined in terms of populations and subpopulations, the options being compared, and the relevant jurisdiction. Cooper and colleagues reviewed 180 HTA reports, of which 42 included decision-analytic models, with the purpose of assessing how evidence was sourced for models (Cooper et al. 2005). The authors concluded that evidence on the main measure of clinical effect was usually obtained using an explicit systematic review. For other types of evidence (e.g. resource use and HRQoL weights), the approach to identifying evidence was rarely transparent. A key component of systematic literature reviews, therefore, is to undertake searches for studies using, for example, bibliographic databases, which are transparent and reproducible, and use explicit selection criteria (Centre for Reviews and Dissemination 2009). The general methods of systematic review are well understood and widely used in relation to effectiveness evidence (Centre for Reviews and Dissemination 2009; Cochrane Collaboration 2011; Institute of Medicine 2011). Table 10.1 summarizes the key standards for systematic review as set out by the Institute of Medicine (Institute of Medicine 2011).

Consistent with the findings of Cooper and colleagues, however, methods for reviewing other types of evidence for economic analysis are less established (Cooper et al. 2005). This relates to the databases to search, the search strategies, quality assessment, and synthesis. Good reviews of the challenges, methods, and available sources

Table 10.1 A summary of the standards for systematic review defined by the Institute of Medicine

Domain	Standard
Initiating a Systematic Review	Establish a team with appropriate expertise and experience to conduct the systematic review
	Manage bias and conflict of interest of the team conducting the systematic review
	Ensure user and stakeholder input as the review is designed and conducted
	Manage bias and conflict of interest for individuals providing input into the systematic review
	Formulate the topic for the systematic review
	Develop a systematic review protocol
	Submit the protocol for peer review
	Make the final protocol publicly available, and add any amendments to the protocol in a timely fashion
Standards for Finding and Assessing Individual Studies	Conduct a comprehensive systematic search for evidence
	Take action to address potentially biased reporting of research results
	Screen and select studies
	Document the search
	Manage data collection
	Critically appraise each study
Standards for Synthesizing the Body of Evidence	Use a prespecified method to evaluate the body of evidence
	Conduct a qualitative synthesis
	Decide if, in addition to a qualitative analysis, the systematic review will include a quantitative analysis (meta-analysis)
	If conducting a meta-analysis: ◆ Use expert methodologists to develop, execute and peer review the meta-analyses ◆ Address the heterogeneity among study effects ◆ Accompany all estimates with measures of statistical uncertainty ◆ Assess the sensitivity of conclusions to changes in the protocol, assumptions, and study selection

Table 10.1 (continued) A summary of the standards for systematic review defined by the Institute of Medicine

Domain	Standard
Standards for Reporting Systematic Reviews	Prepare final report using a structured format
	Peer review the draft report
	Publish the final report in a manner that ensures free public access

for these other evidence sources can be found elsewhere (Paisley 2014). Some specific databases have also been developed which can be helpful for identifying economic evaluation studies and hence non-clinical evidence which might be suitable for models. These include the NHS Economic Evaluation Database (EED) (<http://www.crd.york.ac.uk/crdweb/AboutPage.asp>) and the Tufts Cost-effectiveness Analysis Registry (<https://research.tufts-nemc.org/cear4/>).

Time and other resource constraints will limit the extent to which a given literature search can be exhaustive—for example, not every bibliographic database may be searched. Researchers should, therefore, focus on being able to justify the extent of their searching on the basis that it is reasonable given resource constraints and the resulting studies are representative.

There will also be situations where it can be justified not to use a full systematic review to identify relevant evidence for models. First, some types of evidence may be judged to be very specific to a jurisdiction and, furthermore, there are few sources of that evidence; in which case an exhaustive search for international sources may be considered unnecessary. This situation is particularly true of costs. For example, in the United Kingdom, costs for many items of resource use are taken from a small number of sources based on routinely collected data, such as the Personal Social Services Research Unit's unit costs in health and social care (Personal Social Services Research Unit 2013), NHS Reference Costs (Department of Health Payment by Results Team 2013), and the British National Formulary (Joint Formulary Committee 2013). It may also be true for certain epidemiological evidence—for example, the incidence of particular cancers is collected centrally and routinely in most resource-rich countries, as cancer is a registrable disease. A second reason why a full systematic review may not be needed to estimate particular model parameters is that the results of the analysis are very insensitive to that variation in that parameter. Here 'results' may not just be the mean measure of interest (e.g. the mean incremental cost-effectiveness ratio), but also include, for example, a measure of decision uncertainty and value of further research (see Chapter 11). In such circumstances relatively simple search methods may be justified, such as the use of a single literature database. Although an exhaustive search for all types of evidence for a decision model may not always be necessary, an important

principle that should always be adhered to is the need for transparency in reporting the sources of evidence used (for all parameters) and justification for the selection of particular sources.

As shown in Table 10.1, critical appraisal is an important element of systematic review. There are various guides and checklists to help the assessment of the quality of different sources of clinical evidence, including randomized trials (Schulz et al. 2010) and diagnostic studies (Whiting et al. 2003). There are fewer examples for other types of parameter, although the general methods of economic evaluation covered in this book can help in the assessment of quality relating to HRQoL weights, resource use, and cost. For studies being undertaken for particular decision-making authorities, there may be published methods guides which indicate how the decision-maker views quality in evidence. For example, the methods guide issued by National Institute for Health and Care Excellence (NICE) in the United Kingdom is clear about the type of HRQoL weights it prefers, based on the EQ5D instrument (NICE 2013). More generally, the concept of the 'reference case' is used by NICE to reflect the characteristics of an economic evaluation that need to be present for at least one of the analyses presented to its decision-making committees. This follows the general principles of the reference case set out by a panel of experts established by the US Public Health Service in 1996, the main focus of which was to promote consistency in methods between studies to enhance comparability over time (Gold et al. 1996). The concept of the reference case has also been taken up in methods guidelines for economic evaluations funded by the Bill and Melinda Gates Foundation, generally in low- and middle-income countries (Bill and Melinda Gates Foundation et al. 2014).

10.3.3 Hierarchy of evidence

Systematic review methods in effectiveness research have often included the concept of the hierarchy of evidence. That is, a ranking of research designs reflecting the likelihood that they will provide suitable estimates of effectiveness. The term 'suitable' here generally relates to the study design providing high levels of internal validity (see Chapter 8). For example, Guyatt and colleagues suggested the following hierarchy of evidence for treatment decisions (Guyatt et al. 2000):

1 N-of-1 randomized trial
2 Systematic review of randomized trials
3 Single randomized trial
4 Systematic review of observational studies
5 Single observational study
6 Physiological study
7 Unsystematic clinical observations.

An N-of-1 randomized trial is a study which includes only one patient, with a randomized order in which they receive the experimental or control therapies, and has little role in generating evidence for economic evaluation.

Even in the context of clinical effectiveness alone, such hierarchies can be criticized—indeed, Guyatt and colleagues argued that they should not be considered 'absolute'. One criticism is that the general design of a study is only one consideration; it is also important to consider how the study has been implemented using detailed quality assessment. It may be the case, for example, that a well-designed and implemented observational study can provide a more suitable estimate of effectiveness than a poor-quality trial. A second issue relates to the trade-off between internal and external validity as discussed in Chapter 8. Again, a well-conducted observational study may provide a more appropriate estimate of effectiveness than a trial conducted in a setting unrelated to where the decision is being taken (see Section 10.2).

If the hierarchy of evidence is used to guide systematic reviews for evidence to include in economic evaluation studies, additional problems emerge. First, the concept of 'effectiveness' in an economic evaluation can often be divided into distinct types of parameter when the clinical focus is on changing the rate of particular events (e.g. the rate of heart attacks or strokes in patients with high blood cholesterol). As discussed in Chapters 8 and 9, studies will often seek an estimate of the underlying (or baseline) rate of events with usual care, as well as the relative effectiveness of another intervention compared with the baseline (e.g. a hazard ratio). The measure of relative effectiveness would ideally be taken from a (well-designed and implemented) randomized trial, due to the risk of selection bias in an observational study. However, the estimate of the baseline rate may be more appropriately taken from an observational study. This is because such a design may recruit a more representative sample of patients (at least in a given jurisdiction), may be large enough to look at how baseline risk varies between particular subgroups, and is not susceptible to selection bias for such a parameter. The conventional hierarchies of evidence focus on relative effectiveness, not baseline risks. Ideally the parameters required for a cost-effectiveness model would be estimated using a synthesis of all studies, whatever their design or quality. However, this would require the risk of bias resulting from a lack of internal or external validity to be reflected in the measure of uncertainty associated with the estimates. This could then be expressed in the measure of decision uncertainty and the value of additional research which are concepts considered in more detail in Chapter 11. Although there is some literature on such 'generalized evidence synthesis' (Prevost et al. 2000) and on quantifying bias in clinical evidence (Turner et al. 2009), these are not yet widely used in parameter estimation for economic evaluation.

Secondly, economic evaluations require a range of evidence other than effectiveness. When identifying evidence on HRQoL, resource use, and costs, for example, such hierarchies of evidence can provide little guidance.

10.4 Synthesizing evidence

In seeking relevant evidence for a model-based economic evaluation, the focus will often be on identifying studies described in published literature (e.g. randomized trials and observational studies). Examples of the types of parameters which could be estimated include:

- the rate of cancer progression in a particular type of patient under standard care
- a hazard ratio showing the rate of progression with a new treatment compared to standard care

- the rate of mortality in patients whose cancer has progressed
- the probability of a given side effect with a new treatment
- the mean HRQoL weight associated with the use of the new therapy
- the mean resource use or cost associated with cancer progression.

10.4.1 To synthesize or not to synthesize: the question of heterogeneity

As described above, a key principle in identifying evidence for models is to seek to capture all relevant evidence, or at least a representative selection. For many model parameters, therefore, there will be multiple estimates. An important issue to address in such situations is whether to combine (synthesize) these different sources into a single estimated mean with uncertainty, or to keep them separate. An important concept to understand in considering the appropriateness and nature of any synthesis is *heterogeneity*, which is relevant in several ways in the context of estimating parameters for models. When evidence is taken from several studies, differences in the estimated parameters between studies may reveal themselves. These differences may just be random statistical 'noise' due to their coming from small samples; or they could reflect genuine differences between studies such as the definition of the control group, the countries in which the studies were undertaken, or the types of patients included. There can also be heterogeneity in the definition of end points in studies or in the way outcomes are measured. For economic evaluation to support decisions, the ideal is to be able to disentangle this heterogeneity by explaining its different sources and potentially using estimates of parameters in models that most suitably reflect the nature of the decision—e.g. specific to a particular jurisdiction/locality or to different subgroups of patients which could be distinguished in decisions. By reflecting heterogeneity in parameter estimates it is also possible to explore whether, for example, a particular therapy is more cost-effective in some types of patients, locations, or settings than others.

Chapter 8 covered the use of a single randomized trial or observational study as a vehicle for economic evaluation. One of the strengths of such a study is that the data on the various effects and resource use implications of interventions are collected on individual patients (or participants in non-clinical settings). This provides a means of studying how costs and effects vary between patients, offering a more rigorous means of reflecting uncertainty and heterogeneity. An advantage of modelling studies is that, when there are multiple sources of evidence for a given parameter, these can all be incorporated into the model. However, a careful assessment of heterogeneity, with a view to estimating results by subgroup, is also an important objective. Whether such data come from a single study, several studies, or are used in combination with summary data, access to individual patient data allows heterogeneity in parameter estimates to be carefully explored using appropriate statistical analysis (Espinoza et al. 2014). For a full assessment of the most cost-effective form of management of a given patient group, it is important both to bring to bear all appropriate evidence (suggesting a search for all relevant studies), and also to have access to some or all studies in the form of individual patient data if possible. This can provide a basis for the modelling to establish which intervention is most cost-effective for particular types of patient. Table 10.2 shows an

Table 10.2 Example of cost-effectiveness analysis with detailed analysis of heterogeneity. It shows how the cost per QALY gained from endovascular aneurysm repair (EVAR) compared with open surgery varies with particular patient characteristics: the size of the abdominal aortic aneurysm (AAA), and fitness. The table also shows the effect of some of the uncertain assumptions in the model in the form of alternative scenarios

		Fitness											
		Good				**Moderate**				**Poor**			
		Age (years)											
Scenario	AAA (cm)	70	75	80	85	70	75	80	85	70	75	80	85
Base case	5.5	1 120 563	215 306	96 902	62 817	79 539	53 000	38 006	28 181	27 658	21 442	17 354	14 604
	6.5	Dom	2 918 114	132 053	59 579	91 947	48 990	32 801	24 959	23 267	17 954	14 857	13 131
	7.5	Dom	138 913	63 775	37 757	44 264	29 560	21 816	17 248	15 168	12 313	10 669	9855
Lower cost of follow-up and lower rate of reinterventions	5.5	178 616	76 951	43 055	31 317	33 486	24 572	19 272	15 568	13 108	10 939	9588	8724
	6.5	Dom	338 899	61 553	31 980	40 766	24 227	17 735	14 631	11 615	9666	8670	8264
	7.5	Dom	60 551	31 961	20 810	20 795	15.014	12 002	10 253	7686	6714	6299	6274
Odds ratio of operative mortality 0.25 not 0.35	5 5	129 313	70 922	46 805	33 019	35 001	26 132	19 819	15 627	14 188	11 506	9647	8533
	6.5	346 307	77 388	43 933	28 413	31 102	21 922	16 737	13 499	11 602	9626	8418	7378
	7.5	74 231	38 924	25 945	18 974	18 792	14 233	11 441	9710	7978	6876	6273	6146

(*continued*)

Table 10.2 (continued) Example of cost-effectiveness analysis with detailed analysis of heterogeneity. It shows how the cost per QALY gained from endovascular aneurysm repair (EVAR) compared with open surgery varies with particular patient characteristics: the size of the abdominal aortic aneurysm (AAA), and fitness. The table also shows the effect of some of the uncertain assumptions in the model in the form of alternative scenarios

		Fitness											
		Good			Moderate				Poor				
		Age (years)											
Lower cost of follow-up, lower rate of reinterventions and equal cost of procedures	5.5	107 229	43 213	22 396	15 019	19 277	13 115	9457	7020	7183	5510	4412	3682
	6 5	Dom	181 687	30 406	14 545	22 467	12 305	8252	6258	6069	4619	3776	3307
	7.5	Dom	32 405	15 749	9434	11 436	7602	5561	4361	4001	3192	2724	2483

Key: Dark grey shading, incremental cost-effectiveness ratio (ICER) >£30 001 per QALY or EVAR dominated; light grey shading, £20 001 per QALY <ICER <£30 000 per QALY; unshaded, ICER <£20 000 per QALY. The ICER is the difference in expected costs/difference in expected QALYs. Dominated (Dom) means that EVAR has less expected benefits and higher costs than open repair.

example of a modelling study which looked at the cost-effectiveness of endovascular aneurysm repair compared with open surgery for abdominal aortic aneurysm (Chambers et al. 2009). On the basis of individual patient data from a particular randomized controlled trial (RCT), heterogeneity in cost-effectiveness according to the size of the aneurysm, age, and fitness was assessed. In principle, decision-makers are able to make different decisions regarding the funding of an intervention for different types of patients (Sculpher 2008).

10.4.2 Treatment effects: fixed- and random-effects methods

The methods of evidence synthesis are most widely described in the context of effectiveness data. The key principles and methods developed with such evidence are, however, relevant to all parameters used in economic evaluation.

In economic evaluations of alternative therapeutic interventions, the effectiveness of one intervention compared to another (the *relative effect* or *treatment effect*) is often a major factor in determining cost-effectiveness. It is, therefore, particularly important for these parameters to be identified and estimated carefully. RCTs are generally preferred as a source of estimates of treatment effect due to their internal validity, secured by randomization of patients to alternative interventions (although see comments in Section 10.3.3). As discussed in Chapter 8, however, there can be challenges to the generalizability of treatment effect estimates from RCTs given that patients are usually selected from a wider population, and the observed and/or unobserved characteristics of that sample may be different from those of the wider population.

It is expected that some form of systematic review will be used to establish which estimates are available and meta-analysis will often be necessary to synthesize these estimates for inclusion in a decision model. It is not the purpose of this book to cover meta-analysis in detail; many good texts are devoted to these methods (e.g. Borenstein et al. 2009; Lipsey and Wilson 2001; Sutton et al. 2000). As a brief overview, it is important to outline here two different approaches to meta-analysis: *fixed-effects* and *random-effects*.

10.4.2.1 Fixed-effects meta-analysis

With a fixed-effects meta-analysis there is an assumption that all studies share a common (true) effect; in other words, that the characteristics of the various studies that could influence the estimated effects are the same across studies so they are all estimating the same underlying effect. Between-study variation in the *observed* effect is, therefore, assumed simply random. A fixed-effects meta-analysis seeks to estimate the population effect from a number of observed effects for each study. This is achieved by calculating a weighted mean across studies where the weight for each study is the inverse of the variance for that study.

The assumption of fixed-effects may be open to challenge in some situations because of between-study sources of heterogeneity such as differences in study participants or interventions being administered. As discussed in Section 10.4.1, in such a situation a judgement is needed as to whether the studies identified are so different from each other that they are estimating markedly different underlying effects and should not be synthesized because this may result in misleading findings in the decision model. In such a situation, subgroup estimates of the parameter (and hence of cost-effectiveness

through the decision model) may be more appropriate. One way of implementing this is by using each study estimate separately in the model and presenting a series of different results reflecting the specifics of each particular study (i.e. each study provides a unique subgroup estimate). This effectively puts the onus on the decision-maker to establish which of the studies from which the evidence is taken is closest or more appropriate to their decision problem.

Alternatively, the studies may be different in terms of particular characteristics which can be used as covariables in regression analysis to generate subgroup-specific estimates of particular parameters. An example of this form of 'meta-regression' was a synthesis of trial evidence on the effectiveness of primary angioplasty versus medical management in patients with acute myocardial infarction (MI) (Asseburg et al. 2007). The effectiveness of primary angioplasty—in terms of the risk of mortality, non-fatal MIs, and non-fatal strokes—varied across studies and this could be partially explained by differences between studies in the average time from symptom onset until angioplasty was undertaken. Meta-regression was used to generate estimates of the treatment effect as a function of this time duration. Figure 10.1 shows the results for the mortality outcome 1 month and 6 months after primary angioplasty. Each circle represents a randomized trial, with its size proportional to the sample size in the study. The central bold line is the estimate of the mean relationship between the absolute mortality difference and the additional delay in providing primary angioplasty compared to thrombolytics; the two thinner lines are the 95% credibility intervals. This was also used in a cost-effectiveness model which found that angioplasty was cost-effective for 'delay times' below about 80 minutes (Bravo Vergel et al. 2007).

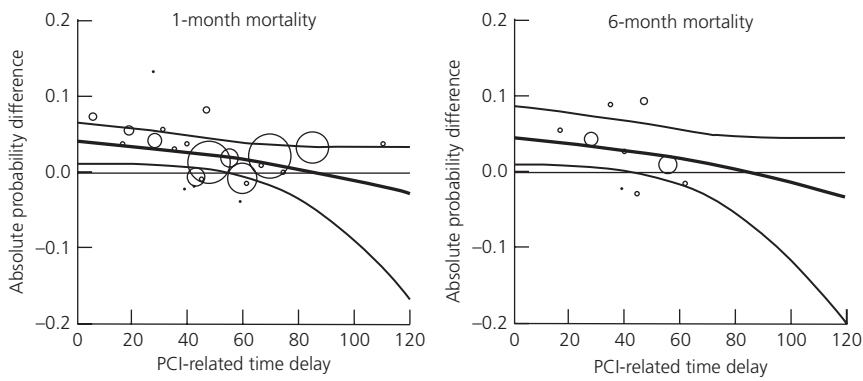


Fig. 10.1 Results of a meta-regression showing the treatment effect on mortality of primary angioplasty relative to thrombolytic treatment. The graphs show means and 95% credibility intervals plotted against the additional time delay to initiating primary angioplasty. Values above the 0.0 horizontal line indicate that angioplasty results in fewer clinical events. Each point represents a trial and the size is proportional to the trial sample size.

Although meta-regression can be a powerful tool for explaining heterogeneity, it should be used carefully given that the effective sample size for such analyses relates to the number of studies, which is often small. This can be a particular problem as the covariables are often average patient characteristics for each study which can lead to 'ecological bias' (Thompson and Higgins 2002). There is also a risk of confounding with meta-regression where an association may actually reflect correlation with another variable not included in the analysis. Further details on methods to explain heterogeneity in synthesis are available elsewhere (Welton et al. 2012). One possible way to use meta-regression is to adjust estimates of a parameter, such as a treatment effect, for the quality or relevance of the study which may be an important source of heterogeneity. This can involve assessment of quality on the basis of individual characteristics (e.g. concealment of treatment allocation) or using a standardized scale covering a range of characteristics and using these as covariables in a statistical model. In principle, the estimates from such a model could be used to parametrize a decision model with 'quality-adjusted' parameter estimates. Juni and colleagues considered the use of quality adjustment using a meta-analysis of trials looking at low-molecular weight heparin compared with standard heparin to prevent postoperative deep vein thrombosis (Juni et al. 1999). Regression analysis was used to assess the importance of 25 different quality scales and individual quality characteristics. The study found that the type of scale could dramatically affect how studies were defined as 'low quality' and 'high quality', and hence the interpretation of the meta-analysis. The authors were critical of the use of composite quality scales given the variation in how they are developed and, in particular, their use to adjust statistically the results of meta-analyses. Rather, they advocated the selection of a small number of relevant quality characteristics and caution with any statistical modelling.

10.4.2.2 Random-effects meta-analysis

In some situations a judgement may be reached that the studies, while not estimating the same underlying effect, are sufficiently common to be meaningfully synthesized. Here random-effects meta-analysis can be considered, which assumes that the true effects being estimated by the studies are not identical but drawn from a common probability distribution. For a random-effects meta-analysis, the weights used to calculate the weighted mean are a function of both a study's within-study variance and variation in the true effect between studies. Although the estimated mean effect will often be similar for fixed- and random-effects meta-analyses, the measure of parameter uncertainty is likely to be different. In general, random-effects analysis generates greater uncertainty in parameter estimates, in particular when this uncertainty appropriately reflects the heterogeneity itself rather than just the mean of the random-effects (Ades and Higgins 2005). When a random-effects meta-analysis is used to estimate a particular parameter for a decision model, the assumption is effectively that it is reasonable to 'average' the estimates across the available studies even though they exhibit some between-study differences (i.e. they are sufficiently similar studies for purposes of decision-making), but that the parameter uncertainty needs to reflect the fact that these studies are measuring slightly different things. The way in which random-effects analyses are implemented in a decision model should reflect our understanding about the similarities between the trials included in the synthesis and the target decision problem.

10.4.3 Network meta-analysis

One area of evidence synthesis for treatment effects that has developed rapidly in recent years and has been increasingly used in economic evaluation is network meta-analysis. As described in Chapters 1, 2, 3, and 4, one of the key principles of economic evaluation is the need for careful selection of options to compare. For example, it would generally not be appropriate to assess the cost-effectiveness of a new treatment for a given condition with one currently available treatment if there are six other interventions that are widely used and there is no evidence as to which existing therapy is the most cost-effective. Therefore, it will often be necessary to compare a number of alternative treatments within a given economic analysis, but all of these will seldom have been compared directly within a randomized trial. So the analysis will need to use a synthesis of treatment effect estimates coming from trials comparing a small number of alternative treatments (usually two) but which can be linked together (networked) to provide comparable overall effect estimates across all interventions. Crucially, this wider approach to synthesis using networks of evidence must not break randomization; that is, all treatment effects being synthesized must be based on comparisons from one or more RCTs. Indeed, even when there is 'direct' evidence from one or more trials comparing all relevant interventions, if studies also exist that compare subsets of treatments, this may call for a network meta-analysis to bring all these sources of evidence together.

10.4.3.1 Indirect and mixed treatment comparisons

Figure 10.2 illustrates the concepts behind network meta-analysis adapted from Sutton et al. (2008). Panel a shows the conventional randomized trial with the lines connecting the circles showing the direct comparison of two treatments of interest (A and B) for the economic evaluation (shown by the A and B circles being filled in). Panel b, however, shows a situation where the two treatments being assessed in the economic evaluation (A and B) have not been compared directly in a trial. In such a situation network meta-analysis would seek one or more other treatments that could facilitate an *indirect* estimate of the treatment effect of A compared to B.

Panel b shows a third treatment, C, which has been compared directly to both A and B in separate randomized trials. Treatment C is not considered relevant to the economic evaluation (indicated by the circle C being unfilled) because, for example, it is not available in the health system for which the analysis is being undertaken. The incorrect way of undertaking this indirect comparison between A and B (sometimes called an 'unadjusted indirect comparison') would to break randomization and simply to compare the absolute effects in treatment A patients (in the AC trial) with the absolute effects in treatment B patients (in the BC trial): as these patients are not distinguished by randomization, the resulting estimate would be potentially biased. The correct approach (the 'adjusted' indirect comparison) would be to retain randomization and to use the relative treatment effects from the AC comparison (which can be defined as d_{AC}) and from the BC comparison (d_{BC}) to estimate d_{AB} (Bucher et al. 1997). This is achieved by using conventional pairwise meta-analysis to estimate d_{AC} and d_{BC} (or taking these directly from a study if only one trial exists for each comparison), and estimating d_{AB} as:

$$d_{AB} = d_{AC} - d_{BC} \tag{10.1}$$

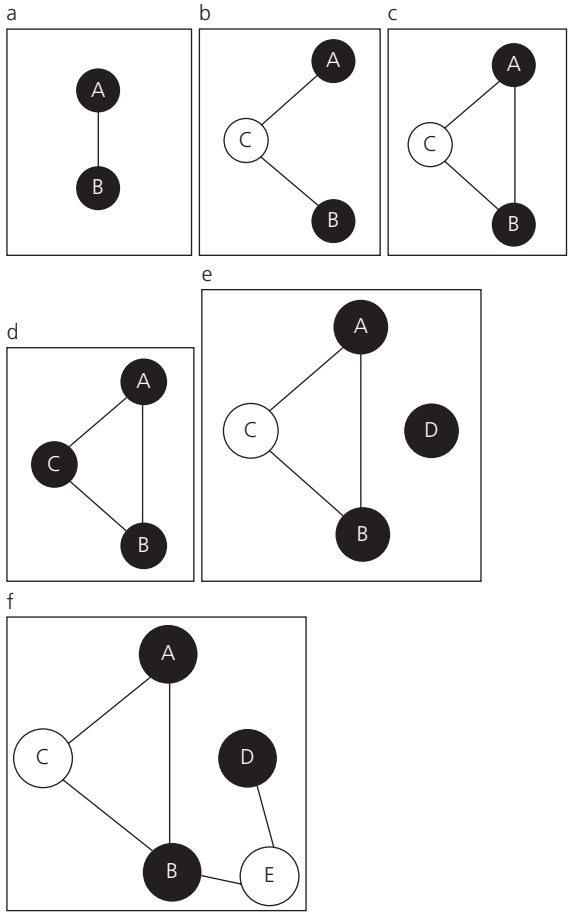


Fig. 10.2 Diagrams showing network meta-analysis using indirect and direct evidence.

Panels c and d in Figure 10.2 show how this method can be generalized to situations when there is not only direct evidence from a trial on treatments A versus B, but also indirect evidence as before (through A versus C, and B versus C). Panels c and d differ in terms of whether treatment C is considered of interest in the evaluation. This type of synthesis is termed *mixed treatment comparison*, and methods have been extensively used to bring these different types of evidence together simultaneously (Welton et al. 2012).

Figure 10.3 provides an example of the type of complex networks of evidence that can be synthesized using mixed treatment comparison. This is based on a study which

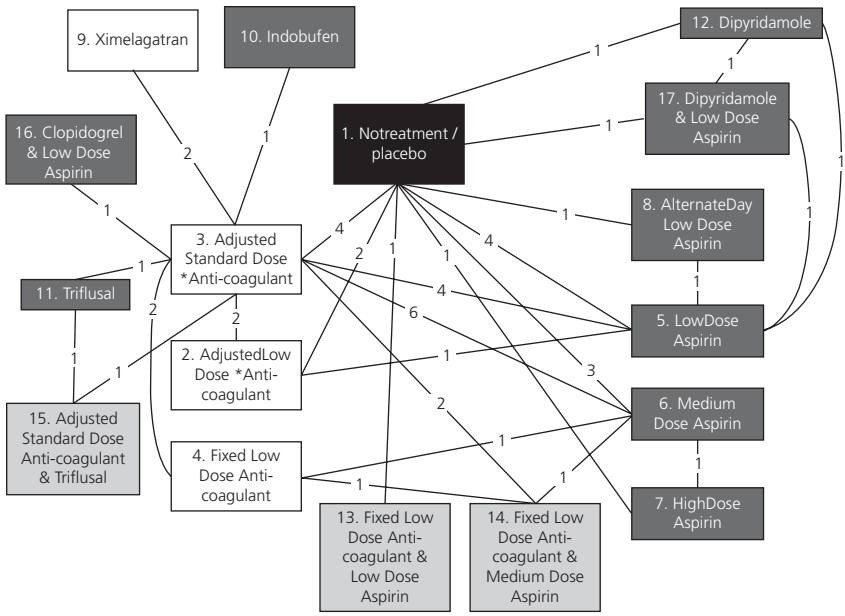


Fig. 10.3 An example of a full network meta-analysis considering alternative treatments for atrial fibrillation.

* In a number of trials clinicians were free in their choice of oral anti-coagulants.

Reproduced with permission from Cooper, N.J. et al, Addressing between-study heterogeneity and inconsistency in mixed treatment comparisons: Application to stroke prevention treatments in individuals with non-rheumatic atrial fibrillation, *Statistics in Medicine*, Volume 28, Issue 14, pp. 1861–81, Copyright © 2009 John Wiley & Sons, Ltd.

looked to compare the effectiveness of different interventions to prevent stroke in patients with non-rheumatic atrial fibrillation (Cooper et al. 2009). Each box (node) shows an option being compared in the synthesis, with the shading representing a class of treatment (white boxes represent anticoagulants, dark grey boxes antiplatelet treatment, light grey mixed, black placebo/no treatment). The lines linking the boxes are trials or pairs of trial arms (with the numbers shown on the lines). This synthesis includes the indirect comparison considered in panel b of Figure 10.2. For example, although options 17 and 8 have not been directly compared in an RCT, there are various common comparators for these interventions (e.g. non-treatment/placebo) which facilitate an indirect comparison. In estimating the comparative effectiveness of, for example, options 9 and 12, the links are more complex but they exist (e.g. via 3 and 1). The choice between fixed- and random-effects methods is also relevant to these analyses.

To return to Figure 10.2, Panel e shows a situation where one of the interventions of interest (D) is not linked to the network. This means that, in order to compare D with the alternatives, it would be necessary to break randomization, thus reducing the internal validity of the comparisons. In some circumstances, however, it may be possible to locate a fifth intervention (E, as shown in panel f) which, although not of interest in the evaluation in its own right (e.g. because it is no longer available), links D to the network and facilitates comparison with the other options without breaking randomization.

10.4.3.2 Decision space and evidence space

An important implication of network meta-analysis is that a network of evidence can consist of *decision space* and *evidence space* (Hoaglin et al. 2011). The former describes the options being compared which are directly relevant to the decision. The latter refers to trials that include these options but may also include other options which may not be related to the decision but which contribute evidence relating to those that are. In panel f of Figure 10.2, for example, the evidence space includes some trials involving interventions not currently in the decision space (interventions C and E). In some situations the inclusion into the network of options which are not related to the decision but which still contribute evidence that may influence estimates of the effectiveness of relevant options would have the potential to make the network very large. In such situations there is a need for careful consideration as to where to place the boundaries of the evidence space within the total network. This needs a judgement regarding the likely impact of adding additional trials into the network which do not include decision-relevant options. In general this impact can be expected to be smaller when such trials are further away in the network from decision-relevant options and the smaller their sample sizes.

10.4.3.3 Critical assessment of network meta-analyses

Although network meta-analyses are now increasingly used to estimate treatment effects for decision-analytic models assessing cost-effectiveness, it is important to be clear about the nature of the evidence lying behind them and the implications for the validity of, and uncertainty in, the estimates. A task force on indirect comparison and network meta-analysis undertaken by the International Society for Pharmacoeconomics and Outcomes Research (ISPOR) outlined the key assessments which have to be made for such analyses (Hoaglin et al. 2011). As with 'pairwise' meta-analysis (see Section 10.4.2), it is necessary to consider heterogeneity between studies due to the characteristics of patients who were randomized or the features of the study such as its design, location or age. The type and extent of heterogeneity, and whether it threatens the validity of a synthesis, is generally a judgement based on clinical guidance. The key issue to consider is whether any differences between the trials are likely to impact on the treatment effect estimates. This means these differences would need to be potential treatment effect modifiers and not just impact on the baseline risk of events in a given study.

For network meta-analyses which include both direct and indirect evidence of particular treatment effects, it is important to assess the degree of consistency between these two sources of evidence. Dias and colleagues offer alternative methods to assess consistency and to present this graphically (Dias et al. 2013). When these analyses suggest inconsistency between direct and indirect estimates, it is important to reconsider all the evidence, and not just that which was shown to be inconsistent as the methods do not indicate which evidence is 'wrong'. This needs clinical input as is it not a statistical issue.

The ISPOR Task Force presents a checklist for assessing network meta-analyses with items relating to search strategies, data collection, the statistical analysis plan, data analysis and reporting (Hoaglin et al. 2011). Readers interested in learning more about network meta-analysis are directed towards Welton et al. (Welton et al. 2012).

10.5 Estimating other parameters for economic evaluation

Most methods guidance regarding parameter estimation relates to treatment effectiveness parameters. This applies, in particular, to the methods of systematic review and evidence synthesis. The results of an economic evaluation may often be driven by estimates of clinical effectiveness, but there are other types of parameter (clinical and economic) that must be estimated for decision models, and these are often central to decision-making about the value of a given intervention. Although often described in terms of estimates of treatment effectiveness, the principles outlined in Sections 10.3 and 10.4 also apply to other types of parameters in models. These general principles include:

- being transparent in identifying evidence for parameter estimation
- being systematic in identifying evidence, in the sense that a given search can be replicated by others
- careful assessment of heterogeneity in parameter estimates
- the need to estimate mean parameter values as well as to characterize uncertainty in those estimates
- appropriate synthesis of estimates from different sources reflecting heterogeneity as far as possible.

10.5.1 Event probabilities

Most decision models rely on probability estimates to characterize the likelihood of specific events over a particular time period. These include the risk of a disease progressing, the likelihood of an adverse effect from a treatment, the probability of a correct diagnosis in patients with the disease in question, and the risk of death over time. These types of parameter estimates can be drawn from a review of previously published studies and from the analysis of individual patient data from single or multiple studies. Section 10.3.3 describes how the effectiveness of an intervention in terms of absolute risk reduction of a particular event such as death is made up of two components: baseline risk (without treatment or with existing therapy) and the effectiveness of a new intervention relative to that baseline. The baseline risk is a probability that may be taken from a randomized trial but could also be taken from an observational study. Sources exist for detailed introductions to probability theory in general (Ross 2013) and in decision analysis in health care in particular (Briggs et al. 2006a; Hunink et al. 2014; Pettiti 1999). Some of the issues that need to be considered regarding estimating probability parameters for models are considered in the following sections.

10.5.1.1 Probabilities in diagnostics

Economic evaluations of diagnostic interventions and strategies are perhaps one of the most challenging areas of the field (Laking et al. 2006; Phelps and Mushlin 1988). This is partly because they can include a range of different aspects of clinical management including testing, screening, establishing prognosis, and monitoring. Complexity is also a challenge as evaluations will often have to compare a very large number of alternative

diagnostic strategies—e.g. a number of tests used in alternative sequences or jointly (remember the group B streptococcus example from Section 9.2.1). Diagnostic models can be complex as they often have to link the results of particular testing strategies with subsequent changes in therapies and ultimate impacts on outcomes. Probabilities are used widely to model the characteristics of diagnostics. The prevalence (or prior probability) of a disease is an important starting point in most diagnostic models: what is the probability that a given disease exists in the particular population to be tested? The accuracy of tests is also expressed in terms of probabilities. For example, a test's sensitivity is the probability that the test will be positive in patients with the disease, and a test's specificity is the probability of the test being negative in patients without the disease. Further reading on the clinical epidemiology of diagnostics is available elsewhere (Newman and Kohn 2009).

10.5.1.2 Probabilities and rates

The terms probabilities (or risks) and rates are often used as if they are synonymous. In fact there are important differences between these concepts which it is important to understand when working with probability estimates in decision modelling. A probability is defined as the likelihood of a particular event over a given period of time and is expressed on a scale running from 0 to 1. In contrast, a rate is defined as the instantaneous potential for an event at any point in time, and runs from 0 to infinity. The two measures are related in that the magnitude of a probability, and how it varies with time, is governed by an underlying rate. Specifically, a rate can be derived using the following expression:

$$r = -[\ln(1 - p)] / t \tag{10.2}$$

where r is the rate (which is assumed to be constant over time), p is the probability over the time period t and ln is the natural log. Although the rate is formally an instantaneous potential, it cannot be observed in this form and is expressed in terms of events per patient per unit of time. By rearranging this equation it is possible to express a probability as a function of a rate (where the latter is again assumed to be constant over time):

$$p = 1 - \exp(-rt) \tag{10.3}$$

where exp is the exponent. This has very practical importance for analysts seeking estimates of probabilities in the literature. This is because probabilities are typically sought to relate to a period defined by the model—for example, the probability of a particular adverse event over the first month of a treatment or transition probabilities in a monthly cycle in a Markov model. If the literature presents estimates of the probability over a different period (say 6 months), it would be inappropriate simply to divide the estimate in the literature (e.g. by 6) to adjust the estimate to the period of a month required for the model. Rather, it is necessary to calculate the underlying rate (which is assumed constant over time), to adjust this to the appropriate time period and then to recalculate the probability for that time period.

As an example, assume that an estimate of a probability of a particular event over 1 year is required for a model. Only one estimate is identified in the published literature, but

only a 5 year probability (estimated at 0.2). It would be incorrect simply to divide 0.2 by 5 to calculate a 1 year probability. Instead, it is necessary, firstly, to recover the underlying rate per patient-year assuming it is constant over time using the expression in equation 10.2 above:

$$\text{Rate per patient-year} = -\left[\ln\left(1-0.2\right)\right]/5 = 0.04463 \tag{10.4}$$

This can then be translated into an annual probability using equation 10.3:

$$1 \text{ year probability} = 1 - \exp\left(-0.04463\right) = 0.043648 \tag{10.5}$$

It can be seen that the correct annual probability is different to a recalculation based on dividing 0.2 by 5 (i.e. 0.4) and to the rate per person-year. Although these differences are small, using an incorrect probability estimate in a model will result in errors. A fuller discussion of the differences between rates and probabilities is available in Fleurence and Hollenbeak (2007) or Miller and Homan (1994).

10.5.1.3 Synthesizing probability estimates

As estimates of the probability of a given event may be available from a number of studies, synthesis may be appropriate as long as it can be assumed that the studies are sufficiently similar to be estimating the same 'true' effect (fixed-effects meta-analysis) or the differences in the 'true' effect being estimated between studies is modest and can be characterized with a probability distribution (random-effects probability distribution). As outlined in Section 10.4.2, it is important to establish whether studies exhibit heterogeneity which is too great to be handled with random-effects analysis and/or which can be explained using covariables and meta-regression.

Most meta-analysis methods focus on treatment effects and are, therefore, comparing variables across, for example, arms of a trial (see Section 10.4.2). Sometimes it might be necessary to synthesis non-comparative probability estimates from a set of studies—these can be called one-variable relationships (Lipsey and Wilson 2001). This might be, for example, prevalence estimates or the risks of side effects from a specific treatment. The methods of meta-analysis are very similar to those used for treatment effects. For example, Lipsey and Wilson present two methods, one based directly on proportions and the other using conversion to the logit scale (Lipsey and Wilson 2001).

10.5.1.4 Probabilities that vary with time: applying survival analysis

For many aspects of decision analysis, a judgement needs to be made regarding whether a simplifying assumption is a reasonable approximation of known reality and, if not, whether it risks the model generating misleadingly inaccurate results. This is often true with respect to probabilities and whether they are constant over time or vary. This may be particularly important in state transition models such as Markov models where transition probabilities can predict a patient's progression through states

representing, for example, disease severity over a large number of years. In some instances the available evidence limits an assessment of whether probabilities are time varying—the example above where summary evidence in a published paper only indicates a 0.2 probability of an event over 5 years is an instance of this.

When there is access to individual patient data from studies it is possible to explore how probabilities change over time. This involves applying the statistical tools for survival analysis to understand how the rate of an event changes over time and using this as the basis to calculate probabilities. Good general texts on survival analysis are available (Collett 1994), and a detailed consideration of how survival analysis is used to estimate time-varying transition probabilities for decision analysis can be found in Chapter 3 of Briggs et al. (2006a). As discussed in Chapter 9, an important aspect of the use of survival analysis to estimate time-varying transition probabilities relates to the frequent need to extrapolate probability estimates beyond the period over which data are observed in studies. It is often necessary to adopt a lifetime time horizon for economic evaluations. This is particularly the case when there is expected to be a difference in mortality between the options being compared and the measure of benefits is life-years or QALYs gained. Even though this may be observed over a short period such as, say, a 3 year follow-up in a clinical trial, considering only differential life-years or QALYs over 3 years may be a significant underestimate of the lifetime benefit. A long-term estimate of life-years and QALYs would need predictions of how mortality probabilities change beyond the clinical study, and how these vary by intervention. This is true not just for the probability of death but can be generalized to any longer-term risks.

As described in Section 9.2.4, various assumptions can be used to achieve this. A feature of modelling studies for economic evaluation over recent years has been the use of parametric survival models to extrapolate long-term probabilities. In a review of technology appraisals in cancer undertaken for the National Institute for Health and Care Excellence (NICE) up until 2009, Latimer found that 32 studies (71% of the sample) used some form of parametric survival model (Latimer 2013a). For this type of modelling the choice of specific survival model is crucial as this determines how the rate of events changes over time. A widely used parametric model for estimating transition probabilities for cost-effectiveness models is the Weibull function which can take a range of shapes depending on the value of the parameters estimated from the available data. For example, the CEA of alternative forms of management for patients with non-ST-elevation acute coronary syndrome described in Section 9.2.6 used a Weibull distribution to estimate the long-term probability of a composite event of death or non-fatal MI, and how that probability varied with time (Henriksson et al. 2008). The model was fitted to trial data over a follow-up period of 5 years and then used to estimate transition probabilities for the Markov model over a patient's lifetime. This also included a series of covariables which modelled how the rate of events varied depending on characteristics of patients when they were first identified (e.g. smoking status and previous diagnoses), and this facilitated various subgroup analyses for the CEA.

The use of parametric survival models as the basis of extrapolation on cost-effectiveness needs to be undertaken with considerable care. The first challenge is the choice of which survival function to use as this can have a large impact on long-term results. It is always necessary to assess how well a range of functions fits the available data

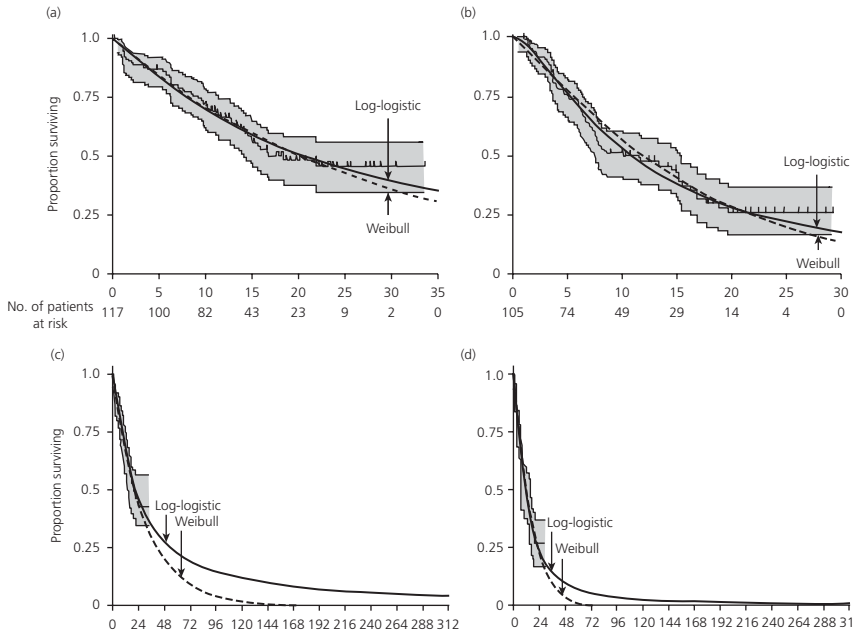


Fig. 10.4 Fitting two alternative parametric functions to trial data as a basis of extrapolation. Kaplan–Meier estimates show the observed trial data (uneven lines) and log-logistic and Weibull functions are shown as smooth lines. Panels a and b show the within-trial period (about 35 months) for two groups of patients. Panels c and d show the longer-term extrapolation over approximately 30 years.

using appropriate statistical 'diagnostics'. However, providing a good fit to the observed short-term data is not a sufficient basis for choosing a function for extrapolation. This is because the different approaches can give widely different long-term estimates even though they may appear to fit the observed data similarly well. This is illustrated in Figure 10.4 which shows the results of alternative approaches to extrapolation using trial mortality data for the management of higher-risk myelodysplastic syndromes (Connock et al. 2011). Log-logistic and Weibull parametric functions were fitted to the trial data—to the azacitidine–best supportive care (BSC) group (Figures 10.4a and 10.4c) and the BSC group (Figures 10.4b and 10.4d). Despite both functions fitting the observed data relatively well over about 3 years (Figures 10.4a and b), there were marked differences in the proportion estimated to remain alive after 30 years (Figures 10.4c and d). It is, therefore, always necessary to select functions which generate plausible long-term projections. Where possible, this process should be informed by other datasets (typically non-trials) which provide longer-term time-to-event estimates than those being used in the survival analysis (Latimer 2013b).

10.5.2 Health-related quality of life weights

Chapter 5 described how HRQoL weights are measured and valued for economic evaluation. Chapter 8 covered issues related to the collection of these data in clinical studies and their use in economic evaluations based on those studies. When decision-analytic models are the vehicle for economic evaluation, there are a number of considerations regarding HRQoL estimates. An important issue is what the decision model structure and approach imply about HRQoL weights. In particular, are the options being compared considered to have a direct and potentially differential impact on HRQoL, or are these effects mediated via treatment effects on clinical events or periods in health states which have consequences for HRQoL? An example of a model with direct HRQoL effects is an analysis of the cost-effectiveness of enhanced external counterpulsation (EECP) compared with no active treatment in patients with chronic stable angina, with outcomes expressed as QALYs (McKenna et al. 2010). Using HRQoL data based on patients' responses to the SF36 instrument in a RCT, the model incorporated the difference between EECP and no treatment in mean change in HRQoL (in terms of preference weights) between baseline and 1 year's follow-up. The model then used elicited parameter estimates from clinical experts to incorporate changes in HRQoL over time and to estimate long-term QALYs (Section 10.5.5 describes expert elicitation in more detail). The structure of this model reflected the fact that EECP was considered to impact directly on symptoms, and hence HRQoL, rather than on clinical events.

An example of a model where HRQoL was incorporated indirectly as a function of clinical evidence is an analysis of salmeterol/fluticasone propionate compared with fluticasone propionate alone in patients with asthma being managed in accordance with clinical guidelines (Briggs et al. 2006b). The model categorized patients into mutually exclusive health states based on control of symptoms using data from an RCT. These states were labelled 'totally controlled', 'well controlled', 'not well controlled but without exacerbation', and 'exacerbation'. The time in these states represented the key source of the difference in effectiveness between the treatments. HRQoL for each state was estimated using regression analysis based on a disease-specific HRQoL instrument collected in the same RCT. This was mapped to preference weights using a published mapping function. QALYs for each treatment were calculated as the sum of the expected time in each state weighted by their respective preference scores. In this example, therefore, the impact of the alternative treatments on HRQoL and QALYs was mediated through patients' mean time in health states.

The second, and more general, issue to consider with HRQoL parameters is to reflect the range of health states and populations within a decision model in identifying appropriate estimates of HRQoL weights. Figure 10.5 was developed by Papaioannou and colleagues to summarize these considerations for a model relating to interventions for osteoporosis (Papaioannou et al. 2010). It reflects the model structure in terms of the full range of health states relating to disease stage, fracture history, fracture type, and non-osteoporotic health states. It also includes a baseline HRQoL level before the fracture which is expected to differ by age and whether or not an individual already has established osteoporosis. The figure also shows the patient populations of interest, indicating that HRQoL weights for the various health states may, in principle, vary

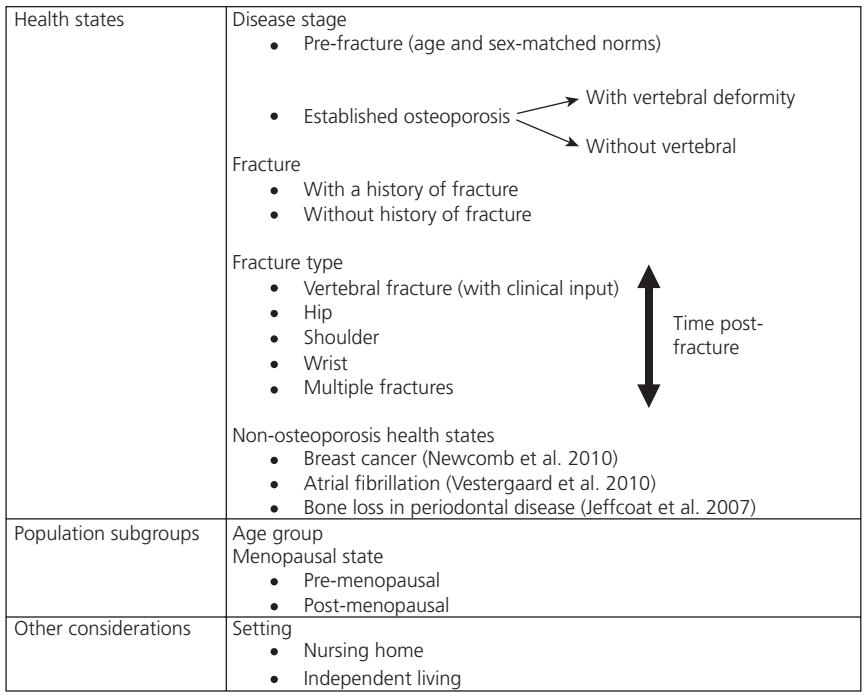


Fig. 10.5 Establishing the need for HRQoL weights for a model. Example of osteoporosis.

Reprinted from *Value in Health*, Volume 16, Issue 4, Papaioannou D. et al., Systematic searching and selection of health state utility values from the literature, pp. 686–95, Copyright © 2013 International Society for Pharmacoeconomics and Outcomes Research (ISPOR), with permission from Elsevier, <http://www.sciencedirect.com/science/journal/10983015>.

according to a woman's age and menopausal status. There is also a consideration of whether HRQoL weights might vary between the different settings in which a patient is located, specifically in a nursing home compared to independent living.

As with other parameters, the principle is that estimates of HRQoL values are identified in a systematic and transparent manner, with clear justification for the estimates selected. This is rarely the case in practice, however: in a review of cost-effectiveness analyses submitted to NICE between 2004 and 2008, 39 submissions (55% of the total sample) took HRQoL estimates from published studies, and only 31% of these were identified through a systematic review (Tosh et al. 2011). A rare example of a systematic review of HRQoL weights was a review and synthesis in the area of breast cancer by Peasgood et al. (2010). The study searched 13 databases, identifying 49 relevant papers and 476 estimates of HRQoL weights. As with any review, an assessment of the appropriateness and quality of the range of studies providing estimates of HRQoL is important. Papaioannou and colleagues suggest a number of quality criteria to assess with HRQoL studies (Papaioannou et al. 2010). These are similar to those that would be

suitable for a clinical study, such as how well the sample matches the target population (e.g. respondent selection and recruitment, and inclusion/exclusion criteria), response rates, loss to follow-up, and missing data. The relevance of the data to the decision being modelled is also important, and would suggest a consideration of the jurisdiction where the data were collected and whether the data reflect any guidelines a decision-maker has issued regarding HRQoL.

A systematic review of HRQoL weights for models would consider a full range of literature, including specific HRQoL estimates from clinical studies in patient populations of interest. There are also published reviews going across a range of disease areas (Bell et al. 2001; Tengs and Wallace 2000). More recently, large general population surveys have been undertaken from which estimates of HRQoL in a range of conditions have been estimated. An example is the nationally representative Medical Expenditure Panel Survey (MEPS) in the United States which, between 2000 and 2002, collected EQ5D data as well as information about medical conditions (Sullivan and Ghushchyan 2006). A sample of 38 678 individuals from the US civilian non-institutionalized population completed the instrument. Statistical modelling was used to estimate the HRQoL decrements associated with a range of medical conditions which were categorized into 693 groups based on three-digit International Classification of Disease-9 (ICD) codes. These models also allowed for other characteristics of individuals which may influence HRQoL, such as age and number of comorbidities. The original analysis used the US valuations of the EQ5D, and the UK valuations were used in a follow-up paper (Sullivan et al. 2011).

The literature holds few examples of the use of meta-analysis to synthesize HRQoL weights when a range of estimates exists. Again, the study by Peasgood and colleagues provides a rare example of the use of such methods (Peasgood et al. 2010). Where there were considered to be sufficient data reflecting reasonable homogeneity in studies, meta-analysis was used, together with meta-regression to control for disease state, valuation method, and source of valuations. The study looked at health states across breast cancer, and categorized these into six groups: screening-related, preventive, adverse events, non-specific, metastatic, and early breast cancer (Peasgood et al. 2010).

Most studies using models focus on specific decision problems relating to particular types of patient, which generally include fewer health states with a limited number of available studies. Together with the difficulty in synthesizing different types of HRQoL measure, this probably explains the limited use of meta-analysis. There will often be more than one estimate of a HRQoL weight for a given health state, however. Therefore, even in the absence of formal methods of synthesis, there remains a need for the analyst to justify the selection of the base-case value. Sensitivity analyses should then be used to assess the implications of the uncertainty in the base-case value (i.e. the precision of the estimate, reflecting variability in the measurement and the sample size); and scenario analysis would be important to consider the impact of using alternative sources of estimates (Chapter 11 considers uncertainty in detail).

One source of HRQoL weights for decision models is a particular RCT or observational study in which relevant data were collected and to which the analyst has access in terms of individual patient data. This is often the case when researchers have collected resource use and HRQoL data within a particular clinical study but have used a decision model as the vehicle for their economic analysis. As described in Chapter 9, there

may very good reasons to do this, including the need to have a wider set of alternative options than those in the clinical study, to extrapolate estimates of costs and benefits beyond the follow-up period of the trial, or to incorporate a wider set of evidence about key parameters than that in the clinical study.

It is always important for researchers to justify their source of parameter estimates, and this remains the case in sourcing HRQoL evidence from a single clinical study. However, if the decision problem addressed by the model is closely related to that implied by the design of the clinical study there may be a strong case to argue that HRQoL data in that study are the best available and synthesis with other evidence is not relevant or justified.

When individual patient data are available for the estimation of HRQoL weights for a decision model, these are often used to reflect an indirect impact of treatment on HRQoL via clinical events or time in states, as discussed earlier in this section. An example of such a study is a CEA of intensive blood glucose and tight blood pressure control in type 2 diabetics using data from the UK Prospective Diabetes Study (UKPDS; Clarke et al. 2002, 2005). Based on the administration of the EQ5D to 3192 patients who remained in the trial when the analysis was undertaken, the HRQoL weights were estimated using multivariable regression analysis. Mean HRQoL was estimated for patients without any complications; in addition, decrements in HRQoL were estimated for a series of clinical events: MI, ischaemic heart disease or angina, stroke, heart failure, amputation, and blindness in one eye. These were then used in the cost-effectiveness model as part of the estimate of long-term QALYs, using data on the incidence of complications from the whole trial.

It is also possible to use statistical analysis to estimate direct treatment effects in terms of HRQoL for use in a decision model. An example of this is the model-based CEA using data from the RITA-2 trial described in Section 10.5.1 (Henriksson et al. 2008). Unlike the UKPDS, EQ5D were collected in all trial patients at specific time points: baseline, 4 months, 1 year, and then annually thereafter. In a similar way to the UKPDS analysis, multivariable regression analysis was developed to estimate the decrements in HRQoL associated with events in the model, but this was based on the history of those events in trial patients at randomization rather than new events in the trial. In addition, the change in HRQoL over 4 and 12 months was estimated, conditional on whether a patient had experienced a non-fatal MI during follow-up. These parameter estimates were used in the model to contribute to an estimate of long-term QALYs for the alternative options depending on the characteristics of patients when decisions about medical management are being taken.

An important challenge for such analyses is to relate patients' responses to HRQoL instruments in a trial to the clinical events they experience. Typically these instruments are administered at baseline and at regular intervals during follow-up, as in the RITA-2 trial. In contrast, clinical events can occur at any point after randomization, and the timing of events may be very different from when patients complete their HRQoL questionnaires. Depending on the size of the sample, the number of times the questionnaire is administered and the number of clinical events, this problem may be overcome by reflecting the difference in time between the event and HRQoL measurement explicitly in the statistical model.

The use of statistical models to estimate HRQoL parameters for decision models is becoming more widespread. Section 8.3.3 considers some of the challenges facing

those undertaking regression analysis of such data (Basu and Manca 2012). The use of regression to map the relationship between disease-specific measures of HRQoL and generic, preference-based measures necessary to estimate QALYs is also very important in parametrizing decision models, and is discussed in Section 5.6.

10.5.3 Resource use and costs

Clearly a range of cost estimates is required for a decision-analytic model focused on economic evaluation. Chapter 7 describes the principles, controversies, and challenges associated with costing. The focus here is on some of the issues associated with identifying evidence and estimating parameters for decision models. As Chapter 7 makes clear, cost estimates are a product of the physical resources associated with an activity and the unit costs/prices used to value the resource use in monetary terms. The principles of this process are those outlined for other parameters elsewhere in this chapter. In particular, there is a need for explicitness in justifying the source of evidence for cost estimates and, where alternative appropriate estimates exist, for this to be reflected in the analysis of uncertainty. Many of the practical issues are similar to those relating to HRQoL weights as discussed in the last section. In particular, there are even fewer examples of formal systematic reviews of cost data and of the synthesis of alternative estimates to inform parameter estimation in decision models. One explanation for this is that, in many health care systems, the unit costs/prices associated with resource use are estimated and published. A good example of this is the list price of pharmaceuticals and medical devices which are generally published by manufacturers and readily available to researchers. A problem with this source of data, however, is the fact that manufacturers often offer discounts to individual health care providers so the list price may not actually be the price paid. Furthermore, the magnitude of discounts is generally commercially sensitive information and is not published, so the actual price is uncertain to the analysis. In an attempt to deal with this issue, a CEA of alternative chemotherapy regimens for patients with poor-prognosis advanced colorectal cancer used prices for the chemotherapy agents paid by one (unidentified) UK hospital in the base-case, and list prices for the product in a sensitivity analysis (Manca et al. 2012). Readily available published prices may also be available for other types of resource use. In the UK, for example, estimates of average unit costs are published for a range of NHS resource items including general practice attendance and nurse visits (Personal Social Services Research Unit 2013). A number of health care systems publish the rates they use to reimburse hospitals and other providers for particular categories of activity—for example, reimbursement of hospitals by Medicare in the United States using diagnostic related groups (see <http://www.cms.gov/Medicare/Medicare.html>). These reimbursement rates may be different from the costs actually incurred by the provider and, depending on the perspective of the analysis, these estimates may need adjusted to reflect this (see Section 7.1.2). The availability of national or regional published data that are suitable for unit costs estimates in modelling limits the need for systematic review. This reflects the fact that, unlike some other parameters in a model such as those relating to efficacy, unit costs are not considered to generalize across jurisdictions (Sculpher et al. 2004), so synthesis of published estimates across locations would generally not make sense.

The process of identifying suitable resource use (as opposed to cost) estimates for a decision model may also benefit from nationally available data. For example, the reimbursement rates paid to hospitals for the in-patient management of a MI should reflect a representative estimate of the duration of in-patient stay and the intensity of care. In which case, this could be used as the basis of estimating a cost to go directly into a model which is modelling the rate of MIs over time, although any significant costs of an MI falling outside the hospital (e.g. primary care) would need to be added. Again, resource use estimates are usually sought which relate to the jurisdiction for which the evaluation is being undertaken, so it would not be appropriate to synthesize estimates from a range of settings. Generally speaking, however, there are likely to be some costs in a model for which resource use estimates are not readily available from nationally representative sources. This may true, for example, of the number of primary care contacts following an MI or the duration of drug therapy following an acute infection. In these circumstances there may be scope for a systematic review of available estimates in the literature although, again, this is likely to be focused on the jurisdiction of interest for the evaluation. Very often, however, there is an absence of data which are sufficiently relevant to the decision problem. In these situations, researchers will need to rely on assumptions to relate representative costs for the health system, such as those taken from reimbursement rates, to the specific parameters they need to estimate for their models.

As for HRQoL weights, researchers may have access to individual patient data on resource use and costs from which relevant parameters for models can be estimated. These could come from prospective clinical studies such as RCTs or administrative databases. For example, in the decision model estimating the cost-effectiveness of alternative forms of management for non-ST-elevation MI using evidence from the RITA-3 trial, the mean costs per patient associated with dying during the initial hospitalization and with non-fatal MI in the year following the event were estimated from resource use collected in the trial (Henriksson et al. 2008). An important advantage of access to data from such studies is that costs can be estimated as a function of patients' characteristics. In the study using RITA-3 data, the impact of a patient's age and clinical characteristics were also considered. In another cardiovascular model, focusing on the cost-effectiveness of an ACE inhibitor in patients with stable angina, individual patient data on resource use were taken from a multinational RCT, although unit costs were taken from UK sources (reflecting the purpose of the analysis to guide UK decisions) (Briggs et al. 2007). The statistical model estimating costs for the decision model, therefore, used a covariable to represent whether the patient was treated in the United Kingdom; this was used to estimate a UK-specific estimate of the background cost of care. Section 8.3.1 discusses the types of statistical models that have been developed to reflect the particular features of resource use and cost data.

10.5.4 Model calibration as a basis of parameter estimation

A method known as model calibration has traditionally been used in situations where particular clinical or epidemiological phenomena cannot be directly observed, so it is not possible to estimate related model parameters directly. An example of such 'unobservables' is the rate of clinical presentation of a disease within a model to estimate

the effectiveness and cost-effectiveness of screening. This is because it is never possible directly to measure the denominator of such a rate in asymptomatic individuals with a disease who are, by definition, unobserved. In such cases, calibration involves finding the value of one or more unobserved parameters which, when used in the decision model, generate model outputs which are reasonably consistent with other sources of data which are not directly used to parametrize the model.

More recently, the use of model calibration has been proposed as a method for informing model input parameter values that are observed, but around which there is significant uncertainty. In this role, model calibration is able to represent likely correlations between model parameters, and reducing the parameter space in which sensitivity analyses are undertaken.

Vanni and colleagues describe seven analytical steps in the use of calibration (Vanni et al. 2011). The first is the decision about the parameters to estimate through calibration, generally all of the parameters that influence the calibration targets identified in step two. The second is the choice of calibration targets; that is the external data against which the models 'goodness of fit' is to be assessed. Given the need for models to relate to the jurisdiction where the decision is being taken, these targets may be estimated by locality. There is no limit to the number of calibration targets, but targets that reflect longer-term outcomes and treatment effects are particularly useful. The third step is to establish the appropriate measure of goodness of fit. These are statistical measures and include least squares and likelihood measures, and it may be appropriate to use various approaches. The fourth step is to specify the method used to identify appropriate values for the input parameters which generate model outputs that are most consistent with the targets. Again, a number of these algorithms exist, and the use of a range of methods may be appropriate. The fifth stage is to define convergence criteria; that is, to establish the acceptable sets of input parameters recognizing that more than one set of input parameters can generate the same goodness-of-fit estimates.

The sixth step is to decide the point when the calibration process is complete (the stopping rule). For example, a simple calibration objective would be where one parameter set is identified which generates model outputs which are close (e.g. within 95% CI) of the calibration target values. Finally, the results of the calibration exercise need to be incorporated into the economic evaluation. Point estimates of the chosen parameter values could be used, but a full uncertainty analysis would need to reflect the uncertainty in (and correlation between) the parameter values which should ideally be produced as part of the calibration exercise. An example of the use of calibration is a CEA of screening for age-related macular degeneration (Karnon et al. 2009). The calibration targets were age- and state-specific clinical diagnosis rates of age-related maculopathy (ARM), visual acuity by ARM state when a patient is identified, and age-specific rates of bilateral 6/60 vision or worse as a result of ARM. The parameters estimated through the process were the rates of clinical diagnosis rates of ARM.

10.5.5 Elicitation of parameter estimates

A key advantage of using a decision-analytic model as the vehicle for economic evaluation is that it provides a framework for informing decisions in the context of evidential uncertainty. This includes helping decision-makers determine the best course of action regarding the need for further research (a topic covered in more detail in Chapter 11).

In some situations there may be no data with which to estimate a particular parameter for an analysis. Getting some initial estimate, together with a realistic view of its uncertainty, is necessary before it is possible to assess the importance of the parameter for the decision and the value of collecting data. In such situations, the use of relevant experts to provide their judgement regarding the magnitude of a given parameter and its uncertainty may be valuable. It is important to emphasize that formal methods for expert elicitation exist, and the process is much more than simply asking for 'best guesses'. A large literature exists on expert elicitation in Bayesian statistics (e.g. O'Hagan et al. 2006) but formal elicitation has been used relatively rarely in economic evaluation in health care.

Bojke and Soares describe a series of steps relating to the use of these methods in the context of developing a decision-analytic model (Bojke and Soares 2014). The first is to establish from whom relevant judgements should be elicited. These would be expected to be relevant experts, generally clinical professionals. It would also be expected that a sample of such experts would be used rather than relying on a single person, although formal guidance on how many is limited.

The second step concerns what quantity to elicit. In principle, this could be any of the groups of parameters discussed in this chapter. In practice these are often probabilities which can be elicited in several ways; for example, asking experts to estimate the probability related to a specific event, the time period expected for a given proportion of patients to experience the event, or the proportion of patients who would have experienced the event over a specific time period.

The third step involves applying specific methods for elicitation. A key aspect is the need to reflect each individual's uncertainty over the quantity concerned, as well as uncertainty across experts. Asking a direct question about the variance of a quantity is generally not advisable, and researchers have used methods including eliciting the credible interval directly (i.e. asking for the range of values considered relevant for a specified level of credibility such as 95%) and the use of a probability grid which shows a histogram of values on which the expert marks a series of crosses to show the likelihood of each value.

The final step synthesizes the elicited quantities across experts. The Delphi method, which encourages experts to reach a consensus about the value of a particular parameter, has been used quite extensively in for this purpose in economic evaluation. An example was a study by Yang and colleagues which used the method to estimate the resource use associated with alternative treatments for schizophrenia (Yang et al. 2009). However, consensus techniques have a number of limitations, the most important of which is that they fail to reflect the uncertainty in parameter estimation across experts. Various mathematical approaches exist, and these generally involve using synthesis methods such as linear pooling and fitting probability distributions to the pooled elicited values. A number of additional factors need to be considered in using formal elicitation methods, including the need to avoid the various biases that can emerge during the process, which can either be associated with the experts (e.g. motivational bias) or with the methods themselves (e.g. relating to the framing of questions). Further details regarding the methods of formal elicitation applied to decision models for economic evaluation can be found elsewhere (Bojke et al. 2010; Bojke and Soares 2014).

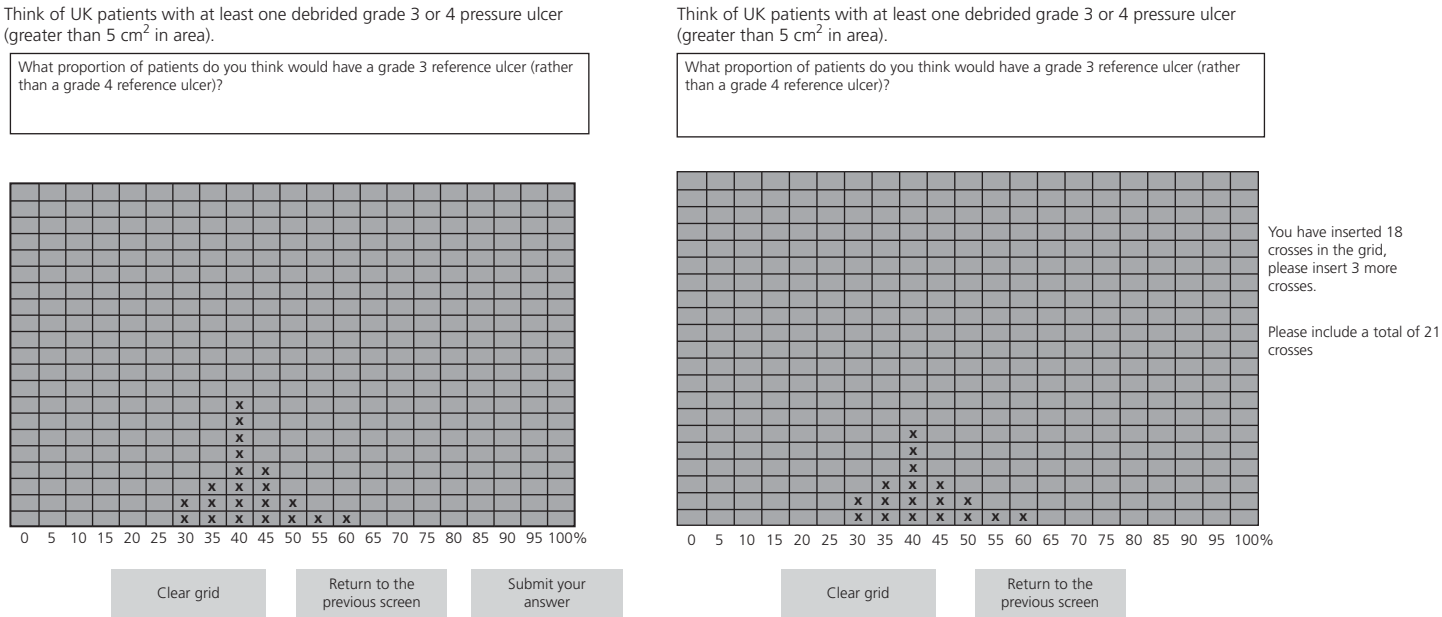


Fig. 10.6 Use of the histogram method to elicit experts' beliefs about the values associated with probabilities relating to severe pressure ulceration.

Reproduced with permission from Soares, M.O., et al, Methods to elicit experts' beliefs over uncertain quantities: application to a cost-effectiveness transition model of negative pressure wound therapy for severe pressure ulceration, *Statistics in Medicine*, Volume 30, Issue 19, pp. 2363–80, Copyright © 2011 John Wiley & Sons, Ltd.

An example of a formal elicitation exercise used the method to estimate a series of transition probabilities (and their uncertainty) for a Markov model to assess the cost-effectiveness of negative pressure wound therapy for severe pressure ulceration (Soares et al. 2011). The study used a sample of 23 nurses with relevant clinical experience. They used the probability grid (histogram) method to elicit a number of probability estimates which were presented as discrete numbers rather than ranges (Figure 10.6). Experts were asked to distribute their 21 allocated crosses by placing more crosses in values they believed the more likely. This effectively means each expert defined a probability distribution over the quantity of interest, reflecting uncertainty in their beliefs. To synthesize the distributions across experts various methods were piloted; linear pooling was adopted as the use of more complex methods had no clear advantages.

10.6 **Conclusions**

Identifying appropriate evidence to populate an analysis is an essential element of any economic evaluation. Defining the process and methods for this purpose cannot be codified in any exact way. In part this is because the time and other resources available to conduct an economic evaluation will vary in different contexts. There is also an inevitable need to make judgements about the evidence to employ, and these can be informed by some key principles. The first is the need to employ evidence that is relevant to the decision context, where this is defined in terms of the patient population(s), comparators, and jurisdiction.

The second principle is the need to identify relevant evidence is an unbiased way. It is rarely possible to guarantee that *all* relevant evidence has been identified, given the challenge of obtaining unpublished research results or those that are difficult to access (e.g. in 'grey' literature). The important thing is to identify evidence which is representative of the entirety and not a biased selection. Central to this is the need to describe the methods used to identify evidence, and this should be done in such a way as to allow others to replicate them.

A third, and related, principle is the need to justify the decisions that are taken on what evidence to use in an analysis. For example, it may be the case that five estimates of a HRQoL weight have been identified in the literature, but there is one which is judged to be the most suitable in terms of the specified decision problem and is, therefore, used in the model. This is entirely acceptable as long as the reasoning for that judgement is explained and the implications for study results are explored. The fourth principle follows and relates to uncertainty analysis. The chance of identifying perfect evidence for an economic evaluation is vanishingly small. Uncertainty in evidence relates to the imprecision of parameter estimates due to finite sample sizes and the extent to which evidence is relevant to the context of the decision (e.g. population(s), jurisdiction, current practice). Careful use of uncertainty analysis is an essential aspect of all economic evaluation and is considered in detail in the next chapter.

10.7 **Exercise**

Read the following paper which is a systematic review of the clinical effects of bariatric surgery in obese individuals:

Buchwald H, Avidor Y, Braunwald E, et al. (2004). Bariatric surgery: a systematic review and meta-analysis. *JAMA*, **292**, 1724–37.

Consider the types of evidence that would be needed to undertake an economic evaluation of bariatric surgery. Assess whether the systematic review provides all the evidence on clinical effects that would be needed for economic analysis; how would any deficiencies be addressed? What other types of evidence would be needed in addition to clinical effects, how could these be identified or generated? You could undertake this exercise in conjunction with the material in Chapter 9 and consider the type of modelling approach you would take for the economic evaluation.

References

Ades, A.E. and Higgins, J.P.T. (2005). The interpretation of random-effects meta-analysis in decision models. *Medical Decision Making*, **25**, 646–54.

Anderson, R. and Shemilt, I. (2010). The role of economic perspectives and evidence in systematic review, in I. Shemilt, M. Mugford, L. Vale, K. Marsh, and C. Donaldson (ed.), *Evidence-based decisions and economics: health care, social welfare, education and criminal justice*, 2nd edition. Chichester: Wiley-Blackwell.

Asseburg, C., Bravo-Vergal, Y., Palmer, S., et al. (2007). Assessing the effectiveness of primary compared with thrombolysis and its relationship to time delay: a Bayesian evidence synthesis. *Heart*, **93**, 1244–50.

Basu, A. and Manca, A. (2012). Regression estimators for generic health-related quality of life and quality-adjusted life years. *Medical Decision Making*, **32**, 56–69.

Bell, C.M., Chapman, R.H., Stone, P.W., et al. (2001). An off-the-shelf help list: A comprehensive catalog of preference scores from published cost-utility analyses. *Medical Decision Making*, **21**, 288–94.

Bill and Melinda Gates Foundation, NICE International, the Health Intervention and Technology Assessment Program, et al. (2014). *Bill and Melinda Gates Foundation Methods for Economic Evaluation Project (MEEP)*. London: NICE International.

Bojke, L., Claxton, K., Sculpher, M., et al. (2010). Eliciting distributions to populate decision-analytic models. *Value in Health*, **13**, 557–64.

Bojke, L. and Soares, M. (2014). Decision analysis: eliciting experts' beliefs to characterize uncertainties, in A.J. Culyer (ed.), *Encyclopedia of health economics*. Amsterdam: Elsevier.

Borenstein, M., Hedges, L.V., Higgins, J.P.T., et al. (2009). *Introduction to meta-analysis*. Chichester: Wiley.

Bravo Vergel, Y., Palmer, S., Asseburg, C., et al. (2007). Results of a comprehensive decision analysis. Is primary angioplasty cost-effective in the UK? *Heart*, **93**, 1238–43.

Briggs, A., Claxton, K., and Sculpher, M. (2006a). *Decision modelling for health economic evaluation*. Oxford: Oxford University Press.

Briggs, A.H., Bousquet, J., Wallace, M.V., et al. (2006b). Cost-effectiveness of asthma control: an economic appraisal of the GOAL study. *Allergy*, **61**, 531.

Briggs, A., Mihaylova, B., Sculpher, M.J., et al. (2007). The cost-effectiveness of perindopril in reducing cardiovascular events in patients with stable coronary artery disease using data from the EUROPA study. *Heart*, **93**, 1081–6.

Bucher, H.C., Guyatt, G.H., Griffith, L.E., et al. (1997). The results of direct and indirect treatment comparisons in meta-analysis of randomized controlled trials. *Journal of Clinical Epidemiology*, **50**, 683–91.

Centre for Reviews and Dissemination (2009). *Systematic reviews: CRD's guidance for undertaking reviews in health care*, 2nd edition. York: Centre for Reviews and Dissemination, University of York.

Chambers, D., Epstein, D., Walker, S., et al. (2009). Endovascular stents for abdominal aortic aneurysms: a systematic review and economic model. *Health Technology Assessment*, **13**(48), 1–234.

Clarke, P.M., Gray, A.M., Briggs, A., et al. (2005). Cost-utility analyses of intensive blood glucose and tight blood pressure control in type 2 diabetes (UKPDS 72). *Diabetologia*, **48**, 868–77.

Clarke, P., Gray, A., and Holman, R. (2002). Estimating utility values for health states of type 2 diabetic patients using the EQ-5D (UKPDS 62). *Medical Decision Making*, **22**, 340–9.

Cochrane Collaboration (2011). *Cochrane handbook for systematic reviews of interventions, version 5.1.0 [updated March 2011]*, ed. J.P.T. Higgins and S. Green. <www.cochrane-handbook.org>.

Collett, D. (1994). *Modelling survival data in medical research*. London: Chapman & Hall/CRC.

Connock, M., Hyde, C., and Moore, D. (2011). Cautions regarding the fitting and interpretation of survival curves. Examples from NICE Single Technology Appraisals of drugs for cancer. *PharmacoEconomics*, **29**, 827–37.

Cooper, N., Coyle, D., Abrams, K., et al. (2005). Use of evidence in decision models: an appraisal of health technology assessments in the UK since 1997. *Journal of Health Services and Research Policy*, **10**, 245–50.

Cooper, N.J., Sutton, A.J., Morris, D., et al. (2009). Addressing between-study heterogeneity and inconsistency in mixed treatment comparisons: Application to stroke prevention treatments in individuals with non-rheumatic atrial fibrillation. *Statistics in Medicine*, **28**, 1861–81.

Cummins, E., Asseburg, C., Suresh Punekar, Y., et al. (2011). Cost-effectiveness of infliximab for the treatment of active and progressive psoriatic arthritis. *Value in Health*, 14, 15–23.

Department of Health Payment by Results Team (2013). *Reference costs guidance for 2013* <https://www.gov.uk/government/uploads/system/uploads/attachment_data/file/217040/2012–3-reference-costs-guidance.pdf> Leeds: Department of Health.

Dias, S., Welton, N.J., Sutton, A.J., et al. (2013). Evidence synthesis for decision-making 4: inconsistency in networks of evidence based on randomized controlled trials. *Medical Decision Making*, **33**, 641–56.

Espinoza, M.A., Manca, M., Claxton, K., et al. (2014). The value of heterogeneity for cost-effectiveness subgroup analysis: conceptual framework and application. *Medical Decision Making*, **34**, 951–64.

Fleurence, R. and Hollenbeak, C.S. (2007). Rates and probabilities in economic modelling. Transformation, translation and appropriate application. *PharmacoEconomics*, **25**, 3–6.

Gold, M.R., Siegel, J.E., Russell, L.B., et al. (1996). *Cost-effectiveness in health and medicine*. New York: Oxford University Press.

Guyatt, G.H., Haynes, R.B., Jaeschke, R.Z., et al. (2000). Users' Guides to the Medical Literature. XXV. Evidence-based medicine: principles for applying the users' guides to patient care. *JAMA*, **284**, 1290–6.

Henriksson, M., Epstein, D.M., Palmer, S.J., et al. (2008). The cost-effectiveness of an early interventional strategy in non-ST-elevation acute coronary syndrome based on the RITA 3 trial. *Heart*, **94**, 717–23.

Hoaglin, D.C., Hawkins, N., Jansen, J.P., et al. (2011). Conducting indirect-treatment-comparison and network-meta-analysis studies: report of the ISPOR Task Force on Indirect Treatment Comparisons Good Research Practices: Part 2. *Value in Health*, **14**, 429–37.

Hunink, M., Weinstein, M.C., Wittenberg, E., et al. (2014). *Decision making in health and medicine: integrating evidence and values*, 2nd edition. Cambridge: Cambridge University Press.

Institute of Medicine (2011). *Finding what works in health care: standards for systematic reviews*. Washington, DC: Institute of Medicine.

Jeffcoat, M.K., Cizza, G., Shih, W.J., Gneco, R., and Lombardi, A. (2007). Efficacy of bisphosphonates for the control of alveolar bone loss in periodontitis. *Journal of the International Academy of Periodontology*, **9**(3), 70–6.

Joint Formulary Committee (2013). *British National Formulary* No. 64. London: BMJ Group and the Pharmaceutical Press.

Juni, P., Witschi, A., Bloch, R., et al. (1999). The hazards of scoring the quality of clinical trials for meta-analysis. *JAMA*, **282**, 1054–60.

Karnon, J., Czoski-Murray, C., Smith, K.J., et al. (2009). A hybrid cohort individual sampling natural history model of age-related macular degeneration: assessing the cost-effectiveness of screening using probabilistic calibration. *Medical Decision Making*, **29**, 304–16.

Laking, G., Lord, J., and Fischer, A. (2006). The economics of diagnosis. *Health Economics*, **15**, 1109–20.

Latimer, N.R. (2013a). Survival analysis for economic evaluations alongside clinical trials-extrapolation with patient-level data: inconsistencies, limitations, and a practical guide. *Medical Decision Making*, **33**, 743–54.

Latimer, N.R. (2013b). *Survival analysis for economic evaluations alongside clinical trials: extrapolation with patient-level data*. NICE Decision Support Unit Technical Support Document 14. <http://www.nicedsu.org.uk/NICE%20DSU%20TSD%20Survival%20analysis.updated%20March%202013.pdf> Sheffield: NICE Decision Support Unit.

Lipsey, M.W. and Wilson, D.S.B. (2001). *Practical meta-analysis*. Thousand Oaks, CA: Sage.

Manca, A., Asseburg, C., Bravo Vergal, Y., et al. (2012). Cost-effectiveness of different chemotherapy strategies for patients with poor prognosis advanced colorectal cancer (MRC FOCUS). *Value in Health*, **15**, 22–31.

McKenna, C., Hawkins, N., Claxton, K., et al. (2010). Cost-effectiveness of enhanced external counterpulsation (EECP) for the treatment of stable angina in the United Kingdom. *International Journal of Technology Assessment in Health Care*, **26**, 175–82.

Miller, D.K. and Homan, S.M. (1994). Determining transition probabilities: confusion and suggestions. *Medical Decision Making*, **14**, 52–8.

Newcomb, P.A., Trentham-Dietz, A., and Hampton, J.M. (2010). Bisphosphonates for osteoporosis treatment are associated with reduced breast cancer risk. *British Journal of Cancer*, **102**(5), 799–802.

Newman, T.B. and Kohn, M.A. (2009). *Evidence-based diagnosis* (Practical Guides to Biostatistics and Epidemiology). New York: Cambridge University Press.

NICE [National Institute for Health and Clinical Excellence] (2013). *Updated guide to the methods of technology appraisal*. London: NICE.

O'Hagan, A., Buck, C.E., Daneshkhah, A., et al. (2006). *Uncertain judgements: eliciting experts' probabilities*. Chichester: Wiley.

Paisley, S. (2014). Searching and reviewing non-clinical evidence for economic evaluation, in A.J. Culyer (ed.), *Encyclopedia of health economics*. Amsterdam: Elsevier.

Papaioannou, D., Brazier, J., and Paisley, S. (2010). *The identification, review and synthesis of health state utility values from the literature*. NICE DSU Technical Support Document 9. Sheffield: NICE Decision Support Unit.

Peasgood, T., Ward, S.E., and Brazier, J. (2010). Health-state utility values in breast cancer. *Expert Reviews in Pharmacoeconomics and Outcomes Research*, **10**, 553–66.

Personal Social Services Research Unit (2013). *Unit costs of health and social care 2013*. Canterbury: PSSRU, University of Kent.

Pettiti, D.B. (1999). *Meta-analysis, decision analysis, and cost-effectiveness analysis*. New York: Oxford University Press.

Phelps, C.E. and Mushlin, A.I. (1988). Focusing technology assessment using medical decision theory. *Medical Decision Making*, **8**, 279–89.

Prevost, T.C., Abrams, K.R., and Jones, D.R. (2000). Hierarchical models in generalised synthesis of evidence: an example based on studies of breast cancer screening. *Statistics in Medicine*, **19**, 3359–76.

Ross, S. (2013). *A first course in probability*, 9th edition. New York: Pearson.

Schulz, K.F., Altman, D.G., Moher, D., et al. (2010). CONSORT 2010 Statement: updated guidelines for reporting parallel group randomised trials. *BMJ*, **340**, 698–702.

Sculpher, M.J. (2008). Subgroups and heterogeneity in cost-effectiveness analysis. *PharmacoEconomics*, **26**, 799–806.

Sculpher, M.J., Pang, F.S., Manca, A., et al. (2004). Generalisability in economic evaluation studies in health care: a review and case studies. *Health Technology Assessment*, **8**(49), 1–213.

Soares, M.O., Bojke, L., Dumville, J.C., et al. (2011). Methods to elicit experts' beliefs over uncertain quantities: application to a cost-effectiveness transition model of negative pressure wound therapy for severe pressure ulceration. *Statistics in Medicine*, **30**, 2363–80.

Sullivan, P.W. and Ghushchyan, V. (2006). Preference-based EQ-5D index scores for chronic conditions in the United States. *Medical Decision Making*, **26**, 410–20.

Sullivan, P.W., Slejko, J.F., Sculpher, M.J., et al. (2011). Catalogue of EQ-5D scores for the United Kingdom. *Medical Decision Making*, **31**, 800–4.

Sutton, A.J., Abrams, K.R., Jones, D.R., et al. (2000). *Methods for meta-analysis in medical research*. Chichester: Wiley.

Sutton, A., Ades, A.E., Cooper, N., et al. (2008). Use of indirect and mixed treatment comparisons for technology assessment. *PharmacoEconomics*, **26**, 753–67.

Tengs, T. and Wallace, A. (2000). One thousand health-related quality of life estimates. *Medical Care*, **38**, 583–637.

Thompson, S.G. and Higgins, J.P.T. (2002). How should meta-regression analyses be undertaken and interpreted? *Statistics in Medicine*, **21**, 1559–73.

Tosh, J.C., Longworth, L.J., and George, E. (2011). Utility values in National Institute for Health and Clinical Excellence (NICE) technology appraisals. *Value in Health*, **14**, 102–9.

Turner, R.M., Spiegelhalter, D.J., Smith, G.C.S., et al. (2009). Bias modelling in evidence synthesis. *Journal of the Royal Statistical Society Series A*, **172**(Part 1), 21–47.

Vanni, T., Karnon, J., Madan, J., et al. (2011). Calibrating models in economic evaluation: a seven-step approach. *PharmacoEconomics*, **29**, 35–49.

Vestergaard, P., Schwartz, K., Pinholt, E.M., Rejnmark, L., and Mosekilde, L. (2010). Risk of atrial fibrillation associated with use of bisphosphonates and other drugs against osteoporosis: a cohort study. *Calcified Tissue International*, **865**(5), 335–42.

Welton, N., Sutton, A.J., Cooper, N.J., et al. (2012). *Evidence synthesis for decision-making in healthcare*. Chichester: Wiley.

Whiting, P., Rutjes, A.W.A., Reitsma, J.B., et al. (2003). The development of QUADAS: a tool for the quality assessment of studies of diagnostic accuracy included in systematic reviews. *BMC Medical Research Methodology*, **3**, 25.

Yang, L., Li, M., Tao, L., et al. (2009). Cost-effectiveness of long-acting risperidone injection versus alternative atypical antipsychotic agents in patients with schizophrenia in China. *Value in Health*, **12**(suppl 3), S66–S69.

Chapter 11

Characterizing, reporting, and interpreting uncertainty

11.1 Some basics

The general issue of whether existing evidence about the cost-effectiveness of a health care intervention is *sufficient* to justify its approval for widespread use can be seen as central to a number of policy questions in many different types of health care systems. For example, decisions about the approval or reimbursement of new drugs are increasingly being made close to their launch. At this point, the evidence base to support their use is least mature and there may be substantial uncertainty surrounding their cost-effectiveness. In these circumstances, further evidence may be particularly valuable as it will lead to better decisions about the use of the drug which, in turn, will improve patient outcomes and/or reduce resource costs.

It is important, therefore, to establish the key principles of how uncertainty in estimates of cost-effectiveness can be quantified and represented. It is also necessary to consider which assessments are needed to decide whether there is sufficient evidence to support recommending the use of an intervention, or what type of additional research might be required and how it might be designed. Decision-makers also need to consider whether the intervention should be recommended but additional evidence sought, or whether its widespread use should be restricted until the additional evidence is available. Such assessments of uncertainty and the value of additional evidence can help to inform the questions posed by coverage with evidence development and managed entry in many health care systems including restricting approval to 'only in research' (Briggs et al. 2010; Claxton et al. 2012; Garrison et al. 2013; Mohr and Tunis 2010; Niezen et al. 2007; Walker et al. 2012).

11.1.1 Uncertainty, variability, and heterogeneity

It is important to make a clear distinction between uncertainty, variability, and heterogeneity at the outset (Briggs et al. 2012; Claxton 2008). *Uncertainty* refers to the fact that we do not know what the expected costs and effects of using an intervention will be in a particular population of patients (i.e. the net benefits of an intervention on average). This remains the case even if all patients within this population have the same characteristics as far as we can observe them. Additional evidence can reduce uncertainty and provide a more precise estimate of the expected costs and effects in the whole population; or within subpopulations that might be defined based on different observed characteristics.

Variability was a term we first encountered in Chapter 9. It refers to the fact that individual responses to an intervention (including costs as well as health effects) will differ within the population, or even in a subpopulation, of patients with the same observed characteristics. This natural variation in response cannot be reduced by acquiring additional evidence about the expected or average costs and effects.

Heterogeneity has been discussed in a number of earlier chapters, and refers to those individual differences in response that can be associated with differences in observed characteristics. In other words, this is the situation where some of the sources of natural variability can be identified and understood. As more becomes known about the sources of variability (i.e. as variability is turned into heterogeneity), the patient population can be partitioned into subpopulations or subgroups. Each of these has a different estimate of the expected effect of the intervention and the uncertainty associated with it. As more sources of variability become known, finer stratification of subpopulations becomes possible which ultimately become individual patients—that is, individualized care (Basu 2011; Basu and Meltzer 2007; Espinoza et al. 2014).

11.1.2 Why does uncertainty matter?

When decisions are made about health care interventions, information about cost-effectiveness is critical (see Chapter 4). However, any assessment of the additional health benefits and additional cost offered by an intervention is uncertain. Therefore, any decision based on expected cost-effectiveness will also be uncertain. This uncertainty arises from a number of sources. First, there is uncertainty in the estimates of inputs or parameters of the type of decision models commonly used to estimate costs and effects (see Chapter 9). There is also a wide range of other sources of uncertainty, which might include the potential bias or relevance of evidence, and the assumptions required when extrapolating effects and costs over time. In the face of this uncertainty (only some of which is generally explicitly characterized and presented in an analysis), a decision-maker must come to a view about whether approval of an intervention is expected to be cost-effective.

If the purpose of the health care system is to improve overall health outcomes using available resources, it is the health benefits and additional costs that are *expected* to occur (i.e. on average) that are of primary concern. For this reason there are strong arguments for basing decisions about use of an intervention on the expected incremental effects and costs rather than applying hypothesis testing and traditional rules of statistical significance to estimates of cost-effectiveness (see Section 11.4). However, making decisions based only on expected effects and costs should not imply that uncertainty is unimportant. Indeed, an assessment of the implications of uncertainty surrounding a decision is an essential part of any decision-making process that is concerned to improve health outcomes given the resource constraints that are faced. Characterizing the uncertainty surrounding estimates of effects and costs is required:

- To provide correct estimates of expected effect and cost. When effects and costs are evaluated using a decision model in which there is a non-linear relationship between inputs and outputs (e.g. a Markov model), the correct calculation of expected effects and costs requires the uncertainty associated with all the inputs to be fully expressed (often using probabilistic sensitivity analysis as described in Section 11.2). Therefore, characterizing uncertainty matters even if a decision-maker only wishes to consider expected cost-effectiveness in their decisions (see Figure 11.1 and Box 11.1).
- To assess the potential value of additional evidence (see Section 11.3).
- To inform the types of evidence that might be needed and how further research might be designed (see Section 11.3).
- To consider the implications that the need for additional evidence might have for a decision to approve a technology that is expected to be cost-effective. That is, to assess whether access to a technology should be restricted until the additional evidence required becomes available (see Section 11.4).

Box 11.1 Uncertainty and expected cost-effectiveness

The lower pane of Figure 11.1 illustrates an input or parameter of a model (θ) that has a non-linear relationship with net benefit (NB). If only a single point estimate of this parameter is used (θ_1), based on its mean or expected value, it will return an estimate of net benefit of NB_1, i.e. the result from a deterministic analysis. However, the value of θ is uncertain and has a probability distribution (illustrated in the upper pane of the diagram), so it is possible that θ will take a value of θ_2, generating NB_2, or a value of θ_3, generating NB_3. Although both these values are equally likely (they could represent the 95% confidence interval for θ), they generate very different values of net benefit. Therefore, when net benefit is evaluated over the possible values that θ might take, and the expected NB averaged over these possible values ($E_\theta(NB,\theta)$) will not be equal to NB_1. Therefore, a deterministic analysis that only uses the point estimates or mean values of parameters will generate a biased estimate of cost-effectiveness. In this example it would overestimate cost-effectiveness and NB.

The only exceptions are models where there is a linear relationship between the inputs (θ) and outputs (NB). This includes multilinear models such as decision trees, but only when the inputs are uncorrelated. All Markov models generate non-linear relationships between inputs and outputs, so a deterministic analysis of these models will provide biased estimates of cost-effectiveness. In some types of more complex models (not Markov models) there is also a non-linear relationship between the variability in parameter values and net benefit. In these circumstances, simulation of both variability and uncertainty would be required to generate unbiased estimates of net benefit (Griffin et al. 2006).

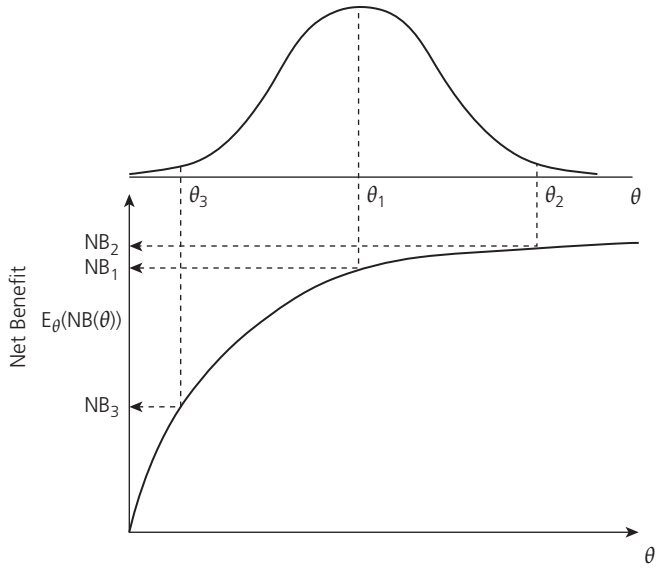


Fig. 11.1 Uncertainty and expected cost-effectiveness.

11.2 **Characterizing uncertainty**

Uncertainty arises from a number of sources, and the terminology used in the literature can be confusing. Different terms are used to refer to the same concept and similar terms are used to refer to very different concepts. A task force set up by the International Society of Pharmacoeconomics and Outcomes Research (ISPOR) and the Society of Medical Decision Making (SMDM) (Briggs et al. 2012) has usefully classified the terms that have been used to refer to four key concepts that are described in Table 11.1. We have already distinguished variability (sometimes called stochastic or first-order uncertainty) from heterogeneity which is discussed further in Section 11.5. Our primary concern here is uncertainty in the expected costs and effects in a specific population (sometimes called second-order uncertainty). There are two broad sources of uncertainty:

- *Parameter uncertainty*—uncertainty in the estimates of the inputs or parameters of the type of decision models discussed in Chapter 9. How to characterize and present this source of uncertainty is discussed in Sections 11.2.1, 11.2.2, and 11.2.3.
- *Structural uncertainty*—the different types of scientific judgements that have to be made when constructing a model of any sort (a decision model or a statistical model). How these sources of uncertainty might be quantified and presented is discussed in Section 11.2.4.

Table 11.1 Uncertainty: concepts and terminology

Preferred term	Concept	Other terms sometimes used	Analogous concept in regression
Stochastic uncertainty	Random variability in outcomes between identical patients	Variability Monte Carlo error First-order uncertainty	Error term
Parameter uncertainty	The uncertainty in estimation of the parameter of interest	Second-order uncertainty	Standard error of the estimate
Heterogeneity	The variability between patients that can be attributed to characteristics of those patients	Variability Observed or explained heterogeneity	Beta coefficients (or the extent to which the dependent variable varies by patient characteristics)
Structural uncertainty	The assumptions inherent in the decision model	Model uncertainty	The form of the regression model (e.g. linear, log-linear)

From Briggs, A.H. et al., *Medical Decision Making,* Model parameter estimation and uncertainty: A report of the ISPOR-SMDM Modelling Good Research Practices Task Force Working Group-6, Volume 32, Issue 5, pp. 722–32, Copyright © 2012. Reprinted by Permission of SAGE Publications.

11.2.1 Deterministic sensitivity analysis

Prior to the wider application of probabilistic sensitivity analysis (see Section 11.2.2), it was common to explore uncertainty in the cost-effectiveness of interventions through a series of one-way or multiway sensitivity analyses (Briggs and Sculpher 1995).

11.2.1.1 One-way sensitivity analysis

In this type of deterministic analysis, single values for each of the input parameters are used to estimate the expected cost, effect, and NB based on their mean values. The parameter values are varied and the effect on model outputs reported. This form of analysis is very simple to understand and easy to implement. It provides a quick way to understand the quantitative relationship between changes in inputs and outputs. One-way sensitivity analysis has its greatest value in developing and reviewing a model. This is because it enables the implications of the structure of the model to be explored, understood, and checked by setting parameters at values which should produce a clearly predicable result (e.g. where costs and effects should be the same or higher/lower for one alternative compared to another).

This type of deterministic analysis has also been used to represent uncertainty by varying parameter values by some specified amount (e.g. plus or minus a proportionate change in the mean value of each parameter) and reporting the impact on cost-effectiveness. Although this indicates how sensitive the model outputs might be to changes in particular inputs, it cannot indicate how uncertain a decision might be. Nor can it indicate which parameters contribute most to this *decision uncertainty,* so this

type of one-way sensitivity analysis is not recommended as a way to represent uncertainty. There are two reasons for this:

- It is changes in the decision rather than model outputs themselves that are important. Sometimes large changes in a model's output will not change which alternative is cost-effective, but in other circumstances small changes in outputs will impact a decision.
- Unless the changes in parameter values are related to the uncertainty in how they were estimated, it is not clear whether the changes in costs and effects are likely or very unlikely. For example, a parameter with low sensitivity (i.e. a large proportionate change in its value leads to a small proportionate change in costs and effects) might be estimated with greater uncertainty. In which case it could have a greater impact on which alternative is cost-effective compared to a parameter with higher sensitivity that is estimated with greater precision and is less uncertain.

An alternative deterministic sensitivity analysis is where each parameter in turn is set at 'extreme but plausible' upper and lower bounds, and the difference in cost and effect over this range is recorded. The parameter can be described as sensitive if the decision about which alternative is cost-effective changes in response to plausible changes in its values. The problem is identifying lower and upper bounds for each parameter that can be regarded as 'extreme but plausible'. If the range used is arbitrarily chosen or based on some implicit assumptions, the results may be misleading. Upper and lower bounds can be justified based on the evidence used to estimate the parameter and the distribution that might be assigned to represent the uncertainty associated with its estimation (e.g. representing a number of standard errors or the 95% confidence interval).

More usefully, this type of analysis can be used to identify the 'threshold' values that parameters would need to take for cost and effects to change sufficiently to alter the decision about which alternative offered the highest expected net benefits (see Table 11.2 and Box 11.2). To identify whether a decision is uncertain and which parameters contribute most to this decision uncertainty, a decision-maker would need to judge whether the threshold values for the parameter are likely or extremely unlikely to occur. It is possible to report the probability that parameters would take values more extreme than their threshold values based on how they were estimated, reflecting the amount and quality of exiting evidence (see Table 11.3 and Box 11.2).

Unless this type of probability information is presented (which requires consideration of appropriate distributions for the estimated parameters), it will not be possible for decision-makers to assess the likelihood of a parameter taking values greater than the threshold in a way that is consistent with the evidence base. Therefore, conducting a deterministic one-way sensitivity analysis that is consistent with the evidence base and can be interpreted appropriately requires very similar information about the appropriate distributions as is required in probabilistic sensitivity analysis (see Section 11.2.2).

Unfortunately, the combined effect of uncertainty in the value of all the parameters cannot be represented or interpreted in a one-way analysis. For example, if a series of one-way sensitivity analyses is conducted, the decision may not appear sensitive to any plausible values of any of the parameters individually. Uncertainty in

Box 11.2 Threshold analysis for parameters

A decision model of the cost-effectiveness of alternative durations of treatment (12, 6, 3 or 1 month) with clopidogrel compared to standard National Health Service (NHS) care was used as a case study in exploring when the National Institute of Health and Care Excellence (NICE) in the United Kingdom should issue recommendations for use only in the context of research (see Claxton et al. 2012 for more details). The results indicate that 12 months of treatment with clopidogrel (Clop12) is expected to be cost-effective and offers the highest expected net benefits, using a cost-effectiveness threshold of £20 000 per quality-adjusted life-year (QALY). The question is whether a decision to approve 12 months of treatment with clopidogrel is uncertain.

Although a series of one-way sensitivity analyses can indicate the effect of parameter values on the costs, effects and expected net benefit of each of the five alternatives, they do not directly help the assessment of what values they must take to change the decision to approve Clop12 and how likely such values might be. A simple summary of the values particular parameters must take to make each of the alternatives cost-effective is more useful. These threshold values for parameters are reported in Table 11.2 and provide some useful information. For example, there are only six parameters (10, 12, 14, 17, 18 and 19) which could possibly take values that would make NHS care without clopidogrel cost-effective.

A judgement about how likely it is that parameters might take values within the threshold ranges in Table 11.2 is also required. The probability that each parameter could take values which would lead to each of the five alternatives being cost-effective are reported in Table 11.3. These probabilities are based on probabilistic sensitivity analysis (see Section 11.2.2) and depend on the cost-effectiveness threshold. Interestingly, the results suggest that many of the parameter do not (alone) contribute to the uncertainty associated to approving Clop12 (e.g. parameters 1–6, 9, 11, 15–16, 20, 22, and 25–27). It is the estimate of relative effect on mortality (parameter 17) that contributes most to the uncertainty associated with Clop12. There is a probability of 0.55 that this parameter will take values where Clop12 will have the highest net benefit. It is also the only parameter which (alone) could take values that would make any of the other alternatives cost-effective. However, this does not necessarily mean that other parameters are unimportant because they may jointly contribute to overall decision uncertainty.

each individual parameter separately may be very unlikely to change a decision, so it might be tempting to conclude that the decision is not uncertain. However, all the parameters are simultaneously uncertain, so when they are considered together there may be considerable decision uncertainty. For this reason, even a well-conducted one-way sensitivity analysis will tend to underestimate the uncertainty surrounding the decision. It can only provide a sufficient, but not necessary, condition for a qualitative conclusion that a decision to approve an intervention that is expected to be cost-effective is uncertain.

Table 11.2 Thresholds associated with parameters (duration of treatment with clopidogrel)

		Parameter	Mean value	Clop12	Clop6	Clop3	Clop1	NHS
Natural history	1	P_die_0.1	0.032	0 to 0.10	0.11 to 0.54	0.54 to 0.63	0.63 to 1	–
	2	P_NFMI_0.1	0.04	0 to 0.14	0.14 to 0.71	0.71 to 0.82	0.82 to 1	–
	3	P_die_1.3	0.022	0 to 0.10	0.10 to 0.55	0.55 to 1	–	–
	4	P_NFMI_1.3	0.004	0 to 0.10	0.10 to 0.7	0.7 to 1	–	–
	5	P_die_3.6	0.023	0.01 to 0.10	0.10 to 1	0 to 0.01	–	–
	6	P_NFMI_3.6	0.011	0 to 0.11	0.11 to 1	–	–	–
	7	P_die_6.12	0.024	0.02 to 1	0 to 0.02	–	–	–
	8	P_NFMI_6.12	0.009	0.005 to 1	0 to 0.005	–	–	–
	9	TP_AC	0.018	0 to 0.06	0.06 to 1	–	–	–
	10	**TP_AD**	**0.072**	**0 to 0.08**	**0.08 to 0.10**	**–**	**–**	**0.10 to 1**
	11	TP_CD	0.188	0.12 to 1	0 to 0.12	–	–	–
	12	**TP_BD**	**0.07**	**0.06 to 1**	**0.04 to 0.06**	**–**	**–**	**0 to 0.04**
Utilities	13	U_Well	0.798	0.29 to 1	0 to 0.29	–	–	–
	14	**U_Well1**	**0.93**	**0.90 to 1**	**0.74 to 0.90**	**–**	**–**	**0 to 0.74**
	15	U_NFMI	0.801	0 to 1	–	–	–	–
	16	U_POSTMI	0.931	0 to 1	–	–	–	–
RE	**17**	**RR_death**	**0.931**	**0 to 0.93**	**0.94 to 0.97**	**0.97 to 0.98**	**0.98 to 0.99**	**1.00 to max**
	18	**RR_NFMI**	**0.71**	**0 to 0.82**	**0.83 to 1.55**	**1.56 to 1.83**	**–**	**1.84 to max**
Costs	**19**	**C_Well**	**2061.5**	**0 to 2690**	**2690 to 5611**	**–**	**–**	**5611 to max**
	20	C_MI_LT	6050	0 to max	–	–	–	–
	21	C_PostMI	2309.7	870 to max	0 to 870	–	–	–
	22	TC_Well_Dead	871.5	0 to 20474	20474 to max	–	–	–
	23	C_t1	895.1	0 to 910	910 to max	–	–	–
	24	C_t2	651.6	630 to max	0 to 630	–	–	–
	25	C_t3	524.2	370 to max	–	0 to 370	–	–
	26	C_t4	434.8	150 to max	–	–	0 to 150	–
	27	C_t5	329.8	0 to max	–	–	–	–

Table 11.3 Probabilities associated with threshold parameter values (duration of treatment with clopidogrel)

		Parameter	Clop12	Clop6	Clop3	Clop1	NHS
Natural history	1	P_die_0.1	1	–	–	–	–
	2	P_NFMI_0.1	1	–	–	–	–
	3	P_die_1.3	1	–	–	–	–
	4	P_NFMI_1.3	1	–	–	–	–
	5	P_die_3.6	1	–	–	–	–
	6	P_NFMI_3.6	1	–	–	–	–
	7	P_die_6.12	0.65	0.35	–	–	–
	8	P_NFMI_6.12	0.91	0.09	–	–	–
	9	TP_AC	1	–	–	–	–
	10	TP_AD	0.83	0.17	–	–	–
	11	TP_CD	1	–	–	–	–
	12	TP_BD	0.85	0.15	–	–	–
Utilities	13	U_Well	1	–	–	–	–
	14	U_Well1	0.94	0.06	–	–	–
	15	U_NFMI	1	–	–	–	–
	16	U_POSTMI	1	–	–	–	–
RE	**17**	**RR_death**	**0.55**	**0.18**	**0.01**	**0.10**	**0.16**
	18	RR_NFMI	0.97	0.03	–	–	–
Costs	**19**	**C_Well**	**0.78**	**0.19**	**–**	**–**	**0.03**
	20	C_MI_LT	1	–	–	–	–
	21	C_PostMI	0.89	0.11	–	–	–
	22	TC_Well_Dead	1	–	–	–	–
	23	C_t1	0.95	0.05	–	–	–
	24	C_t2	0.99	0.01	–	–	–
	25	C_t3	1	–	–	–	–
	26	C_t4	1	–	–	–	–
	27	C_t5	1	–	–	–	–

11.2.1.2 Multiway sensitivity analysis

Multiway sensitivity analysis can be conducted and represented as a two-way threshold analysis (a three-way threshold analysis can be represented graphically but only for two alternative treatments; Briggs et al. 2012). With more than two parameters and two alternatives it becomes very difficult to present. It also becomes especially difficult to interpret correctly as it requires a judgement about the joint probability of two parameters taking values greater than their threshold values simultaneously, and this becomes impossible if some parameters are correlated (Claxton et al. 2005).

Best and worst case scenarios, where all the parameters are set at extreme but plausible values that are favourable or unfavourable, can be constructed. However, these are very difficult to interpret correctly even when the values are based on how the parameters were estimated. This is because the probability of all parameters taking extreme but plausible values simultaneously will be very small indeed, so the results may not be plausible possibilities but very unlikely and extreme events. If a decision is sensitive to such extremes it is not clear whether it should be regarded as uncertain or not.

A well-conducted deterministic sensitivity analysis that can be interpreted appropriately requires the same information and judgements about the type of distribution that would appropriately represent uncertainty in parameter values as a probabilistic sensitivity analysis (see Box 11.2 and Section 11.2.2). However, deterministic sensitivity analysis will provide biased estimates in non-linear models (see Box 11.1), and much of the information about the distribution of individual parameters is not presented or used in the analysis. For example, where parameters are based on statistical estimates from a meta-analysis or regression analysis much of the information about the uncertainty in these estimates, and especially any correlation between parameters, is in essence 'thrown away' (Claxton et al. 2005).

11.2.2 Probabilistic sensitivity analysis

Given the limitations of deterministic sensitivity analyses, probabilistic sensitivity analysis (PSA) is increasingly used to characterize parameter uncertainty (Briggs et al. 2006; Claxton 2008). PSA is recommended in several guidelines for cost-effectiveness analysis (Briggs et al. 2012; CADTH 2006; NICE 2013).

The principles of PSA are intuitive and the process of conducting PSA is graphically illustrated in Figure 11.2. Distributions are assigned to each of the model parameters (inputs) reflecting the evidence available to inform the estimates (see Chapter 10). These distributions are then sampled (often using Monte Carlo simulation which samples at random). Each set of samples from all of the inputs generate a single estimate of expected costs and expected effects and, therefore, expected NB can be calculated by the model. These outputs are recorded and then a new set of possible values for the parameters are sampled. This process of sampling inputs and recording outputs is repeated a large number of times (e.g. 10 000). In this way, the range of values the parameters are likely to take is represented in the range of outputs. As well as providing the correct estimate of expected cost, effect, and NB in non-linear models, the output of this process also provides the proportion of times (the probability) that each alternative is cost-effective (see Section 11.2.3). It offers all the information required to assess quantitatively whether current evidence is sufficient or whether additional research is needed (see Section 11.3).

11.2.2.1 Assigning distributions to parameters

PSA represents parameters as distributions of possible mean values instead of single point estimates in a deterministic analysis (this characterization of parameter values as random variables relies on a Bayesian interpretation of probability—see

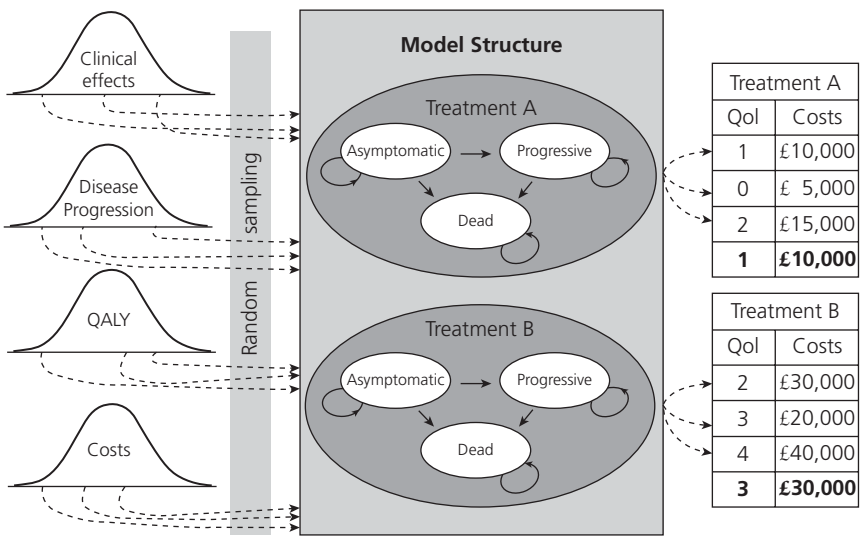


Fig. 11.2 Process of probabilistic sensitivity analysis.

Section 9.1). It may appear that PSA introduces further assumptions about the choice of distribution to represent the uncertainty in the model inputs. However, as we have already seen, interpreting deterministic analysis requires the same, albeit implicit, judgements about distributions and probability to be made. PSA forces the analyst to be explicit, justifying the use of particular distributions on the basis of current evidence and the credibility of any assumptions that might be required. In fact, the choice of distribution is not at all arbitrary if standard statistical methods are followed. The choice should be informed by the nature of the parameter itself, the way the parameter was estimated, and the summary statistics reported, so the statistical uncertainty in its estimation is reflected. As in all statistical analysis, it also requires judgements about the potential bias and relevance of the available evidence (Briggs et al. 2006, 2012).

This means that there will only be a very limited choice of appropriate distributions. Table 11.4 provides a summary of some of the common types of parameter used in decision models, the logical constraints on their values, and candidate distributions, based on the data generating and estimation process. For example, probability parameters are bounded by 0 and 1, so it is inappropriate to specify a distribution that gives a probability to obtaining values outside this range. Similarly it is inappropriate to specify a distribution that gives a probability to obtaining values for costs less than 0. However, it is also important to remember that, with sufficient sample size, we know from central limit theorem that mean values will be normally distributed irrespective of the data generating process (the distribution of the data)—see Section 8.3.1. The problem is that we do not commonly have enough data to safely assume normality.

If parameters are based on results from previously published studies, these commonly report summary statistics (e.g. mean and standard error). When combined with

Table 11.4 Common parameters and candidate distributions

Parameter	Logical constraint	Form of data	Methods of estimation	Candidate distribution
Probability	$0 \leq \theta \leq 1$	Binomial Multinomial Time to event	Proportion Proportions Survival analysis	Beta Dirichlet Lognormal
Relative risk	$\theta > 0$	Binomial	Ratio of proportions	Lognormal
Cost	$\theta \geq 0$	Weighted sum of resource counts	Mean, standard error	Gamma Lognormal
Utility decrement	$\theta \geq 0$	Continuous	Mean, standard error	Gamma Lognormal
All parameters (sufficient data)	Any constraint	Any distribution of data	Mean, standard error	Normal

Adapted from Briggs, A et al., *Decision modelling for health economic evaluation,* p.108, Oxford University Press, Oxford, UK, Copyright © 2006, by permission of Oxford University Press.

information on how they were estimated, this should provide sufficient information to characterize a distribution for possible mean values. For example, where a probability is estimated from a proportion, the beta distribution is the natural choice. However, if the probability parameter is estimated from a logistic regression, then the parameters of interest are the coefficients on the log-odds scale and multivariate normality on this scale would be appropriate. For probabilities estimated from time-to-event data, the parameters would be the coefficients from a survival analysis estimated on the log hazard scale; again, the appropriate assumption would be multivariate normality on this scale.

Some sources of secondary data provide only partial information to inform choice of an appropriate distribution, in which case some assumptions may be necessary. If alternative assumptions might be credible, this provides another source of structural uncertainty (see Section 11.2.4). If the evidence to inform a parameter consists solely of individual patient-level data, then the analyst can use bootstrapping as a non-parametric alternative for describing the distribution of possible mean values (see Section 8.3.1).

11.2.2.2 Correlation

Model parameters may be correlated (related to each other in some way) and, although early examples of PSA often assumed independence between parameters, this is not necessary. For example, where a regression analysis has been used to estimate model parameters, the relationship between them can be informed by the covariance matrix (Briggs et al. 2006). Similarly, methods of evidence synthesis are increasingly used to estimate the relative effect of interventions (see Section 10.4), and these can be used to generate correlated outputs. These correlations can be fully captured by sampling from or directly using the output of such synthesis in the PSA. If there is no evidence from statistical analysis that parameters are correlated it is generally not necessary to impose it.

Of course any logical or structural relationship between parameters (e.g. that transition probabilities in a Markov model must sum to 1) should be reflected in the model structure rather than imposing correlation. In other situations, conditional independence is often assumed. For example, the rates of clinical events on and off treatment will be correlated if independent distributions are assigned to a common baseline event rate and relative treatment effect (see Sections 9.2.5 and 10.3.1). This assumes that relative effect is independent of the baseline rate. In the absence of evidence to test this assumption or to estimate a relationship between relative effect and baseline, this assumption would be another source of structural uncertainty as discussed in Section 11.2.4.

In some circumstances there may be no published estimates or other evidence to inform distributions for some parameters. In these circumstances it would be inappropriate to assign a single assumed value. Rather, a continuous distribution should be assigned to represent the considerable uncertainty about potential mean values over plausible and theoretically possible ranges (uniform and triangular distributions should be avoided (Briggs et al. 2012)). Formal methods to elicit such distributions from experts were discussed in Section 10.5.5, and are discussed further in Section 11.2.4.

11.2.2.3 Computational challenges

The process of conducting PSA requires substantially more computation than a deterministic analysis. This is because the model must be evaluated not once but for every sample of possible parameter values. Given the type of software and processing capacity now available, this generally does not pose serious difficulties for the type of cohort models described in Chapter 9. However, some types of modelling approaches require simulation to obtain a single estimate of expected cost and effects for a single set of parameter values, e.g. individual sampling models (see Section 9.4.6).

The simulation required to estimate costs and effects in these models samples from variability in the parameter values ('stochastic uncertainty' in Table 11.1) rather than the uncertainty in the estimates of their mean values. To characterize parameter uncertainty this type of simulation needs to be repeatedly conducted for the range of possible mean parameter values. In other words, such models must sample from both uncertainty in mean parameter values and then variability given a sampled mean value.

In some complex models there may be a non-linear relationship between the variability in parameter values and net benefit. Therefore, simulation of both variability and uncertainty is required to generate unbiased estimates of expected net benefit (see Box 11.1). Importantly, there are no circumstances where variability 'matters' but uncertainty does not. In other words, if it is important to represent variability then uncertainty also needs to be fully characterized. Therefore, the computational expense of the chosen modelling approach and platform cannot reasonably justify a failure to characterize uncertainty adequately (Griffin et al. 2006).

There are a number of ways in which unavoidable computation expense can be overcome:

- Faster processing and the use of more efficient programming platforms, including improving in the efficiency of how simulated values are generated and used (O'Hagan et al. 2006a).
- Some types of emulators, which are essentially a model of a model, can dramatically reduce the computational burden of any almost any type of non-linear and complex models (Stevenson et al. 2004).
- If an unbiased linear approximation to non-linear aspect of a model can be found, this can also reduce the computational required to generate unbiased estimates of expected costs and effects (Ades et al. 2004).

11.2.3 Representing decision uncertainty

The key questions that an analysis of uncertainty and the way it is presented should seek to inform are:

- Is a decision to approve the intervention which is expected to be cost-effective (i.e. offers the highest expected NB) uncertain, based on current evidence?
- How uncertain is this decision is likely to be?
- Which alternatives might be better?
- What are the potential gains (in NB) of resolving some of the current uncertainty by acquiring additional evidence?

The output of probabilistic analysis can directly address the first three of these questions and provides the information required to address the fourth (see Section 11.3). Therefore, when considering the different ways of presenting the results of PSA, it is important to consider whether they can be interpreted in a way that addresses these questions.

The output of PSA includes an estimate of the expected cost, expected effect, and (given a particular cost-effectiveness threshold) expected NB of each alternative for every simulated sample of mean values of the parameters (see Figure 11.2). This output can be presented in a number of different ways. The most appropriate way to present uncertainty on cost-effectiveness will depend on whether there are multiple alternatives and whether any relevant decision-maker has been willing to specify a range of cost-effectiveness thresholds for how decisions will be made.

11.2.3.1 Scatter plots on the cost-effectiveness plane

When there are only two alternatives, the output of PSA can be used to represent the joint probability distribution on the incremental cost-effectiveness plane. This can be illustrated graphically in several equivalent ways, two of which are ellipses and scatter plots (Van Hout et al. 1994). Scatter plots simply plot each simulated estimate of the expected incremental costs and effects. This is illustrated in Figure 11.3 using simulated output from a PSA of two alternatives (B compared to A in Table 11.5). The circle indicates the expected incremental cost-effectiveness (ICER) of B compared to A (the ICER is a ratio of mean values, *not* an average of simulated ratios—see Section 8.3.1). In this example B is expected to be more effective but more costly than A with an ICER

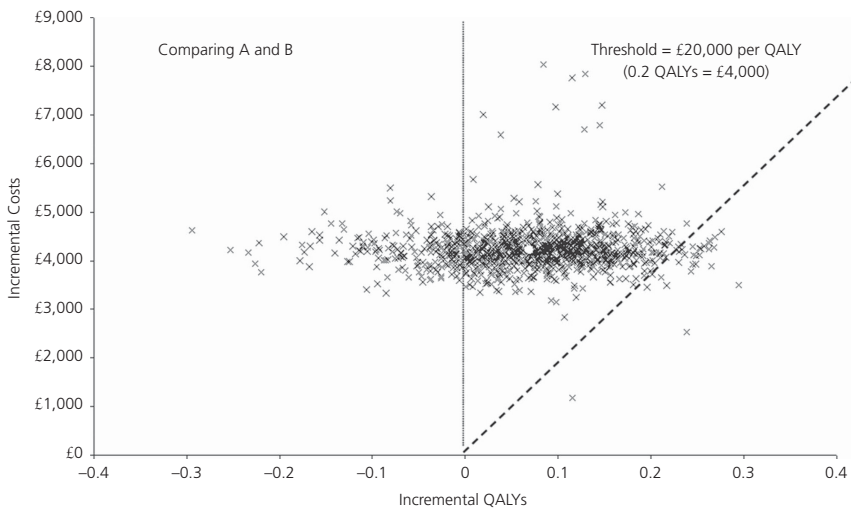


Fig. 11.3 Scatter plot with two alternatives.

Adapted from Springer, *PharmacoEconomics*, Volume 26, Issue 9, 2008, pp 781–98, Exploring Uncertainty in Cost-Effectiveness Analysis, Claxton K., Copyright © 2008, Adis Data Information BV. With kind permission from Springer Science and Business Media.

of £61 680 per QALY. The threshold is represented by the dashed line, here with slope of £20 000 per QALY. Therefore, B is not regarded as cost-effective and A would offer higher expected net benefits. The uncertainty surrounding a decision to reject B in favour of alternative A is uncertain. A decision to reject B on these grounds requires an assessment of the proportion of points which lie below the dashed line (where B is cost-effective and offers higher net benefits than A). There is clearly some decision uncertainty as some points do lie below this line, but the extent of uncertainty is difficult to assess even when there are only two alternatives to consider. There are two reasons for this:

- When represented in two dimensions (or even in three), it is difficult to assess accurately the proportion of points that lie below the line. Therefore, it is challenging to provide an intuitive estimate of the probability that B is, in fact, cost-effective compared to A (i.e. the error probability associated with the decision).
- It is difficult to visualize the impact of alternative cost-effectiveness thresholds in this incremental cost-effectiveness space (i.e. dashed lines of different slopes, where a steeper slope represents a higher threshold).

When there are more than two alternatives, which is commonly the case, scatter plots become impossible to interpret correctly. Figure 11.4 illustrates a scatter plot when alternative C is also considered alongside A and B. Now only the expected cost and effect pairs, rather than the increments, can be plotted. Intervention A remains cost-effective at a threshold of £20 000, and alternative B is extendedly dominated by A and C (see Section 4.4.1). Alternative C has a mean ICER of £24 628 per QALY when compared to

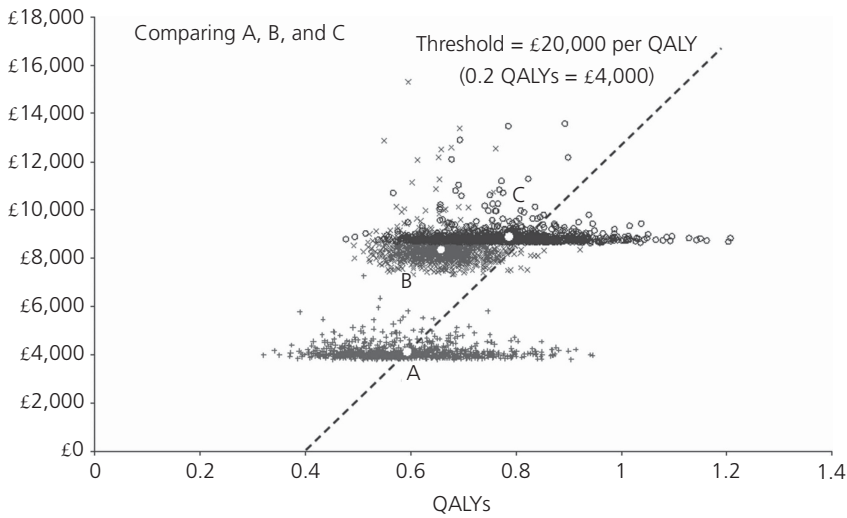


Fig. 11.4 Scatter plot with multiple alternatives.

Adapted from Springer, *PharmacoEconomics*, Volume 26, Issue 9, 2008, pp 781–98, Exploring uncertainty in cost-effectiveness analysis, Claxton K., Copyright © 2008, Adis Data Information BV. With kind permission from Springer Science and Business Media.

A (the non-dominated alternative). But consideration of uncertainty associated with a decision to reject both B and C, because they are not expected to be cost-effective, requires some assessment of the proportion of points lying below the dashed line for interventions B and C compared to A. This is difficult enough, but a critical piece of information is also missing: each of the simulated cost/effect pairs for A, B and C will be correlated by the structure of the model itself. For example, when simulated parameter values result in A having a higher cost, then B or C might also have a higher cost. Without this information it is impossible to assess the probability that A is cost-effective or the probabilities that B or C, respectively, might be cost-effective and offer the highest expected net benefit. This is also true for other summary measures such as confidence ellipses, confidence intervals, or presenting the distribution of net benefit for each of the alternatives (see Section 8.3.2).

11.2.3.2 Cost–effectiveness acceptability curves

These difficulties can be easily overcome by transforming each simulated set of costs and effects for each of the alternatives into expected net benefit, using a cost-effectiveness threshold (see Section 4.3), and recording the number of times each offers the highest expected net benefit. The probability that each alternative is cost-effective is the proportion of times that it has the highest expected net benefit. These probabilities can be calculated (without additional simulation) for a range of cost-effectiveness threshold values, and can be plotted as a cost-effectiveness acceptability curve (CEAC). The CEAC associated with the scatter plot in Figure 11.4 is illustrated in Figure 11.5.

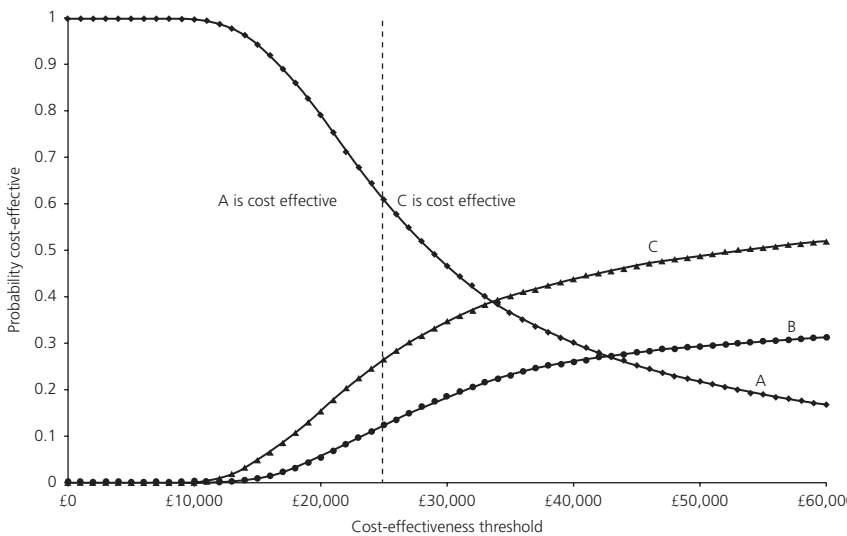


Fig. 11.5 Cost–effectiveness acceptability curve.

Adapted from Springer, *PharmacoEconomics*, Volume 26, Issue 9, 2008, pp 781–98, Exploring uncertainty in cost-effectiveness analysis, Claxton K., Copyright © 2008, Adis Data Information BV. With kind permission from Springer Science and Business Media.

The CEAC was introduced in Section 8.3.2 in the context of cost-effectiveness analysis based on individual patient data from, for example, a randomized controlled trial (RCT), and probabilities were calculated using parametric statistics or non-parametric bootstrapping. At first sight, the CEAC seems much easier to interpret because the probability that A, B, or C is cost-effective can simply be read off for any particular threshold. For example, if the threshold is less than £10 000 there is no uncertainty associated with a decision to reject B and C based on expected cost-effectiveness. At a threshold of £20 000 per QALY, this decision is more uncertain (probability that A offers the highest net benefit is 0.792), and there is a chance that alternative C or B could be cost-effective (probabilities of 0.154 and 0.054 respectively).

However, it is important to note that the alternative with highest probability of being cost-effective may not necessarily be the alternative that is expected to be cost-effective and offers the highest expected NB. This occurs in Figure 11.5 where, for thresholds between £24 628 and £34 000, alternative C is expected to be cost-effective (it has higher expected net benefits than A) but has a lower probability of being cost-effective than A. This might seem counter-intuitive, but it simply reflects the fact that, when C is better than A, it is 'much better', but when A is better than C it is only a 'little bit better' (i.e. the distribution of differences in net benefits is positively skewed, so the mean value is higher than the median). Therefore, presenting a CEAC alone is not enough. It is important to indicate which of the alternatives is expected to be cost-effective as well as its probability. This is indicated by the dashed line on Figure 11.5, or the cost-effectiveness acceptability frontier (CEAF), which is a plot of the probability that the alternative expected to be cost-effective, offering the highest expected net benefit, is cost-effective (Fenwick et al. 2001).

It can be difficult to compare multiple CEAC and CEAFs associated with alternative scenarios, but they have significant advantages over scatter plots and other similar approaches. Importantly, for any number of alternatives, CEACs can:

- indicate whether a decision based on expected cost-effectiveness is uncertain
- quantify the extent of decision uncertainty (the error probability)
- identify which other alternatives offer some possibility of being cost-effective.

However, they do not provide information about the differences in expected net benefits. Nor do they enable an assessment of how much the decision uncertainty 'matters' and whether more evidence is required (see Section 11.3).

11.2.3.3 Other ways to present uncertainty

If a decision-maker is willing to specify one or a range of cost-effectiveness thresholds, then a simple alternative to the CEAC and CEAF is to report expected effects, costs, ICERs, the expected net benefit (expressed in health or money terms), probabilities for each alternative being cost-effective and the error probability for a decision based on expected cost-effectiveness (Claxton 2008). This type of tabular reporting associated with the CEAC in Figure 11.5 is illustrated in Table 11.5 for cost-effectiveness thresholds of £20 000 and £30 000 per QALY.

The results reported in Table 11.5 indicate that A is expected to be cost-effective and offers the highest expected net benefits at a threshold of £20 000, but with a probability

Table 11.5 Presenting uncertainty and cost-effectiveness

	Cost	QALYs	ICER	Threshold = £20 000 per QALY			Threshold = £30 000 per QALY		
				Net Benefit	Probability	P(error)	Net Benefit	Probability	P(error)
A	£4,147	0.593	–	£7,722	0.792	*0.208*	£13 656	0.465	
B	£8,363	0.658	ED	£4,794	0.054		£11 373	0.186	
C	£8,907	0.787	£24 628	£6,827	0.154		£14 695	0.348	*0.652*

Adapted from Springer, *PharmacoEconomics*, Volume 26, Issue 9, 2008, pp 781–98, Exploring uncertainty in cost-effectiveness analysis, Claxton K., Copyright © 2008, Adis Data Information BV. With kind permission from Springer Science and Business Media.

of error greater than 0.2. This uncertainty is primarily in the choice between A and C (there is a very small chance that B will be cost-effective). At a threshold of £30 000 per QALY, C is expected to be cost-effective, but now a decision based on expected cost-effectiveness is much more uncertain (the error probability is greater than 0.6). Again the uncertainty is primarily between A and C. The difference in expected net benefit between the uncertain alternatives starts to give some indication of how much this uncertainty 'matters'. It is important, however, not to over-interpret these comparisons as they do not necessarily indicate the scale of the consequences of uncertainty and the potential gains from resolving it (Claxton et al. 2012).

11.2.4 Characterizing other sources of uncertainty

The uncertainty associated with the estimated parameters for a decision model is only one, and not necessarily the most important, source of uncertainty associated with expected effects and costs. Other sources of uncertainty include the different types of scientific judgements that have to be made when constructing a model of any sort. They might include alternative judgements about the potential bias or relevance of evidence used to estimate parameters, the choice of alternative statistical models to estimate parameters, and the assumptions required in structuring a decision model such as those used to extrapolate effects and costs over time. Insofar as the assumptions and judgements made are the only ones that are plausible, there would be no other sources of uncertainty. However, if other sets of assumptions are plausible, there will be uncertainty between these alternative 'scenarios' as well as within each (i.e. the parameter uncertainty given a particular set of assumptions).

11.2.4.1 Probabilistic scenarios

Assumptions and judgements will be informed by a range of direct and indirect evidence that may be available, the characteristics of the evidence, and the current understanding of the disease process that the models seek to represent. While these considerations may rule out some assumptions and judgements that are inappropriate and widely regarded as implausible, they may not point to a single best alternative. Performing PSA for each set of plausible assumptions (scenarios) will provide estimates of

expected cost, effect, and net benefit given each set of assumptions as well as the associated decision uncertainty. There are two problems with using scenarios in this way to represent structural uncertainty:

- It is not clear how many scenarios should be conducted and how they should be constructed to represent fully all sources of structural uncertainty.
- It leaves the decision-maker to consider multiple scenarios and implicitly to assess the plausibility of each and the impact on cost-effectiveness and decision uncertainty.

Scenarios can, however, be combined explicitly and quantitatively if probabilities are assigned to reflect their plausibility. A simple weighted average based on these probabilities would provide appropriate estimates of expected costs, effects, and net benefit. However, it will not provide a proper assessment of decision uncertainty or its consequences. This requires merging the simulated output from the scenarios using the probabilities representing the plausibility of each (Bojke et al. 2009; Claxton et al. 2012; Price et al. 2011; Strong et al. 2012).

11.2.4.2 Parametrizing structural uncertainty

In many situations, the differences between scenarios can be expressed in the form of an additional parameter in the model. That is, an assumption can be thought of as a missing parameter in the model or a parameter that has been set at an extreme value. For example, there may be no evidence available about whether the benefits from treatment are sustained beyond the period observed in the available clinical trials. Alternative assumptions might be that the benefit from treatment ceases as soon as the trials ended (A), that the benefit from treatment will be sustained but diminishes over time (B), or that the benefits from treatment observed in the trial continue over the remaining lifetime of the patients (C). Instead of presenting three scenarios (A, B, and C), the model could instead include a parameter that described the proportion of the treatment effect that was sustained over time. A range of values for this parameter (0 to 1) would represent the two extreme scenarios (A and C). It now becomes possible to assign a distribution which represents the uncertainty in this key parameter and to conduct PSA in the way described in Section 11.2.2. Therefore, what was previously structural uncertainty can be treated as parameter uncertainty. However, this does require the assignment of a distribution which reflects informed judgements about its possible mean values when evidence to estimate it is not available.

11.2.4.3 Elicitation

Parametrizing structural uncertainty or combining probabilistic scenarios requires explicit quantitative judgements to be made about the value of, and the uncertainty associated with, a parameter for which no direct evidence is available. As described in Section 10.5.5, methods to elicit distributions to represent the beliefs of relevant experts have begun to be applied in cost-effectiveness analysis (Bojke et al. 2010; Bojke and Soares 2014; Grigore et al. 2013; Soares et al. 2011). Elicitation has been used in risk analysis (O'Hagan et al. 2006b), and is especially useful when decisions need to be informed but when data happen to be sparse or of poor quality (Soares et al. 2013).

Importantly elicitation helps to identify what type of additional evidence might be most valuable and what type of additional research should be conducted (see Section 11.3.2) (McKenna and Claxton 2011; Soares et al. 2013). However, the use of formal elicitation poses a number of important questions:

- Who should provide the judgements (which experts or decision-makers)?
- Which particular methods of elicitation should be used?
- How should the quality of judgements be calibrated (tested) and weighted?
- How should judgements from different experts be combined?

Elicitation is also time and resource intensive and, for these reasons, it is more common to present structural uncertainty as a series of probabilistic scenarios. However, this leaves the decision-maker to make the same judgements implicitly that would be required in a more explicit and quantitative analysis.

11.2.4.4 Model averaging

Model averaging does not refer to combining very different decision models by taking some form of weighted average. Rather than combining the results of analyses that may not be appropriate or implausible with those of others, the reason for differences in results should be identified, critically examined, and a preferred analysis that represents any structural uncertainties identified. Combining alternative probabilistic scenarios within a decision model (as described above) has sometimes been described as 'model averaging'. However, it is probably best to reserve the term for the way it has been used in statistics, where alternative statistical models can be applied to data when estimating one or more particular parameter(s). The measures of performance of these statistical models can be used as 'weights' to combine the results of the different statistical models that are possible and credible (Hoeting et al. 1999; Jackson et al. 2009).

11.3 Is current evidence sufficient?

Assessments of cost-effectiveness are inevitably uncertain and, without sufficient and good-quality evidence, decisions about the use of technologies will also be uncertain. There will be a chance that the resources committed by the approval of a new intervention may be wasted if the expected positive net health effects are not realized. Equally, rejecting a new intervention that is not expected to be cost-effective will risk failing to provide access to a valuable intervention if the net health effects prove to be greater than expected. Therefore, if the objective is to improve overall health for both current and future patients, then the need for and the value of additional evidence is an important consideration when making decisions about the use of technologies. This is even more critical once it is recognized that the approval of a new technology for widespread use might reduce the prospects of the type of research that would provide the required evidence being conducted. In these circumstances there will be a trade-off between the net health effects for current patients from early access to a cost-effective technology and the health benefits for future patients from withholding approval until valuable research has been conducted (Griffin et al. 2011).

Research also consumes valuable resources which could have been devoted to patient care, or to other more valuable research priorities. Also, uncertain events in the near or distant future may change the value of the technology and the need for evidence. These events could include prices of existing technologies, the entry of new technologies, other evidence about the performance of technologies, and the natural history of disease (Philips et al. 2008). In addition, implementing a decision to approve a new technology may commit resources which cannot subsequently be recovered if a decision to approve or reimburse might change in the future (e.g. due to new research reports) (Eckermann and Willan 2008). Therefore, appropriate research and coverage decisions will depend on whether the expected benefits of research are likely to exceed the costs, and whether any benefits of early approval or reimbursement are greater than withholding approval until additional research is conducted or other sources of uncertainty are resolved. Methods of analysis which provide a quantitative assessment of the potential benefits of acquiring further evidence allow these types of research and reimbursement decisions to be addressed explicitly and accountably (see Section 11.4).

11.3.1 Value-of-information analysis

The principles of value-of-information analysis have a firm foundation in statistical decision theory. There are closely related concepts and methods in mathematics and financial economics with diverse applications in business decisions, engineering, environmental risk analysis, and financial and environmental economics (Pratt et al. 1995). There are now many applications in health, some commissioned directly to inform policy and others published in specialist as well as general medical and health policy journals (Claxton and Sculpher 2006; Colbourn et al. 2007; Soeteman et al. 2011; Stevenson and Jones 2011; Welton et al. 2008; Yoyota and Thompson 2004). Most commonly these methods of analysis have been applied in the context of probabilistic decision analytic models used to estimate expected cost-effectiveness of alternative interventions. However, the same type of analysis can also be used to extend standard methods of systematic review and meta-analysis (Claxton et al. 2013). Indeed the principles of value-of-information analysis can also be used as a conceptual framework for qualitative assessment of how important uncertainty might be, and the relative priority of alternative research topics and proposals (Fleurence and Meltzer 2013; Meltzer et al. 2011).

Additional evidence is valuable because it can improve patient outcomes by resolving existing uncertainty about the cost-effectiveness of the interventions available, thereby informing treatment choice for subsequent patients. For example, the balance of existing evidence might suggest that a particular intervention is expected to be cost-effective and offer the greatest net health benefits. However, there will be a chance that others are in fact more cost-effective, offering higher net health benefits. If treatment choice is based on existing evidence, then there will be a chance that other interventions would have improved overall health outcomes to a greater extent. In other words, there are adverse net health consequences associated with uncertainty.

11.3.1.1 Expected value of perfect information

The scale of uncertainty can be indicated by the results of probabilistic analysis of a decision-analytic model (see Section 11.2.3 and Table 11.5). The expected

Table 11.6 Calculating EVPI from the results of PSA

How things could turn out	Net Health Benefit (QALYs)			Best we could do if we knew
	Treatment A	Treatment B	Best choice	
θ_1	8	12	B	12
θ_2	16	8	A	16
θ_3	9	14	B	14
θ_4	12	10	A	12
θ_5	10	16	B	16
Average	**11**	**12**		**14**

consequences of this uncertainty can be expressed in terms of net health benefits or the equivalent health care system resources that would be required to generate the same net health effects. These expected consequences can be interpreted as an estimate of the net health benefits that could potentially be gained per patient if the uncertainty surrounding their treatment choice could be resolved. It indicates an upper bound on the expected net health benefits of further research which is also known as the expected value of perfect information (EVPI) (Briggs et al. 2006).

EVPI can be calculated directly from the output of PSA (see Table 11.6 and Box 11.3). More formally, assume there are alternative interventions (j) where the net benefit (NB) of each depends on uncertain parameters that may take a range of possible values (θ). The best decision based on the information currently available would be to choose the intervention that is expected to offer the maximum net benefit ($\max_j \mathrm{E}_\theta \mathrm{NB}(j, \theta)$). For example, this would be 12 QALYs in Table 11.6. If this uncertainty could be fully resolved (with perfect information), the decision-maker would know which value θ would take before choosing between the alternative interventions. They would be able to select the intervention that provides the maximum net health benefit for each particular value of θ (i.e. $\max_j \mathrm{NB}(j, \theta)$). However, when a decision is about whether further research should be conducted, the results (the true values of θ) are necessarily unknown. Therefore, the expected net health benefits of a decision taken when uncertainties are fully resolved (with perfect information) is the found by averaging these maximum net benefits over all the possible results of research that would provide perfect information ($\mathrm{E}_\theta \max_j \mathrm{NB}(j, \theta)$). For example, 14 QALYs in Table 11.6.

The EVPI for an individual patient is simply the difference between the expected value of the decision made with perfect information about the uncertain parameters θ, and the decision made on the basis of existing evidence:

$$\mathrm{EVPI} = \mathrm{E}_\theta \max_j \mathrm{NB}(j, \theta) - \max_j \mathrm{E}_\theta \mathrm{NB}(j, \theta) \tag{11.1}$$

For example, in Table 11.6 the additional net health benefits of resolving uncertainty, or EVPI, is 2 QALYs per patient. This is greater than the additional net benefits offered by alternative B (i.e. 1 QALY). It should be apparent from Table 11.6 that the per-patient EVPI is easily calculated from the simulated output of a PSA. This may also include a quantitative assessment of structural as well as parameter uncertainty as described in

Box 11.3 Calculating EVPI from the results of PSA

The principles of value-of-information analysis and how EVPI can be calculated directly from the results of PSA are illustrated in Table 11.6 for the choice between two alternatives, A and B. The net health benefit of A and B that would result from five possible (randomly sampled) values that the parameters might take (θ_1,.., θ_5) are reported in each row. They represent the results of PSA using Monte Carlo simulation (see Section 11.2.2). Although a PSA would sample many more times from the distributions assigned to the parameters, imagine that there are only five values that the parameters might take (how things might turn out). Since these values are sampled at random they are equally likely, so the expected net benefit of alternative A and B is the average over these five possibilities.

The best decision based on current information would be to choose the alternative that has the greatest expected net benefit, i.e. alternative B which offers expected net benefits of 12 QALYs per patient compared to 11 QALYs for alternative A. However, the decision to choose B is uncertain because B is the best alternative in only three out of the five possibilities (θ_1, θ_3, θ_5), so the probability that B offers the greatest net benefit is 0.6. The error probability associated with this decision is 0.4 because A offers greater net benefits than B in two out of the five possibilities (A offers 16 rather than 8 QALYs for θ_2 and 12 rather than 10 QALYs for θ_4).

If the decision-maker could choose between A and B after it is known which value the parameters would take (i.e. with perfect information), they would be able choose B for values of θ_1, θ_3, and θ_5 but chose A for values of θ_2, θ_4, and achieve the maximum net benefits in the final column of Table 11.6. However, which value the parameters will take is unknown so the expected value of taking this decision with perfect information is the average of these maximum net benefits, i.e. 14 QALYs per patient. The maximum value of additional evidence is the difference between the expected net benefits of the decision taken with perfect information (14 QALYs) and the expected net benefits of the decision using current information (12 QALYs). Therefore, the EVPI is 2 QALYs per patient. It is worth noting that:

- Once the PSA has been conducted, calculating EVPI per patient is very straightforward.
- Additional evidence is only valuable if it might lead to different decisions and therefore improve net benefit.
- The value of acquiring additional evidence about the performance of a new technology may well exceed the value of having access to it based on current information, e.g. the additional value of access to B is 1 QALY but the value of additional evidence about the choice between A and B is 2 QALYs (see Section 11.4.3).

Section 11.2.4. Therefore, the time and effort required to calculate EVPI is negligible once all the evidence has been assembled and the judgements have been made in constructing a model, synthesizing evidence and assigning probability distributions to all the sources of uncertainty.

Once the results of research are available they can be used to inform treatment choice for all subsequent patients. Therefore, the EVPI needs to be expressed for the population of patients that can benefit from it. The population EVPI will increase with (1) the size of the patient population whose treatment choice can be informed by additional evidence; and (2) the time over which evidence about the cost-effectiveness of these interventions is expected to be useful.

11.3.1.2 Time horizons for research decisions

The information generated by research will not be valuable indefinitely. This is because other things change over time which will have an impact on the future value of the information generated by research that can be commissioned today. For example, the prices of the alternative technologies are likely to change over time (e.g. patent expiry of branded drugs and the entry of generic versions of those products). Also, new and more effective interventions become available which will eventually make current comparators obsolete, so information about their effectiveness will no longer be relevant to future clinical practice. Other information may also become available in the future which will impact on the value of the evidence generated by research that can be commissioned today. Finally, as more information about individual effects is acquired through greater understanding of the reasons for variability in patient outcomes, the value of evidence that can resolve uncertainty in expected or average effects for the patient population and/or its subpopulations will decline (see Section 11.5). For all these reasons there will be a finite time horizon for the expected benefits of additional evidence; that is, there will be a point at which the additional evidence that can be acquired by commissioning research today will no longer be valuable.

Specifying a time horizon for a particular research decision is an proxy for a complex, and uncertain process of future changes (Philips et al. 2008). Nonetheless, some judgement is unavoidable when making decisions about research priorities. Some assessment is possible based on historical evidence and judgements about whether a particular area is more likely to see earlier patent expiration, future innovations, other evaluative research, and the development of individualized care. Information is often available about clinical studies that are already planned and under way, as well as future innovations from registered patents and/or phase 1 and 2 trials and licensing applications. For these reasons, an assessment of an appropriate time horizon may differ across clinical areas and specific research proposals. The incidence of patients who can benefit from the additional evidence may also change over time, although this may not necessarily decline as other types of effective health care change competing risks.

11.3.1.3 Research prioritization

Two questions are posed when considering whether further research should be prioritized and commissioned:

- Are the potential expected net health benefits of additional evidence (population EVPI) sufficient to regard the type of research likely to be required as potentially worthwhile?
- Should this be prioritized over other research that could be commissioned with the same resources?

These assessments require some consideration of the costs of different types of research. There is also a need to consider that the time likely to be taken for the proposed research to be commissioned, conducted and reported.

One way to address the question is to ask whether the health care system could generate similar expected net health benefits more effectively elsewhere. This is equivalent to asking whether the costs of the research would generate more net health benefits if these resources were made available to the system to provide health care. For example, estimates of the relationship between changes in UK NHS expenditure and health outcomes suggests that the NHS spends approximately £25 000 to gain 1 life-year and somewhat less than £15 000 to gain 1 QALY (Claxton et al. 2015). Using these estimates, proposed research activities that cost, for example, £2 m could have been used to deliver health care which is likely to save 80 life-years and to generate more than 130 QALYs elsewhere in the NHS. If these opportunity costs of research are substantially less than the expected benefits (population EVPI) then it would suggest that the proposed research is potentially worthwhile.

However, most research funders have limited resources (with constraints relevant to a budgetary period) and cannot draw directly on the other (or future) resources of the health care system. Therefore, even if the population EVPI of proposed research exceeds these opportunity costs, it is possible that other research may be even more valuable. If similar analyses are conducted for all proposals competing for limited research resources, it becomes possible to identify a short list of those which are likely to be worthwhile and then to select those that are likely to offer the greatest value.

It should be noted that the population EVPI represents only the *potential* or *maximum* expected benefits of research that could be conducted. There are two reasons for this. First, regardless of how large the sample size or how well a study is conducted, no research can resolve all uncertainty and provide perfect information. Secondly, there are usually a large number of uncertain parameters that contribute to θ and are relevant to differences in the net benefit of the alternative interventions; and most research designs will not provide information about all of them.

Nonetheless, EVPI provides an upper bound to the value of conducting further research. Therefore, when compared to the opportunity cost of conducting research (e.g. the health equivalent of the resources required), it can provide a necessary condition for a decision to conduct further research while the intervention is approved for widespread use. It also provides a sufficient condition for early approval when approval would mean that the type of further research needed would not be possible or would be too costly to be worthwhile. This could be, for example, because there would be a lack of incentives for manufacturers, or further randomized trials would not be regarded as ethical and/or would be unable to recruit. In these circumstances, the population EVPI represents an upper bound on the benefits to future patients that would be forgone or the opportunity costs of early approval based on existing evidence (see Section 11.4.3).

11.3.2 What type of evidence is needed?

The type of analysis described above indicates the potential value of resolving all the uncertainty surrounding the choice between alternative interventions. However, it would be useful to have an indication of which sources of uncertainty are most important and what type of additional evidence would be most valuable. This can provide an indication of the type of research design that is likely to be required, whether such a study will be possible once a new technology is approved for widespread use, and the sequence in which different studies might be conducted.

11.3.2.1 Expected value of perfect parameter information (EVPPI)

The potential expected benefits of resolving the different sources of uncertainty that determine the net benefit of the alternative interventions can be established using the same principles as calculating EVPI (Ades et al. 2004; Briggs et al. 2006). For example, if the net benefit (NB) of each intervention (j) depends on two (groups of) uncertain parameters (θ_1 and θ_2) that may take a range of possible values, the best decision based on current information is still to choose the intervention that is expected to offer the maximum net benefit (i.e. $\max_j \mathrm{E}_{\theta_2,\theta_1}\,\mathrm{NB}(j, \theta_1,\theta_2)$). If the uncertainty associated with only one of these groups of parameters (θ_1) could be fully resolved (with perfect parameter information), the decision-maker would know which value θ_1 would take before choosing between the alternative interventions. However, the values of the other parameters (θ_2) remain uncertain, so the best they can do is to select the intervention that provides the maximum expected net health benefit for each value of θ_1 (i.e. $\max_j \mathrm{E}_{\theta_2|\theta_1}\,\mathrm{NB}(j, \theta_1, \theta_2)$). Which particular value θ_1 will take is unknown before research is conducted, so the expected net health benefit when uncertainty associated with θ_1 is fully resolved is the average of these maximum net benefits over all the possible values of θ_1, (i.e. $\mathrm{E}_{\theta_1} \max_j \mathrm{E}_{\theta_2|\theta_1}\,\mathrm{NB}(j, \theta_1, \theta_2)$). The expected value of perfect information about parameter θ_1 ($\mathrm{EVPPI}_{\theta_1}$) is simply the difference between the expected value of the decisions made with perfect information about θ_1, and a decision based on existing evidence:

$$\mathrm{EVPPI}_{\theta_1} = \mathrm{E}_{\theta_1} \max_j \mathrm{E}_{\theta_2|\theta_1}\,\mathrm{NB}(j, \theta_1, \theta_2) - \max_j \mathrm{E}_{\theta_2,\theta_1}\,\mathrm{NB}(j, \theta_1, \theta_2) \qquad (11.2)$$

It should be noted that this describes a general solution for non-linear models. However, it is computationally intensive because it requires an inner loop of simulation to estimate the expected net benefit for each value of θ_1 ($\mathrm{E}_{\theta_2|\theta_1}\,\mathrm{NB}(j, \theta_1, \theta_2)$), as well as outer loop of simulation to sample the possible value θ_1 could take. The computational requirements can be somewhat simplified if there is a multilinear relationship between the parameters and net benefit because the inner loop of simulation is unnecessary (if θ_1 and θ_2 are uncorrelated the mean values of θ_2 will provide the correct estimate of $\mathrm{E}_{\theta_2|\theta_1}\,\mathrm{NB}(j, \theta_1, \theta_2)$ in equation 11.2). More efficient methods to estimate EVPPI have been developed and can substantially reduce this computational burden in non-linear models (Brennan et al. 2007; Strong et al. 2014).

11.3.2.2 Informing research design

Identifying which sources of uncertainty are most important and what type of evidence is likely to be most valuable is useful in two respects. It can help to identify the type of

research design that is likely to be required. For example, an RCT may be needed to avoid the risk of selection bias if additional evidence about the relative effect of an intervention is required—see Chapter 8). It can also identify the most important end points to include in any particular research design (Claxton and Sculpher 2006). This type of assessment can also be used to consider whether there are other types of research that could be conducted relatively quickly (and cheaply) before more lengthy and expensive research (e.g. a large RCT) is really needed. That is, it can help to identify the sequence in which different types of study might be conducted (Griffin et al. 2010).

11.3.3 What type of research should be conducted?

Estimates of EVPI and EVPPI discussed in Sections 11.3.1 and 11.3.2 only provide a *necessary* condition for conducting further research. To establish a *sufficient* condition and to identify the optimal research design, estimates of the expected benefits and cost of sample rather than perfect information are required.

11.3.3.1 Expected value of sample information (EVSI)

The same principles of value-of-information analysis which were illustrated in Table 11.6 can be extended to establish the expected value of sample information (EVSI) for a particular research design over a range of possible sample sizes. Now, rather than identifying what choice would be made and what net benefits would result if each of the possible values of the parameter was known with certainty, the question is which choice would be made and what net benefit would result from each of the possible results of the sample. This requires a series of analytical steps to be taken:

- Simulating ('predicting') a possible sample result for each of the sampled parameters.
- Updating beliefs about the value of these parameters with the predicted sample result.
- Identifying which alternative would offer the highest net benefit if that turned out to be the result of the sample.
- Repeating this to represent the range of possible sample results.

The EVSI is simply the average of the maximum expected net benefits across all the predicted sample results (Ades et al. 2004; Briggs et al. 2006). Calculating EVSI can require intensive computation, especially if the relationships between the sampled parameters (end points in the research design) and the net benefit of the alternatives are non-linear. However, there are a number of ways in which this computational burden can be eased (Kharroubi et al. 2011; Strong et al. 2015).

11.3.3.2 Optimal sample size and other aspects of design

To establish the optimal sample size for a particular type of study, EVSI calculations need to be repeated for a range of sample sizes. The difference between the EVSI and the costs of acquiring the sample information is the expected net benefit of sample information (ENBS) or the societal payoff to research. The optimal sample size is

simply the one that generates the maximum ENBS. The same type of analysis can be used to evaluate a range of different aspects of research design. These include which end points to include, which interventions should be compared and the appropriate length of follow-up. The best design is the one that provides the greatest ENBS (Tuffaha et al. 2014; Welton et al. 2014). It can also be used to identify whether a combination of different types of study might be required (Conti and Claxton 2009).

It should be recognized that the costs of research do not only include the resources consumed in conducting it: they also include the opportunity costs falling on the patients enrolled in the research and those whose treatment choice can be informed once the research reports. Therefore, optimal research design will depend, among other things, on whether or not patients have access to the new technology while the research is being conducted and how long it will take before it reports (McKenna and Claxton 2011; Willan and Eckermann 2010). It is also possible to take account of the likelihood of research findings actually being implemented when designing research. For example, if the impact of a study on clinical practice depends on it reporting a statistically significant result this will influence optimal sample size.

11.3.3.3 An iterative approach

Once the results of additional research are available the type of analysis discussed above can be repeated in an iterative process. That is, updating the synthesis of evidence, re-estimating the net benefits of the alternative interventions and updating the value-of-information analysis. This facilitates a consideration of whether the research was indeed definitive (additional research is not worthwhile) or whether more or different types of evidence might be required. Similarly, value-of-information analysis can also provide the analytic framework to consider when to stop a clinical trial, how to allocate patients between the arms of a trial as evidence accumulates (sequential and group sequential designs) and when other types of evidence might become more important as the results of research are reported over time (Pertile et al. 2014).

It should be noted that assessing the value of research with hindsight and whether it has led to a change in clinical outcomes is potentially misleading. This is because research is not the only way to influence clinical practice (see Section 11.4.1), and the findings of research are only one realization of the uncertainty about the potential results that could have been found. For example, the results of a trial might not change a decision about whether or not an intervention is worthwhile (with hindsight the trial realized no value). However, if at the time it was commissioned, there was a chance that the results could have changed the decision and improved net benefit, it was correctly judged to be valuable nonetheless.

11.4 Implications for approval and research decisions

Much economic evaluation is concerned with supporting decisions about whether the adoption of a new intervention is cost-effective compared to alternatives for a given group of patients. There is, however, inevitable uncertainty in knowledge about the relevant disease and the impacts of the range of alternative forms of management. Those interested in interpreting clinical evidence in the context of uncertainty generally use the principles of classical or 'frequentist' inference as their guide: that is, a

set of 'rules'—usually defined in advance of data analysis—to establish whether a null hypothesis can be rejected in favour of an alternative. As described in Chapter 8, these statistical tools are also used by some researchers undertaking economic evaluation in health care (Glick et al. 2014). The link between the use of these statistical methods and the use of economic evaluation to guide decisions with an ultimate objective of improving population health subject to constraints on resources is, however, not clear (Claxton 1999). Rather than applying 'rules' regarding how much uncertainty is 'acceptable', an alternative framework, founded on Bayesian decision theory (Pratt et al. 1995), focuses on expanding the options available to decision-makers and identifying which option is expected to generate the greatest gain in net benefit (Claxton et al. 2012). This can inform an assessment of whether more health might be gained through efforts to implement the findings of existing research or by acquiring more evidence to inform which intervention is most cost-effective.

11.4.1 Value of implementation

Overall health outcomes can be improved by ensuring that the accumulation of research findings are implemented and have an impact on clinical practice. Indeed, the potential improvements in health outcome by encouraging the implementation of what existing evidence suggests is the most cost-effective intervention may well exceed the potential improvements in net health benefits through conducting further research.

The distinction between these two very different ways to improve overall health outcomes is important because, although the results of additional research may influence clinical practice and may contribute to the implementation of research findings, it is certainly not the only, or necessarily the most effective, way to do so. There may be other more effective mechanisms (e.g. more effective dissemination of existing evidence) or policies (e.g. those that offer incentives and/or sanctions to practitioners to undertake cost-effective interventions). In which case continuing to conduct research to influence clinical practice, rather than because there is real value in acquiring additional evidence itself, may be inappropriate. This is because research resources could have been used elsewhere to acquire additional evidence in areas where it offered greater potential net health benefits.

However, the health benefits of conducting further research will only be realized (i.e. health outcomes actually improve and/or resources are saved) if the findings of the research do indeed have an impact on clinical practice. Recognizing that there are very many ways to influence clinical practice is important when considering other policies to improve implementation of research findings instead of, or in combination with, conducting further research (Fenwick et al. 2008; Hoomans et al. 2009; Soeteman et al. 2011). The importance of being able to implement the findings of proposed research might influence its priority if implementation is unlikely, or if influencing clinical practice would require costly research designs (Willan and Eckermann 2010).

11.4.2 Decisions based on the balance of existing evidence?

It should be recognized that restricting attention to whether or not the results of a clinical trial, meta-analysis of existing trials, or the results of a cost-effectiveness analysis offer statistically significant results is unhelpful for a number of reasons: it provides

only a partial summary of the uncertainty associated with the cost-effectiveness of an intervention; and it does not indicate the importance of the uncertainty for overall patient outcomes or the potential gains in net health benefit that might be expected from acquiring additional evidence that could resolve it. More importantly, failing to implement an intervention which is expected to offer the greatest expected net benefits will impose opportunity costs. This suggests that always waiting to implement research findings until the traditional rules of statistical significance are achieved (whether based on frequentist hypothesis testing or on Bayesian benchmark error probabilities) may well come at some considerable cost to patient outcomes and health care system resources (Claxton 1999). However, there are a number of issues that need to be considered before decisions to approve or to reimburse a new technology can be based on the balance of accumulated evidence, that is, expected cost-effectiveness and expected net health benefits:

- If approval or reimbursement means that the type of research required to generate the required evidence is impossible or more difficult to conduct, then the expected value of additional evidence that will be forgone by approval needs to be considered alongside the expected net benefits of implementation (Griffin et al. 2011) (see Section 11.4.3).
- Account must be taken of the consequences of the likelihood that widespread use of an intervention will be difficult or take time to reverse if subsequent research demonstrates that it is not cost-effective. That is, an analysis needs to reflect the opportunity costs associated with the chance that research finds that the intervention is not cost-effective but being unable to implement these findings quickly and to withdraw its use (Eckermann and Willan 2006; Palmer and Smith 2000).
- Sometimes the implementation of an intervention will require the commitment of capital, infrastructure, or training costs which cannot be recovered if research later suggests that the intervention is not worthwhile (Eckermann and Willan 2008). Even where these types of irrecoverable costs are not required, an intervention that offers longer-term benefits but more immediate costs (e.g. any effect on mortality risk), is likely to commit initial losses of net benefit compensated by later expected gains; in other words, costs tend to be committed before the improvement health outcomes are achieved. In these circumstances the approval or reimbursement of the intervention commits irrecoverable short-term opportunity costs for each patient treated. If the uncertainty about the intervention's cost-effectiveness might be resolved in the future due to commissioned research reporting, then it may be better to withhold approval or reimbursement until the research findings are available. This is more likely to be the case when a decision to delay initiation of treatment is possible and is associated with more limited health impacts; for example, in chronic and stable conditions (Claxton et al. 2012; Forster and Pertile 2013; McKenna and Claxton 2011).
- There is a quite natural aversion to iatrogenic effects; that is, the health that is lost through adopting a new intervention. This tends to be of greater concern than the same amount of health being lost through continuing to use existing interventions that are less effective than others available. However, it should be noted that

the consequences for patients, in terms of health forgone, are the same for both types of decisions. Furthermore, this 'aversion' depends entirely on which intervention happened to have diffused into common clinical practice first.

These considerations can inform an assessment of whether more health is expected to be gained by implementing the findings of existing research or by delaying implementation until more evidence is available, or until other sources of uncertainty resolve. Therefore, there are many circumstances where approval or reimbursement should not be based simply on the balance of evidence (i.e. expected cost-effectiveness or expected net benefit) (Eckermann and Willan 2008; Griffin et al. 2011). However, there is certainly no single 'rule' based on notions of the statistical significance. These considerations can be dealt with explicitly and quantitatively within a well conducted value-of-information analysis (Claxton et al. 2012).

11.4.3 Approval with research or approval only in research?

Decisions about the approval or reimbursement of new medical technologies are increasingly being made close to their launch when there may be substantial uncertainty surrounding their cost-effectiveness (Tunis and Pearson 2006). It poses the question of whether a new technology should be approved but additional evidence sought, or whether its widespread use should be restricted until the additional evidence is available. This is, then, a question of whether coverage with evidence development is appropriate and how the entry of new technologies into health care systems ought to be managed (Claxton et al. 2012; Garrison et al. 2013; Walker et al. 2012). To address this, an assessment is necessary of the value to current patients of offering early access to a new technology based on currently available evidence, and the value to future patients of being able to conduct the research which would provide the evidence needed to resolve the uncertainty about whether the new technology is worthwhile (Eckermann and Willan 2006; Griffin et al. 2011).

If a new technology is expected to be cost-effective based on existing evidence there will be value to current patients of early approval (the additional health benefits gained). This assessment is likely to be uncertain, so further research may be worthwhile, which will offer additional net benefits for future patients. If the research can be conducted while the new technology is approved for widespread use, then approving the technology with research would enable gains in net benefit for both current and future patients. If the new technology is not expected to be cost-effective, further research might still be worthwhile if this assessment is uncertain. In these circumstances the technology can be approved for use only in research, with this restriction being reconsidered on the research reports. If decisions to approve can be easily reversed and do not commit the type of irrecoverable costs discussed in Section 11.4.2, this suggests that approval with research (or coverage with evidence development) will be appropriate if a technology is expected to be cost-effective and restricting access to approval only in research if it is not (Chalkidou et al. 2007).

However, the approval of a technology for widespread use may reduce the prospects of conducting the research which can provide the required evidence. For example, if the key source of uncertainty was the relative effect of the new technology, then the type

of randomized trial that might be needed to provide an unbiased estimate is unlikely to be possible once the technology is approved for widespread use. This is because randomization is unlikely to be regarded as ethical and a trial would be unable to recruit informed patients who already have access to the new technology (Miller and Pearson 2008; Pearson et al. 2006). There would also be limited incentives for manufacturers to conduct the research once their technology has unrestricted market access. Unless it is possible to provide the additional evidence needed from the type of observational data that are likely to be available, there will be a trade-off between the additional net benefits of approval based on current evidence and the net benefits to future patients of conducting research which will no longer be possible as a consequence of approval.

Therefore, restricting the approval of a new technology to only in research may well be appropriate even when it is expected to be cost-effective based on current evidence (even if initial approval does not commit irrecoverable costs and could be easily reversed; see Section 11.4.2). For example, a technology with an ICER just equal to the cost-effectiveness threshold might be judged to be (just) cost-effective but will offer no additional expected net health benefits to current patients based on existing evidence (see Section 4.2). There may well be substantial net benefits for future patients from conducting further research which may not be possible if it is approved for widespread use. In these circumstances, withholding approval of a cost-effective technology until valuable research has been conducted would improve health outcomes overall (Claxton et al. 2012; Griffin et al. 2011; McKenna and Claxton 2011).

11.5 Uncertainty, heterogeneity, and individualized care

There are two important ways in which patient outcomes can be improved. One is by acquiring additional evidence to resolve the uncertainty in the expected effects of an intervention. The other is by understanding the sources of variability and dividing the population into finer subgroups where the intervention will be expected to be cost-effective in some but not in others (see Chapters 8 and 10 for further discussion on heterogeneity). However, a greater understanding of heterogeneity also has an impact on the value of additional evidence (Espinoza et al. 2014). As more subgroups are defined, the precision of the estimates of effect is necessarily reduced (the same amount of evidence offers fewer observations in each subgroup). The uncertainty about which intervention is most cost-effective may be reduced in some (e.g. where it is particularly effective or positively harmful), but increase in others. Therefore, the value of additional evidence per patient may be higher or lower in particular subgroups. In the limit, as more sources of variability are observed, the value of additional evidence will fall (the sum across all subgroups of the population). Indeed, if all sources of variability could be observed then there would be no uncertainty at all.

Value-of-information analysis can be applied within each subgroup based on existing evidence (Colbourn et al. 2007; Espinoza et al. 2014). This is useful because it can indicate which types of patient need to be included in any future research design and those who could be excluded. Although the potential value of additional evidence about the whole population is simply the sum of values for each of its subpopulations, the value of acquiring evidence within only one subgroup depends on whether that

evidence can also inform decisions in others. For example, evidence about the relative effect of an intervention in one subgroup might also inform the relative effect in others. On the other hand, evidence about a subgroup-specific baseline risk might not be relevant to others. In principle, these questions of 'exchangeability' of evidence can be informed by existing evidence and ought to be reflected in how it is synthesized and the uncertainties characterized.

There is potential value of conducting additional research which might not resolve uncertainty but, instead, reveal the reasons for variability in outcomes. Such research informs which patient subgroups could benefit most from an intervention, or the choice of the physician and patient in selecting care given the symptoms, history, and preferences (i.e. individualized care) (Basu 2011; Basu and Meltzer 2007). This type of research may be very different from the type of evaluative research that reduces uncertainty about estimates of effectiveness. For example, it might include research into the value of diagnostic technologies and pharmacogenetics, analysis of observational data, and treatment selection as well as novel trial designs which can reveal something about the distribution of outcomes in different subgroups (Basu 2015).

This type of analysis can start to inform decision-makers whether resources should be invested in a range of possible activities. These include providing early access to new technologies; ensuring the findings of existing (or commissioned) research are (or will be) implemented; conducting research to provide additional evidence about particular sources of uncertainty in some (or all) subgroups; and conducting research which can lead to a better understanding of variability in effects. Of course, some combination of these policy choices may well offer the greatest impact on overall health outcomes.

11.6 Concluding remarks

The discussion in previous sections of this chapter has been founded on a health care system which faces some constraints on the growth of health care expenditure. In such systems, additional costs displace other care that would have otherwise generated improvements in health, so costs and health effects can be expressed as net health benefit. However, in all systems, interventions impose costs (or offer benefits) which fall outside the health care system and displace consumption rather than health. In other circumstances health care costs may fall entirely on consumption. As discussed in Chapter 4, if some consumption value of health is specified, then these effects can also be expressed as their health equivalent and included in the expression for net health benefit. Alternatively, impacts on health, health care resources and consumption can also be expressed in terms of the equivalent consumption effects (see Sections 4.3.3 and 4.3.4). Therefore, the methods of analysis outlined in this chapter are not restricted to cost-effectiveness analysis applied in health care systems where health expenditure is constrained and/or where decision-making bodies disregard effects outside that system (see Section 4.5.3). It is just as relevant to an appropriately conducted cost-benefit analysis (i.e. one which accounts for the impact of any constraints on health care expenditure). Equally, the principles of value-of-information analysis can be usefully applied even in circumstances where decision-making bodies are unwilling or unable to explicitly include any form of economic analysis in their decision-making process.

For example, a quantitative assessment of the expected health (rather than net health) benefits of additional evidence is possible by applying value-of-information analysis to the results of standard methods of systematic review and meta-analysis (Claxton et al. 2013).

References

Ades, A.E., Lu, G., and Claxton, K. (2004). Expected value of sample information in medical decision modelling. *Medical Decision Making*, **24**, 207–27.

Basu, A. (2011). Economics of individualization in comparative effectiveness research and a basis for a patient-centered health care. *Journal of Health Economics*, **30**, 549–59.

Basu, A. (2015). Welfare implications of learning through solicitation versus diversification in health care. *Journal of Health Economics*, doi 10.1016/j.jhealeco.2015.04.001.

Basu, A. and Meltzer, D. (2007). Value of information on preference heterogeneity and individualized care. *Medical Decision Making*, **27**, 112–27.

Bojke, L. and Soares, M. (2014). Decision analysis: eliciting experts' beliefs to characterize uncertainties, in A.J. Culyer (ed.), *Encyclopedia of health economics*. Amsterdam: Elsevier.

Bojke, L., Claxton, K., Sculpher, M., et al. (2009). Characterizing structural uncertainty in decision-analytic models: a review and application of methods. *Value in Health*, **12**, 739–49.

Bojke, L., Claxton, K., Sculpher, M., et al. (2010). Eliciting distributions to populate decision analytic models. *Value in Health*, **13**, 557–64.

Brennan, A., Kharroubi, S., O'Hagan, A., et al. (2007). Calculating partial expected value of perfect information via Monte Carlo sampling algorithms. *Medical Decision Making*, **27**, 448–70.

Briggs, A. and Sculpher, M. (1995). Sensitivity analysis in economic evaluation: a review of published studies. *Health Economics*, **4**, 355–71.

Briggs, A., Claxton, K., and Sculpher, M. (2006). *Decision modelling for health economic evaluation*. Oxford: Oxford University Press.

Briggs, A., Ritchie, K., Fenwick, E., et al. (2010). Access with evidence development in the UK: past experience, current initiatives and future potential. *PharmacoEconomics*, **28**, 163–70.

Briggs, A.H., Weinstein, M.C., Fenwick, E.A.L., et al. (2012). Model parameter estimation and uncertainty: A report of the ISPOR-SMDM Modelling Good Research Practices Task Force Working Group-6. *Medical Decision Making*, **32**, 722–32.

CADTH [Canadian Agency for Drugs and Technologies in Health] (2006). *Guidelines for the economic evaluation of health technologies: Canada*, 3rd edition. Ottawa: CADTH.

Chalkidou, K., Hoy, A., and Littlejohns, P. (2007). Making a decision to wait for more evidence: when the National Institute for Health and Clinical Excellence recommends a technology only in the context of research. *Journal of the Royal Society of Medicine*, **100**, 453–60.

Claxton, K. (1999). The irrelevance of inference: a decision-making approach to the stochastic evaluation of health care technologies. *Journal of Health Economics*, **18**, 342–64.

Claxton, K. (2008). Exploring uncertainty in cost-effectiveness analysis. *PharmacoEconomics*, **26**, 781–98.

Claxton, K. and Sculpher, M.J. (2006). Using value of information analysis to prioritise health research: some lessons from recent UK experience. *PharmacoEconomics*, **24**, 1055–68.

Claxton, K., Sculpher, M., McCabe, C., et al. (2005). Probabilistic sensitivity analysis for NICE technology assessment: not an optional extra. *Health Economics*, **14**, 339–47.

Claxton, K., Palmer, S.J., Longworth, L., et al. (2012). Informing a decision framework for when NICE should recommend the use of health technologies only in the context of an appropriately designed programme of evidence development. *Health Technology Assessment*, **16**(46), 1–342.

Claxton, K., Griffin, S., Koffijberg, H., et al. (2013). *Expected health benefits of additional evidence: principles, methods and applications.* Centre for Health Economics Research Paper 83. York: Centre for Health Economics, University of York.

Claxton, K., Martin, S., Soares, M., et al. (2015). Methods for the estimation of the NICE cost-effectiveness threshold. *Health Technology Assessment*, **19**(14).

Colbourn, T., Asseburg, C., Bojke, L., et al. (2007). Preventive strategies for group B streptococcal and other bacterial infections in early infancy: cost-effectiveness and value of information analyses. *BMJ*, **335**, 655–62.

Conti, S. and Claxton, K. (2009). Dimensions of design space: a decision theoretic approach to optimal research portfolio design. *Medical Decision Making*, **29**, 643–60.

Eckermann, S. and Willan, A. (2006). Expected value of information and decision-making in HTA. *Health Economics*, **16**, 195–209.

Eckermann, S. and Willan, A.R. (2008). The option value of delay in health technology assessment. *Medical Decision Making*, **28**, 300–5.

Espinoza, M.A., Manca, M., Claxton, K., et al. (2014). The value of heterogeneity for cost-effectiveness subgroup analysis: conceptual framework and application. *Medical Decision Making*, **34**, 951–64.

Fenwick, E., Claxton, K., and Sculpher, M. (2001). Representing uncertainty: the role of cost-effectiveness acceptability curves. *Health Economics*, **10**, 779–89.

Fenwick, E., Claxton, K., and Sculpher, M. (2008). The value of implementation and the value of information: combined and uneven development. *Medical Decision Making*, **28**, 21–32.

Fleurence, R.L. and Meltzer, D.O. (2013). Toward a science of research prioritization? The use of value of information by multidisciplinary stakeholder groups. *Medical Decision Making*, **33**, 460–2.

Forster, M. and Pertile, P. (2013). Optimal decision rules for HTA under uncertainty: a wider, dynamic perspective. *Health Economics*, **22**, 1507–14.

Garrison, L.P., Towse, A., Briggs, A., et al. (2013). Performance-based risk-sharing arrangements—good practices for design, implementation, and evaluation: report of the ISPOR Good Practices for Performance-Based Risk-Sharing Arrangements Task Force. *Value in Health*, **16**, 703–19.

Glick, H.A., Doshi, J.A., Sonnad, S.S., et al. (2014). *Economic evaluation in clinical trials*, 2nd edition. Oxford: Oxford University Press.

Griffin, S., Claxton, K., Hawkins, N., et al. (2006). Probabilistic analysis and computationally expensive models: necessary and required? *Value in Health*, **9**, 244–52.

Griffin, S., Claxton, K., and Welton, N. (2010). Exploring the research decision space: the expected value of sequential research designs. *Medical Decision Making*, **30**, 155–62.

Griffin, S., Claxton, K., Palmer, S., et al. (2011). Dangerous omissions: the consequences of ignoring decision uncertainty. *Health Economics*, **29**, 212–24.

Grigore, B., Peters, J., Hyde, C., et al. (2013). Methods to elicit probability distributions from experts: a systematic review of reported practice in health technology assessment. *PharmacoEconomics*, **31**, 991–1003.

Hoeting, J.A., Madigan, D., Raftery, A.E., et al. (1999). Bayesian model averaging: a tutorial. *Statistical Science*, **14**, 382–417.

Hoomans, T., Fenwick, E., Palmer, S., et al. (2009). Value of information and value of implementation: Application of an analytical framework to inform resource allocation decisions in metastatic hormone-refractory prostate cancer. *Value in Health*, **12**, 315–24.

Jackson, C.H., Thompson, S.G., and Sharples, L.D. (2009). Accounting for uncertainty in health economic decision models by using model averaging. *Journal of the Royal Statistical Society Series A*, **172**, 383–404.

Kharroubi, S.A., Brennan, A., and Strong, M. (2011). Estimating expected value of sample information for incomplete data models using Bayesian approximation. *Medical Decision Making*, **31**, 839–52.

McKenna, C. and Claxton, K. (2011). Addressing adoption and research design decisions simultaneously: the role of value of sample information analysis. *Medical Decision Making*, **31**, 853–65.

Meltzer, D.O., Hoomans, T., Chung, J.W., et al. (2011). Minimal modeling approaches to value of information analysis for health research. *Medical Decision Making*, **31**, E1–E22.

Miller, F.G. and Pearson, S.D. (2008). Coverage with evidence development: ethical issues and policy implications. *Medical Care*, **46**, 746–51.

Mohr, P.E. and Tunis, S.R. (2010). Access with evidence development: the US experience. *PharmacoEconomics*, **28**, 153–62.

NICE [National Institute for Health and Clinical Excellence] (2013). *Updated guide to the methods of technology appraisal*. London: NICE.

Niezen, M., de Bont, A., Stolk, E., et al. (2007). Conditional reimbursement with the Dutch drug policy. *Health Policy*, **84**, 39–50.

O'Hagan, A., Stevenson, M., and Madan, J.S. (2006a). Monte Carlo probabilistic sensitivity analysis for patient-level simulation models: efficient estimation of mean and variance using ANOVA. *Health Economics*, **16**, 1009–23.

O'Hagan, A., Buck, C.E., Daneshkhah, A., et al. (2006b). *Uncertain judgements: eliciting experts' probabilities*. Chichester: Wiley.

Palmer, S. and Smith, P.C. (2000). Incorporating option values into the economic evaluation of health care technologies. *Journal of Health Economics*, **19**, 755–66.

Pearson, S.D., Miller, F.G., and Emanuel, E.J. (2006). Medicare's requirement for research participation as a condition of coverage: is it ethical? *JAMA*, **296**, 988–91.

Pertile, P., Forster, M., and La Torre, D. (2014). Optimal Bayesian sequential sampling rule for the economic evaluation of health technologies. *Journal of the Royal Statistical Society Series A*, **177**(Part 2), 419–38.

Philips, Z., Claxton, K., and Palmer, S. (2008). The half-life of truth: What are the appropriate time horizons for research decisions? *Medical Decision Making*, **28**, 287–99.

Pratt, J., Raiffa, H., and Schlaifer, R. (1995). *Statistical decision theory*. Cambridge, MA: MIT Press.

Price, M.J., Welton, N.J., Briggs, A.H., et al. (2011). Model averaging in the presence of structural uncertainty about treatment effects: influence on treatment decision and expected value of information. *Value in Health*, **14**, 205–18.

Soares, M.O., Bojke, L., Dumville, J.C., et al. (2011). Methods to elicit experts' beliefs over uncertain quantities: application to a cost-effectiveness transition model of negative pressure wound therapy for severe pressure ulceration. *Statistics in Medicine*, **30**, 2363–80.

Soares, M.O., Dumville, J.C., Ashby, R.L., et al. (2013). Methods to assess cost-effectiveness and value of further research when data are sparse: negative-pressure wound therapy for severe pressure ulcers. *Medical Decision Making*, **33**, 415–36.

Soeteman, D.I., Busschbach, J.J., Verheul, R., et al. (2011). Cost-effective psychotherapy for personality disorders in the Netherlands: the value of further research and active implementation. *Value in Health*, **14**, 229–39.

Stevenson, M.D. and Jones, M.L. (2011). The cost-effectiveness of a randomized controlled trial to establish the relative efficacy of vitamin K1 compared with alendronate. *Medical Decision Making*, **31**, 43–52.

Stevenson, M.D., Oakley, J., and Chilcott, J.B. (2004). Gaussian process modelling in conjunction with individual patient simulation modelling: a case study describing the calculation of cost-effectiveness ratios for the treatment of established osteoporosis. *Medical Decision Making*, **24**, 89–100.

Strong, M., Oakley, J.E., and Chilcott, J. (2012). Managing structural uncertainty in health economic decision models: a discrepancy approach. *Journal of the Royal Statistical Society Series C*, **61**, 25–45.

Strong, M., Oakley, J.E., and Brennan, A. (2014). Estimating multi-parameter partial expected value of perfect information from a probabilistic sensitivity analysis sample: a non-parametric regression approach. *Medical Decision Making*, **34**, 311–26.

Strong, M., Brennan, A., and Oakley, J.E. (2015). An efficient method for computing the expected value of sample information—a non-parametric regression approach. *Medical Decision Making*, doi: 10.1177/0272989X15575286.

Tuffaha, H.W., Reynolds, H., Gordon, L.G., et al. (2014). Value of information analysis optimizing future trial design from a pilot study on catheter securement devices. *Clinical Trials*, **11**, 648–56.

Tunis, S.R. and Pearson, S.D. (2006). Coverage options for promising technologies: Medicare's 'coverage with evidence development'. *Health Affairs*, **25**, 1218–30.

Van Hout, B.A., Al, M.J., Gordon, G.S., et al. (1994). Costs, effects and c/e-ratios alongside a clinical trial. *Health Economics*, **3**, 309–19.

Walker, S., Sculpher, M., Claxton, K., et al. (2012). Coverage with evidence development, only in research, risk sharing, or patient access scheme? A framework for coverage decisions. *Value in Health*, **15**, 570–9.

Welton, N.J., Ades, A.E., Caldwell, D.M., et al. (2008). Research prioritization based on expected value of partial perfect information: a case-study on interventions to increase uptake of breast cancer screening. *Journal of the Royal Statistical Society Series A*, **171**, 807–41.

Welton, N.J., Madan, J.J., Caldwell, D.M., et al. (2014). Expected value of sample information for multi-arm cluster randomized trials with binary outcomes. *Medical Decision Making*, **34**, 352–65.

Willan, A.R. and Eckermann, S. (2010). Optimal clinical trial design using value of information methods with imperfect implementation. *Health Economics*, **19**, 549–61.

Yoyota, F. and Thompson, K.M. (2004). Value of information literature analysis: a review of applications in health risk management. *Medical Decision Making*, **24**, 287–98.

Chapter 12

How to take matters further

12.1 Taking matters further

We hope that, after reading this book, you have a taste for economic evaluation in health care and want to pursue your interests. In this concluding chapter we offer a few practical hints.

12.2 Further reading and key sources of literature

The other major general text on economic evaluation in health care is the report of the work of the Washington Panel by Gold et al. (1996), shortly to be updated. There is also another recent text by Hunink et al. (2014). In addition, a series of workbooks, published by Oxford University Press, go into more detail on various aspects of economic evaluation, including trial-based analyses (Glick et al. 2014), modelling (Briggs et al. 2006), quality of life measurement (Brazier et al. 2007), cost-effectiveness analysis (Clarke et al. 2011), and cost-benefit analysis (McIntosh et al. 2010).

Also, it is important to keep abreast of the literature in applied studies and methodological developments. These are published in a wide range of sources, including several specialist clinical journals in different fields. Journals to consult regularly include the *European Journal of Health Economics, Health Economics*, the *International Journal of Technology Assessment in Health Care, Medical Decision Making, PharmacoEconomics*, and *Value in Health*.

Finally, the Centre for the Evaluation of Value and Risk in Health (CEVR) at Tufts New England Medical Center maintains a registry of economic evaluations in health care (<https://research.tufts-nemc.org/cear4/>). The CEA Registry is a comprehensive database of around 4000 cost-effectiveness studies using quality-adjusted life-years (QALYs), covering a wide variety of diseases and treatments, and is a useful starting point for a literature search for economic evaluations on a particular clinical topic, or studies embodying a particular methodological approach.

12.3 Planning and undertaking an economic evaluation

Most empirical studies are undertaken by a multidisciplinary team, containing the relevant skills in health economics, epidemiology, medical statistics, information science, and relevant clinical expertise (depending on the disease or topic being studied).

The starting point should always be to pose the question 'who needs this study and why?' This is critical because, to a large extent, it will determine the study perspective, the alternative options to be compared and, in some cases, key elements of the methods to be used.

In some situations the audience for the study will be self evident, particularly if it is being conducted to inform a defined reimbursement decision or recommendation, such as to list a drug on a national formulary or to initiate a vaccination or screening programme. Often the decision-making body requesting the study may have published methods guidance, as discussed in Chapter 3. If the study is being performed to guide decisions in general, it is still advisable to have a decision-maker, or decision-makers, in mind, whether these be individual clinical practitioners in the field concerned, the head of a hospital department, or the manager of a health plan.

The reason is obvious. The point of undertaking applied research is that someone will act upon it, so the study needs to be relevant to the setting in which the decision is being taken, reflecting its objectives and constraints (in particular the budgetary responsibilities). The responsibility of the decision-maker regarding delaying decisions, requiring or commissioning research, or negotiating the price of a technology are also important in planning appropriate uncertainty analysis (see Chapter 11). On some occasions it might be useful to go even further and anticipate some of the issues in implementing the results of the research (e.g. the costs of any necessary organizational changes), although we appreciate that many researchers consider this to be someone else's responsibility.

If the decision-making context really cannot be specified, although we think that it almost always can, the only option would be to conduct a study considering a broad range of costs falling on different decision-makers' budgets comparing all the relevant alternatives, in the hope that it will appeal to one or more unspecified decision-makers some time in the future! At any rate, it is always useful to try to specify the study question as precisely as possible, as outlined in Chapter 3 (Section 3.2.1).

Certainly, it will be important to anticipate any publication of the research. This includes being sure that you will be able to meet generally accepted criteria for reporting economic evaluations (e.g. those implied by the checklist in Chapter 3, Box 3.1) and the ethical principles specified by the International Committee of Medical Journal Editors (<http://www.icmje.org>).

12.4 Expanding your network in economic evaluation

One of the best ways of enhancing your knowledge and expertise in economic evaluation is to interact with others in the field. If you work in a large research centre this will be relatively easy, but many analysts working in the field may not have this opportunity. Fortunately, there are a number of scientific societies and associations that hold regular scientific meetings where you will have the opportunity to meet other researchers.

The largest and most broadly based society is the International Society for Pharmacoeconomics and Outcomes Research (ISPOR). This holds one conference in North America and one in Europe every year, with meetings in alternate years in Asia and South America. In addition, the society's website (<http://www.ispor.org>) contains considerable amounts of useful information, including several Good Research Practices Task Force Reports, covering many of the aspects of economic evaluation discussed in this book.

The Society for Medical Decision Making (SMDM) (<http://www.smdm.org>) holds annual meetings in North America and a biannual meeting in Europe. Membership of this society is of particular use to researchers interested in decision-analytic modelling and methodological developments in economic evaluation.

Health Technology Assessment International (HTAi) (<http://www.htai.org>) is also a broadly based society, which has a substantial membership from the organizations and agencies that conduct or use economic evaluations in their decision-making processes. This society holds an annual meeting in different regions of the world and should be of particular interest to researchers concerned with the interface between research and policy-making.

Finally, the main professional associations of health economists also hold conferences that include discussions on economic evaluation, alongside other research areas including health care financing and organization, health care markets and incentives, health system performance, health econometrics, and global health. The main associations are the International Health Economics Association (<http://www.healtheconomics.org>), the American Society of Health Economists (<http://www.ashecon.org>), and the European Health Economics Association (<http://www.euhea.eu>). There are also many national associations of health economists.

12.5 Looking to the future

Through four editions of this book, spanning 28 years, we have documented how the field of economic evaluation in health care has advanced, both in terms of methods and practice. No doubt further advances will be made in the near future. We cannot anticipate those here, nor prevent the book becoming out of date. However, we do hope that, by reading the book, you become quickly conversant with the methods of economic evaluation and make your own contribution to this rapidly developing field.

References

Brazier, J.E., Ratcliffe, J., Saloman, J.A., and Tsuchiya, A. (2007). *Measuring and valuing health benefits for economic evaluation*. Oxford: Oxford University Press.

Briggs, A.H., Claxton, K., and Sculpher, M.J. (2006). *Decision modeling for health economic evaluation*. Oxford: Oxford University Press.

Clarke, P.M., Wolstenholme, J.L., and Wordsworth, S. (2011). *Applied methods of cost-effectiveness analysis in health care*. Oxford: Oxford University Press.

Glick, H., Doshi, J.A., Sonnad, S.S., and Polsky, D. (2014). *Economic evaluation in clinical trials*, 2nd edition. Oxford: Oxford University Press.

Gold, M.R., Siegel, J.E., Russell, L.B., and Weinstein, M.C. (1996). *Cost-effectiveness in health and medicine*. New York: Oxford University Press.

Hunink, M.G., Weinstein, M.C., Wittenberg, E., Drummond, M.F., Pliskin, J.S., Wong, J.B., and Glasziou, P.P. (2014). *Decision making in health and medicine: integrating evidence and values*, 2nd edition. Cambridge: Cambridge University Press.

McIntosh, E., Louviere, J.J., Frew, E. and Clarke, P.M. (2010). *Applied methods of cost-benefit analysis in health care*. Oxford: Oxford University Press.

Author index

S

T

U

V

Subject index

D

T

U