eunethta

EUROPEAN NETWORK FOR HEALTH TECHNOLOGY ASSESSMENT

GUIDANCE DOCUMENT

Practical considerations when critically assessing economic evaluations

Version 1.0, 09 March 2020

European Network for Health Technology Assessment | JA3 2016-2020
www.eunethta.eu

The primary objective of EUnetHTA methodological guidelines is to focus on methodological challenges that are encountered by HTA assessors while performing relative effectiveness assessments of pharmaceuticals or non-pharmaceutical health technologies.

Version	Description of changes	Funding	Date of publication
1.0	First edition	EUnetHTA JA3	09/03/2020

Guideline team

Authoring team	Author / contact person	• Mattias Neyt (KCE, Belgium)
	Co-authors	• Martin Eriksson (SBU, Sweden) • Lidia García-Pérez (SESCS-FUNCANIS, Spain) • Pia Johansson (SBU, Sweden) • Fabienne Midy (HAS, France) • Conor Teljeur (HIQA, Ireland)
Dedicated Reviewers		• Marius Ciutan (NSPHMPDB, Romania) • Silvia Florescu (NSPHMPDB, Romania) • Emer Fogarty (NCPE, Ireland) • Judith Gibbert (IQWIG, Germany) • Anna Kaczorek-Juszkiewicz (AOTMIT, Poland) • Saskia Knies (EUR, Netherlands) • Miriam Luhnen (IQWIG, Germany) • Boglarka Mikudina (NICE, United Kingdom) • Peter O'Neil (NICE, United Kingdom) • Kinga Orzel (AOTMIT, Poland) • Ingrid Rosian (GÖG, Austria) • Anja Schwalm (IQWIG, Germany) • Vassilia Verdiel (NICE, United Kingdom) • Alicja Wojcik (AOTMIT, Poland)
Observers		• none

Acknowledgments

Internal review	• AETSA (Spain) • AOTMIT (Poland) • AQuAS (Spain) • BAG-SNHTA (Switzerland) • EUR (Netherlands) • GÖG (Austria) • IQWIG (Germany) • LBI-HTA (Austria) • NICE (United Kingdom) • OGYEI (Hungary)
Project management	• Mattias Neyt (KCE, Belgium) • Patrice Chalon (KCE, Belgium)
Graphical edition	• Patrice Chalon (KCE, Belgium) • Ine Verhulst (KCE, Belgium)
Internal coordination	• Chantal Guilhaume (HAS, France) • Anna Zaremba (AOTMIT, Poland) • Anna Zawada (AOTMIT, Poland)

Contact

EUnetHTA Secretariat: eunethta@zin.nl

Funding

This guidance document is part of the joint action '724130 / EUnetHTA JA3' which has received funding from the European Union's Health Programme (2014-2020).

Disclaimer

The content of this Guidance document represents a consolidated view based on the consensus within the Authoring Team; it cannot be considered to reflect the views of the European Network for Health Technology Assessment (EUnetHTA), EUnetHTA's participating institutions, the European Commission and/or the Consumers, Health, Agriculture and Food Executive Agency or any other body of the European Union. The European Commission and the Agency do not accept any responsibility for use that may be made of the information it contains.

Copyrights

How to cite this document

EUnetHTA JA3WP6B2-5 Authoring Team. Practical considerations when critically assessing economic evaluations. Guidance document. Diemen (The Netherlands): EUnetHTA; 2020. Available from https://www.eunethta.eu/

Table of contents

Acronyms - Abbreviations

Acronym	Full term
15D	15 Dimension instrument
ADE	Adverse drug event
AdViSHE	Assessment of the Validation Status of Health-Economic decision models
AE	Adverse event
AUC	Area under the curve
BMS	Bare-metal stents
C	Costs
CAD	Coronary artery disease
CADTH	Canadian Agency for Drugs and Technologies in Health
CEA	Cost-effectiveness analysis
CEAC	Cost-effectiveness acceptability curve
CHD	Coronary heart disease
CHEC	Consensus on Health Economic Criteria
CHEERS	Consolidated Health Economic Evaluation Reporting Standards
CoI	Conflict of interest
CONSORT	Consolidated Standards of Reporting Trials
CRD	Centre for Reviews and Dissemination
CRT(-P/D)	Cardiac resynchronization therapy (biventricular pacemakers/biventricular defibrillators)
CUA	Cost-utility analysis
DALY	Disability-adjusted life years
DES	Drug-eluting stents
E	Effects
EED	Economic Evaluation Database
EMA	European Medicines Agency
EQ-5D	EuroQoL-5 dimension
ERG	Evidence Review Group
EU	European Union
EUCTR	EU Clinical Trials Register
EUnetHTA	European Network for Health Technology Assessments
EVPI	Expected value of perfect information
FDA	Food and Drug Administration
FUNCANIS	Fundación Canaria de Investigación Sanitaria (Spain)
GDP	Gross domestic product
GNU	Gnu is Not Unix
HF	Heart failure
HIQA	Health Information and Quality Authority
HIV	Human immunodeficiency virus
HPV	Human papillomavirus
HRQoL	Health-related quality of life
HRT	Hormone replacement therapy
HTA	Health Technology Assessment
HUI	Health Utilities Index
ICD	Implantable cardioverter defibrillators
ICER	Incremental cost-effectiveness ratio

INAHTA	International Network of Agencies for Health Technology Assessment
INB	Incremental net benefit
IQWiG	German Institute for Quality and Efficiency in Health Care
ISPOR	International Society for Pharmacoeconomics and Outcomes Research
ITR	Inverted T-shaped reduction
ITT	Intention to treat
JBI	Joanna Briggs Institute
KCE	Belgian Health Care Knowledge Centre
KM	Kaplan Meier
LYG	Life-years gained
LYS	Life-years saved
MACE	Major adverse cardiac events
MI	Myocardial infarction
MWA	Microwave ablation
mRS	Modified Rankin Scale
nbDMARDs	Non-biological disease-modifying anti-rheumatic drugs
NICE	National Institute for Health and Care Excellence
NHB	Net health benefit
NIH	National Institutes of Health
NIHR	National Institute for Health Research
NMB	Net monetary benefit
NRS	Non-randomised studies
NSPHMPDB	National School of Public Health, Management and Professional Development, Bucharest
NYHA	New York Heart Association
OMERACT	Outcome Measures in Rheumatology Task Force
OS	Overall survival
p*q table	Price and quantities tables
PartSA	Partitioned survival analysis
PBAC	Pharmaceutical Benefits Advisory Committee
PF-LYS	Progression-free life-year saved
PFS	Progression-free survival
PrEP	Pre-exposure prophylaxis
PSA	Probabilistic sensitivity analysis
QALY	Quality-adjusted life years
QoL	Quality of life
RA	Rheumatoid arthritis
RCT	Randomized controlled trial
REA	Relative effectiveness assessment
RoB	Risk of bias
RR	Relative risk
SC	Standard care
SESCS	Servicio de Evaluación del Servicio Canario de la Salud (HTA unit at The Canary Islands, Spain)
SF-6D	Short Form-6 dimension
SG	Standard gamble
TLR	Target lesion revascularisation
TNF	Tumour necrosis factor
TTO	Time-trade-off
VAS	Visual analogue scale

VSR	Vertical scar reduction
WHI	Women's Health Initiative
WHO	World Health Organization
WTP	Willingness to pay
λ	Willingness to pay

List of examples

Summary and table with main points for consideration

If colleagues ask a health economic expert how to start an assessment of an economic evaluation, then often reference is made to the existing national guidelines for conducting such studies, as well as to various checklists for reviewing them. In addition to being aware of the existence of these guidelines and checklists, experience also plays a major role. Addressing a possible problem by giving an example often increases the clarity of critical assessments.

The authors of this guidance document try to pass on some of their knowledge and know how to all stakeholders with an interest in economic evaluations. Instead of limiting this report to a theoretical overview of issues, we have tried to provide an overview of important points for consideration when performing/assessing such evaluations. This overview is supported by a selection of real-world examples. These examples are not used to criticize previous work performed by researchers working in health technology assessment (HTA) bodies, industry, university teams or government institutes. They are not used to replace or overrule national guidelines for economic evaluations. Neither are they used to convince readers about the cost-effectiveness of a specific intervention or whether a specific conclusion is correct. The examples are used from an educational perspective to support those who want to perform or assess an economic evaluation. As mentioned by one of the reviewers after reading the first draft of this document: "*what can be learned in thousands of hours of training, reading and working is disclosed in a very attractive manner. … The idea of inserting examples from the real-world is outstanding.*"

Whatever your background, we hope you as a reader also interpret this guidance not as a criticism of flawed approaches to economic modelling, but rather as a supportive tool for a better understanding, appropriate critical assessment and (re)use of economic evaluations. As part of a Health Technology Assessment, such evaluations can provide support to decision makers when pursuing an accessible health care system which is both financially sustainable and of the highest quality.

A selection of points for consideration

In this report, we examine points for consideration when performing/assessing an economic evaluation. Not all listed points will apply to a single evaluation. Depending on the subject, certain remarks will have more importance and others will not be relevant at all. We don't tell readers when to rely on or ignore the conclusions of a study. How assessors deal with a possible identified problem is another issue and very context-specific. It is also not possible to provide a 'one-size-fits-all solution' and we cannot cover everything. Potentially relevant issues might not be tackled in this document. Technology is very broad, not restricted to pharmaceuticals. This document is a general guidance document, not specifically for one type of intervention. The guidance document might help assessors decide which elements to focus on to be able to judge which evaluations are reliable or where you might ask for specific adjustments to be made. In the selection of issues and provided examples, the authors have tried to find a balance between not being too basic without becoming too technical. In doing so, we hope that the document will be of practical use.

The following table provides an overview of the elements that are discussed in this report, together with a selection of points for consideration. For more information, we refer to the respective parts of the full report.

Efficacy/effectiveness and safety (see 3.1)	• Has all evidence been taken into account to be able to make a balanced evaluation of the treatment effect? • Was the assessment of the efficacy/effectiveness carried out according to current standards? • Was the impact of adverse events on costs and benefits taken into account?
Comparator (see 3.2)	• Be aware of inappropriate exclusion of relevant (possibly more cost-effective) alternatives. • Be aware of incremental cost-effectiveness ratios (ICERs) calculated by comparing with inappropriate alternatives (inclusion of non-cost-effective alternatives).
Subgroup analysis (see 3.3)	• Inappropriate conclusions based on average measures of cost-effectiveness, if the cost-effectiveness of the assessed technologies varies between subgroups. • Identification of subgroup analyses based on non-clinical considerations may be required. • If heterogeneity of the relative treatment effect is not demonstrated between subgroups, an assumption of equivalence is made (same as the relative treatment effect observed in the intention-to-treat (ITT) population). • A subgroup analysis may be rational from a scientific point of view but may not be useful to the decision maker.
Baseline risk of the target population (see 3.4)	• Are differences in the baseline risk for specific events (e.g. rehospitalisation) in the randomized controlled trial (RCT) population versus the real-world population for which a decision is taken considered?
Compliance/adherence and persistence (see 3.5)	• Is adherence and/or persistence of importance for the technology under evaluation? • Consider whether adherence is lower in the target population than in the underlying trial(s). • Has evidence applicable to the research question been used to determine adherence in the target population?
Quality of life (see 3.6)	• Is good quality-of-life (QoL) data lacking? • Be aware of problems with mapping outcomes of disease-specific or generic instruments to utilities when no generic utility instruments has been used.
Intermediate/surrogate versus final endpoints (see 3.7)	• Avoid the use of non-validated surrogate endpoints. • Evidence on (the absence of) a link between surrogate and final endpoints should be taken into account. • Information on other endpoints (e.g. QoL & OS) should also be considered.

Time horizon and extrapolation (see 3.8)	• As a modelled time horizon extends, it is associated with increasing inherent uncertainty. • Especially in long-term models, the extrapolation assumptions on the relative treatment effect are crucial. • The immaturity of the available data creates great uncertainty on the extrapolation.
Discount rate (see 3.9)	• Especially in long-term models, applying different discount rates might have a large impact on results.
Perspective (see 3.10)	• Be aware of the impact that different perspectives might have on the inclusion of items and their valuation.
(Context-specific) costs (see 3.11)	• Cost items, resource use and prices should be reported separately and summarized in a prices * quantities (p*q) table. • Know the financing system in the country for which an analysis is performed to avoid incorrect inclusion of (context-specific) costs. • Beware of applied statistical tests, as patient costs are often skewed. • Adjustments for protocol-driven costs might be needed.
Uncertainty/sensitivity analysis & probability distributions (see 3.12)	• Has the uncertainty around model parameters been presented and is the imprecision clearly linked to an evidence base? • Have sensitivity and scenario analyses been presented to sufficiently explore uncertainty in the model outputs? • Has the proper probability distribution function been selected to incorporate parameter uncertainty? • Has uncertainty been adequately taken into account in the interpretation of the findings?
Model verification and validation (see 3.13)	• Has the model implemented the assumptions correctly (model verification)? • Are the assumptions reasonable and do they reflect reality (model validation)? • Are results consistent with results from other studies and can identified differences be explained?
Transferability of economic evaluation results (see 3.14)	• Transparent reporting is necessary to assess transferability. • Transferability of all major factors (i.e. resource use, unit costs, effectiveness, QoL weights) should be considered.

ICER threshold (see 3.15)	• Cost-effectiveness is not the only criterion to make decisions. Each country might apply different decision making rules and other factors (other than the results of an economic evaluation) might influence the final decision. • In some cases, authors compare with an ICER threshold without any explanation/justification for the selection of the cost-effectiveness threshold (or range of thresholds). Authors might e.g. compare with relatively high ICER thresholds that are not accepted in their country at that moment of time. • Presenting results on the cost-effectiveness acceptability curve (CEAC) might facilitate the interpretation of outcomes (e.g. by allowing the reader to apply different ICER thresholds when there is no explicit threshold included in the national guidelines).
Publication bias of economic evaluations and conflicts of interest (see 3.16)	• Be aware that in general industry-sponsored studies are more likely to report favourable cost-effectiveness results. • Industry-sponsored studies are more likely to report favourable qualitative conclusions than non-profit-sponsored studies. • Be aware that economic evaluations may be subject to publication bias.

1 Introduction

1.1 Background and problem statement

Most European countries have their own specific national guidelines for conducting economic evaluations. An overview of these guidelines is available in the European Network for Health Technology Assessments (EUnetHTA) report “Methods for health economic evaluations - A guideline based on current practices in Europe (May, 2015)”.[1] Having such guidelines is necessary to improve consistency, relevance and transparency. It guides both those who perform and assess economic evaluations, as well as researchers setting up research protocols. For the latter group, it is important to know from the beginning, i.e. when setting up studies, what researchers will need later on (e.g. when they want to perform an economic evaluation to support a reimbursement request). As such, they can avoid losing a lot of time (and money), e.g. at the reimbursement request when economic considerations might be taken into account.

There also exist various guidelines and checklists for the transparent reporting of economic evaluations (e.g. the Consolidated Health Economic Evaluation Reporting Standards (CHEERS)[2] or the Drummond guidelines[3]), methodological quality checklists of economic evaluations (e.g. the Consensus on Health Economic Criteria (CHEC)-list,[4] or the Joanna Briggs Institute (JBI) Critical Appraisal Checklist for Economic Evaluations[5]) or to evaluate the quality of decision models in health technology assessment (e.g. the Good Practice Guidelines for Decision-Analytic Modelling set up by Philips et al.[6]). These guidelines support authors, peer reviewers, editors, policy makers and other stakeholders in identifying all relevant items in an economic evaluation.

During a EUnetHTA workshop[a] on identifying gaps in existing guidelines and opportunities for developing new guidelines, several participants proposed to make a practical guideline on the critical assessment of economic evaluations. Reporting guidelines are very helpful for both researchers writing down the study results of their economic evaluation, as well as assessors identifying the relevant elements when reading such studies. Transparent reporting of the input variables and the assumptions made is necessary to enable a critical evaluation. Nevertheless, reporting guidelines do not say anything about the reliability or relevance of the results for a policy maker in a specific context. Critical assessment is the necessary next step. Existing quality checklists give an overview of the relevant questions that should be asked when reviewing an economic evaluation. They question, for example, whether competing alternatives are clearly described,[4] or whether the study discusses the generalisability of the results to other settings and patient/client groups.[4] However, for assessors and modellers, it is not always clear what the points for consideration are when performing or assessing an economic evaluation. A practical guideline to support this task has been missing, while reliability/quality assurance of the performed economic evaluations

[a] The workshop was organized on 15-16 October, 2016. In preparation of this workshop, a draft overview of existing guidelines, tools and templates that are useful for performing HTA was made. During this workshop, several methodological guidelines were suggested to be developed within EUnetHTA’s work package 6-B2 of Joint Action 3. Not surprisingly, many more topics were proposed then we actually could work out within this work package. Two topics were selected for which there was to a large extent a need to develop transparent guidelines to be able to support researchers, assessors, policy makers and other stakeholders as much as possible in making or (re)using HTA reports: 1) a guideline on the critical assessment of clinical evaluations; and 2) a guideline on the critical assessment of economic evaluations.

are crucial for the adaptation/adoption[b] of previously performed economic evaluations. In complement with existing national guidelines on economic evaluations, reporting and quality checklists, this guidance document provides a non-exhaustive overview of points for consideration when making a critical assessment of economic evaluations. This not only applies to economic evaluations of pharmaceuticals, but also of other interventions such as medical devices, diagnostics, prevention or screening campaigns, etc.

1.2 Objective(s) and scope of the guidance document

This guidance document tries to provide assistance in the critical assessment of relevant elements of an economic evaluation. This is done by providing an overview of the most common points for consideration in economic evaluations for each of the included elements. To enhance understanding, some of the provided points for consideration are supported with (real-world) examples presented in boxes. Where appropriate, reference is made to existing (EUnetHTA) guidelines.

The scope of the guidance document is limited to a non-exhaustive list of elements that are considered to be of importance when reviewing economic evaluations (see Table 1 in section 2). Two very important elements (treatment effect and safety) will be elaborated on in a separate guideline (Critical assessment of clinical evaluations – under construction) and are thus not elaborated on in detail in this report. For these elements, we will only include a limited selection of examples of points for consideration and refer to the more detailed guidelines (under construction). Explaining the standard methodology on how to perform economic evaluations is out of scope of this document. For this, we refer to the national guidelines[c] on economic evaluations and standard handbooks on this topic (e.g. Drummond[7] and Briggs[8]). This guidance document focusses on a non-exhaustive list of points for consideration. It does not tell researchers when results are reliable/useful or how to solve or deal with identified issues. This judgement is part of the critical assessment and might be case-specific.

The purpose of this guidance document is to support better critical assessment of economic evaluations. This is useful for stakeholders that: 1) assess existing economic evaluations (part of a systematic review or a reimbursement request) or 2) perform economic evaluations (e.g. as part of a full HTA report) and 3) other stakeholders interested in economic evaluations. The guidance document might help to identify the most credible existing economic evaluation(s) for a given research question and provide support to decide whether the existing evaluation is sufficient or, if this is not the case, whether the conduct of a de novo economic evaluation is required. It might, for example, also be useful for researchers when setting up research protocols, e.g. to have insights in what is needed to allow the creation of a high-quality economic evaluation. In the end, this guidance should facilitate the

[b] Adaptation: the systematic extraction of relevant HTA information from an existing report (from a whole report or from part of a report); Adoption: making use of the report without making any changes at all (except perhaps translation into your own language). (Source: EUnetHTA. Glossary of HTA Adaptation Terms, 2007, available from: https://www.eunethta.eu/wp-content/uploads/2018/01/Glossary-of-HTA-Adaptation-Terms.pdf).

[c] An overview of these national guidelines is provided in the EUnetHTA guideline “Methods for health economic evaluations - A guideline based on current practices in Europe”.[1] We remark that it is possible that updates of these national guidelines have been published since the publication of this report. The original source should thus be checked to identify the most up-to-date national guideline.

(re)use of good economic evaluations, both by institutions that perform full HTAs and set up economic evaluations themselves as well as by stakeholders that wish to make use of previously published economic evaluations.

1.3 Related (EUnetHTA) documents

Please also see the following related documents:

- Methods for health economic evaluations - A guideline based on current practices in Europe. Methodological Guideline: EUnetHTA; 2015.[1]
- HTA Adaptation Toolkit & Glossary. Version 5, 2011.[9]
- Endpoints used for Relative Effectiveness Assessment: health-related quality of life and utility measures. Methodological Guideline: EUnetHTA; 2015.[10]
- Comparators & comparisons: direct and indirect comparisons. Methodological Guideline: EUnetHTA; 2015.[11]
- Comparators & comparisons: Criteria for the choice of the most appropriate comparator(s). Methodological Guideline: EUnetHTA; 2015[12]
- Endpoints used for Relative Effectiveness Assessment: Clinical Endpoints. Methodological Guideline: EUnetHTA; 2015.[13]
- Levels of evidence: Applicability of evidence for the context of a relative effectiveness assessment. Methodological Guideline: EUnetHTA; 2015.[14]
- Endpoints used in Relative Effectiveness Assessment: Surrogate Endpoints. Methodological Guideline; 2015.[15]

For information on the terminology, key principles and approaches of modelling techniques, we refer to the following books (from a wide range of possible alternatives):

- Methods for the economic evaluation of health care programmes. 4th ed: Oxford University Press 2015.[7]
- Decision Modelling for Health Economic Evaluation: Oxford University Press August 2006.[8]
- Evidence-Based Decisions and Economics: Health Care, Social Welfare, Education and Criminal Justice. 2nd ed: Blackwell 2010.[16]

The methodology, input, assumptions and results should be published transparently to allow a critical assessment. We refer to the CHEERS and Drummond guidelines for standards on transparent reporting, as well as to a selection of critical assessment checklists[de]:

- Husereau D, Drummond M, Petrou S, Carswell C, Moher D, Greenberg D, et al. Consolidated Health Economic Evaluation Reporting Standards (CHEERS) statement. BMJ. 2013 Mar 25;346:f1049.[2]
- Drummond MF, Jefferson TO. Guidelines for authors and peer reviewers of economic submissions to the BMJ. The BMJ Economic Evaluation Working Party. BMJ. 1996 Aug 03;313(7052):275-83.[3] (~Drummond checklist)
- Evers S, Goossens M, de Vet H, van Tulder M, Ament A. Criteria list for assessment of methodological quality of economic evaluations: Consensus on Health Economic Criteria. Int J Technol Assess Health Care. 2005 Spring;21(2):240-5.[4] (~CHEC-list)
- Philips Z, Bojke L, Sculpher M, Claxton K, Golder S. Good practice guidelines for decision-analytic modelling in health technology assessment. Pharmacoeconomics. 2006;24(4):355–371.[21]
- Gomersall JS, Jadotte YT, Xue Y, Lockwood S, Riddle D, Preda A. Conducting systematic reviews of economic evaluations. Int J Evid Based Healthc. 2015 Sep;13(3):170-8.[5] (~the Joanna Briggs Institute (JBI) Critical Appraisal Checklist for Economic Evaluations)

[d] We refer to the publication of Wijnen et al.[17] for a more comprehensive overview of checklists to assess economic evaluations. The supplement of this article contains an overview table including eleven identified checklists.

[e] Some HTA bodies also have their own checklists, e.g.: Checklist for Assessing the Quality of Trial-Based Health Economic Studies. SBU, 2018;[18] Checklist for Assessing the Quality of Health Economic Modelling Studies. SBU, 2018;[19] AOTMiT. Health Technology Assessment Guidelines. Warsaw: AOTMiT; 2016 (p34-37).[20]

2 Methods

During the kick off e-meeting of the project 6B2-5 “Critical assessment of economic evaluations”, a list of ‘elements’ to be addressed in this guidance document was provided by the project leader. This was in the first place based on the elements that need to be presented when reporting on an economic evaluation, making use of the CHEERS guidelines.[2] During and after this meeting, based on the input and experience of the co-authors, this list was expanded. The following table provides an overview of the identified elements. As mentioned in the scope of the report, the first two elements are discussed in a separate guideline (under construction) and are only discussed briefly in this report.

Table 1: Elements suggested for inclusion in this guidance document on critical assessment of economic evaluations

Treatment effect (efficacy/ effectiveness)	Part 3.1.1
Safety	Part 3.1.2
Comparator	Part 3.2
Subgroup analysis	Part 3.3
Baseline risk	Part 3.4
Compliance/adherence	Part 3.5
Quality of life	Part 3.6
Intermediate/surrogate versus final endpoints	Part 3.7
Time horizon & Extrapolation	Part 3.8
Discount rate	Part 3.9
Perspective	Part 3.10
(Context-specific) costs	Part 3.11
Uncertainty/sensitivity analysis & probability distributions	Part 3.12
Model verification and validation (& model sharing)	Part 3.13
Transferability of economic evaluation results	Part 3.14
ICER threshold	Part 3.15
Publication bias of economic evaluations and conflicts of interest	Part 3.16

A systematic review of the literature on the critical assessment of economic evaluations was planned in the CRD (Centre for Reviews and Dissemination) HTA and NHS EED (Economic Evaluation Database) databases, Medline (OVID, both indexed and in-process citations), and EMBASE. However, the results of a first search in OVID (see Annex 1 – Documentation of literature search) performed separately for all elements, were rather disappointing (high number of searched references with only few relevant references identified after going through the title, abstract and keywords). This was mainly due to the non-standardised indexation of methodological literature in this field. Therefore, the co-authors decided to rely on: 1) the results from the OVID search, 2) a search for relevant guidelines from EUnetHTA, HTA organisations (EUnetHTA and International Network of Agencies for Health Technology Assessment (INAHTA) members), International Society for Pharmacoeconomics and Outcomes Research (ISPOR), and other grey literature; and 3) rely on the experience of the involved researchers.

The included points for consideration mentioned in the above table are a non-exhaustive list where the focus lies on the major issues, as identified by the involved experts, rather than points of methodology that only apply in very specific contexts. Examples of points for consideration are gathered through collaboration with health economists from the author and reviewer group and other HTA experts from different HTA institutions. The examples are not used to support or criticize the authors or results of a specific economic evaluation. They are selected from an educational point of view to support the reader of this document.

3 Overview of points for consideration to support the critical assessment of economic evaluations

In what follows, we provide an overview of points for consideration for all the elements that are listed in Table 1. Where possible, we refer to existing recommendations from, for example, EUnetHTA guidelines. Then we give a non-exhaustive list of a number of points for consideration, followed by a number of examples that are presented in boxes.

We recognise that some parts might be more or less technical or elaborated than others. This is based on the personal experience of the authors and the feedback received from the reviewers where for some elements many more points for consideration and examples were cited than for others.

3.1 Efficacy/effectiveness and safety

The reliability and applicability of the results of an economic evaluation depend in the first place on the applied treatment effect and impact of adverse events. In an HTA report, safety and efficacy/effectiveness are evaluated first and provide input for the economic evaluation. The critical assessment of these elements is thus of utmost importance. Another EUnetHTA guideline elaborates on this. We refer the reader to this guideline (under construction) for further details. Nevertheless, in line with the other parts of this guidance document, we refer to a selection of recommendations mentioned in other EUnetHTA guidelines and combine this with a selection of examples presented in boxes.

3.1.1 Efficacy/effectiveness

To be able to support evidence based medicine, all evidence should be available, both published and non-published. Publication and reporting bias should be avoided. Otherwise, evidence biased medicine is performed. We refer to Box 1 to show how an HTA body identified a major problem of publication bias and how they coped with this problem. In Box 2 we provide an overview of studies related to this issue. It is of great importance for the reliability of the results of the economic part of an assessment that a balanced assessment of all clinical evidence is performed since the results of the medical assessment are used as an input for the economic evaluation. Related to the issue of non-published evidence, the EUnetHTA guideline on information retrieval for systematic reviews and health technology assessments on clinical effectiveness[22] states that:

- *"A systematic review should regularly include a search for unpublished literature to identify both unpublished studies, and unpublished data from published studies."*

Randomized controlled trials (RCTs) are considered the golden standard to measure treatment effect. There is a EUnetHTA guideline that focuses on the assessment of the risk of bias of RCTs.[23] More attention is also going to the use of observational data. We refer to the article described in Box 3 comparing outcomes from RCTs and non-randomised studies, underlining the potential misinformation about the estimated treatment effect provided by the latter group of studies.[f] The EUnetHTA guideline on internal validity of non-

[f] This does not exclude that in some cases, it is impossible to perform RCTs, for example, for ethical reasons or because a clear benefit was observed through non-randomized studies. A frequently quoted non-medical example notes that no randomized studies have been performed for parachutes.[24] On the other hand, it is

randomised studies (NRS) on interventions,[26] intended to provide recommendations on the assessment of the internal validity of NRS used for the evaluation of effects of interventions, mentions the following:

- *"As the inclusion of non-randomised studies (NRS) in an HTA report requires large efforts (but often fails to increase the validity of the report's conclusion), the decision to do so should be made only after careful consideration of all advantages and dis-advantages."*
- *"Assessment of risk of bias (RoB) covers at least 5 different types of bias: selection bias (including bias due to confounding), performance bias, detection bias, attrition bias, and reporting bias."*

Bias in head-to-head comparisons and uncertainty linked to evidence relying on indirect comparisons instead of direct comparisons is also an important issue which should be taken into account when interpreting results of economic evaluations. In Box 4, we provide an example of bias in head-to-head comparisons were the outcomes depended on the study sponsor. We also refer to the EUnetHTA guideline on direct and indirect comparisons[11] which states that:

- *"The choice between direct and indirect comparison is context specific and dependent on the question posed as well as the different evidence available. Where sufficient good quality head-to-head studies are available, direct comparisons are preferred as the level of evidence is high. Should substantial indirect evidence be available, then it can act to validate the direct evidence. When there is limited head-to-head evidence or more than two treatments are being considered simultaneously, the use of indirect methods may be helpful."*
- *"An indirect comparison should only be carried out if underlying data from comparable studies are homogeneous and consistent, otherwise the results will not be reliable."*

Under certain circumstances, population-adjusted indirect comparisons may be used to correct for heterogeneity across studies. As with any evidence synthesis methodology, the approach used and underlying assumptions should be clearly reported, and there should be evidence that the chosen approach was appropriate.[27]

Finally, a general recommendation from the EUnetHTA guideline on clinical endpoints is that "both relative and absolute measures should be presented."[13] We refer to Box 5 for an illustration. In economic evaluations, the cost-effectiveness is driven by the absolute benefit. We refer to part 3.4 for more information on the importance of the baseline risk and the impact on the absolute treatment outcomes.

---

also important to point out that several medical examples do not stand the comparison with the parachute example. A study evaluated claims that a medical practice is akin to a parachute. The authors conclude that "*although we found that the parachute analogy is seldom used to describe a medical practice, when it is used it is often inappropriate, incorrect or misused.*"[25] But also in this paper, the authors nuance by mentioning that this "*does not imply that RCTs are always feasible, possible, necessary or ethical.*"[25]

Points for consideration

The input from the assessment of the clinical evidence is indispensable when performing or assessing an economic evaluation. Performing a critical assessment of the medical literature is not in the scope of this guidance document. For a more extensive analysis we refer to another EUnetHTA guideline (under construction). Nevertheless, one general point for consideration can be formulated in this respect:

- Researchers performing or assessing an economic evaluation must make sure that the assessment of the efficacy/effectiveness was carried out correctly. Was all relevant evidence provided by researchers to avoid publication and reporting bias? Were other types of bias assessed? Was appropriate sensitivity analysis performed? Was account taken of the uncertainty associated with indirect evidence? Etc.

Instead of repeating what has already been said in other guidelines, we prefer to refer to these other guidelines (https://www.eunethta.eu/methodology-guidelines/) and only include a number of examples in the following boxes related to the following issues:

- Publication bias (see Box 1 & Box 2)
- Evidence on the relative treatment effect from non-randomized studies (see Box 3)
- Bias in head-to-head comparisons (see Box 4)
- Relative and absolute treatment effect (see Box 5)

Examples

Box 1: The necessity of having all clinical evidence available to be able to make a proper assessment of the treatment effect

As an example, in 2009, the German Institute for Quality and Efficiency in Health Care (IQWiG) made an evaluation of the antidepressant reboxitine. "*To minimise the influence of publication bias and increase transparency, IQWiG requests manufacturers of drugs under assessment to sign a voluntary agreement requiring submission of a list of all sponsored published and unpublished trials; submission of CONSORT*[g] *compliant documents (generally the clinical study reports) on all relevant trials selected by IQWiG; and permission for publication of all previously unpublished relevant data in the assessment report.*"[28] The manufacturer provided a list of all published trials and documents provided to European authorities. Unfortunately, results of unpublished studies were not provided. A literature search brought to light that the drug was tested in at least 16 trials, including about 4600 patients. In contrast, results were only published for about 1600 of these patients.[28] "*There were insufficient data available for the majority of potentially relevant trials and patients. The assessment of the evidence at this point showed that further analysis of the limited data available would probably be seriously biased, as would any deduced conclusions on the proof of benefit or harm from reboxetine.*"[29] Therefore, IQWiG concluded it could not make a meaningful assessment of the drug due to the high risk of publication bias. Later on, the company also provided most of the unpublished information. The following assessment concluded that the drug had no benefit.[28] Assessors should be able to obtain all relevant information from studies in which the intervention was used, both published and unpublished, to make a balanced assessment.

[g] CONSORT stands for Consolidated Standards of Reporting Trials and encompasses various initiatives developed by the CONSORT Group to alleviate the problems arising from inadequate reporting of randomized controlled trials.(source: http://www.consort-statement.org/)

Box 2: The major problem of publication bias

Unfortunately, the example of publication bias in Box 1 is not an isolated case. It is often mentioned that about 50% of research is not published.[30-32] In the European Union (EU), since 2014, the regulation on clinical trials on medicinal products for human use,[33] requires (art. 37 (4)) that:

- "*Irrespective of the outcome of a clinical trial, within one year from the end of a clinical trial in all Member States concerned, the sponsor shall submit to the EU database a summary of the results of the clinical trial.*"

The European Medicines Agency ('the Agency') will manage this EU database. "*In order to ensure a sufficient level of transparency in the clinical trials, the EU database should contain all relevant information as regards the clinical trial submitted through the EU portal. The EU database should be publicly accessible and data should be presented in an easily searchable format, with related data and documents linked together by the EU trial number and with hyperlinks, for example linking together the summary, the layperson's summary, the protocol and the clinical study report of one clinical trial, as well as linking to data from other clinical trials which used the same investigational medicinal product. All clinical trials should be registered in the EU database prior to being started.*"[33]

In the US, the Food and Drug Administration Amendments Act (FDAAA) of 2007 requires researchers to report summary results on ClinicalTrials.gov within one year of the trial's completion.[34] In 2012, a study only looking at results reported at ClinicalTrials.gov found that only 22% of clinical trials had reported summary results.[35] These numbers are an underestimation of all published results since it does not take into account other ways of publishing results such as journal articles and posting results on websites.[30] Several studies were also exempted from the reporting requirements.[h] Nevertheless, the recovery of all relevant information needed to perform a balanced evaluation is not as obvious as it may first seem.

The problem is not only related to industry-sponsored studies. In fact, in 2015, a study looking at the compliance with results reporting at ClinicalTrials.gov found that only 13.4% of trials reported summary results within 12 months after trial completion. A sample review suggested that many (45%) of the industry-funded trials were not required to report results.[i] Nevertheless, the study concluded that "*despite ethical mandates, statutory obligations, and considerable societal pressure, most trials that were funded by the NIH* [National Institute of Health] *or other government or academic institutions and were subject to FDAAA provisions have yet to report results at ClinicalTrials.gov, whereas the medical-products industry has been more responsive to the legal mandate of the FDAAA. However, industry, the NIH, and*

[h] In a comment, the FDA said the study overestimated the non-compliance with data reporting laws, e.g. because "*the analysis included some trials that were completed before the law came into effect, and did not exclude those – such as uncontrolled trials – that are exempt from the reporting requirements. Nor did the authors exclude all trials of unapproved products, which at present are excluded from the law.*"[36]

[i] "*45% of industry-funded trials were not required to report results, as compared with 6% of trials funded by the National Institutes of Health (NIH) and 9% of trials that were funded by other government or academic institutions.*"[37]

other government and academic institutions all performed poorly with respect to ethical obligations for transparency."[37] An unofficial analysis of the NIH is in agreement with this study: "*companies are outperforming their governmental and academic counterparts. On-time reporting rates were 52% for industry, 21% for NIH-based sponsors and 14% for NIH-funded academic sponsors.*"[36] In contrast, a study looking at the non-publication of large (>500 participants) randomised controlled trials (RCTs) that were prospectively registered with ClinicalTrials.gov and completed prior to January 2009 found a higher publication rate than the often mentioned 50% in combination with a higher percentage of industry-sponsored studies that were not reported. The study searched PubMed, Google Scholar, and Embase to identify published trial results. "*Of 585 registered trials, 171 (29%) remained unpublished. These 171 unpublished trials had an estimated total enrolment of 299 763 study participants. The median time between study completion and the final literature search was 60 months for unpublished trials. Non-publication was more common among trials that received industry funding (150/468, 32%) than those that did not (21/117, 18%), P=0.003. Of the 171 unpublished trials, 133 (78%) had no results available in ClinicalTrials.gov.*"[38] More importantly, as mentioned by Glasziou and Chalmers, "*the best predictor of publication seems to be whether the study is "positive" or "negative," which means that the half of the research results we can access is biased. So there is both waste and distortion.*"[31]

Goldacre et al.[39] also analysed compliance rates with the European Commission's requirement that all trials on the EU Clinical Trials Register (EUCTR) post results to the registry within 12 months of completion (final compliance date 21 December 2016). They found that "*of 7274 trials where results were due, 49.5% (95% confidence interval 48.4% to 50.7%) reported results. Trials with a commercial sponsor were substantially more likely to post results than those with a non-commercial sponsor (68.1% v 11.0%, adjusted odds ratio 23.2, 95% confidence interval 19.2 to 28.2); as were trials by a sponsor who conducted a large number of trials (77.9% v 18.4%, adjusted odds ratio 18.4, 15.3 to 22.1). More recent trials were more likely to report results (per year odds ratio 1.05, 95% confidence interval 1.03 to 1.07).*"[39] Half of all trials were thus non-compliant. The research group set up a trials tracker were more up-to-date information can be retrieved (http://eu.trialstracker.net/).

There are many more studies which analysed the publication rate of clinical trials.[40-45] Results varied depending on which type of studies were included (e.g. phase II or III studies), who sponsored the studies, which sources were used to find results (e.g. ClinicalTrials.gov, PubMed, Embase, etc.), which time frame was allowed to publish results, etc. With the exception of one situation, none of the studies provided numbers close to 100%. This high rate was achieved in the UK, where it was noted that, "*98% of the studies funded by the NIHR* [National Institute for Health Research] *Health Technology Assessment Programme have led to the publication of full reports (Ruairidh Milne, personal communication). The programme has achieved this by holding back a proportion of the research grant until a report has been submitted for publication, by chasing authors on a regular basis, and by providing a publication vehicle – Health Technology Assessment – for all trials.*"[46] All efforts to identify (e.g. by searching all relevant studies in trial registries) and retrieve all evidence, inclusive results from non-published studies, should be supported to allow better unbiased estimates of the treatment effect.

Box 3: The (questionable) reliability of evidence on the relative treatment effect from non-randomized studies

It might be tempting to try to use observational data to estimate a treatment effect. These observational data may include larger numbers and reflect reality. However, for estimating a treatment effect this is difficult since there is no comparator group. There exists a danger of misinterpretations. Two of the most well-known examples are the use of digoxin in patients with heart failure and hormone replacement therapy (HRT).

Digoxin is a drug used to reduce symptoms from heart conditions. There have been concerns about the safety of this drug since observational studies reported an increased mortality with digoxin.[47-49] A systematic review and meta-analysis of observational and controlled trial data was performed to find out the impact of this drug on mortality and other outcomes, taking into account the original study design and statistical analysis performed.[50] For all-cause mortality, 41 relevant studies were identified including 999 994 patients and about 4 million patient years of follow-up. Outcomes were assessed according to four subtypes of analysis: unadjusted, adjusted, propensity matched, and randomized. The risk ratios for all-cause mortality were as follows:[50]

- Unadjusted data from observational studies: 1.76 (95% CI: 1.57 to 1.97, P<0.001) (based on 33 observational analyses, n=331 935)
- Adjusted data from observational studies: 1.61 (95% CI: 1.31 to 1.97, P<0.001) (based on 22 observational analyses, n=245 049)
- Propensity matched observational studies: 1.18 (95% CI: 1.09 to 1.26, P<0.001).(based on 13 propensity matched cohort analyses, n=414 604)
- RCTs: 0.99 (95% CI: 0.93 to 1.05, P=0.75)(based on seven RCTs, n=8406).

The analysis showed digoxin had no effect on mortality in comparison with placebo. In fact, the drug was associated with a small but significant reduction in all cause hospital admission across all study types (risk ratio 0.92, 95% CI: 0.89 to 0.95; P<0.001, n=29 525).[50] Based on their analysis, the authors state that "*observational studies that report increased mortality with digoxin use (regardless of statistical methods) were unable to adjust for systematic differences in the type of patients who received digoxin. … studies exhibiting a higher risk of bias*[j] *reported a stronger association with all-cause mortality, highlighting the need to base clinical decisions relating to patient management on high quality data derived from controlled trials, rather than post hoc or observational data.*"[50] In this case, digoxin is particularly prone to prescription bias as clinicians have been trained to use digoxin in sicker patients.[50] When these patients died, there was therefore a true but misleading association between death and digoxin.[53]

In the case of hormone replacement therapy (HRT), at the beginning of the nineties, "*more than 30 published observational clinical studies have addressed postmenopausal hormone use and cardiovascular disease demonstrating*

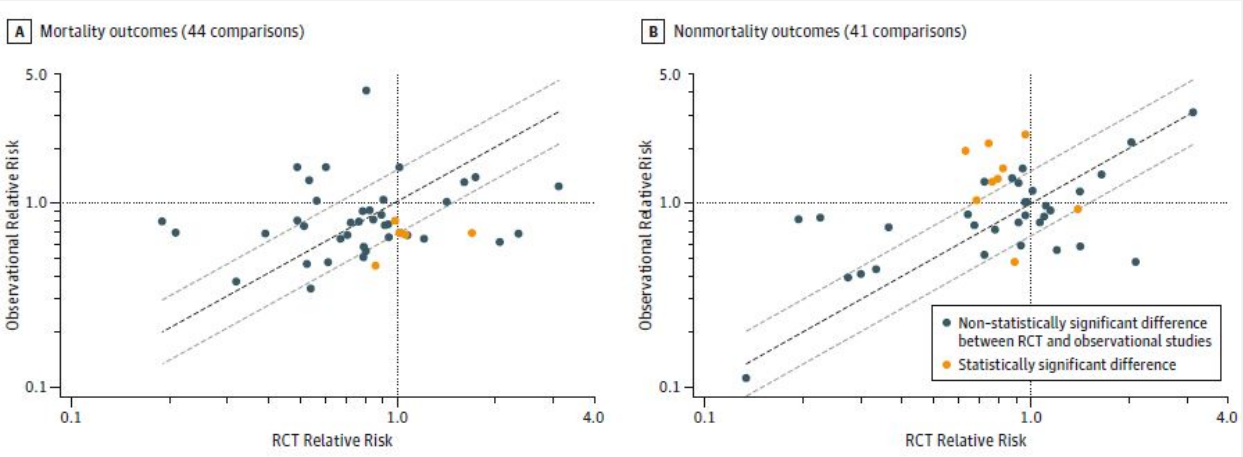
[j] The risk of bias was assessed with the Cochrane Collaboration's risk of bias tool for randomised controlled trials and the risk of bias assessment tool for non-randomised studies (RoBANS).[51, 52]

favourable associations between postmenopausal oestrogen replacement and cardiovascular morbidity, mortality, and risk factors. Meta-analyses of these observational studies suggest that postmenopausal HRT reduces the risk of CAD [coronary artery disease] up to 50%.[54, 55]"[56] The observational information on HRT and cardioprotection was very promising and researchers recommended HRT in healthy women as well as in women with cardiovascular disease and in women with increased risk for this disease.[56] In the US, in 2000, about 2 in 5 women used HRT and 46 million prescriptions were made for Premarin, making it the second most used drug in the US.[57] The Women's Health Initiative (WHI) trial had a component in which 16 608 postmenopausal women (50-79 years) were included between 1993-1998 to assess the major health benefits and risks of commonly used combined hormone preparation in the US. Conjugated equine estrogens and medroxyprogesterone acetate (n=8506) were compared with placebo (n=8102). The trial had a planned duration on 8.5 years but was stopped early, after a mean follow-up of 5.2 years, based on health risks that exceeded health benefits:[58] "*absolute excess risks per 10 000 person-years attributable to estrogen plus progestin were 7 more CHD* [coronary heart disease] *events, 8 more strokes, 8 more pulmonary embolisms, and 8 more invasive breast cancers, while absolute risk reductions per 10 000 person-years were 6 fewer colorectal cancers and 5 fewer hip fractures.*"[58] In contrast with the observational studies, an increase of ischaemic heart disease was thus seen. The authors concluded that this regimen should not be initiated or continued for primary prevention of CHD.[58][k]

Finally, as mentioned in a JAMA viewpoint of Dahabreh and Kent, there is concern that inferences from observational data can lead to poor health care decisions by misrepresenting association for causation.[60] In their article, they present the combined results of 3 studies contrasting the results of propensity score analyses and RCTs, comparing the effect of the same interventions in similar patient populations (Figure 1).[61-63] The figures show no clear pattern of agreement and the authors indicate there is no way of knowing when observational study results are reliable.[60] Methods to analyze the treatment effect based on observational data might improve in the near future. However, at this moment, evidence shows it often provides misinformation.

[k] Researchers tried to identify the conditions for valid observational studies (OS).[59] They studied the WHI data containing "*information on more than 800 possible confounders including information that made it possible to accurately predict HT [hormone therapy] use. It also contained information on factors that might have influenced response to HT. Some of these factors were related to the timing hypothesis (e.g., age, time since menopause, previous HT use, beginning HT after baseline), and some were identified empirically (e.g., blood pressure, previous coronary revascularisation and private medical insurance). Since OS and RCT participants differed with respect to these factors, these factors could have conceivably contributed to differences between the OSs and the RCTs. However, after taking into account all of these confounding factors and stratifying on factors that may have influenced the response to HT, OS and RCT differences remained.*"[59] In their conclusion, the researchers state that they "*did not find that the comprehensive data provided by the WHI were adequate to overcome problems often attributed to OSs. The findings do not imply that most OSs are invalid. They do suggest, however, that given the current methodology, even very good OS datasets may not be adequate to give reliably valid results. … Without better OS methodology there will be underuse or misuse of OSs for comparative effectiveness research.*"[59]

Figure 1: Comparison of propensity score analyses and RCT results from 3 empirical assessments



Scatter plots of results from empirical comparisons of propensity score analyses (y-axis) and corresponding randomized clinical trial (RCT) results (x-axis). Markers denote comparisons between observational and randomized study estimates for the same research question (similar populations, interventions, and outcomes); statistically significant differences ($P < .05$) are shown in orange. The dotted lines indicate lack of effect in RCTs (vertical lines) and observational studies (horizontal lines). Values lower than 1 indicate that the new treatment evaluated in the trial was more effective than the more established treatment; observational study results are expressed in the same way as the corresponding trial results. Markers in the top-right and bottom-left quadrants in each panel indicate agreement between randomized and observational results with respect to the direction of effects. Markers in the top-left and bottom-right quadrants indicate discordant direction of effects between designs. Black dashed diagonal lines indicate the line of identity (perfect agreement) between RCT and observational study results; gray dashed lines demarcate observational study relative risks that are between 0.67 and 1.5 times those produced by the corresponding RCT results. The term "relative risk" is used to denote risk, odds, or hazard ratio estimates, as reported in the 3 empirical analyses contributing data to this figure.

Source: Dahabreh and Kent, JAMA, 2014.[60]

Box 4: Sources of bias that limit the validity of head-to-head comparisons: an example where A is better than B, B is better than C, and C is better than A

Heres and colleagues provide an example where several sources of bias limit the validity of head-to-head comparisons of second-generation antipsychotics for the treatment of schizophrenia.[64] The title of their article is as follows: "*why olanzapine beats risperidone, risperidone beats quetiapine, and quetiapine beats olanzapine: an exploratory analysis of head-to-head comparison studies of second-generation antipsychotics.*" In their study, they analyze the relationship between the study sponsor and the overall outcomes of the trial, as well as different potential sources of bias. They identified 42 reports published before 2004. Thirty studies were sponsored by a pharmaceutical company and included an abstract.[64]

To obtain an objective evaluation, the names and doses of the drugs were masked in the abstracts of the studies. Two physicians not involved in the design of the study were blinded to the study sponsor and independently rated which drug was favored by the overall outcome measures. Two other researchers not blinded to the study sponsor looked for potential sources of bias that could have influenced results in favor of the study sponsor.[64]

Ninety percent of these studies had positive outcomes for the sponsor's drug. Studies including the same drugs but with different sponsors provided contradictory findings. Potential sources of bias identified were related to doses and dose escalation, study entry criteria and study populations, statistics and methods, and reporting of results and wording of findings.[64] One of the most obvious sources of bias was an inappropriate dosage of the comparator drug. For example, one drug (risperidone) was given at too high a dose (10-12mg/day instead of 4-8mg/day) with an increasing risk of extrapyramidal side effects without any gain in efficacy.[65, 66] Or another drug (clozapine) was given in relatively low mean daily doses (<400 mg/day), while at that time doses up to 600 mg/day[67] or even 900 mg/day[68, 69] proved highly efficacious in treatment-resistant schizophrenia.[64] Another important source of bias identified in this study was related to the statistics used in studies with a noninferiority design related to what is considered acceptable for declaring noninferiority and adjustment for multiple testing.[64] Finally, there also appeared to be bias in the reporting and wording of results: "*A complete disclosure of all results of the head-to-head comparison would appear to be mandatory but is not always provided. Results favoring the drug manufactured by the sponsor are often presented in detail, and unfavorable results often are mentioned in a brief sentence at the very end of the report's results section or not mentioned at all.[64, 70, 71]* These different sources of bias were linked to the contradictory overall conclusions of studies comparing the same two antipsychotic drugs but with a different study sponsor.

Box 5: Necessity to publish both relative and absolute treatment effects to support proper interpretation of treatment outcomes

A point for consideration when reporting on benefits and risks is how these outcomes are expressed. Only mentioning the relative impact might be misleading. A good example of reporting benefits and risks is provided[72] in a meta-analysis looking at aspirin in the primary and secondary prevention of vascular disease.[73] Amongst others, the impact on serious vascular events (myocardial infarction, stroke, or vascular death) and major bleeds was reported.

- In primary prevention (six trials: 95 000 individuals at low average risk, 660 000 person-years, 3554 serious vascular events), a 12% proportional reduction in serious vascular events (rate ratio (RR): 0.88, 95% CI: 0.82 – 0.94) and a 54% increase in major extracranial bleeding (RR: 1.54, 95% CI: 1.30 – 1.82) was reported for aspirin in comparison with the control group. This relative impact could be misinterpreted as a higher increase in bleedings versus the reduction in vascular events. However, the reporting of the absolute numbers shows a reduction of 6/10 000 in serious vascular events (0.51% versus 0.57% per year, p=0.0001), and an increase of 3/10 000 in major extracranial bleedings (0.10% versus 0.07% per year, p<0.0001). The authors concluded that "*in primary prevention without previous disease, aspirin is of uncertain net value as the reduction in occlusive events needs to be weighed against any increase in major bleeds.*"[73]
- The study also analyzes trials in secondary prevention (16 trials: 17 000 individuals at high average risk, 43 000 person-years, 3306 serious vascular events). Comparing the treatment effect on major coronary events (non-fatal myocardial infarction (MI) and coronary heart disease (CHD) mortality) between primary and secondary prevention, the proportional reduction in major coronary events seemed to be similar: primary prevention: RR 0.82, 95% CI: 0.75 – 0.90; secondary prevention: RR 0.80, 95% CI: 0.73 – 0.88. The absolute benefit reveals a larger difference: 0.06% per year versus 1.00% per year in primary and secondary prevention, respectively.

There are many more examples, like the conclusion in a NEJM article that "*PSA* [prostate-specific antigen]-*based screening reduced the rate of death from prostate cancer by 20% but was associated with a high risk of overdiagnosis*".[74] The 20% relative reduction in the rate of death from prostate cancer equalled an absolute risk difference of 0.07%.[l][74] Both numbers are mentioned in the abstract of the article. Overdiagnosis and overtreatment are mentioned to be the most important adverse effects of prostate-cancer screening but are not discussed further in detail in the article. Having a clear view on both positive and negative consequences is necessary to make good decisions.

Researchers or the media might mislead the readers by selective reporting of outcomes. To make the benefits seem larger and the harms seem smaller, "*the benefits are presented in relative terms, while the harms or side effects are*

[l] Prostate cancer deaths in the screening group: 214/72 890 = 0.29%; in the no screening group: 326/89 353 = 0.36%. Absolute difference: 0.36% - 0.29% = 0.07%; Relative difference: (0.36%-0.29%)/0.36% = 20%.[74]

presented in absolute terms."[m] Only providing relative numbers runs the risk of being misinterpreted. "*To know the meaning of a reduction in relative risk, you have to know how likely it was to happen in the first place.*"[75] Readers can better interpret the impact if both relative and absolute numbers are mentioned.

3.1.2 Safety

We can define **safety** as the "*substantive evidence of an absence of harm"*.[76] There are several terms related to safety (harms, adverse effect, adverse event, adverse reaction, side effect, complication, etc.). For example, **harms** may be defined as "*the totality of possible adverse consequences of an intervention or therapy; they are the direct opposite of benefits, against which they must be compared"*[76]), while **adverse effect** may be defined as "*a harmful or undesirable outcome that occurs during or after the use of a drug or intervention for which there is at least a reasonable possibility of a causal relation*."[77] To see a list of terms and definitions, please refer to the EUnetHTA guideline 'Endpoints used in Relative Effectiveness Assessment: Safety'.[78] Apart from the many definitions, the harms or adverse effects can be classified in different ways according to frequency, incidence, severity, and seriousness.[78]

Safety can be measured in the framework of different type of studies. "*RCTs may be appropriate for common, anticipated adverse effects, observational studies may be particularly useful for long-term or rare adverse effects, and post-marketing monitoring data may be useful in detecting previously unknown adverse effects.*"[79] All these type of studies have limitations. RCTs may not be generalizable as they tend to exclude patients at higher risk of adverse effects. Their usual short-term follow-up and sample size may reduce the likelihood of appearance of adverse events. Observational studies are very useful for the observation of adverse events as they tend to not suffer from the above RCT limitations. Indeed, "*the lack of evidence of a rare adverse effect is therefore not proof that such an adverse effect is not associated with the intervention of interest*."[79]

Based on a review of guidelines for modellers, Craig et al.[80] concluded that "*it is clear from the available guidance that all relevant outcomes should be included in the economic decision model and there appears to be a general if not clearly stated consensus that this includes adverse effects*" but "*articles contained very little information or guidance of direct relevance to the incorporation of adverse effects in models*". More recent guidelines mention the safety issue although, to our knowledge, none of them dedicates a chapter to its management in economic evaluation.[7, 81, 82] The recommendations by Craig et al. are:[80]

- *"…much clearer and explicit reporting of adverse effects, or their exclusion, in decision models… separate sections on adverse effects should be included in the*

[m] Source: https://www.healthnewsreview.org/toolkit/tips-for-understanding-studies/absolute-vs-relative-risk/

clinical effectiveness and modelling chapters of every technology assessment report.

- *Even when a systematic review of adverse effect data is not feasible, summaries of such data should be presented…"*

The Canadian guidelines for the economic evaluation of health technologies provide the following clear description:[83]

- *"Researchers should be explicit about how the adverse events included in the economic evaluation were identified, and what methods were used to incorporate them. Where adverse events have a negligible impact on health effects, or no impact on costs and resources, it is often appropriate to exclude these events from the model. Where adverse events are not included, a clear justification must be provided.*
- *Adverse events should be incorporated into the model by combining both the health condition and the associated adverse effects. In the case of utilities, the utility for a specific health state can then be adjusted by applying a disutility for an adverse event to allow the utility for the health state with an adverse event to be estimated.[84]*
- *If effects are transitory (i.e., short-term), they should be incorporated through appropriate refinement of the states or events within the model. Where data are available on the prevalence, costs, and disutility associated with each adverse event by intervention, this facilitates greater transparency."*

Points for consideration

Some of the potential problems related to the misuse of safety data in economic evaluations are:[78]

- Not all potential adverse effects associated with the technology under evaluation are identified. Serious, frequent and/or costly adverse effects might not be taken into account (see Box 6). If there are adverse effects omitted from the analysis, the reasons to do so should be explained.
- The relevant effect of some safety issue on health states, adherence/withdraws, subsequent treatments, use of resources, quality of life/disutilities or mortality might not be considered in the analysis.
- The data on adverse effects may come from different sources with different risk of bias. Ideally the safety profile of the technology should be described against the comparator and those clinically significant differences in adverse effects between the technology and the comparator should be considered in the analysis. Observational studies may be particularly useful for long-term or infrequent adverse effects. On the other hand, this may not provide an unbiased source of comparative safety information.
- Safety-related parameters should be tested by means of sensitivity analysis.

Examples

Box 6: Management of adverse drugs events of new biological drugs for rheumatoid arthritis.

Heather et al.[85] published a systematic review of economic evaluations of anti-tumour necrosis factor-α drugs (anti-TNFs) for adult with rheumatoid arthritis (RA). They were interested in how the decision analytic models considered the effects of adverse drug events (ADE) on costs and consequences of the treatment. The anti-TNFs have demonstrated some effectiveness retarding the progression of the disease although they are more expensive than the non-biological disease-modifying anti-rheumatic drugs (nbDMARDs) and are associated with higher risk of serious infections.

These are some of the findings of this systematic review, related to the inclusion of ADE in economic evaluations:[85]

- 34 out of 43 studies included in the systematic review "*did not consider the wider implications of ADEs in the economic models, 16 did not incorporate ADEs in any form. Only four acknowledged the omission. ADEs were implicitly included within an all-cause treatment-discontinuation parameter in 15 studies, and three studies explicitly modelled the early cessation of treatment due to ADEs. The most commonly cited reasons for not comprehensively including the implications of ADEs were a relative paucity of data and a negligible impact on the relative costs and consequences of treatment.*"
- "*Nine studies were critically appraised because they had considered the direct implications of ADEs on health care costs and/or patient HRQoL* [health-related quality of life] *in the economic model.*" … "*There was substantial variation amongst the nine studies in terms of the methods used to incorporate ADEs into the economic models and the associated assumptions made. Differences arose in (i) the specific type of model used, which then influenced how ADEs were parameterised; (ii) the time interval during which ADEs could occur; (iii) the assumptions made regarding the impact of an ADE on the disease and treatment course; and (iv) the extent to which the risk of an ADE was adjusted for distinct patient sub-populations.*"
- "*All nine studies included some estimation of the direct health care costs of treating ADEs, but this was reported with differing degrees of detail. In contrast, only two studies included some consideration of the consequences. Only one study included an estimate of the direct and independent impact of an ADE on patient HRQoL.*"
- "*ADEs were predominantly assumed to preclude treatment continuation in the majority of studies rather than directly affect patient HRQoL or treatment effectiveness.*"
- "*Data informing the incidence, cost, and consequences of ADEs were drawn from a myriad of sources. Incidence-related data were predominately abstracted from secondary sources including clinical trial reports, published observational studies, and, in three instances, drug package inserts.*"

The authors of the review discussed some of the issues that arise in the treatment of safety in economic evaluations. For example, they highlight that recent economic

evaluations have omitted the evidence on ADE, despite the fact that *"evidence to suggest that rates observed in clinical trials are generally lower than those seen when the drug is used in clinical practice", "the evidence base on the safety of anti-TNFs has significantly improved with the establishment and maturation of national biologics registers in, for example, the UK, Italy, Germany, the Netherlands, and Sweden"* and that *"results from long-term observational studies show an elevated risk of ADEs from anti-TNFs when used in a clinical real-world setting.*"[85]

Heather et al.[85] concluded that *"the findings contradict recommendations in current national UK guidelines[86] and also in the reference case for economic evaluations of drug treatments for RA* [rheumatoid arthritis] *proposed by the Outcome Measures in Rheumatology Task Force (OMERACT)[87] , which explicitly state the need to consider the impact of 'adverse effects,' 'adverse events,' and 'toxicity'."*

Extra information

- Critical assessment of clinical evaluations. Methodological guideline. Diemen: EUnetHTA; In preparation.
- Institute for Quality and Efficiency in Health Care (IQWiG). Process of information retrieval for systematic reviews and health technology assessments on clinical effectiveness Methodological Guideline; 2017.[22]

3.2 Comparator

Based on the results of a review of national guidelines for economic evaluations, the EUnetHTA guideline for methods for health economic evaluations recommends that:

- "*the comparator(s) reflect the most relevant alternative intervention(s) used in clinical practice and that the choice of comparators should be clearly presented and justified.*"[1]

The first recommendation of the EUnetHTA guideline on criteria for the choice of the most appropriate comparator(s) states that:

- "*Under ideal circumstances the comparator for a REA* [Relative effectiveness assessment] *applicable across European countries should be the reference treatment according to up to date high-quality clinical practice guidelines at European or international level with good quality evidence on the efficacy and safety profile from published scientific literature, and with an EU marketing authorisation or another form of recognised regulatory approval for the respective indication and line of treatment.*"[12]

However, up-to-date evidence-based clinical practice guidelines are not always available (e.g. for rare diseases) and not all interventions recommended in clinical practice guidelines are necessarily reimbursed. Researchers should thus look further and also consider current standard practice or the reimbursed alternatives, which in some cases might be different from the optimal care described in practice guidelines. Expert opinion or patients' view might also be helpful in identifying relevant comparators. Of course, in first instances, national guidelines on the choice of the comparator should be respected. In what follows, we reflect on several points for consideration in relation to the choice of comparator.

Points for consideration

Some of the potential problems related to the comparator included in economic evaluations are:

- A description of the comparator (e.g. standard care) might be missing or be vague. A clear description should also be available for the intervention under evaluation.
- The choice of comparator will critically determine the relative cost-effectiveness of the technology and the relevance of the assessment to the decision-makers.[88]
- the choice of the comparators included in an economic evaluation may be context-specific and depend on national guidelines (see final paragraph of this part).
- (Inappropriately) excluding a relevant comparator (with possibly a better cost-effectiveness – see Box 7 and Box 8). Researchers should not only think about interventions used in routine practice, but also other new interventions that might replace current practice, (evidence-based) off-label use,[n] less intensive treatment/screening intervals (see Box 8), etc.
- (Inappropriately) comparing with an alternative with an unfavourable cost-effectiveness (see Box 7). The incremental cost-effectiveness ratio[o] (ICER) should be calculated against the last comparator on the efficiency frontier[p] (excluding alternatives that are dominated or extendedly dominated). In case of inappropriate exclusion of this comparator and inclusion of another comparator with a worse cost-effectiveness, the ICER of the intervention will incorrectly be improved.
- It is possible that standard of care is not cost-effective (or that the cost-effectiveness has not been assessed previously) but is used as a comparator since it is routinely used. In such cases, if possible (e.g. based on the presence of reliable evidence), it

[n] National guidelines should be checked to see whether (evidence-based) off-label use can be applied in the reference case or a scenario analysis. For example, the Belgian guidelines indicate "*off-label used pharmaceutical products can be used as valid comparators in a pharmacoeconomic evaluation if evidence is available about the clinical safety and efficacy of the off-label use, e.g. from government sponsored trials.*"[89] In Ireland, "*technologies that do not have marketing authorisation (or CE mark for medical devices) for the indication defined may also be considered for the comparator if they are part of established clinical practice for that indication. Where such an unlicensed technology is used as the comparator, the evidence of efficacy and safety included in the assessment must be relevant to the unlicensed use.*"[88] Also in Poland, off-label drugs can constitute "current medical practice" and can be officially reimbursed and be a suitable comparator for economic evaluations.(personal communication with reviewer from AOTMIT related to the Polish HTA guidelines[20])

[o] Remark: The cost-effectiveness results may be reported in two equivalent measures: the incremental cost-effectiveness ratio (ICER) or the incremental net benefit (INB) expressed in monetary terms (net monetary benefit – NMB) or in health terms (net health benefit – NHB). In this document, no preference is expressed about these measures. Remarks/examples that refer to ICERs also apply to NMB or NHB. We refer to the annexes (part 5.2) for some further information on the INB approach.

[p] "*The efficiency frontier is the line on the cost-effectiveness plane connecting the non-dominated treatment alternatives. It can be constructed as follows: 1. Exclude interventions that are dominated by other interventions with lower costs and greater therapeutic benefits. 2. Exclude extendedly dominated alternatives, which means that linear combinations of other strategies can produce the same (or greater) benefit at lower (or the same) cost. 3. For the remaining alternatives, calculate the cost effectiveness by comparing each strategy with the next more costly and more effective intervention.[8, 90]*"[91]

is important to also consider other comparators which are more cost-effective as relevant alternatives.

- If several therapies are not cost-effective versus standard care, then basing conclusions on a comparison between these therapies without mentioning the comparison versus standard care does not provide full information to decision makers and might result in misleading conclusions (see last paragraph of Box 7).
- When the comparator is not the same across the target population, the estimation of the cost-effectiveness should be calculated per subgroup of patients since results on the population level might be difficult to interpret (e.g. in aortic stenosis, some patients are inoperable (→ optimal medical treatment as comparator), while others are operable (→ surgery as a comparator)).
- Relevant differences in the treatment pathway after the point of randomization should be taken into account. Researchers should consider whether the introduction of a new intervention would replace the existing treatment(s), whether the existing treatment(s) would still be used if the new intervention has failed, whether the introduction of the new intervention has an influence on the follow-up of patients, etc.
- Identifying potentially relevant alternatives is usually part of the clinical section of an HTA. These alternatives might be identified through a systematic review of the medical literature, clinical practice guidelines, expert opinion, patients experience and perspective, sales figures, etc. Alternatives for which no evidence is available are difficult to evaluate in an economic evaluation.

Examples

Box 7: Problems related to the inappropriate ex- or inclusion of alternatives

As an illustration, we refer to the Belgian Health Care Knowledge Centre (KCE) report in which the cost-effectiveness of cardiac resynchronization therapy (CRT) was assessed in patients with New York Heart Association class III/IV heart failure.[92] Two different types of CRT devices were available: biventricular pacemakers (CRT-P) and biventricular defibrillators (CRT-D). Both interventions were included in the COMPANION (Comparison of Medical Therapy, Pacing and Defibrillation in Heart Failure) trial[93] and compared with optimal pharmacological therapy (OPT). Comparing CRT-D with OPT resulted in an average ICER of about €25 600 per quality-adjusted life year (QALY) (Table 2). However, the ICER of CRT-P was much better at €11 200/QALY and, being an alternative to the patients included in this trial, should thus become CRT-D's comparator. Doing so, CRT-D's ICER was about €56 600/QALY, making results much less optimistic and possibly influencing the reimbursement recommendation/decision. In this example, exclusion of CRT-P as a relevant treatment alternative would have been wrong and would have incorrectly improved CRT-D's ICER from €56 600/QALY to about €25 600/QALY. This is also shown in Figure 2 where the slope of the dotted line (i.e. the incorrect comparison) is flatter than the slope of the solid line (the correct comparison on the efficiency frontier).

Table 2: Cost-effectiveness of CRT-D in comparison with OPT or CRT-P

	CRT-P versus OPT	***CRT-D versus OPT*****	**CRT-D versus CRT-P**
IC	€14 745	*€45 624*	€30 879
IE (QALY)	15.77 months	*22.32 months*	6.55 months
ICER (€/QALY)*	€11 219/QALY	*€25 639/QALY*	€56 615/QALY

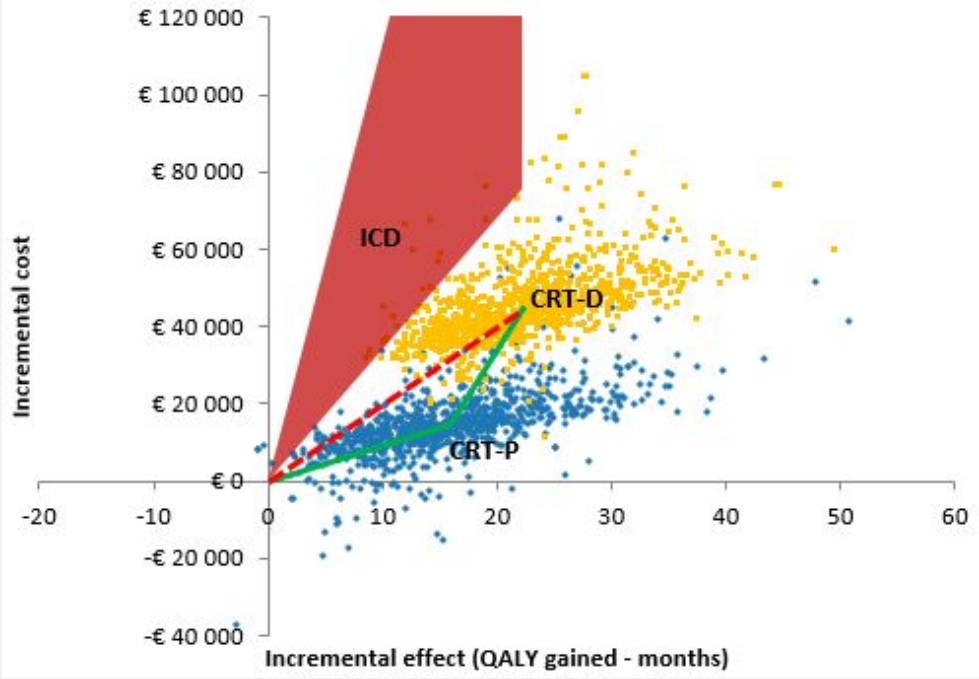
Source: Van Brabandt et al., KCE, 2010.[92]
CRT-D: CRT-D Cardiac Resynchronisation Therapy, combined with ICD; CRT-P: CRT-P Cardiac Resynchronisation Therapy, combined with Pacing; IC: incremental cost; ICD: implantable cardioverter defibrillators; ICER: incremental cost-effectiveness ratio; IE: incremental effect; OPT: Optimal Pharmaceutical Therapy; QALY: quality-adjusted life year.
** Only the mean estimates from the probabilistic analysis are shown in the above table. For more details, we refer to the original publication.*
*** This is the inappropriate comparison if CRT-P is also considered to be a relevant treatment alternative.*

This example can also be used to illustrate the pitfall when comparing an intervention with an alternative that is not situated on the efficiency frontier and has an unfavourable cost-effectiveness. This is methodologically incorrect. If CRT-P was taken out of the comparison and CRT-D was compared with an implantable cardioverter-defibrillator (ICD – see the red zone in Figure 2), the ICER would be too optimistic. The authors support their point with a non-medical simplistic example: A Porsche Panamera, which is a four-seater, can mistakenly be considered cost-effective (even cost-saving) if compared with a Ferrari for the transport of a family with 2 adolescent children since the Ferrari is more expensive and the two children would not even fit in the car. However, the Ferrari itself is not a cost-effective alternative and using it as a comparator mistakenly results in a favourable ICER for

the Porsche. If other relevant cost-effective ways of transport were included, the Porsche's ICER would be much higher (possibly even dominated).[91]

This simplified example can be extrapolated to the comparison of e.g. expensive biologics which are compared with other expensive biologics, while their cost-effectiveness compared to standard care has not yet been demonstrated. For example, in a study evaluating the use of tumour necrosis factor-alpha (TNF-a) inhibitors, adalimumab and infliximab, for Crohn's disease, the authors explain why it is important not to compare only biologicals with each other. This would only be relevant "*where both adalimumab and infliximab have been first justified as maintenance therapies versus standard care (SC). Where one or both maintenance therapies are not cost-effective versus SC, this comparison provides no information to decision-makers.*"[94]

Figure 2: An illustration of the impact of (not) working on the efficiency frontier (A)

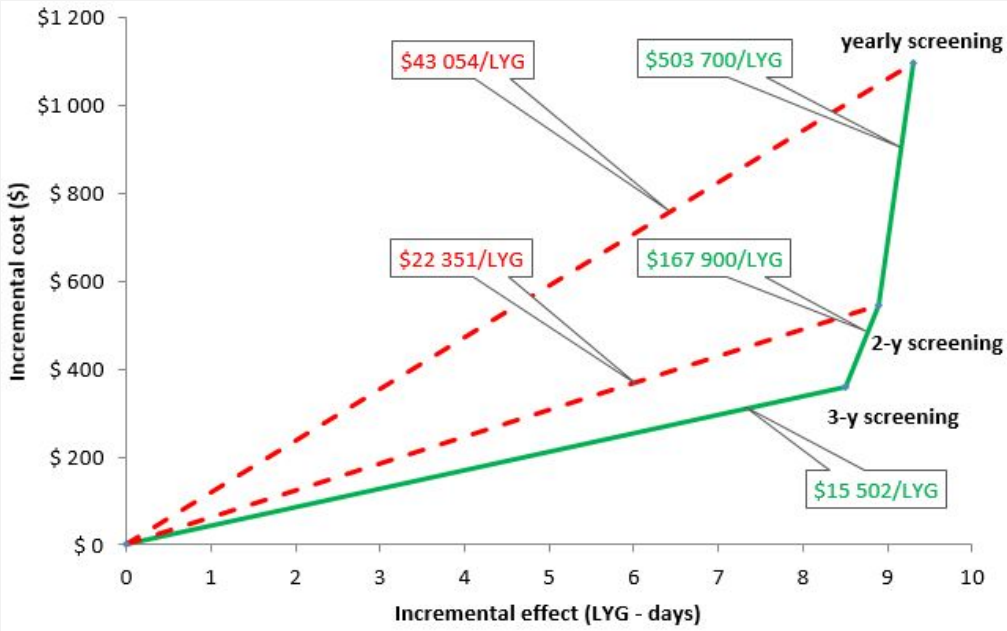


Source: Neyt and Van Brabandt, Pharmacoeconomics, 2011.[91]
CRT(-P/D): cardiac resynchronization therapy (biventricular pacemakers/biventricular defibrillators); ICD: implantable cardioverter defibrillators.
The red area represents the estimated ICER (€71 400/QALY (95% CI 40 200, 134 600)) calculated in a previous report.[95, 96]
The green line shows the comparison on the efficiency frontier (CRT-P versus OPT: €11 200/QALY; CRT-D versus CRT-P: €56 600/QALY). The red line (CRT-D versus OPT: €25 600/QALY) incorrectly excludes CRT-P as a relevant alternative.

Box 8: Problems related to the inappropriate exclusion of alternatives

Another example, previously used in an educational paper[97] can be used to demonstrate the large influence of excluding the relevant comparator. In one of the scenarios in the underlying paper of Eddy,[98] the cost-effectiveness of one-, two- or three-yearly cervical cancer screening was calculated. Figure 3 presents the results. The screening interval influences the ICERs. In compliance with good practice of economic evaluations, calculations are made on the efficiency frontier and lowering the interval from yearly to 2- or 3-yearly screening improves the ICER from about $504 000 per life-year gained (LYG), to $168 000/LYG to $15 500/LYG. If the author had compared with no screening arguing this is standard practice, the ICERs would mistakenly have been too optimistic: $43 000/LYG and $22 400/LYG for yearly and 2-yearly screening, respectively. We note that the choice of screening frequencies in the model should not be arbitrary, the explored alternatives should be supported by clinical evidence.

Figure 3: An illustration of the impact of (not) working on the efficiency frontier (B)



Source: Based on Briggs, Pharmacoeconomics, 2000.[97]
LYG: life-years gained. The dotted red lines exclude the comparison with the previously most cost-effective alternative. The green full lines are situated on the efficiency frontier.

Finally, we recognise that the choice of the comparators included in the economic evaluation may be context-specific and depend on national guidelines. For example, the standard care might be different between countries or the guidelines related to the inclusion of off-label use might be different. For example, in the UK, off-label/unlicensed technologies can be considered as comparators by the National Institute for Health and Care Excellence (NICE)

if they are part of established clinical practice.[q] Also in Belgium, the guidelines allow the inclusion of off-label use when this is supported by evidence on its clinical efficacy and safety.[89] As a result, economic evaluations can include an off-label evidence-based shorter treatment schedule of trastuzumab (Herceptin®) for the treatment of breast cancer,[99, 100] or the off-label bevacizumab (Avastin®) as an evidence-based and scientifically appropriate comparator for ranibizumab (Lucentis®) in the treatment of wet age-related macular degeneration.[101] Whether or not these off-label alternatives are included in an economic evaluation might have a large influence on results, conclusions and recommendations.

Extra information

- Comparators & comparisons: Criteria for the choice of the most appropriate comparator(s). Methodological Guideline: EUnetHTA; 2015[12]

3.3 Subgroup analysis

The EUnetHTA guideline for methods for health economic evaluations recommends:

- "*to perform subgroup analyses in the economic analysis when there is a clinical rationale to believe that the cost-effectiveness of the assessed technologies may vary between subgroups. It is important that the choice of subgroups is clearly justified and described.*"[1]

Furthermore, the EUnetHTA guideline on clinical endpoints mentions that appropriate adjustment should be considered for multiple hypothesis testing.[13]

Points for consideration

- Measures of cost-effectiveness for the overall study population may lead to incorrect treatment recommendations, if the cost-effectiveness of the assessed technologies varies between subgroups.[102]
- The cost-effectiveness of an intervention will be different for subgroups if the relative treatment effect differs. However, the heterogeneity of the absolute treatment effect is also of importance. In this context, it is important to consider other factors such as sociodemographic characteristics (e.g. age, sex, social class) or clinical characteristics (e.g. baseline risk (see part 3.4) or disease severity).[103]
- Economic evaluations may require the specification of some subgroups based on non-clinical considerations, such as heterogeneity in treatment costs (e.g. when the dose is weight dependent, cost of events is comorbidity dependent or the cost of an intervention is localisation dependent) or heterogeneity in health value (e.g. when perceived impact of event is experience dependent).[103]

[q] The Guide to the methods of technology appraisal 2013 (section 6.2.4) states that: "*The Appraisal Committee can consider as comparators technologies that do not have a marketing authorisation (or CE mark for medical devices) for the indication defined in the scope when they are considered to be part of established clinical practice for the indication in the NHS. Long-standing treatments often lack a sponsor to support the licensing process. Specifically when considering an 'unlicensed' medicine, the Appraisal Committee will have due regard for the extent and quality of evidence, particularly for safety and efficacy, for the unlicensed use.*"[86]

- If subgroups are defined on the basis of heterogeneity in the relative treatment effect, be attentive that subgroup analyses follow methodological standards (ideally pre-specified in the study protocol with rationale for expected subgroup effects and statistically powered).
- If subgroups are defined based on other considerations, ensure that the assumption regarding relative treatment effect between the subgroups is founded on an assumption of equivalence (i.e. the modelled treatment effect is the same as the relative treatment effect observed in the ITT population).
- In most cases, subgroup analyses are exploratory and should be interpreted cautiously (e.g., subgroup sizes often too small to detect moderate differences, unless included in sample size calculations). Post-hoc results cannot be regarded as confirmatory. In Box 9 an illustration is provided of authors warning for the danger of misinterpretation of (false-positive) findings by including the results of subgroup analyses for the astrological birth signs.

Examples

Box 9: The validity of subgroup analysis – significance dependent on the astrological birth sign

In the ISIS-2 trial, between March 1985 and December 1987, 17 187 patients entering 417 hospitals after the onset of suspected acute myocardial infarction (MI) were randomised.[104] Patients could receive streptokinase (1-hour intravenous infusion of 1.5 MU), aspirin (160mg/day for one month), both active treatments, or neither. A 2x2 factorial study design was used in which half of all patients were randomized to receive streptokinase or placebo and half of all patients were also randomized to receive aspirin or placebo. One of the outcomes was the effect of these treatments on vascular mortality during the first 5 weeks. In comparison with placebo, both streptokinase and aspirin individually significantly reduced vascular deaths during the first 5 weeks with a 25% (95% CI: 18-32) and 23% (95% CI: 15-30) reduction in the odds of death. The combination of both drugs even had a greater effect with a 42% reduction (95% CI: 34-50).[104]

The authors also presented a figure with results for subgroup analyses of the odds of vascular deaths in the first 5 weeks. The first results are presented for the astrological birth sign gemini/libra versus other birth signs. For the latter group, the treatment effect of aspirin remains positive with a 28% significant reduction, while for people born under the astrological sign of gemini or libra a non-significant increase of vascular mortality was observed. The authors remark that "*it is clear that the best estimate of the real size of the treatment effect in each astrological subgroup is given not by the results in that subgroup alone but by the overall results in all subgroups combined. … 'Lack of evidence of benefit' just in one particular subgroup is not good 'evidence of lack of benefit'.*"[104] When there is little evidence of any real heterogeneity, more weight should be given to the overall results.[104]

Fayers and King[105] discuss this Lancet article explaining the danger of false-positive findings if subgroup analyses are performed. They clarify the inclusion of the star signs subgroup by stating that "*The Lancet was keen to include what seemed like clinically relevant subgroup findings. The authors agreed, with one proviso—namely, that the journal allowed the star-sign groups to appear first, simply to underline for readers the reliance they might put (or not) on the validity of these*

analyses."[105] They also summarize several guidelines related to performing and reporting subgroup analysis. The first two bullets relate to 1) the factors for subgroups, and the rationale for subgroup analyses, which "*should have been formally prespecified in the protocol. The credibility of subgroup analyses is improved if confined to the primary outcome and to a few predefined subgroups, on the basis of biologically and/or psychologically plausible hypotheses.*"[105] And 2) "*Factors for subgroups should have been assessed before randomisation.*"[105] For more information, we refer to the original article and underlying references.

Finally, we remark that in certain cases, a subgroup analysis, which has been predefined in the protocol of an HTA report, can only be conducted using post-hoc subgroup analyses from clinical trial data. The risk of bias from one pre-specified biologically and/or psychologically plausible subgroup analysis is very different from the risk of bias if an unknown number of subgroup analyses have been performed and only a selection of results have been reported.

3.4 Baseline risk of the target population

Outcomes can be summarised and presented in absolute or relative terms. The EUnetHTA guideline on clinical endpoints states that "*despite the advantages of absolute measures, they are of limited generalisability due to their dependence on the baseline values. It would be inappropriate, for example, to extrapolate published absolute measures from one population to another population with a different baseline value. Pooling absolute measures in a meta-analysis is highly problematic due to fact that the variation in baseline values is not accounted for.[106] By extension, where data are presented without a subgroup analysis it is feasible to apply relative effects to different subgroups with the understanding that baseline values will vary by subgroup and that any interaction between subgroup characteristics and treatment effect is ignored. It is not possible to make such a generalisation using absolute measures.*"[13] One of the conclusions of this guideline is that:

- "*Absolute measures are useful to clinicians as they provide a realistic quantification of treatment effect which is meaningful for treatment evaluation and prognosis. However, due to the dependence of absolute measures on baseline risk, relative measures are more generalizable across studies.*"[13]
- It is recommended that "*both relative and absolute measures should be presented.*"[13]

The EUnetHTA guideline on applicability of evidence for the context of a relative effectiveness assessment also recommends that you should:

- "*Describe general characteristics of enrolled populations, how this might differ from target population, and effects on baseline risk for benefits or harms.*"[14]

The authors of this guideline also state that "*exploring sources of heterogeneity is a key issue to assess whether observed differences in treatment effects can be explained by trial-level characteristic. For example, patient age or other patient characteristics may influence the baseline risk so that treatment effect may be over- or underestimated when applying the results of a trial to other patients.*"[14]

Points for consideration

Potential problems related to the baseline risk in economic evaluations are:

- Not being aware of the possible differences in the baseline risk for certain events in a specific population (e.g. selected population in an RCT) versus the population in the economic evaluation and/or the general population to which the decisions of policy makers apply. Not adjusting for such differences in baseline risk might have a large impact on the (modelled) absolute treatment effect and the related ICER calculations.
- It is possible that a statistically significant effect is shown on a relative outcome while the absolute effect is not clinically relevant. Publishing both the relative and absolute treatment effect might support the proper interpretation of treatment outcomes (see Box 5).

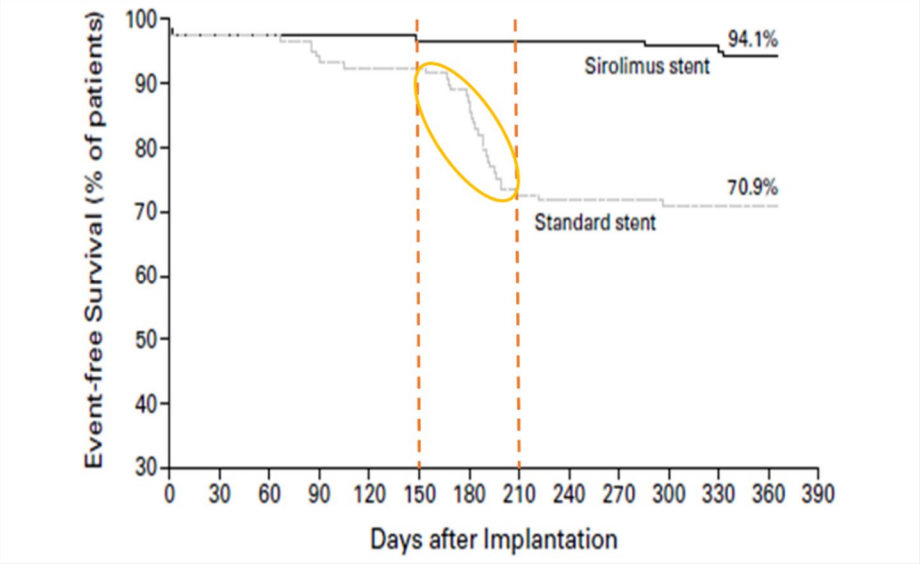
(Meta-analyses of) RCTs are generally acknowledged to provide the highest level of evidence for the treatment effect of medical interventions.[107] Nevertheless, cost-effectiveness is driven by the absolute treatment effect, which is the combination of both relative treatment effect and baseline risk for specific events[108] (see Box 10 and Figure 5). The baseline risk for specific events (like mortality, (re-)hospitalisations, etc.) might be very different in the trial versus the real-life population under consideration. For example, the population included in the RCTs might not be representative for the population eventually getting the intervention. Or there may be an increased number of follow-up examinations (driven by protocol) in the RCT population compared with the real-life population leading to an increased number of hospitalisations. Inappropriate application of the absolute treatment effect from RCTs may result in unrealistic estimates of an intervention's benefit for the real-world target population if the baseline risk of events in this target population differs significantly from the baseline risk in the RCT population. To illustrate this with a hypothetical simple example: if a trial shows that an intervention reduces the one-year mortality rate in a specific indication from 20% to 12%, but in reality, the one-year mortality rate in this indication with the current treatment is only 5%, then, of course, you cannot avoid an absolute 8% (or in other words: 8 percentage points) of deaths. Similarly with relative outcomes: if a trial shows a baseline event rate occurring in 30% of the population in the comparator arm and the intervention has a relative risk (RR) of two, then we see the event in 60% of the population in the intervention arm. If our baseline event rate in the real-world population was 55% then simple application of the same RR would give us an event in 110% of the population after introduction of the intervention, which is of course not possible. Applying incorrect or unrealistic absolute benefits in economic evaluations will of course result in non-reliable cost-effectiveness outcomes.

Examples

Box 10: Adjustment for baseline risk and its influence on the (modelled) absolute benefit

In an HTA report comparing drug-eluting stents (DES) and bare-metal stents (BMS), the authors performed a review of the economic literature and identified rather opposite results: some authors indicate that DES may be cost-effective or even cost-saving in specific patients, while others mention DES is not cost-effective with ICERs of about 200 000 Canadian dollar per QALY gained.[109, 110] One of the most important determining variables for the ICER, next to the price difference of DES and BMS, was the baseline repeat revascularisation rate with BMS. The authors note that this risk for a re-intervention using BMS ranges from 5% to 14% in registries[110] and is much smaller than reported in RCTs (up to 30%). They remark that this might be due to the influence of protocol-driven angiographic follow-up in RCTs, which are mandated to assess in-stent restenosis. The influence of this protocol-driven follow-up was also demonstrated in another study where the baseline risk for both MACE (major adverse cardiac events) and TLR (target lesion revascularisation) with BMS was about 12% lower in absolute numbers without angiographic follow-up.[111] The following figure shows the potential impact of such protocol-driven follow-up in one of the trials comparing DES and BMS: the largest increase of events in the BMS group takes place when the angiography is performed, i.e. around 180 days after implantation.[112] If such a follow-up investigation is not performed in real-world, and the patient has no symptoms, there might be no clinical need for revascularisation.

Figure 4: potential influence of protocol-driven follow-up on the baseline risk of events



Source: Morice et al., NEJM, 2002.[112]

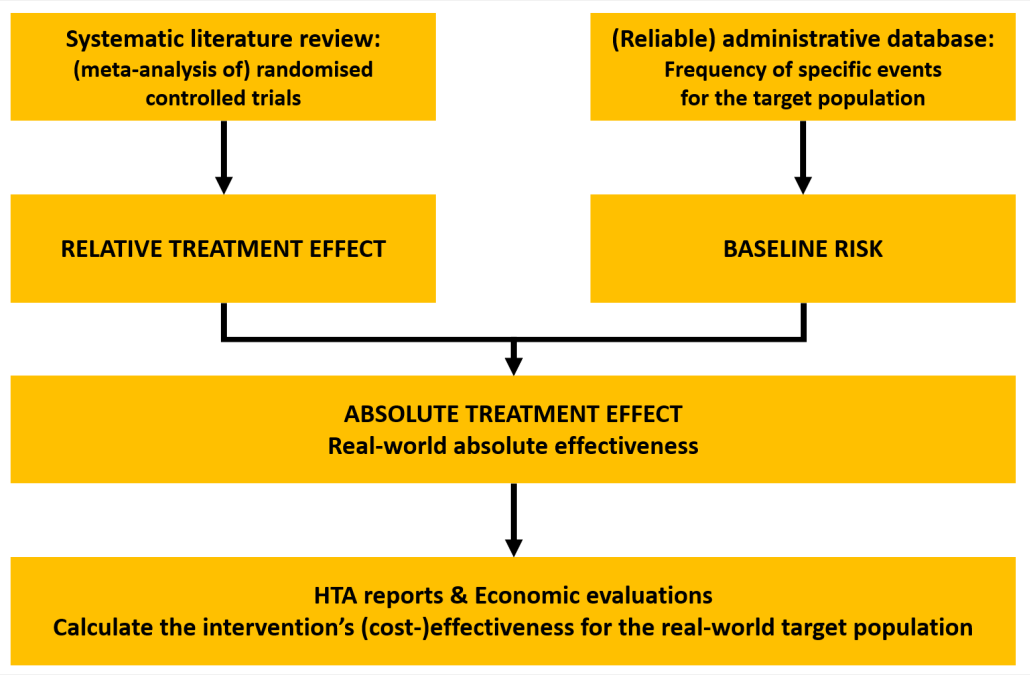
The problem of the baseline risk has also been discussed in the context of the GRADE system providing a framework for assessing confidence in estimates of the effect.[113] For economic evaluations, this is of major importance since an identical relative risk reduction in combination with an artificially higher baseline risk inflates estimates of absolute risk reduction (and vice versa), which drives cost-

effectiveness calculations.[109] If in the above example, the risk for repeat revascularisation is lower under real-world circumstances in the BMS comparator group, the absolute reduction in repeat revascularisation and thus the potential gain of using DES will be smaller than RCTs might indicate.

Next to performing pragmatic RCTs, another approach to check and handle this problem is to combine the strengths of both observational and RCT data (Figure 5).[114] Reliable administrative or register data can provide an estimate of the real-world baseline risks for specific events under usual circumstances. In combination with the relative treatment effect from well-performed RCTs this results in an estimate of the absolute benefit for the relevant target population. Applying this approach, one must remain cautious about the validity of the assumption of a constant relative treatment effect.[114] We note that it can be a challenge to find up-to-date information on the baseline risk for a particular event at the national level. If such information is available, e.g. from administrative databases, the reliability must be checked.

This approach was applied in the above example of DES versus BMS.[109] Belgian administrative data on about 11 500 patients with BMS showed that the number of repeat hospitalisations was on average almost 15% after one year of which less than half (43%) were due to restenosis.[110, 115] This implies a real-world baseline risk for this population of less than 7% for repeat hospitalisations due to restenosis with BMS. Next to the protocol-driven angiographic follow-up, this might also be due to e.g. differences in eligibility criteria in the RCT versus real-world selection of patients. Nevertheless, whatever the cause might be for this lower baseline risk, modelling an incremental benefit that is higher than the real-world baseline risk is unrealistic and should be avoided. Including an adjustment for the real-life baseline risk as presented in Figure 5 may avoid such implausible modelling.

Figure 5: The influence of the baseline risk on the absolute treatment effect - (Appropriate) use of specific sources to support economic evaluations.



Based on: Neyt et al., Health Policy, 2012.[114]

Extra information

- Endpoints used for Relative Effectiveness Assessment: Clinical Endpoints. Methodological Guideline: EUnetHTA; 2015[13]
- Levels of evidence: Applicability of evidence for the context of a relative effectiveness assessment. Methodological Guideline: EUnetHTA; 2015.[14]

3.5 Compliance/adherence and persistence

The concepts of compliance and adherence generally refer to a patient completing a treatment regimen as set out by a health care professional. Compliance and adherence are often considered synonyms, however 'compliance' implies a passive role for the patient, where they do or do not follow the instructions of their clinician. 'Adherence', on the other hand, is considered a measure of the extent to which a patient's behaviour coincides with the advice of their clinician. This document will use the phrase 'adherence' in preference to 'compliance'.[116] ISPOR has defined adherence as the extent to which a patient acts in accordance with the prescribed interval, and dose of a dosing regimen.[117] Adherence implies a binary response – a patient does or does not adhere – and does not capture how long a patient follows the advice. The duration may be captured by the concept of 'persistence', which reflects the duration of time from initiation to discontinuation of therapy.[117] Therefore, adherence and persistence are both relevant when considering the extent to which a patient completes treatment.

The concepts of adherence and persistence apply not just to pharmaceuticals, but can be extended to most technologies. For complex care pathways, adherence can potentially encompass a range of steps. In a cancer screening programme, for example, steps could include an initial screening test, confirmatory testing, and follow-up care.

Adherence and persistence are not an issue for all technologies. For example, if a treatment is entirely completed in a hospital setting under clinical supervision, it may be reasonable to assume that patients for whom treatment is initiated are fully adherent for the duration of treatment.

From an economic evaluation perspective, adherence and persistence could be important as failure to complete treatment as intended can independently impact on both clinical and economic outcomes. Poor adherence may diminish beneficial clinical outcomes for patients, but equally may reduce the incidence of treatment-related adverse effects. Indeed, poor adherence may be directly due to a patient experiencing adverse outcomes of treatment, such as nausea. Poor adherence can also have direct economic implications due to medicine wastage. For short-term treatments, at an individual patient level poor adherence and persistence may mean that all or most of the costs are accrued (for example, for a short course of medicine) but few or none of the benefits. With longer-term treatments, proportionately lower costs may accrue, particularly if treatment is discontinued early on. Economic implications also arise indirectly through changes in clinical outcomes.

Failure to persist with a treatment can have important implications with respect to chronic conditions, where efficacy may be measured using surrogate markers (e.g., blood pressure, A1C, lipid levels) and long-term reductions in morbidity and mortality are estimated under the assumption of persistence with treatment.

Typically the reporting of adherence in randomised controlled trials is poor, and the methodology used is inconsistent across trials.[118, 119] Obtaining relevant data for economic models may therefore be challenging. The relevance of adherence and

persistence to economic evaluation is the extent to which the evaluated population will use the intervention as intended.

It is acknowledged that there may be limited data to support assumptions about adherence, and that the importance of those assumptions will be highly context specific. As such, when critically appraising an economic evaluation one must consider the context of the intervention and indication to determine whether adherence is likely to be an issue, and whether it has been adequately addressed in the evaluation.

Points for consideration

Some of the potential issues relating to the (non-)inclusion of adherence in economic evaluations are:

- Is there a reason to believe that adherence and/or persistence are important for the technology under evaluation? It may be important because there is evidence to suggest issues with poor adherence and or persistence for that treatment, or because poor adherence would plausibly have a substantive impact on cost-effectiveness. The most likely situation is that adherence has not been explicitly considered but it is likely to impact on cost-effectiveness. In that case, a judgement needs to be made as to whether the omission of adherence/persistence will substantively bias the results.
- The assumption that adherence in the modelled population will be the same as in the underlying trials of efficacy/effectiveness may not hold. If adherence in the target population is lower than observed in the trials then outcomes may be biased (for both effectiveness and safety).
- If the treatment pathway can be disaggregated into a number of steps, each of which requires adherence, is the evidence of outcomes linked to adherence to all steps in the pathway? It is important that the estimates of adherence correspond to the same pathway as the evidence of clinical effect.
- The impact of poor adherence should be quantified based on appropriate evidence. It is possible that such information is not available. If available, it should be clear if the impact of poor adherence is based on the trials used to estimate clinical effectiveness or if it has come from other sources (Box 11). The impact of adherence and persistence might be addressed in a sensitivity analysis.
- A dose-response relationship may exist such that the effectiveness of the treatment is correlated with the quantity of treatment received. In relation to persistence, a dose-response relationship could be significant. For example, completing half the course of a medicine may be more effective than taking a quarter. If such a relationship is assumed, then there must be clear evidence of how the dose-response relationship was determined, and the source of the persistence data (Box 11).
- There are wide-ranging factors affecting adherence depending on the treatment under evaluation,[120] including: age, sex, socio-economic status, ethnicity, and poly-pharmacy. Where patient subgroups are being modelled, adherence may be different from that across the entire patient population. That is, adherence could, for example, be associated with age such that older patients have greater adherence and younger patients have poorer adherence than the population average.

Examples

Box 11: Differing assumptions about adherence and the presence of a dose-response relationship

In a review of economic evaluations of gender-neutral school-based human papillomavirus (HPV) vaccination programmes for children, 28 studies were identified.[121] The available vaccines were originally licensed on the basis of a three-dose schedule, but this was subsequently revised to a two-dose schedule for those aged less than 15 years. The change in doses was also associated with a change in timing from doses at one, three and six months to doses at one and six months. Increasing the time lapse between doses may plausibly reduce adherence. Twenty-two studies used a three-dose schedule in the base-case analysis, while four used a two-dose schedule and two used both two- and three-dose schedules in the base-case.

Four studies were identified that explicitly referred to a dose-response relationship that could be linked to adherence, and reported separate figures for coverage (the percentage of children who presented for the first vaccination) and adherence (the percentage that completed the full schedule of doses) (Table 3). For the other 24 studies it was assumed that there was no efficacy unless the full vaccination schedule was received.

Table 3: Adherence to HPV vaccination

Study	Dose schedule	Coverage	Adherence	Efficacy (relative to full schedule)	
				With 1 dose	With 2 doses
Bresse (2014)[122]	3	65%	80%	23%	45%
Haeussler (2015)[123]	3	90%	Not stated	25%	50%
Largeron (2017)[124]	2	16-56%	90%	0%	100%
Mennini (2017)[125]	2	71%	90%	0%	100%

Source: Health Information and Quality Authority (HIQA). Health technology assessment (HTA) of extending the national immunisation schedule to include HPV vaccination of boys: Draft report for public consultation. Dublin: HIQA; 2018.[121]

The impact of adherence on efficacy was justified based on data in only one of the four studies, where the data used were in relation to a hepatitis B vaccine (Bresse[122]).

Only two[122, 125] of the four studies referred to adherence in their univariate sensitivity analysis, and the exploration was only in terms of the percentage adherent to the full schedule of doses, and not on the impact on efficacy.

In this example, there is limited variability in the assumption regarding adherence. Local data relevant to the technology under evaluation is preferable. In this case that

has only been used by one study.[125] In the absence of local data, there is a preference for relevant international data – which has been used by a second study.[122] Failure to state the assumed value creates challenges in determining the applicability of the results to other settings.

The impact on efficacy is potentially quite important when considering cost-effectiveness in this example, and yet none of the four studies reported a univariate sensitivity analysis of efficacy. Only one study used referenced efficacy data, albeit for a different vaccine. Those data were available to the other studies, but they used assumptions in preference that were not tested in a sensitivity analysis. It is important that the impact of including adherence data, particularly when it is based on expert opinion or assumptions, is adequately explored. The impact of adherence is also important for costs - in the absence of capital costs the total cost of delivering the vaccination programme is directly related to the size of the cohort: a 10% reduction in adherence implies a 10% reduction in the cost of the programme.

Extra information

- Levels of evidence: Applicability of evidence for the context of a relative effectiveness assessment. Methodological Guideline: EUnetHTA; 2015.[14]

3.6 Quality of life

The EUnetHTA guideline for methods for health economic evaluations recommends that:

- Results should be presented in terms of both a cost-effectiveness analysis (CEA) and a cost-utility analysis (CUA).[1]
- the primary outcome measure(s) should be presented where appropriate as natural units (including life-years) and as QALYs.[1]

The health-related quality of life (HRQoL) aspects of the QALY are captured in a HRQoL weight, expressed as utilities. Based on the review of guidelines used by EUnetHTA partners, the EuroQoL-5 dimension (EQ-5D) questionnaire is the most commonly recommended instrument for the derivation of HRQoL weights, although other instruments are also mentioned (e.g. Health Utility Index (HUI), Short Form-6 dimension (SF-6D) or 15 Dimension instrument (15D)).[1] As for all applied questionnaires, the EUnetHTA guidelines also state that "*documentation of the validity, reliability, responsiveness and acceptability of the HRQoL instruments used in REA should be provided.*"[10]

The EUnetHTA guidelines on HRQoL generally recommend both the complementary use of disease- or population-specific and a generic HRQoL measure to adequately capture the impact of a disease on daily life.[r] For countries that require an economic evaluation to

[r] The EUnetHTA guideline on HRQoL mentions that "*the purpose of the REA and the policy context determine the best practice guidelines for HRQoL measurement in the context of REA.*"[10] "*HRQoL can be measured for different purposes. The choice of the HRQoL instrument (generic versus disease-specific, utility versus profile measure) used will depend on the objective of the measurement. For cost-utility analyses, for instance, a utility measure is needed. For informing patients or clinicians, disease-specific HRQoL measures may be preferred over generic measures because they might capture better the specific impact of the disease and its intervention.*"[10]

support a health technology reimbursement application (or another health care decision), this guideline recommends that they should:

- require data emerging from the administration of a generic utility instrument in the clinical trial(s).[10]

This is relevant in all cases when there is a need for the calculation of QALYs. Of course, the validity, reliability, and responsiveness of the generic utility instrument need to be taken into account. Gathering QoL data with a generic utility instrument (also called preference-based instruments) in clinical trials does also not prevent researchers from using data from other sources (e.g. utility values from a bigger and more generalizable cohort than the select population in a clinical trial or including disutilities retrieved from another source when adverse events (AEs) occur). However, researchers should avoid situations where no relevant QoL information is gathered, given that this information may be required to perform reliable economic evaluations.

Points for consideration

Some of the potential pitfalls and important issues to consider related to utility values in economic evaluations are:

- Applying hypothetical (non-evidence based, e.g. based on expert opinion or non-comparative observational studies) utility weights in the economic evaluations because no generic utility instrument is used in the underlying trials. Modelling of utilities without supporting evidence for the applied values might be problematic (unless in e.g. cases were utility values do not seem to have a high impact on the ICERs or when results are very positive under pessimistic assumptions and vice versa).
- Extracting utility values for separate treatment arms from different sources that might even have used different indirect (e.g. EQ-5D-3L, EQ-5D-5L, SF-6D, SF-36, etc.[s]) or direct methods (e.g. time trade-off (TTO), visual analogue scale (VAS) or standard gamble (SG)) of valuation or applied different tariffs (e.g. from different countries[t]) for the same indirect instrument is accompanied by the necessary uncertainty. This is due to both the lack of a direct treatment comparison, as well as the possible differences in utility values if another instrument, method or tariff is used,[126-128] and thus might not be a good estimate of the incremental treatment effect on QoL.

[s] Next to the differences in questions, also the time window differs across questionnaires. For example, the EQ-5D questionnaires ask to "*describe your own health state today*", while this is "*during the past 4 weeks*" in the SF-36 questionnaire or "*during the past week*" in the EORTC QLQ-C30 questionnaire.

[t] We remark that in journal articles often the mean and confidence interval of utility values are published (e.g. in the two treatment arms, before the intervention and at pre-specified points in time after the intervention), applying a tariff from a specific country. It is difficult to apply the tariffs from another country, unless researchers have access to the patient-level data.

- Mapping the outcomes of disease-specific questionnaires to outcomes from a generic utility instrument. Some countries[u] accept mapping if no other data are available. Finding the optimal mapping equation is not straightforward and entails an extra level of uncertainty. Possible manipulation of results through mapping is illustrated in Box 12. Scenario analyses is recommended if mapped utilities are used in cost-utility analyses.[129, 130] As mentioned by Longworth and Rowen, "*mapping can provide a route for linking outcomes data collected in a trial or observational study to the specific preferred instrument for obtaining utility values. In most cases, however, it is still advantageous to directly collect data by using the preferred utility-based instrument and mapping should usually be viewed as a "second-best" solution.*"[131] This is in line with the EUnetHTA recommendations stating that "*mapping of disease-specific or generic instruments to preference-based instruments to obtain utility values is generally not recommended for REA. Authorities should encourage researchers to always include a preference-based instrument in their clinical trial protocol in order to avoid the need for mapping.*"[10][v]
- Similarly, making a (non-evidence based) link between an intermediate/surrogate endpoint and QoL should be interpreted with caution (see example in Box 15).
- Not adjusting for quality of life due to a lack of such data in cases where there are differences in QoL between the treatment arms, e.g. when a specific treatment prolongs life at the expense of the patient's quality of life. In case no utility outcomes were measured and mapping or hypothetical utilities were applied, the robustness of outcomes should be tested in sensitivity analysis and results should be interpreted with caution.
- Assuming a utility of 1 (i.e. perfect health) for patients without any adverse event will very probably overestimate the quality of life for these patients since, in a sample of patients, the average utility will never/most likely not equal 1. This is also the case for a sample of the general population. In case the population of interest without adverse events is similar to the general population (i.e. without any disease-specific symptoms which would have an impact on HRQoL), age-adjusted utility values of the general population could be used instead of applying a utility value of 1.
- Similarly, in the case of e.g. extrapolations to a longer time horizon, it is also important to check whether it is necessary to adjust the utilities for ageing.
- If disutilities are applied for adverse events, the duration of the adverse events should be taken into account.
- When applying multiple disutilities to the same patient at the same time, researchers should carefully look at how these disutilities are taken into account (e.g. as an additive function or in a multiplicative way) and judge whether the chosen approach is justified.

[u] Czech Republic, England, Ireland, Italy, Norway, Scotland, and CatSalut in Spain.[1]

[v] We refer to the ISPOR Good Practices for Outcomes Research Task Force Report on mapping for further information on this topic.[132]

Examples

Box 12: Possible impact on cost-effectiveness results linked to mapping of disease-specific or generic instruments to generic utility instruments

In an HTA report on cardiac resynchronisation therapy (CRT) for patients with chronic heart failure (HF) that are receiving optimal medical treatment,[92] underlying trials traditionally encode the functional status of patients with HF by means of the New York Heart Association (NYHA) classification. The ranking is as follows:[w]

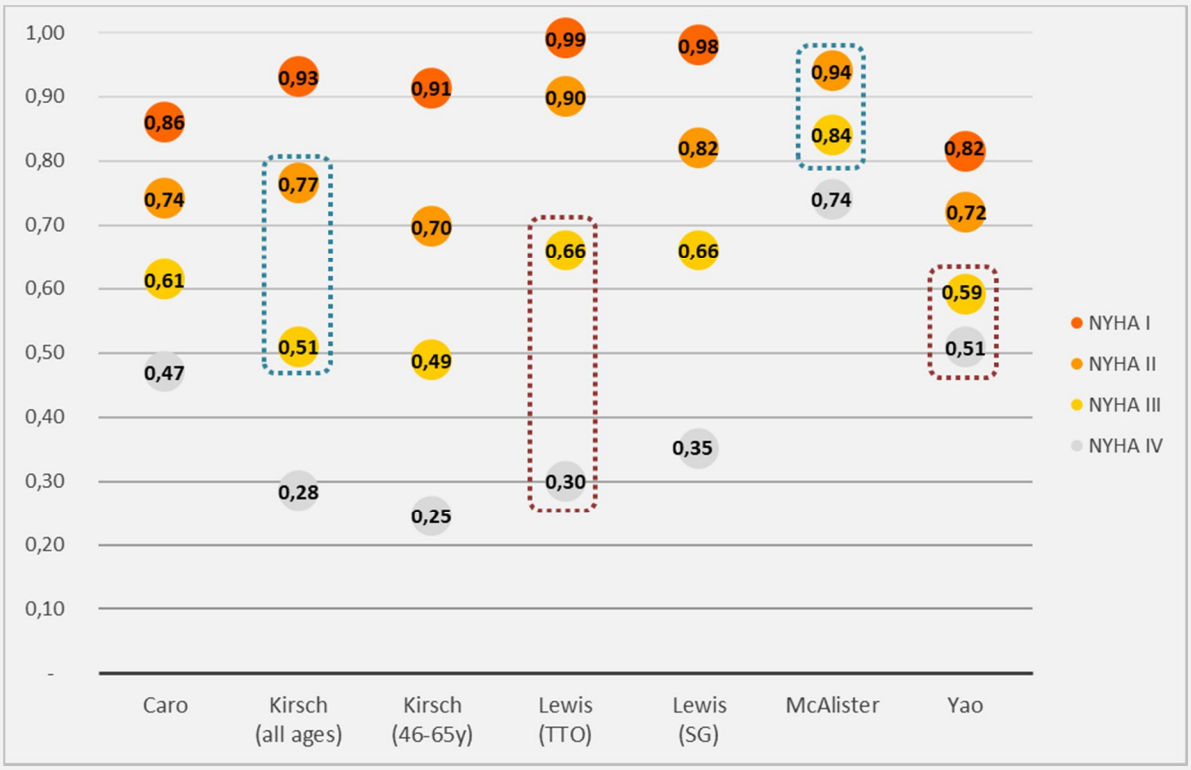
- Class I: No limitation of physical activity. Ordinary physical activity does not cause undue fatigue, palpitation, dyspnea (shortness of breath).
- Class II: Slight limitation of physical activity. Comfortable at rest. Ordinary physical activity results in fatigue, palpitation, dyspnea (shortness of breath).
- Class III: Marked limitation of physical activity. Comfortable at rest. Less than ordinary activity causes fatigue, palpitation, or dyspnea.
- Class IV: Unable to carry on any physical activity without discomfort. Symptoms of heart failure at rest. If any physical activity is undertaken, discomfort increases.

If no generic utility instrument is used, mapping may be considered. However, as shown in Figure 6, different sources linking the NYHA classification to a utility value can be identified. This might introduce a possible manipulation in the direction of a preferred outcome. If the authors prefer a big difference in utilities between NYHA III and IV, they can refer to the study of Lewis et al.[133] In the opposite case, authors might refer to Yao et al.[134] Similarly for differences between NYHA II and III classification: high utilities with small differences are achieved when referring to the study of McAlister et al.,[135] while bigger differences are noticed in the study population of Kirsch et al.[136] In this case, this uncertainty comes on top of the poor agreement in differentiating NYHA class II and III patients.[x] Researchers and assessors should be aware that such an approach is prone to bias. Authors should justify the choice of the utilities used in the reference case and should include the other sources in sensitivity analyses unless there is suitable justification for their exclusion. In such cases, performing a systematic search strategy for QoL information is recommended.

[w] Source: http://www.heart.org/HEARTORG/Conditions/HeartFailure/AboutHeartFailure/Classes-of-Heart-Failure_UCM_306328_Article.jsp

[x] Assigning a NYHA class II or III is very subjective. Inter-operator comparisons on NYHA class II and III patients gave a result that was little better than chance with an agreement between independent assessors of about 55%.[137-139]

Figure 6: Different published utility estimates linked to living with severities of heart failure (NYHA I-IV)



Source: based on figure 8.2 from a KCE HTA report.[96] In the study of Caro et al.,[140] utilities are calculated using a mapping equation linking outcomes from the disease-specific Minnesota Living with Heart Failure questionnaire (MLWHF) to EQ-5D scores. In the studies of Kirsch et al.[136] and Lewis et al.,[133] a time-trade-off (TTO) method is applied. The latter study also applies a standard gamble (SG) approach. This is also the case in the analysis of McAlister et al.[135] Finally, Yao et al.[134] use estimates from the EQ-5D questionnaire to assign utility scores to the NYHA classes.

Extra information

- Endpoints used for Relative Effectiveness Assessment: health-related quality of life and utility measures. Methodological Guideline: EUnetHTA; 2015.[10]

3.7 Intermediate/surrogate versus final endpoints

A **clinical endpoint** is a “direct measure of how a patient feels, functions or survives”.[141] A **surrogate endpoint** is an endpoint that is intended to replace a clinical endpoint of interest that cannot be observed in a trial - it is a variable that provides an indirect measurement of effect in situations where direct measurement of clinical effect is not feasible or practical.[15] A surrogate endpoint may be a biomarker (e.g. blood sugar level) or an intermediate clinical endpoint (e.g. progression-free survival or response rate).

In most countries, the preferred measures for economic analyses are life-years and QALYs.[1] The use of surrogate endpoints in health technology assessment is controversial because not every surrogate endpoint has a demonstrated association with clinical benefits for the patient. A surrogate endpoint may be considered valid if it is sensible, measurable, interpretable and highly accurate in predicting the clinically relevant endpoint. The acceptability of a surrogate endpoint in supporting effectiveness of an intervention is mostly based on its biological plausibility and empirical evidence.[15]

The EUnetHTA guideline for surrogate endpoints[15] recommends that:

- "*the REA should be based whenever possible on final patient-relevant clinical endpoints (e.g. morbidity, overall mortality).*"
- "*In the absence of evidence on a final patient-relevant clinical endpoint that directly measures clinical benefit, both biomarkers and intermediate endpoints will be considered as surrogate endpoints in REA if they can reliably substitute for a clinical endpoint and predict its clinical benefit.*"
- "*If surrogate endpoints are used for REA, they should be adequately validated: the surrogate-final endpoint relationship must have been demonstrated based on biological plausibility and empirical evidence. …*"

In addition, the EUnetHTA guideline on clinical endpoints[13] used for relative effectiveness assessment states that:

- "*If progression-free survival (PFS) is used as an endpoint there should be sufficient independent evidence to demonstrate that this is associated with overall survival. …* [y]
- *Overall survival is the gold standard for demonstrating clinical benefit and as such should be used where possible. …*
- *In the metastatic setting, data on PFS alone is insufficient and should be coupled with quality of life assessment and survival data, the maturity of which will be considered on a case by case basis.*"

For more information and recommendations on this topic, please refer to the EUnetHTA guidelines.[13, 15]

Points for consideration

Some of the potential problems related to the (non-)inclusion of final endpoints in economic evaluations are:

- Translation of a surrogate endpoint to final endpoint without supporting evidence on the statistical association between both (e.g. PFS to overall survival (OS) or QoL) (see Box 13 - Box 15).
- Use of a disease-specific outcome such as PFS or progression-free life-years saved (PF-LYS) instead of life-years saved or gained (LYS or LYG) or QALYs without solid justification (see Box 16). Available information on overall survival and/or QoL should not be ignored.
- Validation of the surrogate endpoint in a different population/disease or for a different technology, or based on an insufficiently large database.

[y] We note that, ideally, researchers not only report the direction of the link but also its size.

- Old or preliminary evidence on the relationship between surrogate and final endpoint, as the evidence for this relationship may change over time.
- Lack of or insufficient explanation of the quantitative methods used to translate surrogate endpoints into final endpoints.
- Lack of or insufficient sensitivity analysis reporting the effect of the surrogate endpoint in the results.

Note: The use of surrogate endpoint may be acceptable in some cases such as very slowly progressive diseases or rare diseases, where long term clinical outcomes and/or big samples are not available. However, the validity of a surrogate outcome depends on empirical evidence, not on the size of the target population. Therefore, also in these cases, the use of surrogate endpoint should be justified and discussed by the authors.

Note: It is worthwhile to mention that a non-negligible proportion of approved oncology drugs enter the market without mature overall survival (OS) data and/or quality of life data. Davis et al.[142] looked at 48 cancer drugs approved by the European Medicines Agency (EMA) for 68 indications between 2009 and 2013. The authors found that "*at the time of market approval, there was significant prolongation of survival in 24 of the 68 (35%). … Out of 44 indications for which there was no evidence of a survival gain at the time of market authorisation, in the subsequent postmarketing period there was evidence for extension of life in three (7%) and reported benefit on quality of life in five (11%).*"[142] The authors concluded that "*most drugs entered the market without evidence of benefit on survival or quality of life.*"[142] The above mentioned points for consideration are thus applicable to many cancer drugs entering the market.

Examples

Box 13: The link between the surrogate endpoint 'progression-free survival' and the final endpoint 'overall survival': the case of cancer treatments

There is a lot of literature on the relationship between surrogate and final endpoints in cancer.[143-145] Intermediate endpoints are widely used in the economic evaluation of new treatments for advanced cancer in order to estimate overall survival (OS).[146] The evidence supporting the relationship between progression-free survival (PFS) or time to progression (TTP) and OS varies by cancer type and is not always consistent within one specific cancer type.[143, 147] Some authors even state that the acceptability of progression free survival does not have the same impact in adjuvant or metastatic disease, that is, PFS appears acceptable in the adjuvant setting, but PFS alone is insufficient in the metastatic setting.[145] Next to providing evidence of an underlying link between the surrogate and final endpoint, several authors coincide in recommending that economic evaluations using PFS as a surrogate endpoint of OS should explain how the relationship has been quantified in the model and include a sensitivity analysis to explore the uncertainty related to this relationship.[147, 148]

Fischer et al. identified different type of studies regarding the relationship between PFS and OS.[143] They did not identify any review finding PFS an appropriate surrogate of OS. On the contrary, they found one review concluding that PFS is not an appropriate surrogate[145] and one review concluding that it depends on particular factors.[149] The first one by Prasad et al. tried "*to identify and evaluate trial-level meta-analyses of randomized clinical trials quantifying the association between a surrogate endpoint and overall survival in medical oncology. Trial-level correlations test whether treatments that improve the surrogate endpoint also improve the final endpoint and are widely considered the strongest evidence to validate a surrogate endpoint.*"[145] Unfortunately, there was only a low correlation and it was concluded that the evidence supporting the use of surrogate endpoints in oncology was limited.[145] In the metastatic setting, the study identified 8 meta-analyses examining whether gains in PFS predict overall survival in metastatic breast cancer. Six reported low correlation; 1 reported medium correlation; and only 1 reported a strong correlation.[145]

Box 14: An example of the lack of relationship between progression-free survival and overall survival: bevacizumab for metastatic breast cancer

The E2100 trial[150] was a single, open-label randomised controlled trial that found statistically significant increase in median PFS for the combination bevacizumab + paclitaxel compared to only paclitaxel in patients with untreated metastatic breast cancer (median PFS: 11.8 versus 5.9 months, P<0.001). However, median OS was not statistically significantly different between the two groups (median OS: 26.7 versus 25.2 months, P = 0.16). No significant differences in the mean change in QoL scores from baseline, measured with the Functional Assessment of Cancer Therapy–Breast (FACT-B) questionnaire, were found either.[150] Two placebo-controlled, randomized trials, AVADO[151] and RIBBON-1[152], were designed to validate the findings of E2100. These trials found statistically significant improvement in PFS and response rate in the groups receiving bevacizumab in addition to chemotherapy. But the OS data showed no statistically significant differences between arms in both studies.

The Food and Drug Administration (FDA) in the USA granted accelerated approval of bevacizumab based on the improvement in PFS in E2100. Later the FDA revoked the indication of bevacizumab in metastatic breast cancer when the AVADO and the RIBBON-1 trials "*failed to verify AVASTIN's clinical benefit.*"[z] The European Medicines Agency (EMA), on the contrary, stated that paclitaxel plus bevacizumab prolonged PFS without negative effects on overall survival and "*concluded that the benefit-risk balance for this combination treatment remains positive.*"[aa] At present, bevacizumab in combination with paclitaxel is indicated for first-line treatment of adult patients with metastatic breast cancer in Europe.

A recent review[153] found three economic evaluations estimating the cost-effectiveness of bevacizumab in metastatic breast cancer.[154-156] The three economic evaluations used the E2100 trial results on OS as the source for effectiveness (median OS: paclitaxel: 25.2 months; bevacizumab+paclitaxel: 26.7 months).

Instead of making use of PFS and making an indirect link to OS, it is better to directly model with the available survival data. Dedes et al. was the first economic evaluation published.[154] In their study from the perspective of the health care system in Switzerland, the authors made use of OS data and calculated an ICER of €189 427 per QALY (euros of 2008). The authors clearly discussed the limitations of their study, referring not only to E2100[150] but also to the AVADO[151] and RIBBON-1[152] trials that were not finished at the moment: "*the study is based on the efficacy and safety data of one single randomised trial and such results usually differ from what is seen in routine clinical practice. Further phase III trials with bevacizumab in breast cancer are underway studying other combination treatments (e.g. docetaxel and anthracyclines) but could not be included in this study as efficacy data are still*

[z] https://www.gpo.gov/fdsys/pkg/FR-2012-02-27/pdf/2012-4424.pdf

[aa] European Medicines Agency. Questions and answers on the review of Avastin (bevacizumab) in the treatment of metastatic breast cancer. 16 December 2010. http://www.ema.europa.eu/docs/en_GB/document_library/Medicine_QA/2010/12/WC500099939.pdf

immature. However, the interim results do not show any hints that assumptions and results of this study would go in a wrong direction". With the latter, the authors wanted to indicate that they did not expect better/lower ICERs based on the interim results from the newer trials than the one they had estimated. Indeed, unfortunately, published results of these trials were less positive and would lead to worse ICER estimates. This was confirmed by the study of Montero et al.[155] Using the data from the other two trials in the sensitivity analysis, they estimated even higher ICERs than in the E2100 base case, as expected, given the poorer results obtained in the AVADO and RIBBON-1 trials in comparison to the results in E2100.[155]

If the authors of these economic evaluations would have used PFS instead of OS or make an indirect link between PFS and OS, the results would have been too optimistic. This underlines that results of studies based on a potential relationship between PFS and OS (or another indirect link that is not supported by evidence) should be used with caution.

Box 15: The link between the surrogate endpoint 'progression-free survival' and 'quality of life' or 'quality-adjusted life years'

In 2019, an article was published by Hwang and colleagues[157] in which the association between progression-free survival and patients' QoL in cancer clinical trials was analysed. The authors performed a retrospective study of phase III clinical trials of drugs for advanced or metastatic solid tumors published between 2010 and 2015. They identified 352 phase 3 trials and QoL data was available for 147 clinical trials. Based on their study results, "*the association between PFS and improvement in global quality of life was weak (r = 0.34; AUC [area under the curve] = 0.72), as was the association between PFS and improvement in any domain of quality of life. In conclusion, PFS benefit was not strongly correlated with improvements in patients' quality of life, and, despite the palliative intent of treatments in the advanced/metastatic setting, the availability of quality of life data from clinical trials of cancer drugs was poor.*"[157]

In economic evaluations, a link between PFS and QoL improvements might be assumed without providing supporting evidence or even when there is evidence contradicting this assumption. As an example, in an HTA report assessing bevacizumab in the treatment of ovarian cancer, a systematic literature review of economic evaluations was performed.[158] At that time, none of the underlying RCTs provided utility estimates per treatment arm. All the economic evaluations made assumptions regarding the impact of bevacizumab treatment on QoL. Some authors modelled a decrease in QoL due to more adverse effects, while others modelled an improvement through prolonged PFS. In case of the latter, the argument was provided that "*cancer survivors whose disease recurs have a worse HRQoL in most indices than those who remain disease-free[159] and the factor causing most distress among cancer patients (and therefore impacting on HRQoL) has been found to be the fear of disease progression[160].*"[161] On the other hand, "*new treatments that increase PFS may not be of sufficient value to patients with advanced-stage cancer unless accompanied by tangible quantity or quality of life advantages. Any symptom relief that patients gain from treatment resulting in tumor shrinkage or stabilization must be balanced against the toxic effects that drug therapy itself creates.*"[162] In this case, both the HTA report and the Evidence Review Group (ERG) Report commissioned by the NIHR HTA Programme on behalf of NICE found that "*some women receiving bevacizumab has a statistically significant but clinically small detriment in global QoL but no HRQoL data are presented … .*"[163]

Also in the case of bevacizumab treatment for breast cancer, improvement in progression-free survival did not translate to a better OS or QoL. In contrast, more AEs were noticed.[164, 165]

Modelling an improvement in QoL through PFS alone should thus be considered with caution. Following the EUnetHTA guidelines on HRQoL, for the calculation of QALYs, this problem could be avoided by considering to include a generic utility instrument and timely publishing of the results instead of mapping a surrogate endpoint to utilities (see part 3.6).

Box 16: A surrogate endpoint used directly for the comparison of alternatives: 'progression-free life years saved' instead of 'life-years saved' or 'quality-adjusted life years'

Smith et al.[166] assessed concurrent and adjuvant chemoradiation with gemcitabine/cisplatin in patients with cervical cancer. In this study, the consequence used as denominator in the ICER estimate was not life years or QALYs gained or saved but progression-free life-year saved (PF-LYS). According to their conclusions radiation and gemcitabine/cisplatin for patients with stage IIB to IVA cervical cancer had an ICER of $97 799 per PF-LYS.

Most HTA bodies and health economists prefer outcomes to be expressed in final endpoints such as LYG or QALYs gained.[1] PF-LYS is not the same and might be difficult to interpret. If the progression is postponed at the expense of worse QoL, e.g. due to (severe) adverse events linked to the treatment, then focusing on PF-LYS might be misleading and must be considered with caution.

Extra information

- Endpoints used for Relative Effectiveness Assessment: Clinical Endpoints. Methodological Guideline: EUnetHTA; 2015.[13]
- Endpoints used in Relative Effectiveness Assessment: Surrogate Endpoints. Methodological Guideline; 2015.[15]

3.8 Time horizon & extrapolation

3.8.1 Time horizon

The EUnetHTA guideline (May 2015) for methods for health economic evaluations recommends that:

- "*the time horizon for the reference case analysis should be sufficiently long to reflect all relevant differences in costs or outcomes between the technologies being compared. The choice concerning any alternative time horizon for the reference case analysis should be clearly justified and described.*"

This is in line with recommendations from other non-EU HTA bodies like the Australian Pharmaceutical Benefits Advisory Committee (PBAC)[167] or the Canadian Agency for Drugs and Technologies in Health (CADTH)[168] and the USA panel on CEA.[169]

Points for consideration

The applied time horizon might have an important impact on the ICER.

- Costs and effects should be modelled for the same time horizon.
- Time horizon should be specified in light of realistic elements (long term data, history of disease) and not solely be based on simulations. The plausibility of the simulated lifetime horizon has to be verified. "*A lifetime horizon relates to the life expectancy of the relevant patient population. Inputs that are not realistic will result in a model predicting an implausible duration of outcomes or survival and, thus, an implausible lifetime time horizon. The assessment of plausibility should also apply to how the model extrapolates the curves to reach this time horizon.*"[167]

- "*As a modelled time horizon extends – in absolute terms and relative to available data – it is associated with increasing inherent uncertainty. Therefore, economic claims based on models with very extended time horizons and predominantly extrapolated benefits will be less certain and are likely to be less convincing to the PBAC.*"[167]
- Parameters related to incremental costs and benefits may vary across simulations but be held at its current value over the time horizon within a simulation. This simplification supports the analysis of uncertainty regarding what we know now, rather than what is unknown about future events (e.g. how will population mortality rates look like in the future? Will the treatment cost for a specific adverse event be higher or lower in the future?).

3.8.2 Extrapolation

Extrapolation is necessary when the time horizon of cost-effectiveness analysis extends beyond the period for which observed data are available. For example, extrapolation will be necessary if the analysis follows patients to 20 years when the longest available trial follow-up is three years. The key consideration underlying extrapolation relates to the transition between health states over a time horizon not captured within the clinical trials that provide the data on initial treatment effect. The transition probabilities may be estimated by extrapolation from the trial evidence, such as is often the case for survival, or may be developed using other data sources such as observational studies or registry data.

The EUnetHTA guideline (May 2015) for methods for health economic evaluations recommends that:

- "*When data are extrapolated beyond the duration of the clinical trials, all assumptions need to be clearly presented and analyzed using different scenarios.*"

Other recommendations of importance for the critical assessment of economic evaluations include:

- The "*clinical and biological plausibility"* of extrapolations should be assessed and that "*alternative scenarios should also be routinely considered to compare the implications of different methods for extrapolation of the results.*"[86]
- "*Derive extrapolations of data where necessary; explain and justify methods used, and prepare alternatives for sensitivity or scenario analyses.*"[167]

In relation to survival functions, to extrapolate beyond the trial follow-up using the observed trial data, a model must be developed that predicts the time to event using some curve. Commonly used functional forms include Kaplan-Meier, Weibull, exponential, log normal and gamma. A complication is that the observed data are typically censored: that is, trial participants are followed for a pre-specified period of time (e.g., 12 months) and it is likely that not all will have experienced the outcome of interest in that time. Survival can be modelled using either a survival function or a hazard function. Both are related and estimated using the available trial data, but may be supported by different assumptions. Functions may be based on parametric or non-parametric models, and may assume constant or time-dependent relationships between survival in the treatment and comparator arms. It is critical that the goodness-of-fit of the applied models is adequately explored, and that the plausibility of the assumptions underlying the choice of model is clearly stated. This is particularly relevant where there is a substantial difference between the length of trial follow-up and the time horizon of the economic model. For example, the proportional

hazards model assumes that the relative hazard of events is constant over time. A selected function might represent a good fit to the observed data but generate long-term predictions that are not credible.

In some cases outcomes are not reported as time to event, so it is only known how many patients experienced the outcome at a single specific time point (e.g., at 6 months follow-up). In these situations there are many methods of estimating longer-term transition probabilities. Data to support the development of transition probabilities may come from a wide variety of sources such as observational studies and registry data. A key consideration is whether the data sources are applicable to the target population in the evaluation, and whether the data are consistent and plausible given the trial data used to estimate clinical effectiveness. It is also important to note whether probabilities have been estimated from rates, and whether the appropriate steps have been taken to convert rates to probabilities. Rates reflect the instantaneous potential for an event to occur while a probability gives the likelihood that the event will occur over a specific period of time.

Points for consideration

Cost-effectiveness can depend on the extrapolation of outcomes from trial data: the model extrapolates an observed treatment effect, perhaps with a relatively short follow-up (e.g., two years), over a longer-term time horizon (e.g., lifetime).

- Where extrapolation is used in an evaluation, it is critical that the underlying assumptions and data sources used to extrapolate beyond the trial time horizon are clearly described.
- The impact of different extrapolation scenarios should be addressed in sensitivity analyses. It should be avoided that only extrapolation scenarios are modelled that derive an effect on benefits or harms from non-statistically significant study results. (see Box 17-Box 18)
- A key consideration is the assumed persistence of treatment effect beyond the trial time horizon (i.e., whether it remains constant, declines or increases). The assumptions of persistence must be justified, specifically in terms of:
    - Expected effect after treatment completion (for example, on the basis of the mechanism of action), and
    - The maturity of the available data, which may be problematic for early data.
- Where survival data have been extrapolated, the model should be checked to ensure it is appropriate given the trial data. Where proportional hazards are used, there must be justification that it is appropriate beyond the time horizon of the observed data.
- Where numerous models have been fitted to survival data, two or more may have similar goodness-of-fit measures. While multiple models may produce a similar fit to the observed data, they may generate quite different predictions for extrapolations. In these situations the consequences of the choice of model should be considered.
- It should be clear that the extrapolation analysis adequately capture uncertainty. If, for example, survival is being modelled, the analysis should ideally incorporate the uncertainty around the curve that has been fitted to the data.
- The results of extrapolation should be presented with sufficient sensitivity or scenario analyses to understand the consequences of the extrapolation assumptions. For evaluations that incorporate survival analyses, the choice of

model can affect the estimated cost-effectiveness, and ideally, the results would be given for justifiable alternate choices of survival model.

- In partitioned survival analysis (PartSA),[bb] the plausibility and validity of the modelled survival curves should be checked. For example, the PFS curve cannot be higher than the OS curve (see Box 19).[cc]

Box 17: Extrapolation of time to event from trial data

The choice of approach to modelling survival can have a substantial impact on the estimated cost-effectiveness of an intervention. As an example, Gerdtham and Zethraeus evaluated the cost-effectiveness of enalapril relative to standard therapy in the treatment of congestive heart failure patients.[170] Data were available from the main trial with mean 0.515 years follow-up, and a ten-year follow-up study. The evaluation fit different survival functions using the initial trial data and then compared the estimated cost-effectiveness using those predictive models against the results using the actual ten-year follow-up. Four survival functions were tested: gamma, lognormal, Weibull and exponential. The models were tested as null hypotheses against the alternative of a generalised gamma model. The Weibull model performed best at predicting actual survival, while the gamma and lognormal models were best at predicting the difference in survival between the intervention and comparator groups (see Table 4). In the example used in the study, the models that were best at predicting actual survival were poor at predicting the difference in survival, and vice versa. In the paper it is stated that the Weibull and the exponential models were rejected against the gamma model at the 1 percent level of significance. In many evaluations, the difference in events will be more important than the actual events, as the relative difference will generally represent the incremental benefit.

[bb] "*In the PartSA approach, state membership is determined from a set of non-mutually exclusive survival curves. The PartSA approach uses an overall survival (OS) curve to estimate the proportion of people alive over time directly and may include a statistical extrapolation beyond the time horizon of the original study depending on the requirement to model a lifetime horizon and the maturity of the available data. … OS may be further disaggregated or "partitioned" into different health states to allow these health states to have different HRQoL and cost implications. Within PartSA models, there is a survival curve for each health state that describes time from model start (i.e. patient entry in to the model) to transiting to any health state that is further along the sequence. This means that the survival curves do not represent mutually exclusive estimates of state membership.*"(source: http://scharr.dept.shef.ac.uk/nicedsu/wp-content/uploads/sites/7/2017/06/Partitioned-Survival-Analysis-final-report.pdf)"

[cc] This is possible in partitioned survival analysis models as PFS and OS are estimated and modelled independently and do not therefore reflect the structural dependency between endpoints.

Table 4: Estimated survival and incremental cost-effectiveness of enalapril relative to standard therapy in the treatment of congestive heart failure

Estimation method	Mean survival time (days)			Incremental cost-effectiveness ratio (SEK/LYG)
	Standard	Enalapril	Difference	
True survival data	502.26	767.28	265.01	18 387
Modelled survival				
Gamma	865.87	1145.88	280.01	17 402
Lognormal	844.03	1117.11	273.07	17 844
Weibull	555.63	737.25	181.61	26 830
Exponential	405.11	539.00	133.88	36 395

Source: Gerdtham & Zethraeus (2003).[170]
LYG: life-year gained; SEK: Swedish Krona.

An important point to note is the distinction between models being good at predicting observed survival or the difference in survival between groups. In the above example, the models that are good at predicting the difference in survival resulted in an accurate estimate of the true incremental cost-effectiveness. However, those models over-estimate survival time for both the intervention and comparator arms. While this appears not to impact on the cost-effectiveness, it could have more significant implications for the impact on organisational resources. An implementation plan based on an overestimate or underestimate of the expected events could lead to inefficient deployment of resources.

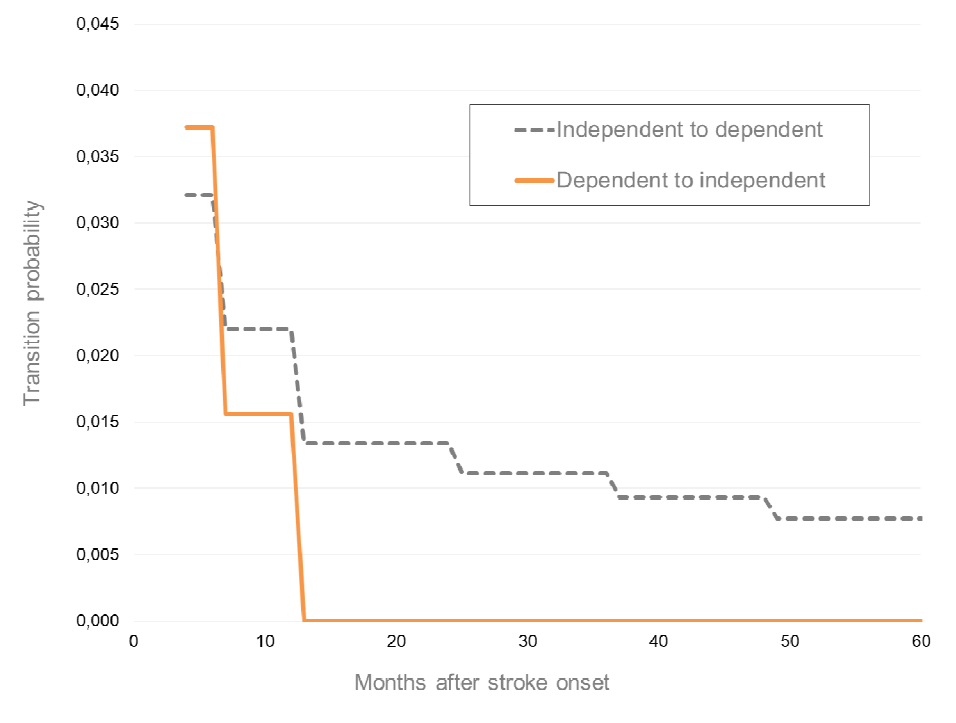
Box 18: Extrapolation without time to event data

Aside from the impact on mortality, acute large-artery ischemic stroke can have profound implications for a person's functional status. Use of new-generation mechanical thrombectomy devices has been consistently shown to increase the proportion of patients achieving good functional outcomes. In the randomised controlled trials evaluating mechanical thrombectomy, functional status was measured at 90 days after stroke onset using the modified Rankin Scale (mRS). Functional status is summarised as independent (mRS 0 to 2), dependent (mRS 3 to 5), or dead (mRS 6). A small proportion of patients may be functionally dependent at 90 days but regain independence after that point. Some patients may become functionally dependent after 90 days having been independent at 90 days. The supporting trials only report outcomes at 90 days and therefore an economic model can either be structured to assume no change in functional status over the longer-term, or else use additional data sources to model longer-term transition probabilities.

An economic evaluation of mechanical thrombectomy with intravenous thrombolysis compared with intravenous thrombolysis alone modelled outcomes to five

years.[171] The model used a decision tree for the first 90 days after acute ischemic stroke. The model then followed patients to five years with monthly cycles using a three-state Markov model (functionally independent, functionally dependent, and dead). Transition probabilities for the Markov model were derived from the Oxford Vascular Study supplemented by a calibration process. Transition probabilities were defined as fixed for periods of months (see Figure 7).

Figure 7: Transition probabilities between functional categories for patients after acute large-artery ischemic stroke



Source: based on data presented in Xie et al. (2016).[171]

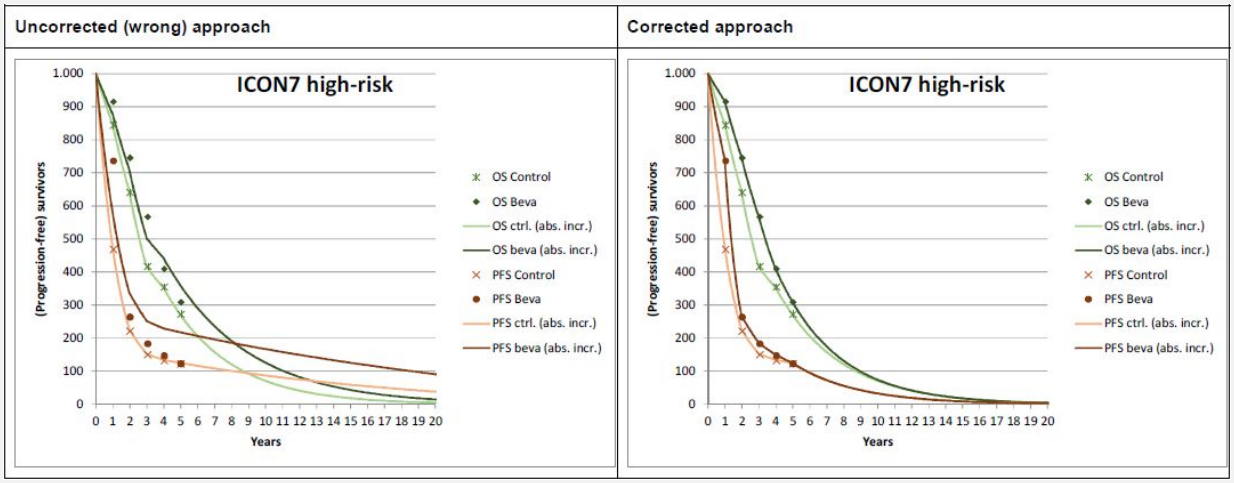
The evaluation estimated the cost-effectiveness of mechanical thrombectomy in Canada while the Oxford Vascular Study collected data on patients in nine general practices across Oxfordshire in the UK. A key consideration is whether the patients in the Oxford Vascular Study are representative of patients that experience acute large-artery ischemic stroke in Canada. In exploring the applicability of the data one must consider the risk factors for stroke and subsequent recovery, severity of stroke, and the treatment pathway for stroke patients. It should be emphasised, in this example, the study team used the data from the Oxford Vascular Study as a starting point, and adapted the data using a calibration exercise that utilised Canadian data.

In the cost-effectiveness model, the transition probabilities were not explicitly varied, and hence it is unclear what influence the transition probabilities had on the estimated cost-effectiveness. A different choice of source data or alternative approach to calibration may have had a significant impact on the cost-effectiveness, but in the absence of a sensitivity analysis, it is not possible to determine the potential impact. The example highlights the challenges in extrapolating long-term outcomes when only short-term point-in-time outcomes are available, and the need to explore the impact of assumptions.

Box 19: Validation of extrapolation in partitioned survival analysis

The Belgian Health Care Knowledge Centre (KCE) published an HTA report studying the safety, efficacy and cost-effectiveness of bevacizumab in the treatment of ovarian cancer.[158] A cost-utility analysis was performed reflecting the results of the identified relevant trials: GOG-0218, GOG-0218 stage IV subgroup, ICON7, ICON7 high-risk subgroup, OCEANS, and AURELIA. The researcher had no access to the underlying individual data and used the published Kaplan-Meier (KM) curves to extract data at specific points in time. Different extrapolation scenarios linked to the Belgian age- and sex-specific mortality rates were applied. As part of the validation of the model, a visual inspection of the model was performed by comparing the modelled OS and PFS curves with points extracted from the published Kaplan-Meier (KM) curves, as well as the position of the OS and PFS curves in the long-term extrapolation phase of the model.[158] Figures were published to support this validation exercise. In case of the ICON7 high-risk subgroup model, for the comparator arm, the modelled OS and PFS curves coincided with the extracted points from the original KM-curve with a five-year follow-up period. However, modelling the OS and PFS survival curves for the intervention group applying a constant hazard ratio didn't provide a good fit with the points extracted from the original KM-curves (Figure 8, left panel). Furthermore, in the extrapolation phase, independent modelling of OS and PFS curves resulted in these curves crossing, which is of course not possible. Therefore, the authors included corrections to better fit with the observed evidence. First, instead of modelling through the hazard ratios, the extracted points from the published KM-curves were used for both treatment arms. Second, where PFS-curves crossed, it was assumed that the curves further coincided and followed the same trend during the extrapolation phase as the OS curves (Figure 8, right panel). As such, model results reflected the underlying evidence and PFS and OS curves followed logical restrictions in the extrapolation phase.

Figure 8: Visual validation of modelled outcomes – unsatisfactory fit/mistakes and performed corrections



Source: Neyt et al.:[158] Figure 50 in the original report.
Left: the uncorrected approach, which was not used further in the economic evaluation. Right: the corrected approach which was used to calculate results. The lines represent the modelled OS and PFS. The indicated points represent the extracted data at fixed points in time from the published KM-curves.
We note that in this case, the assumption of proportional hazards does not hold as the KM curves converge at 5 years, which would justify dismissing the "uncorrected" approach even without the visual validation. Furthermore, statistical tests can also be performed to see whether the proportional hazards assumption holds. Nevertheless, a visual inspection might help to avoid making modelling assumptions that do not reflect the underlying evidence or are implausible.

Extra information

- Survival analysis for economic evaluations alongside clinical trials - extrapolation with patient-level data: Decision Support Unit, ScHARR, University of Sheffield; 2013.[172]
- Partitioned survival analysis for decision modelling in health care: a critical review: Decision Support Unit, ScHARR, University of Sheffield; 2017.[173]

3.9 Discount rate

The EUnetHTA guideline on methods for health economic evaluations states that the impact of discounting in economic evaluation is often substantial.[1] Table A21 in this EUnetHTA guideline[1] provides an overview of discount rates for costs and effects in 24 countries.[dd] These guidelines should be followed in country-specific economic evaluations. In sensitivity analysis, different discount rates are applied and it is recommended that both the discounted and undiscounted results are shown.

[dd] We remark that it is possible that an update of these national guidelines was performed since the publication of the EUnetHTA report on methods for health economic evaluations.

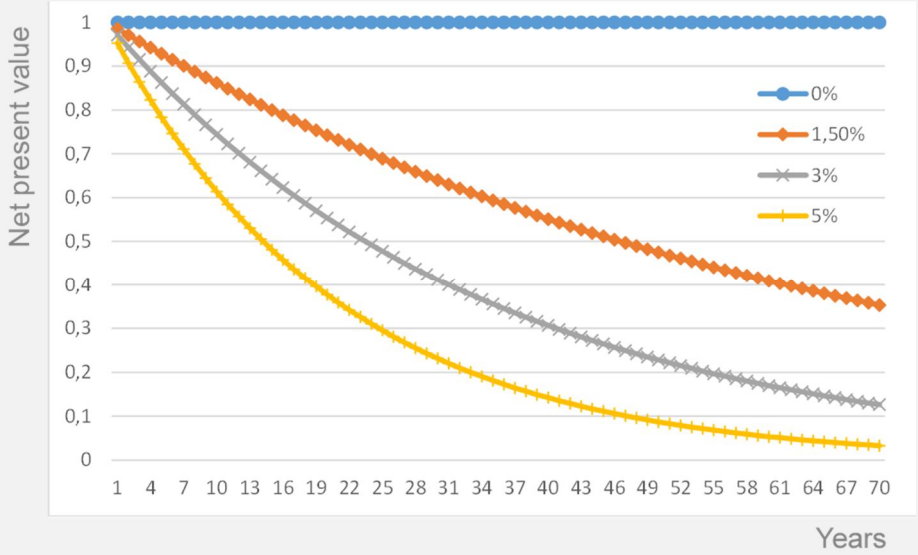
Points for consideration

- The discount rates stated in the national guidelines for economic evaluations should be applied.
- Be aware of the possible large impact of different discount rates, especially in long-term models (see Box 20).
- When looking at the outputs of an economic model, one needs to consider the results in the context of the discount rate used. Where an alternative rate or rates are provided in a sensitivity analysis, it should be determined whether the choice of alternative discount rates substantively affects the decision and, if it does, consider what the decision is likely to be based on the local discount rate.
- Separate reporting of discounted and undiscounted results might facilitate comparisons of these results from studies using different discount rates.

Box 20: The impact of the discount rate on life years

In this box, an example is provided of the impact of applying a discount rate on the net present value of future life years. Assume a person at the age of 12 years with a life expectancy of 70 years. If future life years are not discounted, then this results in a net present value that equals this life expectancy. Applying a discount rate of 1.5%, 3% or 5% reduces this value to 43.15, 29.12 and 19.34 years, respectively. Figure 9 presents the net present value of every year applying four different discount rates. The surface under the curve reduces very fast when applying higher discount rates. As an example, the value of one life-year gained after 20 years equals a value of 1, 0.74, 0.55 or 0.38 when applying a discount rate of 0%, 1.5%, 3% or 5%, respectively. Differences in the discount rate might thus have a considerable influence on the incremental impact on future costs and effects, and thus the ICER.

Figure 9: The impact of applying a discount rate on the net present value of future life years



The net present value (NPV) depends on the applied discount rate. For a life expectancy of 70 years, this NPV equals 70y (0%), 43y (1.5%), 29y (3%) and 19y (5%).

3.10 Perspective

The EUnetHTA guideline on methods for health economic evaluations states that "*the chosen perspective of an economic evaluation is a key element in defining which costs and consequences are included in the analysis.[108] For instance, the choice of perspective affects the way direct and indirect costs are handled (e.g. productivity losses). The most comprehensive perspective is the societal perspective, where all relevant costs and outcomes of the technologies have to be identified, measured and valued, no matter on whom these costs and consequences fall. Other possible perspectives include those of a specific institution, individual patients, or the target group for a specific technology. A health care perspective means that all costs and outcomes related to the health care sector are included in the analysis.*"[1] The perspective is thus specified on both the effect[ee] and cost side, not only on the cost side.

[ee] This concerns, for example, whether or not to add the impact on the caregiver utilities. In a systematic review of literature on spillover effects on caregivers and family members, Alzheimer's disease and other types of dementia were the most frequent focus.[174] However, most Alzheimer's disease/dementia cost-utility analyses incorporated spillover costs, often as caregiver time costs, but considered spillover health impacts less often.[175] If considered relevant, it is important to try to take this into account when setting up research protocols in order to gather reliable information on these spillover health impacts.

National guidelines should be followed when performing an economic evaluation. Based on an overview of these guidelines, the EUnetHTA guideline on methods for health economic evaluations recommends the following:[1]

- "*Economic evaluations should at minimum be conducted from a health care perspective. However, several countries require a societal perspective. Presenting the use of resources as related to other sectors of society may increase the usefulness of the analysis to more EUnetHTA partners. Regardless of perspective taken, it is recommended that the use of resources is presented in as detailed a manner as possible. For example, if a societal perspective is used, indirect costs should be presented separately*."

Points for consideration

- The choice of the perspective depends on the decision problem (e.g. to inform a reimbursement request to the government versus an investment decision for a hospital). The in- or exclusion of costs or effects may depend upon this chosen perspective. Either failure to consider relevant consequences or wrongly taking into account irrelevant consequences may potentially introduce bias of unknown direction.
- The perspective defines the perimeter of consequences of health interventions to consider. It should be transparent which outcomes and costs are studied (e.g. only within or also outside the health care sector), whose outcomes (e.g. only for the patient or also for the caregiver or society) and which costs are studied (e.g. are only costs for the government included or also patient's co-payments and other costs). For example, in the Belgian guidelines for economic evaluations, "*the identification, measurement and valuation of costs should be consistent with the perspective of the Belgian health care payers.*"[89] In this guideline, 'health care payers' refers to both the patients, the federal government and the communities.
- For researchers, e.g. if a research protocol is drawn up for a clinical trial, it is important to think about the potential incremental impact on both costs and effects of using an intervention in comparison with a comparator. As such, researchers might think about how to measure the impact on these elements (e.g. if researchers set up an RCT and indicate there might be a high impact on productivity, then it is important they think about how to measure this impact on this variable and include this in their research protocol). Timely involvement of an expert with knowledge of economic evaluations can ensure that the right information is collected. Having such information at one's disposal might support the conduct and improve the quality of an economic evaluation when the trial is finished.
- The choice of the perspective can strongly influence ICER estimates and the probability that an intervention is considered cost-effective (see Box 21). Presenting results separately for different perspectives is especially relevant in cases where the results, conclusions and recommendations might depend on the applied perspective (e.g. results depend on the in- or exclusion of the impact on productivity or inclusion of costs generated during the gained life years).
- Researchers might "*claim that their studies take a societal perspective, but instead their articles only consider a health care payer perspective.*"[176]

Box 21: Example of a study including three different perspectives

In an economic evaluation alongside a trial,[177] two breast mammaplasty techniques were compared. The trial randomized 255 patients to either vertical scar reduction (VSR) or inverted T-shaped reduction (ITR). The study tried to find out whether VSR was more cost-effective than ITR in patients undergoing breast reduction mammaplasty over a 1-year period based on an RCT in a Canadian center. The authors only considered this short-term time horizon since they did not anticipate any additional health changes or costs beyond one year. The study considered three perspectives:

- the Ministry of Health including direct costs to the health care service,
- the patient including costs incurred by the patient (transportation-related costs, cost of babysitter or housekeeper, and medical supplies not provided or reimbursed), and
- society including all costs comprising productivity costs (time lost from work and activities for the patient and caregiver).

The clinical effectiveness was measured with the Health Utilities Index Mark 3, a generic utility instrument. This patient-reported questionnaire was completed 1 week preoperatively (baseline) and at 1, 6, and 12 months postoperatively.[177] The utilities were used to calculate QALYs by assuming a linear interpolation between the study time points. On the cost side, surgery-related costs including pre- and post-operative costs for follow-up and management of complications were included. Costs incurred by the patient included time lost from paid and nonpaid work (for both themselves and the caregiver) related to their breast reduction surgery and out-of-pocket expenses, as recorded by the patient. The human capital approach was used for the valuation of productivity costs. Costs were reported in 2012 Canadian dollars. Both one-way and probabilistic sensitivity analyses were performed.

A non-significant difference in QALYs was calculated: 0.87 QALYs (95% CI: 0.84 – 0.90) versus 0.86 QALYs (95% CI: 0.83 – 0.89) for ITR and VSR, respectively. From the Ministry of Health perspective, the costs for both procedures were almost the same (ITR: $3090 versus VSR: $3107). However, ITR was more expensive than VSR from the patient (ITR: $9017 versus VSR: $7874) and societal perspective (ITR: $12 107 versus VSR: $11 002). According to the authors, if they applied a threshold of $50 000 per QALY gained, then the probability that VSR was cost-effective was 29.3, 68.2, and 66.9 percent from the Ministry of Health, patient, and societal perspective.[177]

This example shows that results might be different according to the applied perspective and that it is very important to mention which costs are included in which perspective (e.g. are patient's costs included in a health care payer perspective or not).

3.11 (Context-specific) costs

As in the previous section on the perspective, here too the national guidelines on costs must be followed. Based on an overview of these national guidelines, the EUnetHTA guideline for methods of economic evaluations recommends that:[1]

- "*All direct health care costs should be included in the main analysis. It is also recommended to present costs borne by other sectors of the society, e.g. indirect costs, in an additional analysis when relevant.*"[1]
- "*To facilitate adjustments of costs to local settings, it is recommended that the use of resources is clearly presented in natural units, e.g. hospital days or physician visits.*"[1]
- "*To convert costs to the most recent price year by using relevant indices, the index used and the original price year should be clearly indicated.*"[1]

Costs are one of the major components of cost-effectiveness analyses and thus heavily affect the results of evaluations. The incremental costs can include several types of costs, for instance: intervention costs that are directly related to the studied interventions, cost savings that arise as a result of the effectiveness of the interventions, cost increases due to adverse effects related to the studied intervention, cost increases or savings from follow-up treatments, etc. Costs and resource consumption that are common to all the interventions being compared may be excluded from the economic analysis if they are equal in terms of quantity, timing, and duration. Costs that are equal for both the intervention and comparator group do not influence the ICER calculation.

Points for consideration

- In cost calculations and cost reporting, the analytical steps *Identify*, *Quantify* and *Value* are particularly useful. Identify the cost items, quantify the resource use and value the resources by prices. Costs are actually an index of resource use and prices, so a p*q table (prices * quantities) provides important information for critical assessment of results.
- In trial-based economic evaluations, the distinction between intervention costs and other types of costs often becomes blurred as total patient costs are accumulated. A disaggregated reporting of the types of costs included supports critical assessment. In model-based studies, the different cost types including references to their sources should be reported.
- Trials are usually powered to detect differences in clinical outcomes. Researchers should be aware that extracting cost data from studies that are not powered to detect cost differences might be susceptible to outliers that distort the average cost.
- Protocol-driven costs might not reflect the costs that would occur in standard practice, so corrections for such costs might be needed in trial-based economic evaluations (see Box 22).
- The cost items to be included should be determined by the chosen perspective. But administrative rules differ between nations, as health care is organized differently across countries. It is thus possible that costs that fall on health care authorities in one country are paid by some other entity, such as municipality or employers, in another country.
- Health care costs as well as reimbursement rates can be quite different between countries (see Box 23), as well as between regions in the same country, between settings in the same country (e.g. hospital versus primary care), between different types of patients, etc.
- Be aware of possible differences in financing systems between countries when gathering cost information (see Box 24). Differences in financing might also impact the clinical pathway and related costs.

- Patient costs are often skewed, with the majority reporting low costs while a few patients carry very high costs. The assumed parametric statistical distributions in common statistical tests might not reflect this asymmetry, which might lead to incorrect confidence intervals and p-values.
- It is important to verify that all important incremental elements have been sufficiently taken into account. For example, have severe adverse effects, which may not be common but can cause high costs (and impact QoL), been satisfactorily included (see part 3.1.2).
- Costs can comprise fixed (e.g., capital expenditure, salaries, building maintenance) and variable (e.g., medication, diagnostic and therapeutic supplies) components. In the context of health care, and particularly hospital care, a large part of costs accruing are typically fixed in nature. An intervention that achieves efficiency gains, such as through reduced patient length of stay, may have little or no impact on fixed costs over the short term. In reviewing an economic evaluation, one should be cognisant of how fixed and variables costs may be impacted by an intervention.

Examples

Box 22: Protocol-driven costs

Patient cost data for economic evaluations is frequently collected alongside clinical trials, in so-called piggyback economic evaluations. Even though the collection of data is convenient, the differences in objective between clinical trials and cost-effectiveness analyses are known to lead to several drawbacks.[178] Clinical trials are designed primarily to study efficacy, i.e. effectiveness under optimal conditions with no biases stemming from patient population, clinical setting, or clinical management. Economic evaluations, on the contrary, seek to study effectiveness, including all biases that might occur in standard clinical practice.

Protocol-driven costs are one of the drawbacks with trial-based economic evaluations. In the conduct of clinical trials, the protocols are set up to prescribe all the procedures that the trial patients are subjected to. These protocols frequently stipulate more monitoring of patients than in standard care. Consequently, trial patients undergo more frequent check-ups that are performed by more qualified personnel in more specialist settings, than standard care patients do. The trial patient health care costs are therefore often inflated. The standard way of adjusting for protocol-driven costs is to seek to identify and deduct those costs or cost items that are believed to be excessive. It might, however, be difficult to fully identify the exaggerated costs. A possibility might be to introduce a standard practice arm in trials,[178] to perform a pragmatic trial, or to compare the trial data with real-world observational data. This last option should be used with caution, however, due to e.g. the possible non-comparability of populations. Another approach to identify protocol-driven costs might be to ask clinical expert as to which costs would be incurred in clinical practice. However, this approach also has its limitations: data derived from the health care system might provide a better estimate than expert's opinion, the input by experts could be based on too small a sample of opinions, there might be geographical variations in practice between centres in one country so figures could be misleading, etc. On the other hand, the efforts to identify and exclude protocol-driven costs also need to be balanced with the potential impact these costs might have on the results and conclusions of an economic evaluation.

Box 23: Differing health care costs in EU countries: the *Health*Basket

The Drummond textbook[7] includes a comprehensive chapter on costs and cost analyses. One example cited in the book is an EU project on health care costs in nine European countries, named the *Health*BASKET, which was reported in a supplement to an article in the journal Health Economics.[179] This study highlights the different health care costs in EU countries. While much more results are presented in the original report, the following text provides information from the cataract surgery sub-study.[180]

The standardised treatment was described in a vignette:

"*A male patient, 70–75 years old, has consulted a hospital clinic/ophthalmologist's office due to blurred vision. After clinical assessment, a diagnosis of senile cataract is made and the patient is placed on the operating list.*"[180]

A standardised form was used to collect the cost data. Costs for diagnostic procedures, drugs, labour provided by different categories of professionals, medical devices, and overheads were collected and divided into pre-surgery, surgery and post-surgery activities. Data from 41 providers from nine countries were obtained, the majority from hospital outpatient departments.

The reimbursements received for the procedure differed considerably between countries and was markedly higher than the actual costs in most countries. These total costs for the treatment in the nine countries varied threefold; between €318 in Hungary and €1087 in Italy (price level year 2005, see Table 5). The costs of the replacement eye lenses varied, unexpectedly, less across the countries. Regression analyses on the variations in total costs showed that the personnel time used for the procedure as well as the wage levels were important predictors of total costs. Accounting practices also seem to vary between countries, as the share of overheads of the total costs amounted to around 45% in three countries and around 10% in two others (Table 5). This example also highlights the importance of transparent reporting of resource use to allow interpretation and potential adjustment of cost information.

Table 5: Costs for cataract surgery in nine European countries, in Euros 2005 and percentage of total.

	Denmark	England	France	Germany	Hungary	Italy	The Netherlands	Poland	Spain	Mean (S.D.)
Reimbursement rates	1440		1530–1578	597–1322	551–554	968–1436	1041	558–564		
Total costs	602	623	909	741	318	1087	500	473	611	714 (311)
Lens		135 (22%)	139 (15%)	175 (24%)	195 (61%)	160 (15%)	106 (17%)	136 (28%)	217 (35%)	157 (57)
Direct labour	100 (17%)	165 (27%)	406 (45%)	214 (29%)	82 (26%)	165 (15%)	351 (58%)		127 (21%)	221 (151)
Overheads	270 (45%)	293 (47%)	139 (15%)	141 (19%)	20 (6%)	362 (33%)	121 (20%)	241 (45%)	69 (11%)	178 (158)
Other	233 (39%)	29 (5%)	225 (25%)	211 (28%)	21 (7%)	399 (37%)	27 (5%)	148 (28%)	198 (28%)	175 (149)

Source: Fattore & Torbica (2008).[180]

Box 24: Be aware of the financing system in different countries

In the context of performing an HTA report in Belgium, a researcher gathered costs for hospitalizations based on the invoices. However, in Belgium, the cost per hospital day on the invoice does not reflect the 100% per diem price.[89] In this country, the financing of non-medical hospital activities (i.e. capital expenditures for housing and medico-technical facilities, hotel function, nursing care, etc.) is paid by the so-called "budget of financial means". The payment of the budget of financial means to a hospital contains two parts: a fixed part and a variable part. The fixed part (about 80%) is paid by the sickness funds on the basis of monthly advances (the so-called provisional twelfths). The variable part (about 20%), is paid via an invoice, according to the number of admissions and the number of nursing days for the general hospitals, and exclusively according to the number of days for the other hospitals. The invoice is submitted by the hospitals to the sickness funds for all patients enrolled in a sickness fund. The amounts per admission and per nursing day are hospital-specific and also depend on the type of hospital stay (e.g. acute, burned, elderly, psychiatric, palliative and chronic disease care). People not being aware of this financing system can make big mistakes when looking at the invoices to estimate the costs of a specific hospitalization. As such, they would miss the part paid by the provisional twelfths and underestimate the 100% per diem price which might have a big impact on the incremental costs, ICER calculations, conclusions and recommendations.

Note that the codes per admission and nursing days are the only ones recorded in the Belgian administrative database of Minimal Financial Data (Minimale Financiële Gegevens – Résumé Financier Minimum, MFG–RFM). Only counting for these costs means that the fixed part (about 80%) of the budget of financial means is forgotten and thus underestimates hospital stay costs. Fortunately, the amounts per admission and per nursing day are published as excel files on the RIZIV–INAMI website, together with the 100% per diem prices.[ff] These per diem prices are reported per hospital and per type of hospital stay. A non-weighted mathematical average or a weighted (according to the different levels of activities of the hospitals) can be used to derive the average 100% per diem hospitalization price.

Extra information

- Methods for health economic evaluations - A guideline based on current practices in Europe. Methodological Guideline: EUnetHTA; 2015[1]
- Methods for the economic evaluation of health care programmes. 4th ed: Oxford University Press 2015.[7]

3.12 Uncertainty/sensitivity analysis & probability distributions

The EUnetHTA guideline for methods of economic evaluations recommends that:[1]

- "*Based on the results of the current review of guidelines used by EUnetHTA partners and the recommendations in the HTA Core Model, uncertainty should be explored in sensitivity analyses. To be in accordance with the majority of the countries' guidelines, deterministic as well as probabilistic sensitivity analyses should be conducted.*"

[ff] The Excel files are available here: http://www.riziv.fgov.be/nl/themas/kost-terugbetaling/door-ziekenfonds/verzorging-ziekenhuizen/Paginas/verpleegdagprijzen-ziekenhuizen.aspx#.VSbaNWf9m70

- "*For parameter uncertainty, various guidelines recommend probabilistic sensitivity analysis (PSA) (Austria, Belgium, Croatia, England, Finland, France, Germany, Hungary, Ireland, Italy, The Netherlands, Norway, Poland, Scotland, Slovakia, and Spain).*"[1]

Economic models are simplified representations of the true course of a disease and treatment pathways. The intention is to include the key elements that ensure the representation is sufficiently accurate to reflect what happens in reality. Models are used to make predictions about what will happen if those processes and pathways are altered by the application of different health technologies. The outputs of the model are affected by both the parameter values used (e.g., transition probabilities, costs), and the structure of the model and the extent to which it accurately represents the true processes and pathways.

Uncertainty around model input (as the cause of uncertainty of outputs) can be considered within three main concepts:[181]

- Stochastic uncertainty – random variability between otherwise identical patients. It reflects the translation from a population-level probability to an event at an individual level.
- Parameter uncertainty – the parameters that are used in the model (e.g., probability of disease progression) are known with imprecision.
- Structural uncertainty – the impact on outputs of the assumptions used to develop the economic model (see Box 25).

The framework of Briggs et al.[181] also specifies heterogeneity – variability between patients that can be attributed to patient characteristics – as a distinct form of uncertainty. Heterogeneity arises where differences between patients can partly be explained by characteristics (e.g., age, sex, genetic predisposition). This represents real differences between individual patients. Models tend to focus on population-level uncertainty rather than individual-level uncertainty. It may be feasible to carry out separate cost-effectiveness analyses for defined patient subgroups if there are appropriate data to generate relevant parameter values specific to each subgroup. For example, if the magnitude of the treatment effect is associated with age, then it may be possible to run the model for different age groups which can then be presented separately or aggregated into a weighted average across groups. Disaggregated results can enable reimbursement decisions for specific subgroups.

Stochastic uncertainty can be an issue in individual-level models, as it reflects the random nature of an event happening in an individual given the probability of it occurring within a group. In cohort models, this variability is ignored as probabilities are not converted into individual events, but only considered at a group level. Where stochastic uncertainty may be an issue, it is essential that a sufficient number of iterations/simulation runs are used to generate stable estimates of the summary model outputs and minimise Monte Carlo error.[181]

Parameter uncertainty is the most widely considered and addressed source of uncertainty in models – it captures the imprecision in our knowledge of parameters.[8] It can be seen, for example, in the confidence intervals associated with an estimate of treatment effect. Typically parameters are defined by statistical distributions and are allowed to vary within those distributions across many simulations. A model should ideally be fully probabilistic, with all parameters being allowed to vary and summary outcomes calculated as the mean

across many simulations, sometimes presented with associated uncertainty around the estimate. Some models are presented as deterministic, with the main output being calculated with all parameters set at a mean (or best guess) value. That approach can ignore the impact of skewed distributions or correlations between variables if implemented incorrectly. The impact of parameter uncertainty is often explored through one-way, or univariate, deterministic sensitivity analyses where one parameter is set at lower and upper bounds while all other parameters are set at their mean values. The analysis gives an indication of the impact of individual parameter uncertainty on the main outcomes, such as the incremental costs or benefits and the incremental cost-effectiveness ratio. Similarly, multi-variate scenario analyses can be used. A critical element of accounting for parameter uncertainty is that uncertainty is adequately captured in the definition of the statistical distribution. If a parameter is given a spuriously precise distribution, then it will appear to contribute little to uncertainty.

To create a workable economic model, a variety of assumptions are generally made, such as how patients move between disease states.[182] These assumptions guide the structure of the model, and changing any of the assumptions will change the model structure and, potentially, generate different results. The extrapolation of long-term costs and outcomes is also surrounded by uncertainty. Exploring structural uncertainty can highlight where some assumptions may have a substantive impact on model outputs, and this is usually carried out using scenario analyses.

Uncertainty in model outcomes can be explored and reported in a number of ways (see Box 26).[183] The aforementioned univariate sensitivity analysis is typically illustrated with a tornado plot (see Figure 15). A scatter plot of the cost-effectiveness plane is a basic representation that shows the spread of uncertainty around cost-effectiveness (see Figure 12). A cost-effectiveness acceptability curve (CEAC) summarises the impact of uncertainty by showing the probability that an intervention is cost-effective for a range of willingness-to-pay thresholds (see Figure 13 and Figure 14). The expected value of perfect information (EVPI) and the associated value of information measures are used to provide the decision-maker with an indication of the expected costs of uncertainty and the value of collecting additional information to eliminate or reduce uncertainty. Some or all of these approaches should be reported in an economic evaluation to clearly state the sources and impacts of uncertainty on the model outcomes.

Some models are subject to calibration (see part 3.13) to ensure that certain outputs, such as disease incidence, match observed values. The calibration process may involve adjusting individual parameter values or else generating parameter sets that result in plausible outputs. Calibration of a poorly-structured model may result in implausible parameter values. Similarly, attempts to fit outputs too closely to observed data (and thereby ignoring the impact of structural uncertainty) may generate parameter values that are spuriously precise or implausible. The calibration process should be described in sufficient detail to be able to consider its impact on uncertainty in model outputs.

An economic model is intended to support decision making by providing evidence regarding the efficiency or value of an intervention relative to one or more alternatives. From the point of view of the decision-maker, a critical question is whether the uncertainty affects the decision. In other words, is the policy outcome dependent on particular assumptions or choices of parameter values in the model? An assessment of uncertainty should therefore include sufficient exploration to determine whether the decision is sensitive to particular assumptions or parameter values, and to provide reassurance about the validity or plausibility of those choices. An outcome of an evaluation can be to state that further

evidence needs to be gathered regarding specific assumptions or parameters to reduce uncertainty.

Points for consideration

Some of the potential issues related to uncertainty in economic evaluations are:

- Are confidence intervals presented for key parameters and is the imprecision clearly linked to an evidence base/the best available evidence? It should be possible to identify not only the mean values for parameters, but also the upper and lower bounds for statistical distributions used in the modelling process. Those bounds should be plausible and adequately reflect uncertainty in the parameter value. Parameters defined with very narrow confidence intervals may be a cause for concern.[gg]
- The statistical distribution used for a parameter should be appropriate so that the parameter cannot take on implausible values (e.g., a probability of greater than 1 or a negative cost). This is done by modelling specific variables with the appropriate probability distributions. For example, a beta distribution for probabilities or utilities, gamma distribution for costs, a lognormal distribution for relative risks or hazard ratios, etc.
- The information from outliers should not be ignored. For example, the distribution of cost data is often highly skewed. Where possible, has the economic evaluation included the appropriate distribution with the correct mean value (and not e.g. the median cost estimate)?
- Does the evaluation include a sensitivity analysis that is sufficiently comprehensive to identify parameters or assumptions that contribute to uncertainty in the model outputs? Sometimes only a restricted set of parameters are included in the sensitivity analysis.
- Have correlations between parameters been taken into account in the sensitivity analysis? While it is unusual for models to incorporate explicit correlations between parameter values, when they are included it is essential that those correlations are also appropriately reflected in any univariate sensitivity analysis or scenario analyses.
- Have the model outputs (e.g., incremental costs and benefits, incremental cost-effectiveness ratio, net monetary benefit) been presented in a manner that reflects uncertainty? Outputs may be presented with associated 95% confidence intervals, or the ICER may be presented in terms of the probability of being below some defined willingness-to-pay threshold.
- Have the key deficiencies in available data and assumptions been highlighted and discussed? Setting out with the knowledge that a model is an imperfect representation of reality, researchers should be forthcoming in highlighting any issues contributing to uncertainty.

[gg] For example, it may be questioned why a uniform distribution is used in which the average is randomly changed by +/- 5% if there is evidence that the spread around the average cost is much wider.

- Are suggestions made as to how decision uncertainty may be reduced, for example through further data collection to improve the precision of parameter estimates or plausibility of model assumptions?

Examples

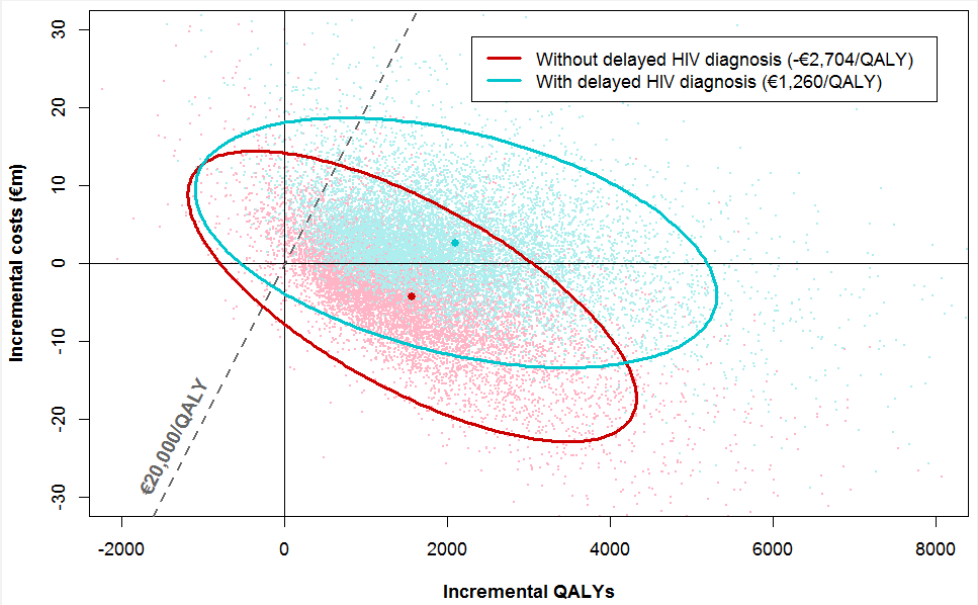
Box 25: Example of exploring structural uncertainty in an economic evaluation

An economic evaluation was carried out to estimate the cost-effectiveness of a pre-exposure prophylaxis (PrEP) for populations at substantial risk of sexual acquisition of HIV (human immunodeficiency virus).[184] The intention is that people at high risk of acquiring HIV infection take PrEP either on an ongoing basis or on an event-based basis to reduce the risk of HIV infection. In the economic model, members of the cohort could be in one of five mutually exclusive states: high risk and taking PrEP; high risk and not taking PrEP; medium and low risk (not on PrEP); having HIV; and dead. An important aspect of the model was the transition probabilities which were subject to substantial uncertainty.

In the model, when those who transitioned to the HIV state immediately began treatment. Costs and benefits were both discounted at 5% per annum. In the base case, those in the HIV state incurred an annual cost of €10 200 for treatment. In reality, although 39% of those who acquire HIV are diagnosed in the first year after infection, 7% are diagnosed at least 9 years after initial infection. By assuming immediate diagnosis, the model began counting HIV treatment costs at the point of infection. This assumption biased the model in favour of the intervention as later diagnosis would be subject to delayed costs and increased discounting. A second version of the model was set up to incorporate data on delayed diagnosis to determine the effect of this structural assumption on the estimated cost-effectiveness.[184]

In this example, the intervention was on average dominant when delayed HIV diagnosis was not taken into account (ICER = -€2704/QALY) (see Figure 10). When the model was reconfigured to take delayed diagnosis into account, the intervention was no longer dominant but was still highly cost-effective (ICER = €1260/QALY). The structural assumption did not affect the decision in this case. In a different context, if a structural assumption changes the policy assumption then it would become important to evaluate whether the assumption was sound and if it was justified. For this example, the estimates for delayed diagnosis came from ten-year-old international literature and it was felt, based on expert clinical opinion, that testing and diagnosis was now much faster. Hence the structural assumption had some justification, although it was considered pragmatic to test the potential impact of that assumption.[184]

Figure 10: Cost-effectiveness plane under different structural assumptions



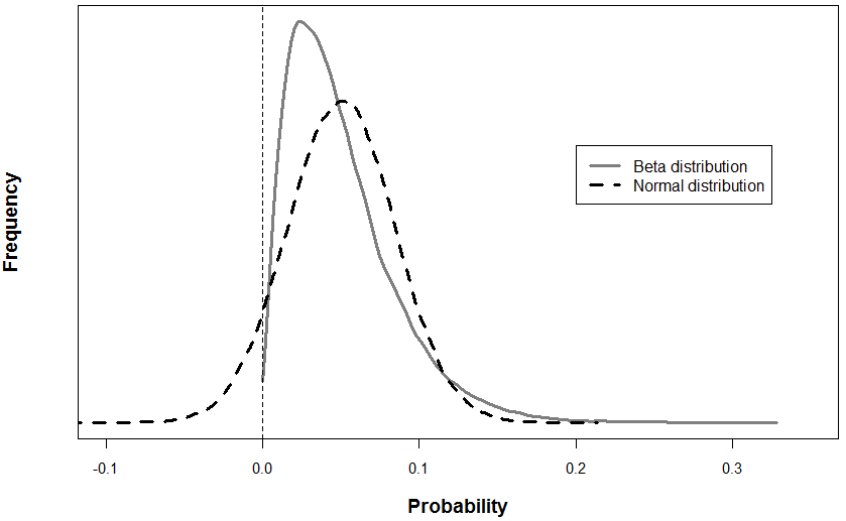
Source: HIQA (2019).[184]

Box 26: Example of conveying uncertainty in an economic evaluation

An economic evaluation was carried out to examine the cost-effectiveness of mechanical thrombectomy for the management of acute ischaemic stroke.[185] Clinical efficacy was estimated using a meta-analysis of nine randomised controlled trials (RCTs). The primary outcome was reported as the modified Rankin Scale (mRS), measuring increasing disability on seven levels running from 0 (no symptoms at all) to 6 (dead). A hybrid decision tree and Markov model was used to simulate a cohort of patients eligible for the intervention. The model included three health states: functionally independent (mRS 0 – 2), functionally dependent (mRS 3 – 5), or dead (mRS 6). The model was fully probabilistic and included 44 parameters defined by statistical distributions. There is only one comparator: the current standard of care. As a cohort model was used, stochastic uncertainty was not considered.

In defining parameters as statistical distributions, care was taken to select appropriate distributions. Probabilities were defined as beta distributions which are upper and lower bound to be between 0 and 1. Costs were defined as log normal distributions which must be positive and are right skewed. By way of example, one probability relating to patient transfer had a mean of 0.05 (95% confidence interval: 0.006 to 0.135) and was defined using a beta distribution. Had the parameter been defined by a normal distribution, the 95% confidence interval (-0.016 to 0.117) would have encompassed negative values which are not possible (Figure 11). Had a normal distribution been used and artificially truncated to be between 0 and 1, then the mean would have risen to 0.055 and been above the intended value.

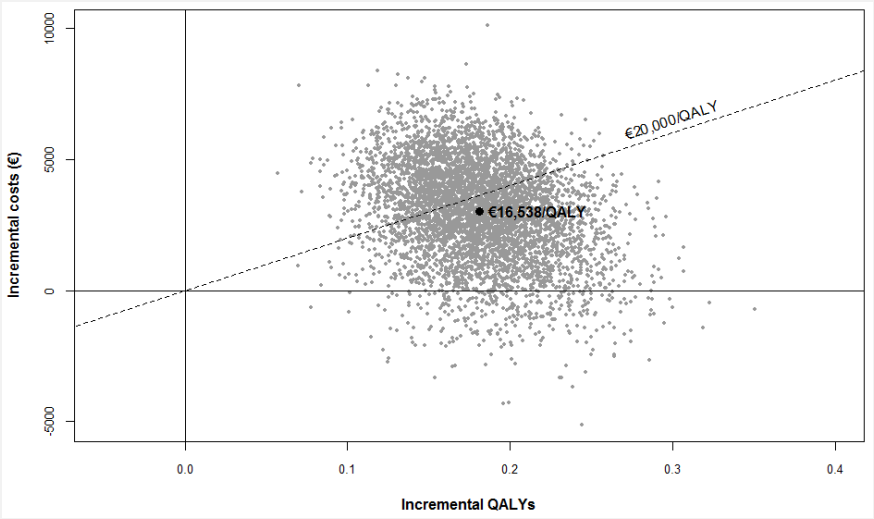
Figure 11: Impact of incorrect choice of statistical distribution



Source: HIQA (2017).[185]

The first display of uncertainty is generally captured by a plot of the cost-effectiveness plane showing incremental costs and benefits across a number of simulations (Figure 12). In this case, the summary incremental cost-effectiveness ratio (ICER) is shown, as well as a line representing a specific willingness-to-pay (WTP) threshold. It can be seen that the summary ICER is below the WTP threshold, but that many simulations (40.9%) generate ICERs above €20 000/QALY. It can also be seen that in some simulations (6.5%) the intervention is cost-saving. From a decision-making perspective, 93.5% of simulations show that the intervention will be more costly and more effective than the current standard of care. If the points for individual simulation span numerous quadrants of the cost-effectiveness plane, then it might raise concerns about what might happen if the intervention would be introduced.

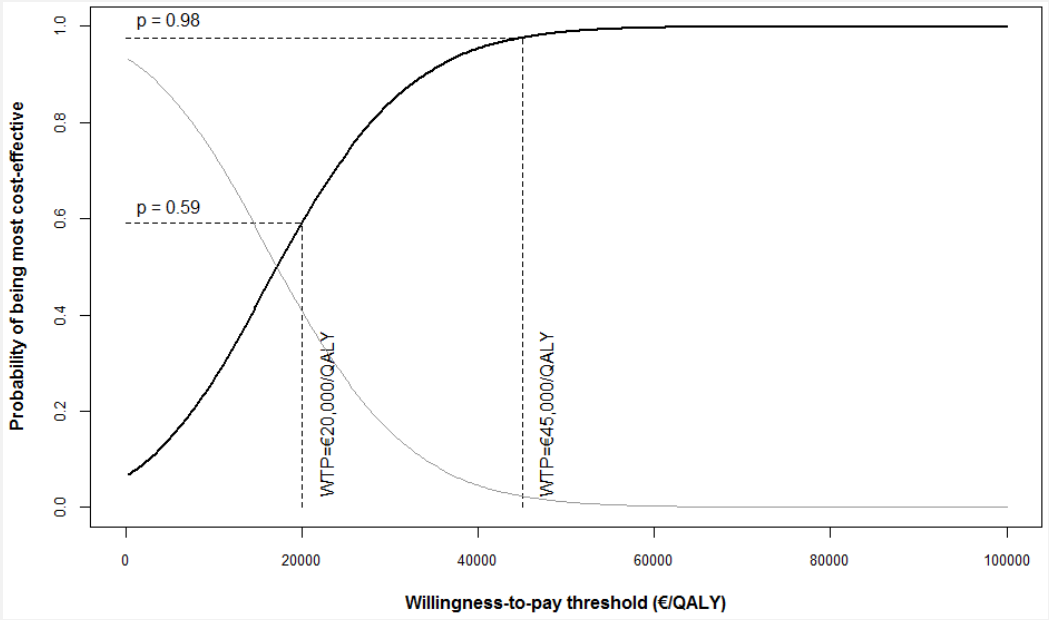
Figure 12: Cost-effectiveness plane



Source: HIQA (2017).[185]

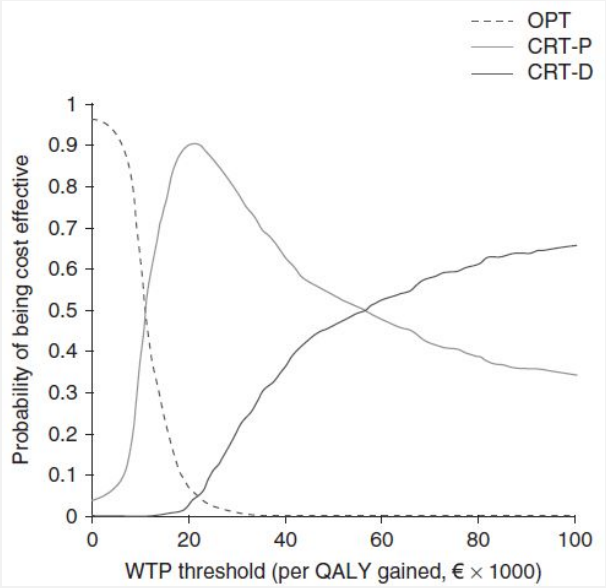
Given the uncertainty around the ICER, a next step is often to determine the probability that the interventions under assessment are cost-effective at different willingness-to-pay thresholds. In this example, that question is simplified by the fact that there is only one comparator. At a willingness-to-pay threshold of €20 000/QALY, there is a probability of 0.59 that mechanical thrombectomy is the most cost-effective option (Figure 13). When there are multiple comparators, the cost-effectiveness acceptability curve (CEAC) contains many overlapping lines and the likelihood of being cost-effective might vary across the range of WTP-values. While it is very difficult to estimate on the cost-effectiveness plane the probability of an intervention being cost-effective for a range of WTP-values, this is much easier on the CEAC. For example, Figure 14 presents the CEAC derived from the results presented on the cost-effectiveness plane in Figure 2 (see Box 7). When comparing to multiple interventions on the CEAC, care must be taken when interpreting results that comparisons are not being made to cost-ineffective comparators (see part 3.2).

Figure 13: Cost-effectiveness Acceptability Curve comparing two interventions



Source: HIQA (2017).[185]

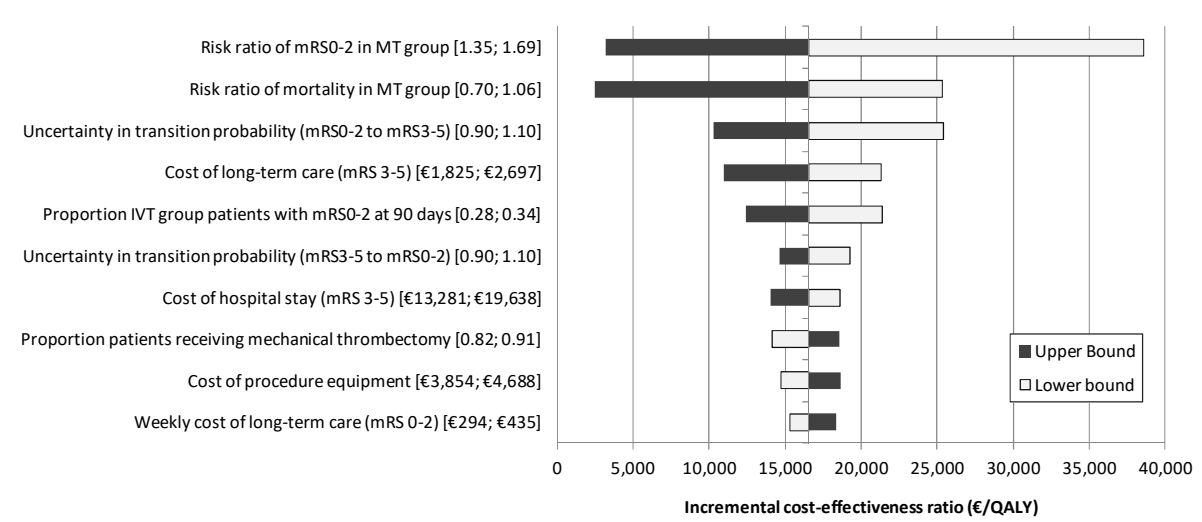
Figure 14: Cost-effectiveness Acceptability Curve comparing three interventions



Source: Neyt et al.[91]
CRT(-P/D): cardiac resynchronization therapy (biventricular pacemakers/biventricular defibrillators); OPT: optimal pharmacological therapy.
"The CEACs show that OPT is the preferred option if the willingness to pay (WTP) for a QALY gained is less than €11 000. Above this threshold, CRT-P is most probably the best alternative, with a probability of about 90% at a threshold of about €21 000 per QALY gained. If the WTP is more than €30 000, the probability that OPT is chosen is almost nil. This WTP has to increase to more than €56 000 per QALY gained for CRT-D to have a probability of >50% of being considered a cost-effective alternative."[91]

While plots of overall model outputs give indications of decision uncertainty, a univariate sensitivity analysis can shed light on which parameters may be contributing most to that uncertainty and which may really have an influence on the decision. Besides presenting results in table format, a common approach to displaying the results of a univariate sensitivity analysis is the tornado plot (see Figure 15). In the most common approach, the plot shows how much a summary outcome such as the ICER will change if a single parameter is set at its upper and lower bounds (sometimes based on the 95% confidence intervals) while all other parameters are set at their mean values. Ideally, the plot enables a distinction between a parameter being set at its lower and upper bound, so that the direction of effect can be seen. In Figure 15, for example, the first seven parameters result in a lower ICER when set at the upper bounds, while the remaining three parameters increase the ICER when at their upper bounds.

Figure 15: Tornado plot of univariate sensitivity analysis



Source: HIQA (2017).[185]

A point to note is that if all parameters were included in the plot, it could become cumbersome or difficult to interpret. In the above example, parameters were only included if variation resulted in at least a 10% change in the ICER. At a WTP threshold of €20 000/QALY, there are five parameters that individually would result in an ICER exceeding the WTP threshold if set at their lower bounds. From a decision making point of view, these parameters should be subject to additional scrutiny to determine either whether uncertainty can be reduced or if the mean estimates used might be inaccurate. According to the tornado graph, the main risk ratio parameters are derived from a systematic review and meta-analysis of the available RCTs, and a cumulative meta-analysis shows that more recent trials have had a limited impact on the estimated magnitude of treatment effect. The next most influential parameter is the uncertainty around the transition probability for moving from functional independence to dependence in the years after the initial stroke incident.

Consideration of structural uncertainty requires identification of potentially influential assumptions made during model development. Often those assumptions are used to underpin simplifications in the structure, making the problem tractable in terms of data requirements or computational burden. In the current example, a simplification was to have three health states when seven could have been modelled based on the mRS (modified Rankin Scale) data available. The simplification was driven by a lack of available data on differences in costs and transition probabilities by individual mRS levels. Indeed, the only data available by individual mRS level was on utilities. To test the impact of this assumption, the model was run with utilities by individual mRS level. The ICER changed from €16 538/QALY to €11 593/QALY. From the perspective of the decision maker, the simplification to three health states leads to a conservative estimate of cost-effectiveness but does not change the decision. In the event that model simplification had changed the decision then it may have been necessary to seek additional data to support a more complex model.

Heterogeneity may arise when specific subgroups of the population may be sufficiently different that distinct parameter values may be warranted. In this case, the therapeutic time window is different for the intervention and standard care: the

treatment window for intravenous thrombolysis is 4.5 hours after stroke onset, whereas mechanical thrombectomy has been successfully used up to 12 hours after stroke onset. The model could, in theory, have simulated those two populations separately.

Extra information

- Methods for health economic evaluations - A guideline based on current practices in Europe. Methodological Guideline: EUnetHTA; 2015[1]

3.13 Model verification and validation (& model sharing)

To assess how good a model is, we must ascertain whether the model implements the assumptions correctly (model verification) and whether the assumptions which have been made are reasonable and reflect reality (model validation).

- Verification is concerned with the technical accuracy of the model and should identify "*programming errors, data entry errors, and logical inconsistencies in the model specification.*"[186]
- Validation is concerned with the structure, content and predictive accuracy of the model.

The HTA Core model specifies "*to fully evaluate how the results of a model should be used, model users would need to be able to know how well the model predicts the outcome(s) of interest. To be able to do this, the model needs to be reported in a transparent way and validated.*" "*Validation relates to the methods of evaluating how accurate a model is in making relevant predictions or abstracting from a complex reality.*" "*… validation is recommended in cases where it is possible, e.g., using a relevant data set.*"[187]

The following verification and validation exercises should be explored:[hh]

- **Face validity**: Does a model structure, its assumptions, input parameter values and distribution and output values and conclusions, make sense and can be explained at an intuitive level?
- **Internal validity (technical verification)**: Has the model been implemented correctly?
- **Cross model validation**: Does the model achieve similar results with other models that were independently developed, but aimed at estimating the same outcomes?

[hh] Other typologies have been proposed in the literature.

- Eddy et al; (1985)[188]: [1] First-order validation requires expert concurrence; [2] Second-order validation compares the model predictions with data used to estimate the model parameters; [3] Third-order validation compares the model prediction with "other" observed data, i.e. data not used in the model construction; [4] Fourth-order validation compares pre-implementation model predictions with observed events post-implementation.
- Vemer and al. (2016)[189]: [1] conceptual validation, [2] data validation, [3] Computerized model validation, [4] Operational validation.

- **External validation**: How can we compare the outputs of the model with actual outputs provided by external sources (not used in the model)? If a source for future events is available, Eddy et al.[190] define “predictive validation”.

This is in line with recommendations from e.g. KCE,[89] NICE,[86] PBAC,[167] HAS,[191] IQWIG,[192] AOTMiT,[20] or CADTH.[168]

Points for consideration

In general:

- Model validation and verification can be performed by the modeller or by an external assessor. In case of the latter, having access to the model facilitates the validation and verification. Several possibilities and levels of model sharing are possible. We refer to part 5.3 in the annexes for further information on model sharing.
- Verification and validation don’t stop at performing a test. Any significant inconsistencies or discordance should be considered and, where possible, the source of the difference is identified or an explanation is provided.
- ”*Model traces for the proposed medicine and its comparator provide a clear depiction of the implications of the model. They can inform the face validity of the model logic, computerisation and external validity.*”[ii][167]
- The outcomes of the model should reflect the underlying evidence. For example, if evidence shows no impact on overall survival over a specific time period, the outcomes of the model should reflect this.
- Where appropriate and applicable, an adjustment is made (calibration phase). It should be clear how a model was calibrated (e.g. what goodness of fit measure was used) and which parameters were adjusted. It should also be checked whether the calibration exercise does not result in implausible values (e.g. are parameters kept within plausible ranges).
- If adjustment is not required or is inappropriate, key factors that could compromise the validity of the model are identified and the potential direction and/or potential magnitude of bias induced are defined, to help decision makers to determine the results’ applicability to their decision. Sensitivity analysis should be performed to support this task.
- “*No matter how many validations are done, there will inevitably be uncertainty about some aspects of a model (…) Sensitivity analysis is an important complement to validation.*”[190]

[ii] “*Use traces to track patients through the model and demonstrate that the logic of the model is correct. Present traces representing the proportions of the cohorts in each health state over time, and the cumulative sum of the undiscounted costs and outcomes (e.g. QALYs) over time. If applicable, state the number of events over time where patient-relevant events occur within a health state. Comment on whether each of the model traces is logical – for example, ensure that any traces of overall survival converge to zero at or before the time horizon of the model.*“[167]

Points for consideration for face validity

- The assessment of face validity is mostly qualitative in nature.
- Look at information about experts' contribution to model development, to determine the degree of expertise about the clinical or care pathway of interest and the potential impact of the interventions.
- Look at the schedule: the face validity should be challenged early and iteratively throughout the analysis.
- Check face validity of model results by comparing with the identified evidence in the clinical part of the HTA or with clinical or patient experts who know about the disease and treatment under consideration.
- As models simplify reality there may be inconsistencies with medical knowledge (e.g., simplifying the subsequent treatment pathway), which do not necessarily invalidate the model.[190] As an example, see McCabe:[193] "*mapping out comprehensive treatment pathways for each individual adverse effect from treatment may not alter the results of the model over and above a simpler representation, and may only confuse the decision maker through its added complexity.*" Model users have to "*determine whether the model has been properly simplified, oversimplified, or undersimplified for a particular problem.*"[190] Making a relatively simple model that reflects the underlying evidence is preferred above making a very complex model for which no data are available. It is important that the model structure is able to reflect the incremental differences in costs and effects of the intervention under consideration in comparison with the relevant comparator(s).

Points for consideration for internal validity

- As a model increases in complexity, the possible sources of error become more important. In the ideal situation, a researcher who is not directly involved in the model development could perform a formal model quality control with well-known methods (i.e. double-programming, tests of the mathematical logic of the model, checking for errors between parameters and the sources, testing repetitions to check mathematical calculations, etc.). If such dual quality control is not performed, the modeller himself should perform sufficient tests to check the validity of the model (see next point).
- Both the modeller and the user of an economic evaluation could do the following:
    - Verify that the model is able to reproduce its input (outcomes before extrapolation are consistent with data sources used in the model).
    - Look for counterintuitive results of the internal validation which might reflect either errors or new insights which must be explored and explained.
    - If the electronic version has been shared (see Annex 3 – Model sharing), some simple analyses can lead to the identification of design deficiencies (i.e. allocate extreme or zero values for different parameters or change the input value and examine whether the change in the output values was expected).
    - Check outcomes of the univariate deterministic sensitivity analyses to identify possible model errors.

Points for consideration for cross-model validity

- "*Validation of models against the results or behaviour of other models is a technique which should be used with care as both may be invalid in the sense that they both may not represent the behaviour of the real system accurately.*"[194]
- To identify inconsistencies or differences is not sufficient and may lead to wrong conclusions if no explanation is sought.
    - The underlying cause of the difference between models may be related to different structures, assumptions and parameters.
- A high degree of dependency among models will reduce the value of cross-validation.

Points for consideration for external validity

- To compare model traces with corresponding empirical data is an essential step of the validation process.
- When comparisons are done between model predictions and actual data, it is important to check that the external reference:
    - is not coming from the same source of data used to populate the model;
    - is sufficiently comparable;
    - is valid, otherwise the model "*would be validated against the sort of flawed estimates that it was designed to replace.*"[193]
- Relevant real-world data are often not available and external validation involves a tension between using data to improve the accuracy of parameter estimates and retaining it for use in validating the model.[193]
- Systematic research of real-world data on the intermediate and final endpoints of the model is needed.
- The external validation concerns both the intervention as well as the comparator arm(s). Both intermediate and final endpoints are involved in this validation exercise.

Box 27 provides some information about the AdViSHE (Assessment of the Validation Status of Health-Economic decision models) tool that helps researchers in their validation efforts. Box 28 provides an example where the fitted survival curves are compared to the original published KM survival curve.

Examples

Box 27: The AdViSHE tool

"*Assessment of the Validation Status of Health-Economic decision models (AdViSHE) is a validation-assessment tool in which model developers report in a systematic way both on validation efforts performed and on their outcomes. Subsequently, model users can establish whether confidence in the model is justified or whether additional validation efforts should be undertaken. In this way, AdViSHE enhances transparency of the validation status of HE models and supports efficient model validation.*"[189]

The AdViSHE tool consists of 13 questions, divided into five parts, each covering an aspect of validation:[189]

Part A: Validation of the conceptual model (2 questions based on face validity and cross validity of conceptual model)

Part B: Input data validation (2 questions based on face validity of input data and model fit testing when input parameters are based on regression models)

Part C: Validation of the computerized model (4 questions based on: the external review of the computerized model by modelling experts; extreme value testing; testing of traces; individual sub-modules testing).

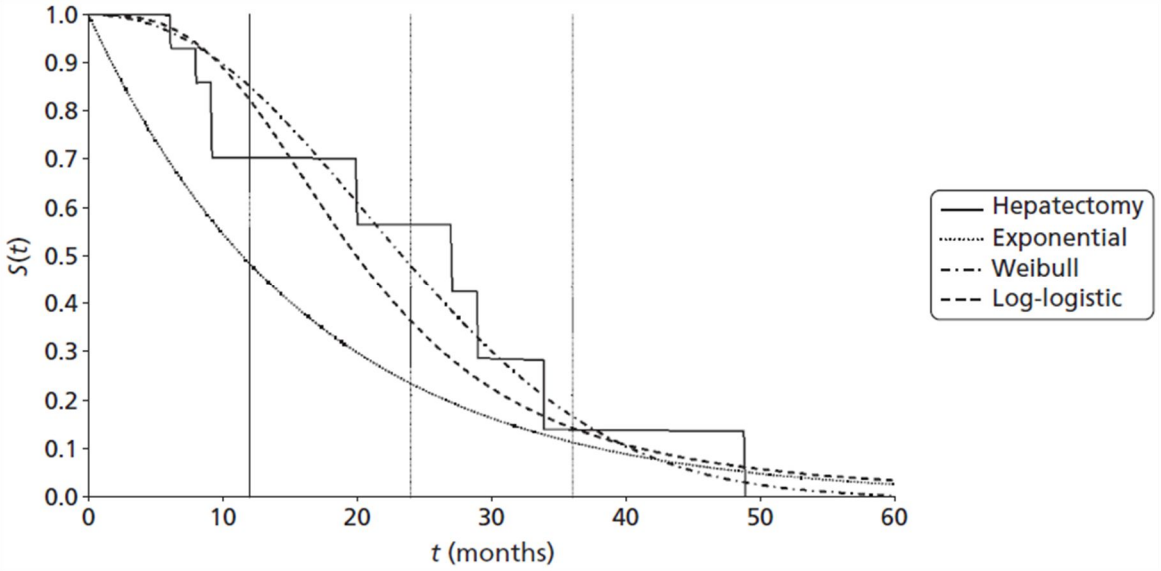
Part D: Operational validation (4 questions based on: face validity and cross validity testing of model outcome; validation against outcomes using alternative input data; validation against empirical data).

Part E: Other validation techniques (1 question to describe any other validation techniques performed).

Box 28: Validation of published versus modelled survival curves

In an HTA report, Loveman et al.[195] studied the clinical effectiveness and cost-effectiveness of ablative therapies in the management of liver metastases. Overall survival curves for patients undergoing surgical resection and microwave ablation (MWA) were extracted from a study by Shibata and colleagues.[196] The KM survival curve for the surgical resection arm was published together with different fitted survival functions (see Figure 16). The authors indicate the Weibull and log-logistic survival functions appear to give the closest fit to the overall survival curves. However, the authors also transparently report both the observed survival at 1, 2 and 3 years and the model predictions, as well as the mean overall survival from the published KM curve and modelled survival functions (see Table 6). The authors correctly note that "*both the Weibull and log-logistic functions overestimate early survival (up to 1 year) and underestimate later survival (from 2 years). The exponential and log-logistic functions estimate lower survival with surgical resection at each reported time point, while the observed data has a higher survival with surgical resection at 3 years.*"[195] Also the modelled mean overall survival shows flaws. While the estimates are similar for the surgical resection arm when comparing the KM estimate and the estimate from the Weibull survival function, the Weibull survival function reports lower than expected values for the MWA arm. The exponential model substantially underestimates OS for both treatment arms. In the Log-logistic model, OS is also underestimated in both arms but the incremental difference is closest to the trial results.[195] The authors indicate that "*none of the survival functions provides good predictions of survival for the reported time periods.*"[195] In this case, an alternative approach could have been to model the survival curve by reflecting the survival exactly as reported at specific points in time (e.g. at year 1, 2 and 3) for both the MWA and surgical resection arm.

Figure 16: Kaplan-Meier survival estimates and fitted survival functions



Source: Loveman et al.(2014):[195] Figure 7 in the original report presents the Kaplan–Meier survival estimates from the Shibata and colleagues (2000)[196] trial showing Weibull and alternative model fit.

Table 6: Survival at 1, 2 and 3 years and mean overall survival (observed vs. modelled)

Year	Trial report		Weibull model		Exponential model		Log-logistic model	
	MWA	Surgical resection	MWA	Surgical resection	MWA	Surgical resection	MWA	Surgical resection
Year 1	71%	69%	82%	83%	48%	40%	80%	78%
Year 2	57%	56%	43%	47%	23%	16%	34%	31%
Year 3	14%	23%	14%	18%	11%	8%	13%	12%
Mean OS (months)	24.8	24.7	22.9	24.1	15.4	12.7	21.7	21.1

Source: Loveman et al.(2014):[195] the numbers are copied from table 42 and 43 of the original report. Table 42 of the original report presents a comparison of observed survival at 1, 2 and 3 years against model predictions. Table 43 of the original report presents the mean overall survival from Kaplan–Meier and modelled survival functions.
MWA: microwave ablation; OS: overall survival.

Extra information

- McCabe C, Dixon S. Testing the validity of cost-effectiveness models. Pharmacoeconomics. 2000 May;17(5):501-13.[193]
- Vemer P, Corro Ramos I, van Voorn GA, Al MJ, Feenstra TL. AdViSHE: A Validation-Assessment Tool of Health-Economic Models for Decision Makers and Model Users. Pharmacoeconomics. 2016 Apr;34(4):349-61.[189]
- Eddy DM, Hollingworth W, Caro JJ, Tsevat J, McDonald KM, Wong JB, et al. Model transparency and validation: a report of the ISPOR-SMDM Modeling Good Research Practices Task Force--7. Value Health. 2012 Sep-Oct;15(6):843-50[190]

3.14 Transferability of economic evaluation results

Transferability is the possibility to apply the results from an economic evaluation[jj] carried out in a specific decision-making context into another setting. More formally, it has been defined as "*the ability to extrapolate results obtained from one setting or context to another.*"[197] Similar concepts are generalisability[kk] of results and variability in methods and data,[199,

[jj] To avoid misunderstandings, we notice that we are referring here to the results of the economic evaluation (e.g. the calculated ICERs) and not to the results of the decision-making process of the policy makers, which can be influenced by many other factors.

[kk] The same terms might be used with a similar but slightly different meaning. For example, ISPOR's Good Research Practices Task Force wrote a guideline on the transferability of economic evaluations across jurisdictions and applied the following working definitions: "*economic evaluations were generalizable if they applied, without adjustment, to other settings. On the other hand, data were transferable if they could be adapted to apply to other settings.*"[198]

200] while the terms applicability and adaptation more frequently refer to the full HTA results including effectiveness, cost-effectiveness and ethical and social aspects.[9, 201]

Transferability was originally seen as a simple yes-no question (Are the results of this particular economic evaluation correct for my setting?)[202] but has evolved into checklists with several questions that aim to help investigators to identify the parameters that are more prone to differ between the original and new setting.[9, 203, 204][ll] Some checklists include a scoring system, where individual studies are assessed on various aspects and an overall index score is obtained.[203, 206] More recent texts, however, stress that transferability is better seen as a process.[198] The important differing parameters are first identified. Then the implications in terms of changes in ICER in the new setting are either discussed thoroughly or the parameters are replaced in a new, frequently model-based, economic evaluation.

Within a previous EUnetHTA project, an HTA adaptation toolkit has been set up as an aid to HTA agencies in the adaptation of HTA reports from one setting into another. This toolkit also contains a list of relevance, reliability and transferability questions to ask when considering the adaptation of information and/or data on economic evaluations. We refer to this toolkit for an overview of these questions (see box 10 in Section 5.4 of the HTA adaptation toolkit, available on https://www.eunethta.eu/eunethta-hta-adaptation-toolkit/).

Points for consideration

- It is seldom possible to directly compare the results from economic evaluations from different settings. Quantitative syntheses of the outcomes of identified economic evaluations, such as calculating the average cost per QALY or more formal meta-analytic techniques, are therefore not recommended.[207] When reporting the results of systematic reviews of economic evaluations, it is more useful to perform a critical assessment of the studies and to focus on the one or two studies that are considered most relevant for the actual decision-maker. Transferring existing evaluations or setting up a de novo economic evaluation can be the next step.
- The ICER result of cost per health effect is unlikely to be identical in two different settings. The important issue is whether the recommended policy decision will be different. This depends on the magnitude, estimated or merely discussed, of the changes in the ICER for the new setting but also on the cost-effectiveness threshold in the setting.
- Health care costs, more precisely the unit costs, are often seen as the most important source of non-transferability. However, other factors in health care, such as practice patterns and diagnostic techniques might be just as important.[208] The effectiveness of the technology investigated is affected by differences in population epidemiology such as baseline risks (see part 3.4), which thus also affects the cost-effectiveness. An example of a systematic literature search to identify cost-effectiveness studies performed on the same drug but in different countries with possible differences in costs, resource use and/or effectiveness is provided in Box 29.

[ll] A review of transferability checklists is found in Goeree et al, 2011.[205]

- The HRQoL weights used for QALY estimates might also differ between settings. Actually, even the weights from the commonly used instrument EQ-5D has been shown to differ between populations.[209][mm]
- It is impossible to assess transferability and it is difficult to transfer results to another setting if economic evaluation data and methods lack transparency.

Examples

Box 29: Example of differences in costs, resource use and/or effectiveness in economic evaluations of specific drugs

A study from Barbieri et al. sought to compare the ICERs of economic evaluations of specific drugs in Western Europe.[210] The authors performed a systematic literature search to identify cost-effectiveness studies performed on the same drug but in different countries. The studies were divided into three groups, depending on which data differed between the pairs of cost-effectiveness analyses. For 19 studies, the only difference was in unit costs (called type C in Table 7), for 8 studies the resource use and unit costs differed (type RC), and for 17 studies the effectiveness also differed between the paired studies (type REC).

It turned out that the results were similar among 36% of the studies where only unit costs differed, among 25% of those with differing resource use and costs, and among 18% of the studies where all three factors differed. Thus, in their sample, for around 80% of the studies where all major data sources differed (type REC), the cost-effectiveness results varied between the pairs. The main reason for the variability was total costs (including resource use and unit costs). In half of the studies where total costs differed (type RC), the main reason for variability was the resource use, not the unit costs. And in the group where only the unit costs varied (type C), the drug cost was the main reason for variability in about a third of pairs.

The authors also identify some systematic differences between the countries. Drugs were less often reported cost-effective in the two countries with higher prices (Germany and the UK) and more often cost-effective in France. The overall impression is, however, that there are very few systematic patterns between the countries, which makes it difficult for a decision-maker to predict in what way the result in country A would differ from that in country B.

[mm] Adjusting the QoL weights to a standard value set from another country is only possible if the health states are available at the patient level. This is often not possible as published clinical trial results on HRQoL are often only available as the average QoL (and CI) for the distinct treatment arms at different points in time.

Table 7: Main reasons for variations in cost-effectiveness estimates among countries in the study of Barbieri et al.

	Type C (n = 19)	Type RC (n = 8)	Type REC (n = 17)
Clinical data	Not applicable	Not applicable	3 (18%)
Total costs (including resource use and unit costs)	Not applicable	See below	8 (46%)
Both clinical data and total costs	Not applicable	See below	3 (18%)
Resource use	Not applicable	4 (50%)	Not clear
Drug costs and other unit costs	See below	1 (12.5%)	Not clear
Both resource use and unit costs	Not applicable	1 (12.5%)	Not clear
Drug costs	6 (32%)	Not clear	Not clear
Other unit costs	6 (32%)	Not clear	Not clear
Similar results	7 (36%)	2 (25%)	3 (18%)

Type C, comparisons based on same effectiveness data and resource use for all countries, with different unit costs.
Type RC, comparisons based on same effectiveness data for all countries, with different resource use and unit costs.
Type REC, comparisons based on different effectiveness data, resource use and unit costs for all countries.
Type NI, comparisons were excluded because of lack of information. These were comparisons with detailed data reported only for one country, but with negligible information given about the cost-effectiveness ratios for other countries.

Source: Table 1 in Barbieri et al, 2005.[210]

Extra information

- HTA Adaptation Toolkit & Glossary: EUnetHTA; 2011.[9]

3.15 ICER threshold

The meaning of the ICER threshold and the interpretation of the ICER are explained in every handbook of economic evaluation of health technologies. The EUnetHTA guideline on Methods for health economic evaluations summarizes some concepts:[1]

- *"The ICER represents the estimated difference in costs between the intervention and the comparator divided by the estimated difference in effect between the intervention and the comparator. In an example where the effect is measured in life years, the estimated ICER could be reported as the cost per life-year gained. If the effect is measured in QALYs, the estimated ICER would be reported as the cost per QALY gained."*
- *"Whether a technology can be referred to as 'cost effective' depends on its relation to the 'decision-makers' willingness-to-pay' or the 'societal willingness-to-pay' for an additional unit of health outcome, or a so-called 'ICER threshold' or 'cost-effectiveness threshold'. (...) If the estimated ICER is higher than the threshold, the technology is not considered to be cost effective and hence allocation of resources to this technology would be unlikely to increase economic efficiency in health care. (...) For some decision-making authorities, the ICER threshold may vary between technologies or diseases, depending on characteristics of the technology or disease that are not necessarily directly reflected in ICER estimates (...) it is rare that the decision-making authorities have explicit thresholds."*
- *"The cost-effectiveness acceptability curve (CEAC) shows the probability that an intervention is cost-effective compared to its comparator or comparators, at different cost-effectiveness thresholds. The vertical axis of the diagram represents the probability that the intervention is cost-effective and the horizontal axis represents different CE thresholds.* [in case of two alternatives] *The curve shows the percentage of the simulated ICERs in the CE plane that are lower than any specific threshold."* In

case of more than two alternatives, the curve shows the probability that an intervention is considered cost-effective for a range of CE thresholds.

When there is no dominance (i.e. the intervention is both more effective and less costly) of an alternative over the other, the estimation of the ICER is needed.[nn] When the outcome is expressed in e.g. life-years or QALYs gained the interpretation requires a threshold to compare with. Such a threshold can be used across indications. However, most countries have no explicit threshold for making decisions. The National Institute for Health and Care Excellence (NICE) in England and Wales explicitly reports a range of £20 000 to £30 000 per QALY as their threshold,[86] if none of the special criteria outlined in recent amendments are met.[211] In other countries, there are recommended or frequently used/cited thresholds, but not formally adopted.[212] For the World Health Organization (WHO) "*interventions that avert one DALY* [disability-adjusted life-year] *for less than average per capita income for a given country or region are considered very cost-effective; interventions that cost less than three times average per capita income per DALY averted are still considered cost-effective; and those that exceed this level are considered not cost-effective.*"[213] DALYs are not equivalent to QALYs. Nevertheless, this criteria has been adopted by some countries, while it cannot be used in others. For instance, in 2012 Poland decided to set the cost-effectiveness threshold of three times the per-capita gross domestic product (GDP) per QALY gained for reimbursing new medicines, that is, 130 002 zloty approximately (€30 500).[214][oo] Applying the WHO threshold to the United Kingdom would result in a threshold value of approximately £89 000, which is questionable since this would not match the system's 'ability to pay'.[215]

There are different ways to set a cost-effectiveness threshold.[212, 216, 217] In this chapter we do not present, recommend or criticize any of them. We want to draw your attention to some issues when reading the interpretation of the results in an economic evaluation. For example, taken from an overview of threshold values for cost-effectiveness in health care by KCE, we can highlight some ideas:[218]

- *"The ICER threshold value is not a static value but changes over time (...) The ICER threshold value is the result of a health maximisation model that applies to a specific context (fixed budget, country), at a specific moment in time and under specific conditions."*
- *"The ICER threshold value is subject to uncertainty and variability. Therefore, the ICER threshold value is not a single value but a range of values. This is important for the kind of conclusions that can be drawn from cost-effectiveness analyses."*

[nn] We remark that presenting the incremental costs and incremental effects is needed, even when dominance is observed, e.g. to present the size of the QALYs gained and cost differences.

[oo] The value of this threshold is updated every year according to the GDP per capita (x3) provided by the National Statistical Office (https://stat.gov.pl/sygnalne/komunikaty-i-obwieszczenia/lista-komunikatow-i-obwieszczen/obwieszczenie-w-sprawie-szacunkow-wartosci-produktu-krajowego-brutto-na-jednego-mieszkanca-w-latach-2013-2015-na-poziomie-wojewodztw-nts2-i-podregionow-nts3,281,4.html).

- *"The units in which the costs and health effects are expressed are important for the interpretation of the ICER threshold value: an ICER threshold value of €30 000/QALY is different from an ICER threshold value of £30 000/LYG."* Both the currency used in the numerator and outcome parameter used in the denominator are of importance. Transferring explicit or implicit ICER threshold values to other outcomes such as progression-free life-years saved (PF-LYS) is not correct.
- *"In all countries decision making is not solely based on cost-effectiveness considerations. The technology is assessed based on efficiency criteria together with other criteria. In the presence of high ICERs, those other criteria become more important."*

Points for consideration

- Cost-effectiveness is not the only criterion to make decisions. It is not just because an intervention has an acceptable ICER that it will be reimbursed and vice versa. Other elements like the uncertainty around the estimates, the budget impact and budgetary context, the degree of unmet medical need, etc. also influence the reimbursement decision.
- There is often no explanation/justification for the selection of the cost-effectiveness threshold (or range of thresholds) (see Box 30). An explanation/justification for the selection of the applied threshold should be given. Readers should pay caution if authors refer to the ICER of an intervention that received a positive reimbursement decision since it is possible that the economic criterion has been ignored/overruled in this decision (e.g. because it was a very small population with a very severe disease and high unmet medical need).
- It might be more difficult to interpret or discuss the results in function of different cost-effectiveness thresholds if no CEAC is presented (see Box 31).
- Conclusions on efficiency cannot be made in the light of any threshold in countries which do not have an explicit threshold. To refer to relatively high ICER thresholds that are not accepted in their country might result in too optimistic conclusions (see Box 32).

Examples

Box 30: Reporting the incremental cost-effectiveness ratio (ICER) in the absence of an agreed ICER-threshold

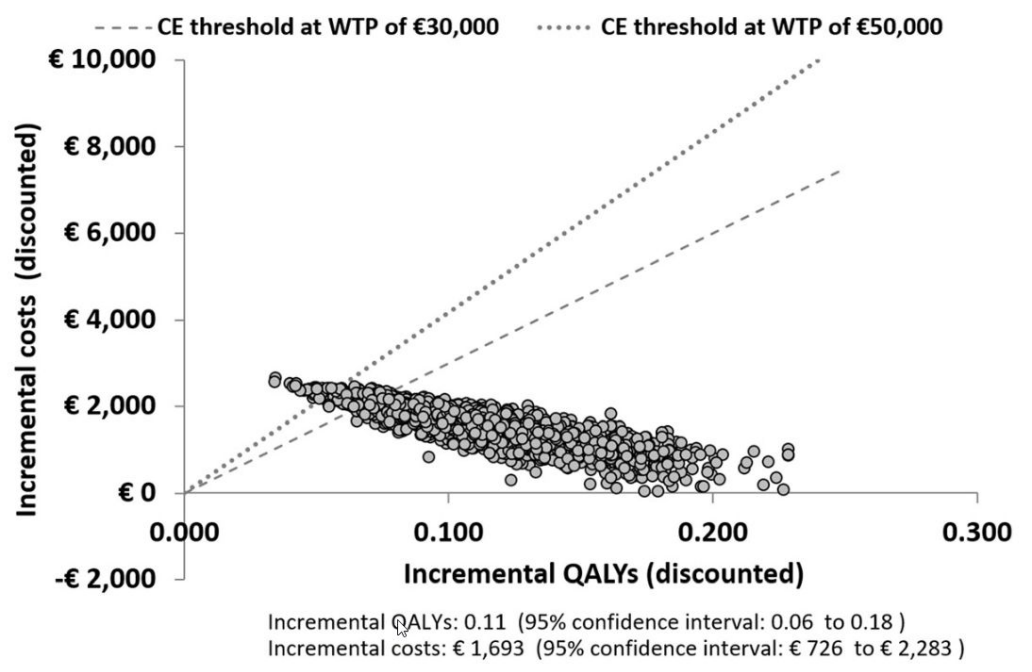
There are many examples of economic evaluations using thresholds without an explanation of its origin. Roze et al. for example, assumed for their economic evaluation in Denmark a commonly accepted willingness-to-pay threshold of €30 000 (approx. DKK225 000) per QALY gained.[219] There is no explanation in the paper of the origin of this €30 000 per QALY.

The same figure is found in the literature in Spain. Catalá et al. conducted recently a systematic review of cost-utility analyses in Spain where they found that 56.5% of studies mentioned the hypothetical threshold of €30 000 per QALY.[220] This figure comes from a review, published in 2002, of the economic evaluations in Spain from 1990 to 2001, where it was found that "*all technologies with a cost-effectiveness ratio lower than €30 000 euros (5 million pesetas) per LYG were recommended for adoption by the authors.*"[221] Since then, having no other criteria in Spain, most of authors have used that figure in their analysis. The systematic review by Catalá also found that 41.3% of studies included the CEAC "*to contrast the results of the analyses against an arbitrary efficiency threshold.*"[220]

Schmidt et al. for example, estimated the cost-effectiveness of palivizumab to prevent respiratory syncytial virus versus placebo in children with congenital heart disease from the societal perspective in Spain.[222] They concluded that *"palivizumab prophylaxis was shown to be a cost-effective health care intervention according to the commonly accepted standards of cost-effectiveness in Spain (ICER below the threshold of €30 000 per QALY)"* but did not cite any reference for this 'commonly accepted' threshold. Nevertheless, the authors commented the results of the probabilistic sensitivity analysis, that is, the graphic representations (cost-effectiveness plane and CEAC) and the probability of acceptance for different thresholds:

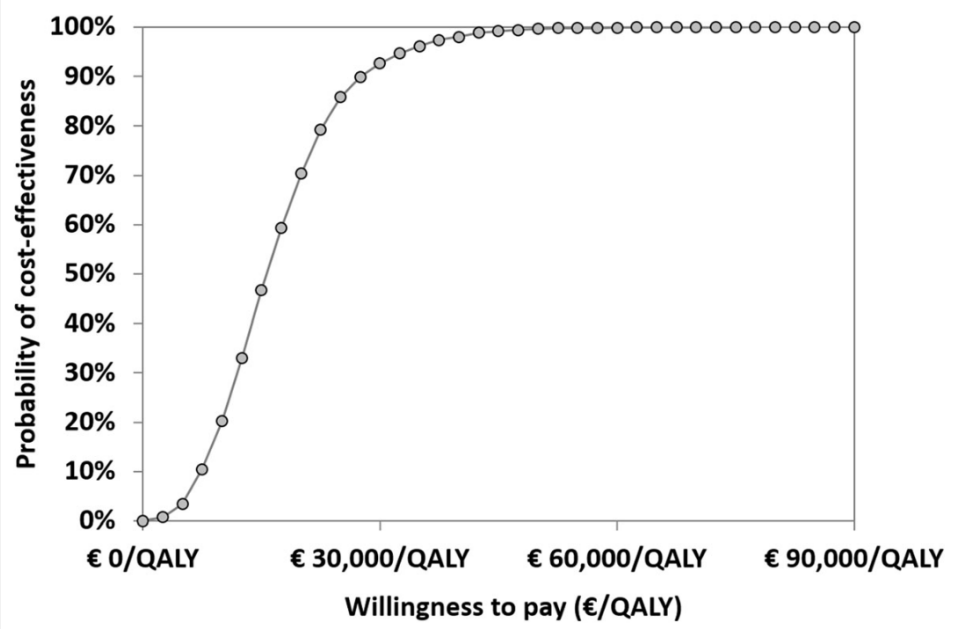
*"Probabilistic sensitivity analyses demonstrated that the probabilities of palivizumab prophylaxis being cost-effective at a threshold of €30 000 per QALY, €50 000 per QALY and €100 000 per QALY were 92.7%, 99.6% and 100.0%, respectively. Results of all simulations (100%) fell in the upper right quadrant of the CE plane, denoting both positive incremental QALYs and costs. Simulation results are shown in [*Figure 17*] (scatter plot of incremental results) and [*Figure 18*] (cost-effectiveness acceptability curve)."*[222] (see figures below). The CEAC provides information for different thresholds allowing the decision makers to interpret the results in absence of an explicit ICER threshold. In the absence of an explicit threshold, researchers should include similar analyses, to aid decision makers and to fully contextualise their results.

Figure 17: Example of indicating a cost-effectiveness threshold on the cost-effectiveness plane



Source: Figure 3 in Schmidt et al. (2017).[222]

Figure 18: Example of a cost-effectiveness acceptability curve



Source: Figure 4 in Schmidt et al. (2017).[222]

A more recently published study by Vallejo et al. estimated the threshold for Spain following accepted methodologies. They estimated a threshold range between €22 000 and €25 000 per QALY and concluded that "*these values are below the cost-effectiveness threshold figure of €30 000 commonly cited in Spain.*"[223] This figure estimated by Vallejo et al. has not been endorsed by the Ministry of Health yet but it is the first time that the threshold has been estimated in Spain. This reminds

us that the threshold could/should change over time and the importance of reporting and explaining the use of a specific threshold (or its absence) as readers may not know the current situation in each country.

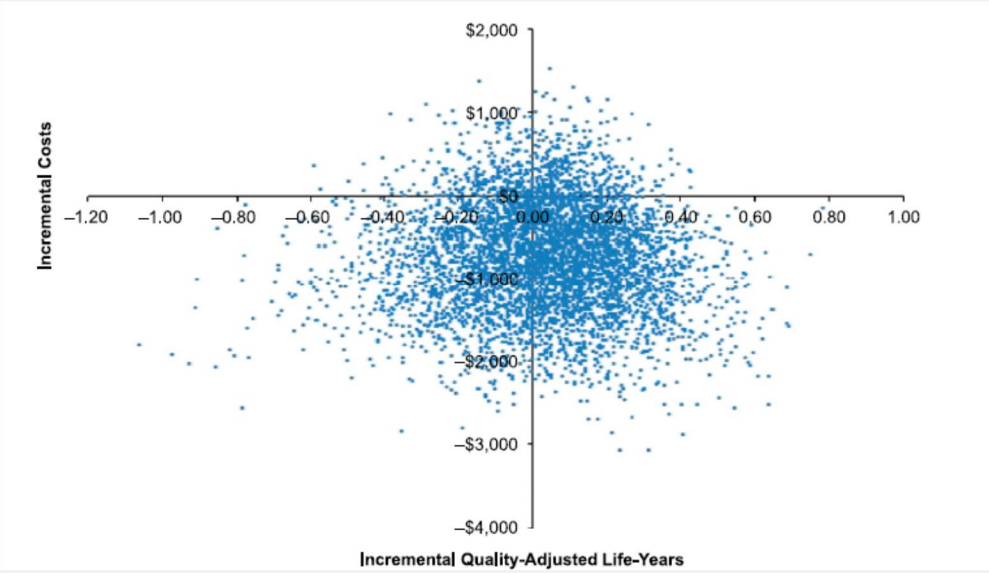
Box 31: Vague interpretation of incremental cost-effectiveness ratios and cost-effectiveness thresholds

Roth et al. evaluated the cost-effectiveness of a biopsy-based, quantitative, multiplex, 8-protein, in situ imaging prognostic assay (ProMark®) to provide prognostic information and inform treatment decisions compared with standard guideline-based care in patients with early stage prostate cancer.[224] They conducted their analysis from the payer perspective in the USA. In the methods section, it is said that the authors "*evaluated the cost-effectiveness at willingness-to-pay thresholds ranging from $10 000 to $150 000 per QALY.[225-227] This range reflects the implied willingness to pay for cancer treatments in the U.S. and is consistent with values used in prior analyses.[226, 228, 229]*"

In the USA there is no formally adopted threshold. It seems that $50 000 per QALY is still the figure most commonly cited, although $100 000 per QALY is also referenced by a number of authors or even recommended.[225] This ambiguity could be used by authors to interpret their results with vagueness.

In Roth et al., the results in the base case concluded that the 8-protein assay is dominant in terms of $ per QALY (0.04 more QALYs and $700 less costs compared with usual care). However, this conclusion is based on the point estimate and the probabilistic sensitivity analysis shows a more complex reality: *"In the probabilistic sensitivity analysis, we found that, relative to the usual care strategy, the 8-protein prognostic assay strategy decreased cost in 86.9% of simulations and increased QALYs in 58.3% of simulations [*Figure 19*]. In addition, probabilistic sensitivity analysis demonstrated that the 8-protein assay strategy is likely to be cost-effective across willingness to pay thresholds ranging from $10 000 to $300 000 per QALY."*[224]

Figure 19: The cost-effectiveness plane



Source: Figure 3 from Roth et al.[224]

The interpretation of the results would not be complete if it is only based on the point estimate. Fortunately, the CE plane is presented. The simulations occupying the four quadrants (Figure 19) tell us that the results are quite uncertain. Having 58.3% of the simulations in the first (north-east) and second (south-east) quadrant of the cost-effectiveness plane (i.e. the intervention is more effective than the comparator) also means that 41.7% of the simulations indicate worse results. Focusing on the point estimate and concluding the intervention is dominant without further nuance would contradict with the uncertainty around the treatment effect.

Box 32: Making conclusions more optimistic by comparing cost-effectiveness results with very high 'incremental cost-effectiveness ratio'-thresholds

Some authors refer to previous reimbursement decisions to extract the willingness-to-pay. However, economic considerations (both cost-effectiveness and budget impact) are only part of this decision and it is possible that rational economic considerations were ignored. Other arguments may have supported the reimbursement decision, e.g. unmet need, innovative nature of the intervention, employment arguments, etc. Referring to ICER thresholds from previous decisions where economic considerations may have been left out of the discussion runs the risk of extracting too high ICER threshold values and systematically reimbursing interventions that do not offer sufficient value for money for society.[pp] Such a

[pp] Vice versa, referring to ICER values from interventions that were not reimbursed can also lead to an underestimation of ICER thresholds. The non-consideration of economic aspects can be based on e.g. an opinion that the intervention is not regarded as innovative or as a priority for the decision makers while it might provide good value-for-money.

situation should be avoided and in the majority of cases, the ability of society to pay for health gains should be taken into account in order to support decisions in favour of an accessible, high-quality and financially sustainable health care system.

As an example, in an HTA report on bevacizumab in ovarian cancer, a systematic review of economic literature was performed.[158] In contrast to NICE's explicit ICER threshold, a wide range of willingness-to-pay values was used in the identified economic evaluations. In several cases, calculated ICERs might be presented as acceptable by comparing them with relatively high and non-well justified ICER threshold values. The danger exists that such relatively high ICER thresholds do not reflect the willingness/ability-to-pay in a context of limited resources. In the next paragraph, we present an overview of the variety of stated willingness-to-pay threshold values that were identified in the economic evaluations of bevacizumab in ovarian cancer.

Barnett refers to "*traditional willingness-to-pay thresholds of $50 000 to $100 000 per QALY - originally established in the dialysis literature several decades ago*" and questions whether this threshold is "*outdated and should be raised to reflect our current economy and practice patterns*".[230] Cohn et al. initially refer to the traditional $50 000 per life-year saved threshold and remark that "*despite the controversy … this convention was used to guide the interpretation of the model rather than to conclude that one intervention should or should not be used.*"[231] In their updated analysis,[232] the traditional $50 000/QALY threshold has been replaced by a $100 000/QALY value. This value is also referred to by Lesnock.[233] Mehta et al.[234] increase this to a societal willingness-to-pay ICER threshold of $150 000/QALY, i.e. roughly three times the US GDP per capita. The authors also refer to another study describing that "*for clinical oncologists, minor effectiveness of intervention is considered of good value. Hence in clinical practice, oncologists demonstrate a willingness-to-pay threshold of $300 000/QALY[226]*". Chan et al. refers to "*a maximum ICER threshold of $200 000 per life-year saved to consider an intervention as a value at which most health care systems approve new therapeutic options*" and mentions that "*the ICER of [bevacizumab] in this study cohort appears comparable to costs in colorectal cancer, but lower than either breast or lung cancer, both of which were found to be more than $200 000.*"[235] Duong et al. note that "*Canada has no official cost-effectiveness threshold that determines the willingness-to-pay of the public health care system. However, many of the oncology therapies currently funded have ICERs well above CAD100 000 per QALY*".[236] Chappel et al.[237] applies the $100 000 threshold value for their disease-specific outcome of progression-free life-year saved as if 'life years', 'QALYs' and 'progression-free life-years' outcomes are comparable. Finally, a manufacturer's submission to NICE goes even further stating that the "*discussion on the threshold to be used in the US [is] still ongoing, and some might consider that the intervention is cost-effective if below US$500 000/QALY.*"[238]

Using ICER threshold values that are too high rather equates to ignoring the economic argument in decision making. When there is no explicit ICER threshold value and decision makers consider the presented ICER threshold too high to apply systematically, presenting the results on the CEAC provides a good alternative. This allows decision makers to interpret the results applying their own willingness/ability-to-pay instead of the (possibly unrealistic) values that might be stated by authors of an economic evaluation.

Extra information

- Methods for health economic evaluations - A guideline based on current practices in Europe. Methodological Guideline: EUnetHTA; 2015.[1]
- Threshold values for cost-effectiveness in health care. Brussels: Belgian Health Care Knowledge Centre (KCE); 2008.[218]
- Cost-Effectiveness Thresholds in Health Care: A Bookshelf Guide to their Meaning and Use. CHE Research Paper 121. 2015.[216]

3.16 Publication bias of economic evaluations and conflicts of interest

Conflict of interest (CoI) can be described as a situation in which someone risks not being able to make a fair decision because they will be affected by the result. Three sources of CoI should be considered when critically assessing economic evaluations: authors affiliations (if authors are affiliated with the company the CoI is potentially high instead of low), authors received funding from a company, or other forms of CoI, such as the authors having developed the intervention.[18, 19] Publication bias relates to *"the publication or non-publication of research findings, depending on the nature and direction of the results."[51]* Assessors, policy makers, and other stakeholders need to be aware of the potential influence of publication bias and CoI on study results, conclusions and recommendations.

Points for consideration

Some of the potential problems related to CoI and publication bias are:

- Industry-sponsored cost-effectiveness studies have been found to be more likely to report favourable conclusions,[239-242] and less likely to report unfavourable conclusions for the sponsor's product, compared to studies with other sources of finance.[240, 243, 244] Studies funded by industry are also more likely to report lower cost-effectiveness ratios.[245-248][qq] This could be a sign of bias due to conflict of interest, publication bias or other reasons, such as selective financing of studies where the result is more likely to be favourable. Another explanation is that the industry in an early development phase discontinue the development of products they deem economically unattractive.
- Studies have shown that there might be a tendency to select values for input variables that favour the sponsor's product (see Box 33).[249, 250]
- Trial-based economic evaluations can sometimes suffer from publication bias influenced by the clinical outcomes of the study. A study on randomized trials on enhanced care for depression showed that effect size was almost twice as large in studies with a concurrent economic evaluation compared to those without.[251] This indicates that authors might be more willing to publish also an economic evaluation if the clinical outcomes are better. Another study found that trials that intend to conduct an economic analysis are less likely to report economic data than effectiveness data. Furthermore, if economic results do appear, they are published at a later time. The

[qq] In one of the studies this relationship was not statistically significant when controlling for other factors such as methodological quality.[246]

authors behind the trials stated different reasons behind the non-publication of data. Among these was the intervention being ineffective and indifference to economic data.[252]

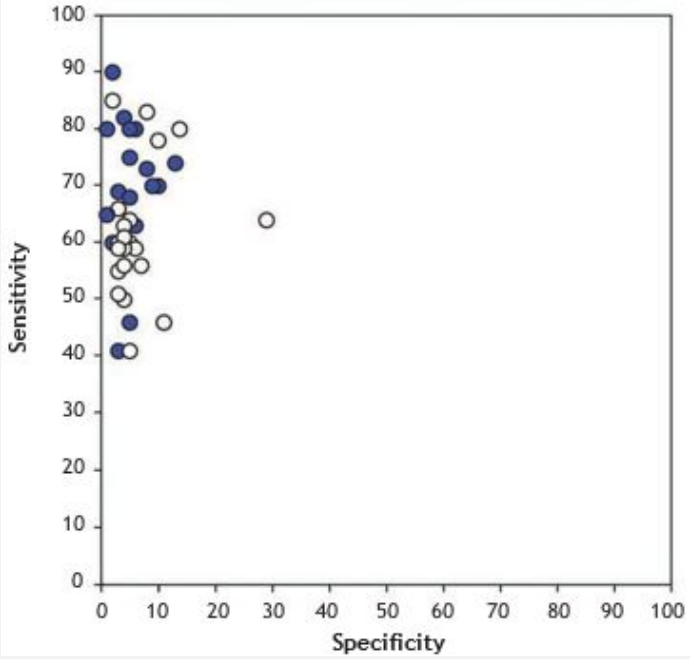
- Economic evaluations are also vulnerable to publication bias in the health-outcomes data available for modelling, resulting from publication bias in the clinical literature (see part 3.1).[253]
- We must be careful that efforts to eliminate bias do not lead to the introduction of new bias. The critical assessment of identified studies is preferable to the complete disregard of studies by authors with a possible CoI.

Examples

Box 33: Differing diagnostic test results in studies with and without manufacturer involvement

Polyzos et al.[249, 250] searched for cost-effectiveness studies where at least one strategy involved the Papanicolaou (Pap) test, a widely used screening test for cervical cancer. They identified 88 studies performing an economic evaluation of a new technique in comparison with the Pap test. The assumed sensitivity of the Pap test was lower (mean: 60% versus 70%, $p<0.001$) in studies with manufacturer-affiliated authors, manufacturer funding or manufacturer-related competing interests versus studies without. The assumed specificity of the Pap test did not differ between trials with manufacturer involvement compared to those without (see Figure 20). The suggested interpretation is that unfavourable assumptions for the comparator arm is used to enhance the cost-effectiveness ratio of the new competing technology.[249, 250] In such cases, a critical assessment of the clinical evidence is needed to judge whether the input values used in the economic evaluation are appropriate. In such situations, decision makers may require scenario analyses to determine the extent of uncertainty generated by varying the model input values.

Figure 20: Assumed sensitivity and specificity of the Pap test in studies with and without various types of manufacturer involvement



Source: Polyzos et al., 2011[250]
Filled circles: Estimates for cost-effectiveness analyses in which authors were not affiliated with, funded by or in conflict of interest in relation to the manufacturer.

Empty circles: Estimates for cost-effectiveness analyses in which at least one author was affiliated with, funded by or in conflict of interest in relation to the manufacturer.

Extra information

- Checklist for Assessing the Quality of Health Economic Modelling Studies. Assessment of methods in health care - A handbook. Version 2017:1 ed: SBU 2018:B8:1-4.[19]
- Checklist for Assessing the Quality of Trial-Based Health Economic Studies. Assessment of methods in health care - A handbook. Version 2017:1 ed: SBU 2018:B7:1-4.[18]

4 Conclusion and main recommendations

This guidance document has been developed to support researchers when performing economic evaluations and provide backing to assessors when assessing such evaluations. This guidance document provides an overview of the various elements in an economic evaluation and a related non-exhaustive list of possible points for consideration. In some cases, these have been further elaborated with an example to make a bridge between theory and practice. In this way, we hope that researchers and assessors will become more familiar with this subject and will have more confidence when assessing existing economic evaluations.

We recommend that researchers not only look at the outcomes of economic evaluations but also critique the inputs and the applied methods that lead to these outcomes. Depending on the results of such an assessment, the results can be used further or may be required to be updated or adjusted to meet the expectations of decision makers. When a result can be considered reliable or when an adjustment is required is very context- and case-specific. We leave this judgement up to the assessor. We hope that this guidance document can support all involved parties when performing or evaluating economic evaluations and thus stimulate the (re)use of economic evaluations in decision-making processes.

5 Annexes

5.1 Annex 1 – Documentation of literature search

5.1.1 Keywords

For this guidance document, no traditional 'PICO'-search is possible as usually is applied in a 'classical' HTA report. Several articles discussing some of the elements mentioned in Table 1 were identified in an attempt to try to extract relevant search terms. However, based on the Medline indexation of the first four references[114, 200, 254, 255] that were considered relevant, no specific (standard) indexing terms could be identified.

A general search strategy was performed combining three groups of search terms. The first group included terms related to HTA, economic evaluations and modelling and were combined with the Boolean operator 'OR' (see Table 8). The second group of search terms referred to guidelines, methodology, reproducibility and validity and were also combined with 'OR'. The third group of search terms referred to one of the specific elements mentioned in Table 1. These three groups of search terms were then combined with the Boolean operator 'AND'.

Table 8: Elements suggested for inclusion in this guidance document on critical assessment of economic evaluations

HTA, economic evaluations and modelling (\OR)	Guidelines, methodology, reproducibility and validity (\OR)	Elements
Technology Assessment, Biomedical/ health technology assessment.mp. systematic review.mp. Cost-Benefit Analysis/ economic evaluation.mp. cost-effectiveness analys$.mp. cost-utility analys$.mp. cost-minimization analys$.mp. cost-consequences analys$.mp. models, economic/ (Modelling or modeling).mp	guideline/ Methodology.mp. Evidence-Based Medicine/ec, mt [Economics, Methods] Reproducibility of Results/ Valid$.mp.	See details in Table 9 - Table 16

The search was initially performed in Medline OVID (Table 9 - Table 23) by a researcher from KCE. For every element (comparator, subgroup analysis, baseline risk, etc.), one of the co-authors was designated to check the selection and results of this search strategy. Unfortunately, the results of the search strategy were rather disappointing: a lot of references were searched with a low number of relevant articles being identified. Most

relevant documents, like EUnetHTA and ISPOR guidelines were not identified through this search. The group of authors (KCE, HAS, SBU, SESCS-FUNCANIS, HIQA) decided not to extend the search to other databases (e.g. EMBASE). The search was thus only used to extend the list of points for consideration and to identify supporting examples (see Boxes in the main document). The main source of information was thus the existing guidelines that were identified on websites (EUnetHTA, websites of HTA bodies and ISPOR) and the experience of the authors, complemented by the comments of the external reviewers. The emphasis was on HTA bodies that are part of the EUnetHTA collaboration. As this document is intended for guidance and is not prescriptive, the review of existing guidelines was not systematic, but rather intended to identify areas of best practice and of common issues in the conduct of economic evaluations.

For transparency, the search strategy is reported in the following part, including the name of the database and interface, date of search, search terms used, and hits per search term.

5.1.2 Search strategy

Table 9: Literature search – comparator

Date	11 October 2017		
Database	Medline OVID (Ovid MEDLINE(R) without Revisions 1996 to September Week 4 2017)		
Search Strategy	1	Technology Assessment, Biomedical/	6484
	2	health technology assessment.mp.	2469
	3	systematic review.mp.	73120
	4	Cost-Benefit Analysis/	55130
	5	economic evaluation.mp.	5798
	6	cost-effectiveness analys$.mp.	7283
	7	cost-utility analys$.mp.	1842
	8	cost-minimization analys$.mp.	387
	9	cost-consequences analys$.mp.	46
	10	models, economic/	7749
	11	(Modelling or modeling).mp.	174558
	12	1 or 2 or 3 or 4 or 5 or 6 or 7 or 8 or 9 or 10 or 11	310566
	13	guideline/	11348
	14	Methodology.mp.	156111

	15	Evidence-Based Medicine/ec, mt [Economics, Methods]	4980
	16	Reproducibility of Results/	317164
	17	Valid$.mp.	445783
	18	13 or 14 or 15 or 16 or 17	810459
	19	12 and 18	44475
	20	comparator.mp.	5352
	21	19 and 20	179
Note	Combining (12 and 20 = 1014 hits) 'OR' (18 and 20 = 660 hits) resulted in 1495 hits, which was considered too much from a practical point of view.		

Table 10: Literature search – subgroup analysis

Date	20 October 2017		
Database	Medline OVID (Ovid MEDLINE(R) without Revisions 1996 to October Week 2 2017)		
Search Strategy	1	Technology Assessment, Biomedical/	6494
	2	health technology assessment.mp.	2478
	3	systematic review.mp.	73768
	4	Cost-Benefit Analysis/	55283
	5	economic evaluation.mp.	5830
	6	cost-effectiveness analys$.mp.	7310
	7	cost-utility analys$.mp.	1857
	8	cost-minimization analys$.mp.	389
	9	cost-consequences analys$.mp.	46
	10	models, economic/	7778
	11	(Modelling or modeling).mp.	175301
	12	1 or 2 or 3 or 4 or 5 or 6 or 7 or 8 or 9 or 10 or 11	312115
	13	guideline/	11372

	14	Methodology.mp.	156547
	15	Evidence-Based Medicine/ec, mt [Economics, Methods]	4988
	16	Reproducibility of Results/	318108
	17	Valid$.mp.	447518
	18	13 or 14 or 15 or 16 or 17	813214
	19	12 and 18	44719
	20	subgroup analys$.mp.	19677
	21	19 and 20	296
Note	/		

Table 11: Literature search – baseline risk

Date	11 October 2017		
Database	Medline OVID (Ovid MEDLINE(R) without Revisions 1996 to September Week 4 2017)		
Search Strategy	1	Technology Assessment, Biomedical/	6484
	2	health technology assessment.mp.	2469
	3	systematic review.mp.	73120
	4	Cost-Benefit Analysis/	55130
	5	economic evaluation.mp.	5798
	6	cost-effectiveness analys$.mp.	7283
	7	cost-utility analys$.mp.	1842
	8	cost-minimization analys$.mp.	387
	9	cost-consequences analys$.mp.	46
	10	models, economic/	7749
	11	(Modelling or modeling).mp.	174558
	12	1 or 2 or 3 or 4 or 5 or 6 or 7 or 8 or 9 or 10 or 11	310566
	13	guideline/	11348

	14	Methodology.mp.	156111
	15	Evidence-Based Medicine/ec, mt [Economics, Methods]	4980
	16	Reproducibility of Results/	317164
	17	Valid$.mp.	445783
	18	13 or 14 or 15 or 16 or 17	810459
	19	12 and 18	44475
	20	baseline risk.mp.	1805
	21	(baseline adj2 adjust$).mp.	5575
	22	Risk Adjustment/	2755
	23	20 or 21 or 22	9943
	24	19 and 23	131
Note	Combining (12 and 23 = 721 hits) 'OR' (18 and 23 = 1031 hits) resulted in 1621 hits, which was considered too much from a practical point of view.		

Table 12: Literature search – compliance and adherence

Date	30 October 2017		
Database	Medline OVID (Ovid MEDLINE(R) without Revisions 1996 to October Week 3 2017)		
Search Strategy	1	Technology Assessment, Biomedical/	6497
	2	health technology assessment.mp.	2482
	3	systematic review.mp.	74105
	4	Cost-Benefit Analysis/	55377
	5	economic evaluation.mp.	5846
	6	cost-effectiveness analys$.mp.	7333
	7	cost-utility analys$.mp.	1864
	8	cost-minimization analys$.mp.	392
	9	cost-consequences analys$.mp.	47

	10	models, economic/	7794
	11	(Modelling or modeling).mp.	175796
	12	1 or 2 or 3 or 4 or 5 or 6 or 7 or 8 or 9 or 10 or 11	313039
	13	guideline/	11377
	14	Methodology.mp.	156828
	15	Evidence-Based Medicine/ec, mt [Economics, Methods]	4993
	16	Reproducibility of Results/	318540
	17	Valid$.mp.	448612
	18	13 or 14 or 15 or 16 or 17	814807
	19	12 and 18	44839
	20	Compliance/	2318
	21	Medication Adherence/	13206
	22	20 or 21	15524
	23	19 and 22	112
Note	The search term 'Medication Adherence' was chosen instead of ‘adherence’ to exclude other meanings of this term (‘Advance Directive Adherence’ or ‘Guideline Adherence’).		

Table 13: Literature search – quality of life

Date	18 October 2017		
Database	Medline OVID (Ovid MEDLINE(R) without Revisions 1996 to October Week 1 2017)		
Search Strategy	1	Technology Assessment, Biomedical/	6488
	2	health technology assessment.mp.	2475
	3	systematic review.mp.	73458
	4	Cost-Benefit Analysis/	55206
	5	economic evaluation.mp.	5816
	6	cost-effectiveness analys$.mp.	7295

	7	cost-utility analys$.mp.	1850
	8	cost-minimization analys$.mp.	388
	9	cost-consequences analys$.mp.	46
	10	models, economic/	7765
	11	(Modelling or modeling).mp.	174936
	12	1 or 2 or 3 or 4 or 5 or 6 or 7 or 8 or 9 or 10 or 11	311366
	13	guideline/	11370
	14	Methodology.mp.	156316
	15	Evidence-Based Medicine/ec, mt [Economics, Methods]	4981
	16	Reproducibility of Results/	317659
	17	Valid$.mp.	446715
	18	13 or 14 or 15 or 16 or 17	811921
	19	12 and 18	44604
	20	"Quality of Life"/	138460
	21	quality of life.mp.	214202
	22	20 or 21	214202
	23	19 and 22	2858
	24	19 and 20	1565
Note	Combining (12 and 20 = 9031 hits) 'OR' (18 and 20 = 20549 hits) resulted in 28015 hits, which was considered too much from a practical point of view. Going through 2858 references (line 23) was also considered too much from a pratical point of view. Therefore, only the MeSH term for quality of life was considered (see line 24).		

Table 14: Literature search – intermediate and surrogate endpoints

Date	20 October 2017
Database	Medline OVID (Ovid MEDLINE(R) without Revisions 1996 to October Week 2 2017)

Search Strategy	1	Technology Assessment, Biomedical/	6494
	2	health technology assessment.mp.	2478
	3	systematic review.mp.	73768
	4	Cost-Benefit Analysis/	55283
	5	economic evaluation.mp.	5830
	6	cost-effectiveness analys$.mp.	7310
	7	cost-utility analys$.mp.	1857
	8	cost-minimization analys$.mp.	389
	9	cost-consequences analys$.mp.	46
	10	models, economic/	7778
	11	(Modelling or modeling).mp.	175301
	12	1 or 2 or 3 or 4 or 5 or 6 or 7 or 8 or 9 or 10 or 11	312115
	13	guideline/	11372
	14	Methodology.mp.	156547
	15	Evidence-Based Medicine/ec, mt [Economics, Methods]	4988
	16	Reproducibility of Results/	318108
	17	Valid$.mp.	447518
	18	13 or 14 or 15 or 16 or 17	813214
	19	12 and 18	44719
	20	surrogate.mp.	31566
	21	intermediary.mp.	5221
	22	intermediate.mp.	143142
	23	20 or 21 or 22	179068
	24	19 and 23	854
	25	surrogate endpoint.mp.	622
	26	surrogate end point.mp.	411

	27	surrogate outcome.mp.	295
	28	intermediate endpoint.mp.	103
	29	intermediate end point.mp.	84
	30	intermediate outcome.mp.	292
	31	intermediary endpoint.mp.	4
	32	intermediary end point.mp.	3
	33	intermediary outcome.mp.	8
	34	25 or 26 or 27 or 28 or 29 or 30 or 31 or 32 or 33	1783
	35	19 and 34	37
Note	Two very relevant references[145, 148] were already identified in the list of 37 references. There also exists a EUnetHTA guideline[15] on this topic. Therefore, it was preferred to just look at the 37 potential relevant references and not to go through the list of 854 references identified in line 24.		

Table 15: Literature search – time horizon and extrapolation

Date	18 October 2017		
Database	Medline OVID (Ovid MEDLINE(R) without Revisions 1996 to October Week 1 2017)		
Search Strategy	1	Technology Assessment, Biomedical/	6488
	2	health technology assessment.mp.	2475
	3	systematic review.mp.	73458
	4	Cost-Benefit Analysis/	55206
	5	economic evaluation.mp.	5816
	6	cost-effectiveness analys$.mp.	7295
	7	cost-utility analys$.mp.	1850
	8	cost-minimization analys$.mp.	388
	9	cost-consequences analys$.mp.	46
	10	models, economic/	7765
	11	(Modelling or modeling).mp.	174936
	12	1 or 2 or 3 or 4 or 5 or 6 or 7 or 8 or 9 or 10 or 11	311366
	13	guideline/	11370
	14	Methodology.mp.	156316
	15	Evidence-Based Medicine/ec, mt [Economics, Methods]	4981
	16	Reproducibility of Results/	317659
	17	Valid$.mp.	446715
	18	13 or 14 or 15 or 16 or 17	811921
	19	12 and 18	44604
	20	extrapolat$.mp.	20589
	21	time horizon.mp.	1904
	22	20 or 21	22403
	23	19 and 22	558

Note	Combining (12 and 22 = 3488 hits) 'OR' (18 and 22 = 3090 hits) resulted in 6020 hits, which was considered too much from a practical point of view.

Table 16: Literature search – uncertainty

Date	30 October 2017		
Database	Medline OVID (Ovid MEDLINE(R) without Revisions 1996 to October Week 3 2017)		
Search Strategy	1	Technology Assessment, Biomedical/	6497
	2	health technology assessment.mp.	2482
	3	systematic review.mp.	74105
	4	Cost-Benefit Analysis/	55377
	5	economic evaluation.mp.	5846
	6	cost-effectiveness analys$.mp.	7333
	7	cost-utility analys$.mp.	1864
	8	cost-minimization analys$.mp.	392
	9	cost-consequences analys$.mp.	47
	10	models, economic/	7794
	11	(Modelling or modeling).mp.	175796
	12	1 or 2 or 3 or 4 or 5 or 6 or 7 or 8 or 9 or 10 or 11	313039
	13	guideline/	11377
	14	Methodology.mp.	156828
	15	Evidence-Based Medicine/ec, mt [Economics, Methods]	4993
	16	Reproducibility of Results/	318540
	17	Valid$.mp.	448612
	18	13 or 14 or 15 or 16 or 17	814807
	19	12 and 18	44839
	20	Uncertainty/	8784

	21	19 and 20	233
Note	/		

Table 17: Literature search – model verification and validation

Date	20 October 2017		
Database	Medline OVID (Ovid MEDLINE(R) without Revisions 1996 to October Week 2 2017)		
Search Strategy	1	Technology Assessment, Biomedical/	6494
	2	health technology assessment.mp.	2478
	3	systematic review.mp.	73768
	4	Cost-Benefit Analysis/	55283
	5	economic evaluation.mp.	5830
	6	cost-effectiveness analys$.mp.	7310
	7	cost-utility analys$.mp.	1857
	8	cost-minimization analys$.mp.	389
	9	cost-consequences analys$.mp.	46
	10	models, economic/	7778
	11	(Modelling or modeling).mp.	175301
	12	1 or 2 or 3 or 4 or 5 or 6 or 7 or 8 or 9 or 10 or 11	312115
	13	guideline/	11372
	14	Methodology.mp.	156547
	15	Evidence-Based Medicine/ec, mt [Economics, Methods]	4988
	16	13 or 14 or 15	172490
	17	12 and 16	12192
	18	Reproducibility of Results/	318108
	19	model verification.mp.	115
	20	18 or 19	318210

	21	17 and 20	683
Note	If the search term 'Valid$.mp.' was added, too many references were identified (e.g. adding 'Valid$.mp.' to line 18 with 'OR' provides 663 072 hits. In combination with line 17 (with 'AND'), this already results in 2190 references.		

Table 18: Literature search – generalisability

Date	20 October 2017		
Database	Medline OVID (Ovid MEDLINE(R) without Revisions 1996 to October Week 2 2017)		
Search Strategy	1	Technology Assessment, Biomedical/	6494
	2	health technology assessment.mp.	2478
	3	systematic review.mp.	73768
	4	Cost-Benefit Analysis/	55283
	5	economic evaluation.mp.	5830
	6	cost-effectiveness analys$.mp.	7310
	7	cost-utility analys$.mp.	1857
	8	cost-minimization analys$.mp.	389
	9	cost-consequences analys$.mp.	46
	10	models, economic/	7778
	11	(Modelling or modeling).mp.	175301
	12	1 or 2 or 3 or 4 or 5 or 6 or 7 or 8 or 9 or 10 or 11	312115
	13	guideline/	11372
	14	Methodology.mp.	156547
	15	Evidence-Based Medicine/ec, mt [Economics, Methods]	4988
	16	Reproducibility of Results/	318108
	17	Valid$.mp.	447518
	18	13 or 14 or 15 or 16 or 17	813214
	19	12 and 18	44719

	20	generalizability.mp.	5787
	21	generalisability.mp.	984
	22	20 or 21	6766
	23	19 and 22	321
Note	/		

Table 19: Literature search – transferability

Date	20 October 2017		
Database	Medline OVID (Ovid MEDLINE(R) without Revisions 1996 to October Week 2 2017)		
Search Strategy	1	Technology Assessment, Biomedical/	6494
	2	health technology assessment.mp.	2478
	3	systematic review.mp.	73768
	4	Cost-Benefit Analysis/	55283
	5	economic evaluation.mp.	5830
	6	cost-effectiveness analys$.mp.	7310
	7	cost-utility analys$.mp.	1857
	8	cost-minimization analys$.mp.	389
	9	cost-consequences analys$.mp.	46
	10	models, economic/	7778
	11	(Modelling or modeling).mp.	175301
	12	1 or 2 or 3 or 4 or 5 or 6 or 7 or 8 or 9 or 10 or 11	312115
	13	guideline/	11372
	14	Methodology.mp.	156547
	15	Evidence-Based Medicine/ec, mt [Economics, Methods]	4988
	16	Reproducibility of Results/	318108
	17	Valid$.mp.	447518

	18	13 or 14 or 15 or 16 or 17	813214
	19	12 and 18	44719
	20	transferability.mp.	2066
	21	19 and 20	75
Note	/		

Table 20: Literature search – model sharing

Date	20 October 2017		
Database	Medline OVID (Ovid MEDLINE(R) without Revisions 1996 to October Week 2 2017)		
Search Strategy	1	model sharing.mp.	23
Note	No other search terms were added with 'AND' since only few references (23) were identified applying the term 'model sharing.mp.' The search was extended by using the following terms: - model*[Title] AND (free[Title] OR open[Title] OR share[Title] OR sharing[Title] OR available[Title]) AND cost-effectiveness - "open source" AND cost-effectiveness		

Table 21: Literature search – ICER threshold

Date	20 October 2017		
Database	Medline OVID (Ovid MEDLINE(R) without Revisions 1996 to October Week 2 2017)		
Search Strategy	1	Technology Assessment, Biomedical/	6494
	2	health technology assessment.mp.	2478
	3	systematic review.mp.	73768
	4	Cost-Benefit Analysis/	55283
	5	economic evaluation.mp.	5830
	6	cost-effectiveness analys$.mp.	7310
	7	cost-utility analys$.mp.	1857
	8	cost-minimization analys$.mp.	389

	9	cost-consequences analys$.mp.	46
	10	models, economic/	7778
	11	(Modelling or modeling).mp.	175301
	12	1 or 2 or 3 or 4 or 5 or 6 or 7 or 8 or 9 or 10 or 11	312115
	13	guideline/	11372
	14	Methodology.mp.	156547
	15	Evidence-Based Medicine/ec, mt [Economics, Methods]	4988
	16	Reproducibility of Results/	318108
	17	Valid$.mp.	447518
	18	13 or 14 or 15 or 16 or 17	813214
	19	12 and 18	44719
	20	ICER threshold.mp.	21
	21	willingness to pay threshold.mp.	529
	22	WTP threshold.mp.	68
	23	20 or 21 or 22	617
	24	19 and 23	55
Note	/		

Table 22: Literature search – publication bias

Date	30 October 2017		
Database	Medline OVID (Ovid MEDLINE(R) without Revisions 1996 to October Week 3 2017)		
Search Strategy	1	Technology Assessment, Biomedical/	6497
	2	health technology assessment.mp.	2482
	3	systematic review.mp.	74105
	4	Cost-Benefit Analysis/	55377
	5	economic evaluation.mp.	5846
	6	cost-effectiveness analys$.mp.	7333
	7	cost-utility analys$.mp.	1864
	8	cost-minimization analys$.mp.	392
	9	cost-consequences analys$.mp.	47
	10	models, economic/	7794
	11	(Modelling or modeling).mp.	175796
	12	1 or 2 or 3 or 4 or 5 or 6 or 7 or 8 or 9 or 10 or 11	313039
	13	guideline/	11377
	14	Methodology.mp.	156828
	15	Evidence-Based Medicine/ec, mt [Economics, Methods]	4993
	16	Reproducibility of Results/	318540
	17	Valid$.mp.	448612
	18	13 or 14 or 15 or 16 or 17	814807
	19	12 and 18	44839
	20	Publication Bias/	4085
	21	19 and 20	112
	22	4 or 5 or 6 or 7 or 8 or 9	59012

	23	20 and 22	32
Note	The references identified in line 21 mainly referred to publication bias in the medical literature. Since this guideline is focussed on economic evaluations, we limited the search to terms related to economic evaluations (line 22).		

Table 23: Literature search – conflict of interest

Date	30 October 2017		
Database	Medline OVID (Ovid MEDLINE(R) without Revisions 1996 to October Week 3 2017)		
Search Strategy	1	Technology Assessment, Biomedical/	6497
	2	health technology assessment.mp.	2482
	3	systematic review.mp.	74105
	4	Cost-Benefit Analysis/	55377
	5	economic evaluation.mp.	5846
	6	cost-effectiveness analys$.mp.	7333
	7	cost-utility analys$.mp.	1864
	8	cost-minimization analys$.mp.	392
	9	cost-consequences analys$.mp.	47
	10	models, economic/	7794
	11	(Modelling or modeling).mp.	175796
	12	1 or 2 or 3 or 4 or 5 or 6 or 7 or 8 or 9 or 10 or 11	313039
	13	guideline/	11377
	14	Methodology.mp.	156828
	15	Evidence-Based Medicine/ec, mt [Economics, Methods]	4993
	16	Reproducibility of Results/	318540
	17	Valid$.mp.	448612
	18	13 or 14 or 15 or 16 or 17	814807
	19	12 and 18	44839

	20	"Conflict of Interest"/	7973
	21	19 and 20	20
	22	4 or 5 or 6 or 7 or 8 or 9	59012
	23	22 and 20	92
Note	Similar as with the search for publication bias, line 21 mainly referred to CoI in the medical literature. Since this guideline is focussed on economic evaluations, we limited the search to terms related to economic evaluations (line 22).		

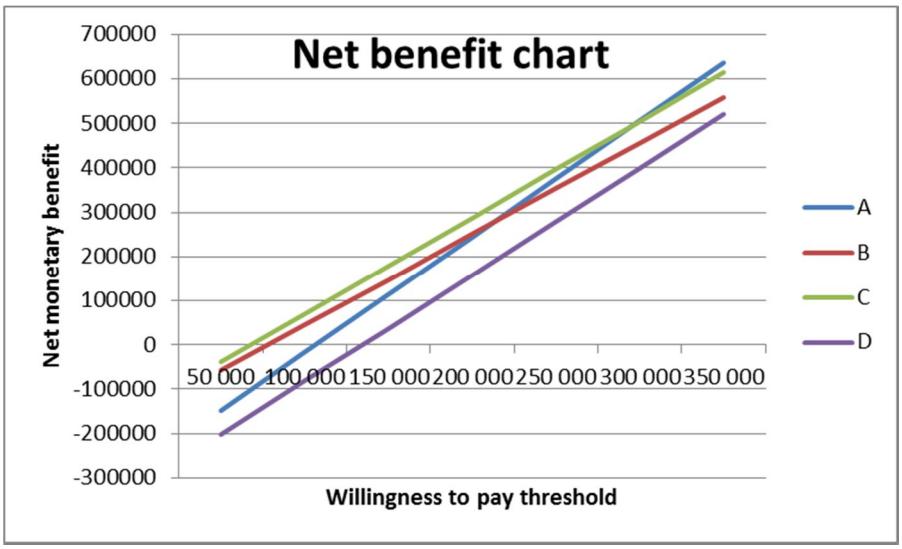
5.2 Annex 2 – Incremental cost-effectiveness ratios and incremental net benefit

The cost-effectiveness results may be reported in two equivalent measures: the Incremental Cost-Effectiveness Ratio (ICER) or the Incremental Net Benefit (INB) expressed in monetary terms (NMB) or in health terms (NHB). The net benefit framework transforms cost (C) and effect (E) into a linear function (with λ being the willingness to pay):

- Net monetary benefit (NMB) = ΔE * λ – ΔC
- Net health benefit (NHB) = ΔE - ΔC / λ

The ICER is the most consistently used measure. However, a number of criticisms have been formulated since 1998,[256] specifically targeted at the analyses of uncertainty. The net benefit approach was proposed in response to this criticism[257] and avoids the problem of interpreting ICERs of simulations with the same sign but in opposite quadrants of the CE-plane. Calculating CEAC also becomes much easier with the NB approach.[8] In the NB approach, the determinist result is reported as a linear function specific to each treatment option, allowing an immediate incremental interpretation depending on the willingness to pay. In Figure 21, intervention C has a positive incremental NMB up to €300 000/QALY, indicating that the intervention is cost-effective compared with all comparators B, C and D, up to this willingness-to-pay threshold.

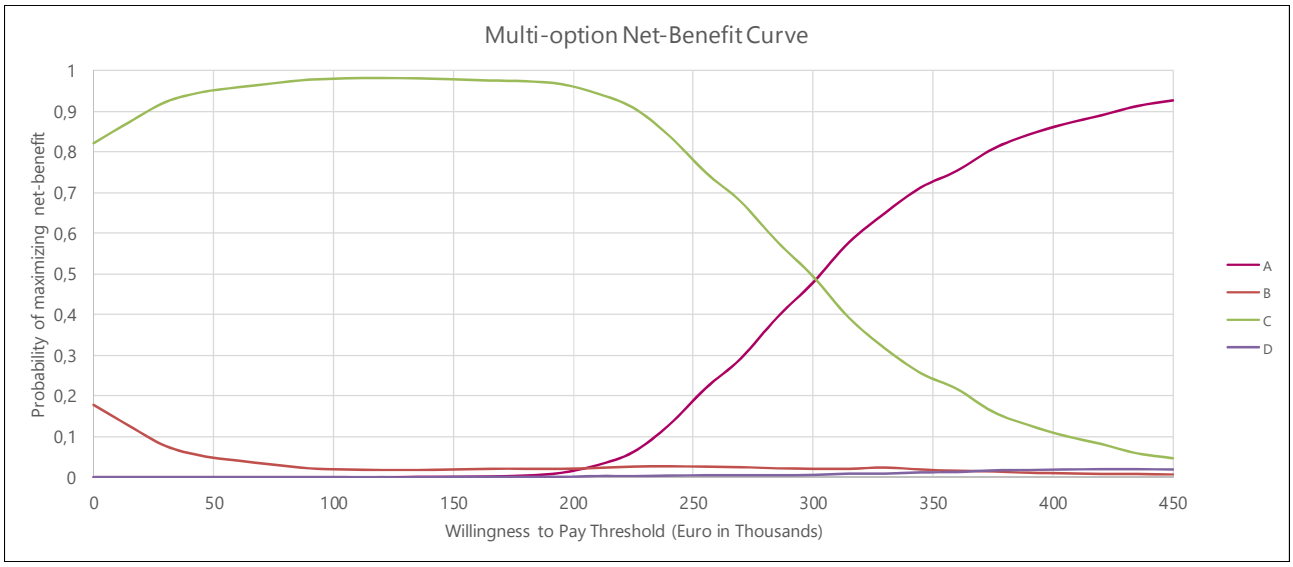
Figure 21: Net benefit chart (deterministic analysis)



Source: Anonymous example received from HAS.
Interpretation: "*(1) Each line represents a treatment option, (2) intercept at y-axis equals the cost of each option, (3) slope represents the effectiveness (steeper = more effective), (4) option with the highest NMB at a given WTP is the most cost-effective, (5) the CE frontier is the solid line at the top of the chart, (6) the incremental NB between any set of options is the vertical distance between the lines, (7) the ICER is where two lines on the frontier intersect.*"[258]

The probabilistic analysis involves a multi-option net benefit curve (or cost-effectiveness acceptability curve), which summarizes the uncertainty in estimates of cost-effectiveness (Figure 22). The curves indicate the probability that an intervention is cost-effective, that is the intervention maximizes the net benefit for a range of λ values.

Figure 22: multi-option net benefit curve (probabilistic analysis)



Source: Anonymous example received from HAS.
Interpretation: The probabilistic results show that intervention C has a probability of more than 50% of being cost-effective up to a €300 000/QALY threshold. The probability of being a cost-effective intervention is more than 80% for a willingness to pay under €250 000/QALY.

5.3 Annex 3 – Model sharing

In the field of economic evaluations of health technologies few models are shared.[259, 260] There are also no registries of economic models, which could help guarantee intellectual property.[260] A 'short piece of research' undertaken to know researchers' opinions on providing and using open source models found that 97% (34 out of 35) of participants said that "*it would be occasionally or very beneficial to have access to the code when reviewing existing models"[259]* and 63% "*had experienced challenges in accessing full details of a health economic model"[259]* (lack of willingness to share among other issues). Among the strategies to encourage the sharing of models suggested by the respondents were the collaboration between stakeholders to *develop open platforms or libraries to encourage sharing* (examples are journals, HTA agencies or consultants) or changes to regulatory processes, that is, *potentially integrated open access processes across organisations and formal request processes by health authorities*.[259]

Transparency and validation are linked and both are needed for gaining the confidence of decision makers, peer reviewers, or just readers.[190] According to the Report of the ISPOR-SMDM Modeling Good Research Practices Task Force-7 on Model Transparency and Validation, "*transparency refers to the extent to which interested parties can review a model's structure, equations, parameter values, and assumptions*" and it serves two purposes.[190] On the one hand, non-technical documentation (non-quantitative description of the model) should be accessible to any reader.[190] On the other hand, "*every model should have technical documentation, written in sufficient detail to enable a reader with the necessary expertise to evaluate the model and potentially reproduce it. The technical documentation should be made available openly or under agreements that protect intellectual property, at the discretion of the modelers.*"[190]

The ISPOR-SMDM report[190] highlights the conditions and limitations related to the provision of technical documentation:

"1. Access should be provided in a way that enables protection of intellectual property. Building a model can require a significant investment in time and money; if those who make such investments had to give their models away without restriction, the incentives and resources to build and maintain complex models could disappear.

2. While not mandatory, an increasing number of journals request that authors state whether full technical documentation is available to readers, and if so, under what terms. Technical documents may be placed in appendices or made accessible by other means. (…)

3. Because most multiapplication models change over time – expanded and updated to incorporate new information and advances in health care technologies – technical documents should be updated periodically.

4. Equations and detailed structure will mean little to readers without the necessary technical background. (…) Furthermore, it is very difficult to understand how accurate a model is simply by examining its equations. (…) Providing the code does not solve this problem unless the reader has the time and resources to actually implement it (…). Provision of code … would also threaten the protection of intellectual property. Some of these limitations can be addressed by giving readers access to the model or to a version applicable to a particular analysis. Even enabling readers to specify inputs and receive outputs of a model without releasing a full copy of it can provide useful information about how the model functions. (…)

Providing such access can be very expensive, including the cost to build the copy and interfaces and support to ensure that the model is used and interpreted accurately. (…)"

More recently, the debate has been reopened with a series of commentaries and point-counter-points published in Medical Care.[261-264]

Cohen et al. present the advantages of open source models:[263] it will enhance the credibility and the value of health economic analyses as the reproducibility is critical for scientific acceptance; it will facilitate the complete evaluation and understandability of models; it will have ancillary benefits by making the research more amenable for adaptation and innovation (…).[263] Moreover, these authors say that "*other fields have moved toward open publication of computer models, and health economics should avoid falling behind*", and that "*moving toward open publication will present challenges, but we believe that the benefits of increased scientific credibility and utility, particularly for health policy and clinical practice decisions, will certainly outweigh the harms.*"[263]

Padula et al. present the unintended consequences of open source models.[264] They highlight the intellectual property rights of the modellers and warn for the potential model misuse by inexperienced modellers. They propose two main solutions: "*licensing system of open source code such that the model originators maintain control of the code use and grant permissions to other investigators who wish to use it*" and "*teaching of cost-effectiveness analysis so that providers and other professionals are familiar with economic modeling and able to conduct analyses with open source code.*"[264]

5.3.1 Several levels of model sharing

In general, no one should use another modeller's model without permission/acknowledgement as we must respect their intellectual property rights. There are several levels of access when sharing a model. In the extreme, you have not sharing at all versus free access to the model (see Box 34). Other possibilities might be free access under non-commercial licenses which can be provided without or with previous registration (see Box 35 and Box 36, respectively), free access to a restricted model (see Box 37), or access under a commercial license (see Box 38). In the following examples, we provide cases of model sharing with different levels of access and transparency.

Examples

Box 34: A model with code files freely available (in two different software) with a request for acknowledgement

Sullivan et al.[265] published a paper with the details of an economic model in chronic pain. This paper focused on the methods more than on the results of the model. Four files are included as supplementary materials, two Microsoft Excel files and two files in the form of R code. The paper (and the supplementary material) is freely accessible as it was published in The European Journal of Health Economics under Open Access. It is a good sign to find models as supplementary materials attached to free papers although these online files are not indexed and consequently difficult to discover.[260]

The paper includes a note on the future use of the model code: *"MundiPharma International encourages free access and adaptation of the model code available as supplementary material. They request that future adaptations or applications make the following statement in the model code and publications. ''This model has been based on a Reference Case model in chronic pain as originally developed by MundiPharma International (Cambridge, UK) doi:10.1007/s10198-015-0720-y.''"*[265]

According to the authors:[265]

- "*The de novo model structure can be potentially used as a 'reference case' for future economic models for pain therapy and to guide future practice. To help with accessibility and applicability to different country settings, an effort has been made to make the model fully flexible and transparent, with the open-source code (in the programming language R) being provided as supplementary material. This is intended to allow other researchers to easily adapt and apply the model to further progress the development of health economic models in pain therapy.*"

In their conclusion, the authors state that they "*hope that the open-source reference model structure, as reported in this article, can act as the initial step in the development of a more consistent and transparent reference point for the development and assessment of future economic models in pain.*"[265]

Box 35: A model freely available under the GNU General Public License

Prakash et al.[266] published their open-source microsimulation model to estimate the cost-effectiveness of various colorectal cancer (CRC) screening strategies. The Colon Modeling Open Source Tool (CMOST) has been implemented in Matlab and is freely available under the Gnu is Not Unix (GNU) General Public License at https://gitlab.com/misselwb/CMOST. "*The GNU General Public License is a free, copyleft license for software and other kinds of works. (…) The GNU General Public License is intended to guarantee your freedom to share and change all versions of a program -- to make sure it remains free software for all its users… .*" (https://www.gnu.org/licenses/gpl-3.0.en.html)

The authors justify the need for its open-source model:[266]

"*…many new countries with different CRC epidemiology and health-care costs will be implementing CRC screening programs. Thus, there is an immediate requirement for an open source tool that is transparent, easily accessible, and adaptable for addressing highly relevant clinical and health economy questions.*"

"*…our model is publicly available under a GNU General Public License. This will enable independent reproduction of predictions and advancement of the model and its implementation.*"

"*…publication of all details of our microsimulation will enable scrutiny and a detailed discussion regarding all relevant aspects of CRC simulations. We hope that future extensions of CMOST will help increasing the validity of simulation results and further improve the in silico design of CRC screening strategies.*"

Box 36: A model with access restricted to those registered previously

Vataire et al.[267] designed an open-source model (a discrete event simulation model) to estimate health outcomes and costs associated with fictitious treatment strategies in different groups of patients with major depressive disorder. The model was implemented by means of Scilab (www.scilab.org), an open-source mathematical software package, and the code is available at https://www.open-model-mdd.org/. You must be a registered member to download the model (source codes and documentation), post comments, share input, share modifications of the codes or share a new version of the model. The authors state that:[267]

- *"This approach aims at transparency, at facilitating the use of the model by researchers from academia, health technology assessment agencies, or industry, and at enabling other researchers to contribute to the development of the model, for example, by sharing enhancements in the programs or by providing new input data."*
- "*By choosing to make it open-source and freely available on the Internet, we hope to foster the research community to develop, implement, and share new data and functions to populate, enhance, and validate the model.*"

Box 37: A model with access to input forms and outputs but not to codes

Coyle et al.[268] published the development and results of an economic model (Markov model) to estimate the cost-effectiveness of smoking cessation strategies, the EQUIPTMOD. This model and some accompanying documents (including a technical manual) are freely available to download from the website of this European project: http://www.equipt.eu/deliverables/. The files (Excel files) are user-friendly and are oriented to be used by the public, selecting their own country and other features according to your research question, and introducing inputs to receive in the end the outputs of the model. There is a video to guide the users (https://www.youtube.com/watch?v=FXOlewnzdGY). There is a web version of the model as well (http://roi.equipt.eu/).

The code is not available as the aim of the model is to be a "*decision support tool available for policy makers and not a research tool available for academics to*

conduct full-fledged cost-effectiveness analysis of a single intervention. However, given the existing assumptions, one could conduct such an analysis via the interface."[269]

Box 38: The case of diabetes mellitus: several models to choose from

An overview of diabetes models identifies 19 unique models finding that a clear, descriptive short summary of the model was often lacking.[270] Here we present two very different models from the point of view of accessibility and transparency.

One of the most cited models is the Core Diabetes Model (CDM) (http://www.core-diabetes.com) by the multinational company IQVIA. The CDM simulates clinical outcomes and costs for cohorts of patients with type 1 or type 2 diabetes mellitus. The outputs of the model include life expectancy, quality-adjusted life expectancy, direct costs, productivity losses, cumulative incidence and time to onset of complication, etc. It comprises 17 inter-dependent sub-models, using Markov modelling with tracker variables running in yearly cycles, with a maximum time horizon of 50 years. The model is "*an internet application linked to a mathematical calculation model and structured query language (SQL) database sited on a central server*" that "*operates with an executable code linked to a user front-end*". "*The justification for this centralized approach is improved version control, security and consistency. (…) Given the complexity of the IQVIA Health CDM, a compiled code is the only practical solution.*"(http://www.core-diabetes.com) Since the first validation of the model in 2004,[271] the model has undergone several updates and re-validations. The CDM is not free with prices varying depending on the package of services acquired (annual licenses, single-project licenses, licenses combined with consulting support, training, etc. As a result of not being free accessible, it is not completely transparent, but charging for its use keeps the model updated and technically maintained.

The other example is the PROSIT Disease Modelling Community. PROSIT is an international scientific open source development community for health economic disease models in medicine. The homepage is hosted at GECKO Institute for Medicine, Informatics and Economics of Heilbronn University (https://www.prosit.de). This project is formed by members from different countries and is open for new co-workers. More than 80 people have contributed to the modelling. The aim of this community is to develop transparent open source health economic disease models for diabetes mellitus and contribute to the credibility of the economic models in this field.[272] The PROSIT are Markov models to represent diabetes and their complications: myocardial infarction, stroke, retinopathy, nephropathy, diabetic foot syndrome, and hypoglycemia. They are developed in the open source sheet software OpenOffice Calc hosted in an Internet platform where documentation (including a Technical Handbook) is available for download by the public. It is needed to register and apply for an account to be able to contribute to the models. The models are available under the GNU Public License Version 2. At present, the site warns the readers that models published before 2017 are prototypes.

5.4 Annex 4 – Glossary

Term	Definition	Source
Absolute treatment effect	SEE Absolute risk reduction	
Absolute risk reduction	A measure of treatment effect that compares the probability (or mean) of a type of outcome in the control group with that of a treatment group, [i.e.: Pc- Pt(or µc- µt)]. For instance, if the results of a trial were that the probability of death in a control group was 25% and the probability of death in a treatment group was 10%, the absolute risk reduction would be (0.25 - 0.10) = 0.15. (See also number needed to treat, odds ratio, and relative risk reduction.)	HTA 101 glossary
Adverse effect	An adverse event for which the causal relation between the drug/intervention and the event is at least a reasonable possibility. The term 'adverse effect' applies to all interventions, while 'adverse drug reaction' (ADR) is used only with drugs. In the case of drugs an adverse effect tends to be seen from the point of view of the drug and an adverse reaction is seen from the point of view of the patient.	Cochrane glossary (https://community.cochrane.org/glossary)
Adverse event	An adverse outcome that occurs during or after the use of a drug or other intervention but is not necessarily caused by it.	Cochrane glossary
Biomarker	A biological molecule found in blood, other body fluids, or tissues that is a sign of a normal or abnormal process, or of a condition or disease. A biomarker may be used to see how well the body responds to a treatment for a disease or condition. Also called molecular marker and signature molecule.	https://www.cancer.gov/
Care pathway	A care pathway is a complex intervention for the mutual decision-making and organisation of care processes for a well-defined group of patients during a well-defined period. Defining characteristics of care pathways include: • an explicit statement of the goals and key elements of care based on evidence, best practice, and patients' expectations and their characteristics; • the facilitation of the communication among the team members and with patients and families; • the coordination of the care process by coordinating the roles and sequencing the activities of the multidisciplinary care team, the patients and their relatives; • the documentation, monitoring, and evaluation of variances and outcomes, and • the identification of the appropriate resources. The aim of a care pathway is to enhance the quality of care across the continuum by improving risk-	Schrijvers et al., 2012[273] referring to Vanhaecht et al., 2007[274]

Term	Definition	Source
	adjusted patient outcomes, promoting patient safety, increasing patient satisfaction, and optimizing the use of resources.	
Clinical endpoint	An event or other outcome that can be measured objectively to determine whether an intervention achieved its desired impact on patients. Usual clinical endpoints are mortality (death), morbidity (disease progression), symptom relief, quality of life, and adverse events. These are often categorized as primary (of most importance) endpoints and secondary (additional though not of greatest interest) endpoints.	HTA 101 glossary
Conflict of interest	A situation in which the private interests of a person contributing to an assessment influence the quality or the results of the assessment or the accuracy of the data.	http://htaglossary.net/
Cost-effectiveness plane	The cost-effectiveness plane is used to visually represent the differences in costs and health outcomes between treatment alternatives in two dimensions, by plotting the costs against effects on a graph. Health outcomes (effects) are usually plotted on the x-axis and costs on the y-axis.	www.yhec.co.uk/glossary
Cost-effectiveness acceptability curve	A curve illustrating the probability that a given option is efficient on the basis of the value assigned to an additional quality-adjusted life year (QALY).	http://htaglossary.net/
Deterministic sensitivity analyses	Probabilistic sensitivity analysis (PSA) is a technique used in economic modelling that allows the modeller to quantify the level of confidence in the output of the analysis, in relation to uncertainty in the model inputs. There is usually uncertainty associated with input parameter values of an economic model, which may have been derived from clinical trials, observational studies or in some cases expert opinion. … In the probabilistic analysis, these parameters are represented as distributions around the point estimate, which can be summarised using a few parameters (such as mean and standard deviation for a normal distribution). … In a PSA, a set of input parameter values is drawn by random sampling from each distribution, and the model is 'run' to generate outputs (cost and health outcome), which are stored. This is repeated many times (typically 1000 to 10 000), resulting in a distribution of outputs that can be graphed on the cost-effectiveness plane, and analysed.	www.yhec.co.uk/glossary
Discount rate	The interest rate used to determine the present value of future costs and benefits.	http://htaglossary.net/

Term	Definition	Source
Dominance	The superiority of an option that entails lower costs than another option and has benefits equal to or greater than the other option, or that entails costs equal to those of another option and has greater benefits than the other option.	http://htaglossary.net/
Dominated	The opposite of dominant (SEE Dominance), i.e. an option that entails higher costs than another option and has benefits equal to or lower than the other option, or that entails costs equal to those of another option and has lower benefits than the other option.	
Dose response relationship	The relationship between the quantity of treatment given and its effect on outcome. In meta-analysis, dose-response relationships can be investigated using meta-regression.	Cochrane glossary)
Effectiveness	The extent to which a specific intervention, when used under ordinary circumstances, does what it is intended to do. Clinical trials that assess effectiveness are sometimes called pragmatic or management trials.	Cochrane glossary
Efficacy	The extent to which an intervention produces a beneficial result under ideal conditions. Clinical trials that assess efficacy are sometimes called explanatory trials and are restricted to participants who fully co-operate.	Cochrane glossary
Efficiency frontier	A curve formed by the incremental cost-effectiveness or cost-utility ratios in a graphical representation of the non-dominated comparators.	http://htaglossary.net/
Extended dominance	In the comparison of mutually exclusive programmes, the situation where one option has a higher incremental cost-effectiveness ratio than a more effective alternative.	http://www.ncpe.ie/glossary/
Generic utility instrument	Generic HRQoL instrument associated with a reference set of utility values (see utilities). Generic measures cover dimensions that are considered important for HRQoL in general, while disease- or population specific measures particularly focus on dimensions that are affected by a specific disease or population.	EUnetHTA guideline – Health-related quality of life and utility measures
Gnu is Not Unix	GNU is an operating system and an extensive collection of computer software. GNU is composed wholly of free software, most of which is licensed under the GNU Project's own General Public License	Wikipedia

Term	Definition	Source
Health-related quality of life	The measures of the impact of an intervention on patients' health status, extending beyond the traditional measures of mortality and morbidity to include dimensions such as physiology, function, social life, cognition, emotions, sleep and rest, energy and vitality, health perception and general life satisfaction. Note: Some of these elements are also called health status, functional status or quality-of-life measures.	http://htaglossary.net/
Heterogeneity	In a systematic review, the variability of or differences in the selected studies. Note: A distinction is sometimes made between "statistical heterogeneity" (differences in the reported effects) and "methodological heterogeneity" (differences in study design with regard to the key characteristics of the subjects, interventions or outcome evaluation criteria). Statistical tests of heterogeneity are used to determine whether the observed variability in study results effect size is greater than the variability that can be expected to occur by chance. However, these tests have low statistical power.	http://htaglossary.net/
Incremental cost-effectiveness ratio	The additional cost of the more expensive intervention compared with the less expensive intervention, divided by the difference between the effects of the interventions on the patients (the additional cost per QALY, for example).	http://htaglossary.net/
ICER threshold value	Benchmark for ICERs (incremental cost-effectiveness ratios) to assess an intervention's cost-effectiveness. Interventions with an ICER below the ICER threshold value are considered cost-effective, interventions with an ICER above the ICER threshold value are not cost-effective.	Bilcke et al. (2011)[183]
Intermediate (clinical) endpoint	A non-ultimate endpoint (e.g., not mortality or morbidity) that may be associated with disease status or progression toward an ultimate endpoint such as mortality or morbidity. They may be certain biomarkers (e.g., HbA1c in prediabetes or diabetes, bone density in osteoporosis, tumor progression in cancer) or disease symptoms (e.g., angina frequency in heart disease, measures of lung function in chronic obstructive pulmonary disease). (See also biomarker and surrogate endpoint)	HTA 101 glossary
Licenced indications	SEE Licensing	
Licensing	A marketing authorisation for medicines which meet standards of safety, quality and efficacy.	http://htaglossary.net/

Term	Definition	Source
Mapping	A set of methods where one outcome measure (e.g. health-related quality of life weights) are statistically predicted from one or more other measures. (e.g. linking disease-specific data to a generic HRQoL measure in order to assign utility values generated with the generic instrument to the disease-specific health state descriptions.	http://www.ncpe.ie/glossary/ and EUnetHTA guideline – Health-related quality of life and utility measures
Net Monetary Benefit	Net monetary benefit (NMB) is a summary statistic that represents the value of an intervention in monetary terms when a willingness-to-pay threshold for a unit of benefit (for example a measure of health outcome or QALY) is known. The use of NMB scales both health outcomes and use of resources to costs, with the result that comparisons without the use of ratios (such as in ICER). NMB is calculated as (incremental benefit x threshold) – incremental cost. Incremental NMB measures the difference in NMB between alternative interventions, a positive incremental NMB indicating that the intervention is cost-effective compared with the alterative at the given willingness-to-pay threshold. In this case the cost to derive the benefit is less than the maximum amount that the decision-maker would be willing to pay for this benefit.	http://www.yhec.co.uk/
Off-label use	Off-label use is the use of a medicinal product for another indication, another patient group, another dose, dose interval or by another route of administration than indicated in the package insert	Strauss, 1998[275]
Open source models	A decentralized software-development model that encourages open collaboration	Wikipedia
Post-hoc analysis	Statistical analyses that were not specified before the data was seen. … Post hoc analysis that is conducted and interpreted without adequate consideration of the multiple testing problem is sometimes called data dredging by critics because the statistical associations that it finds are often spurious.	Wikipedia
Preference-based instrument	SEE Generic utility instrument	
Price and quantities tables	A table detailing the price (p) and quantities (q) of cost items used within an economic evaluation.	SBU
Probabilistic sensitivity analysis	Probabilistic sensitivity analysis (PSA) is a technique used in economic modelling that allows the modeller to quantify the level of confidence in the output of the analysis, in relation to uncertainty in the model inputs. There is usually uncertainty associated with input parameter values of an economic model, which may have been derived from clinical trials, observational studies or in some cases expert opinion. … In the probabilistic analysis, these	http://www.yhec.co.uk/glossary

Term	Definition	Source
	parameters are represented as distributions around the point estimate.	
Protocol-driven costs	the resource use /…/ 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	Drummond et al, 2015, page 273 [7]
Publication bias	A bias due to studies being published based on the nature and direction of their results. Example: A study with statistically significant results favouring the intervention of interest may have a greater likelihood of being published. http://htaglossary.net/publication+bias	HTA Glossary.net
Quality of Life	SEE health-related quality of life	
Reference case analysis	In order to enhance consistency in economic evaluations, guidelines might require to perform a 'reference case', including the essential elements for each economic evaluation together with the most appropriate methodology. Additional analyses are allowed, but should be distinguished from the results of the reference case analysis. Variations to the reference case should be justified and well-argued.	KCE report 183[89]
Relative efficacy/effectiveness assessment	Assessment of the efficacy/effectiveness compared with alternative treatment(s).	Kleijnen et al. Value in Health, 2012[276]
Relative risk reduction	A type of measure of treatment effect that compares the probability of a type of outcome in the treatment group with that of a control group, i.e.: (Pc- Pt) ¸ Pc. For instance, if the results of a trial show that the probability of death in a control group was 25% and the probability of death in a control group was 10%, the relative risk reduction would be: (0.25 - 0.10) ¸ 0.25 = 0.6. (See also absolute risk reduction, number needed to treat, and odds ratio.)	HTA 101 glossary
Relative treatment effect	SEE Relative risk reduction	
Sensitivity (or scenario) analysis	Sensitivity analysis is used to illustrate and assess the level of confidence that may be associated with the conclusion of an economic evaluation. It is performed by varying key assumptions made in the evaluation (individually or severally) and recording the impact on the result (output) of the evaluation.	www.yhec.co.uk/glossary
Societal Willingness to Pay	SEE Willingness to pay	

Term	Definition	Source
Surrogate endpoint	An indicator that, while not being of direct interest for the patient, may reflect important outcomes. Note: For example, blood pressure is not of direct clinical interest to the patient, but is often used as an evaluation criterion in clinical trials because it is a risk factor for stroke and heart attacks. An intermediate outcome is often a physiological or biochemical marker that can be quickly and easily measured, and that is considered to have great predictive value. It is often used when observation of clinical outcomes requires long follow-up. Note: Intermediate outcome is not a synonym for surrogate endpoint. However, an intermediate outcome can become a surrogate endpoint if it is easier to measure than a clinical criterion or if there is a statistical relationship between the occurrence of the clinical outcome indicator and the occurrence of the surrogate endpoint, or if there is a relationship allowing for prediction of the effect of the factor studied on the clinical indicator, on the basis of the observed effect on the surrogate endpoint.	HTA Glossary.net
Target population	The target population of a study is the broad group of people that researchers are examining.	www.focr.org/target-population
Time-trade-off	A method for determining preference between two health states for different lengths of time, to estimate how many years of life a person is prepared to sacrifice to improve his/her health status. Note: For chronic states, the options are the reference health state for time t followed by death, or perfect health for a shorter time followed by death. For temporary states, the options are the reference health state for time t followed by an explicitly specified outcome (usually health), or a worse health state for a shorter time followed by the same outcome.	HTA Glossary.net
Utility	In health economics, a 'utility' is the measure of the preference or value that an individual or society gives a particular health state. It is generally a number between 0 (representing death) and 1 (perfect health). The most widely used measure of benefit in cost-utility analysis is the quality-adjusted life year, which combines quality of life with length of life. Other measures include disability-adjusted life years (DALYs) and healthy year equivalents (HYEs).	www.nice.org.uk/Glossary/

Term	Definition	Source
Willingness to Pay	The maximum amount that a person is willing to pay: (a) to achieve a good health state or particular outcome, or to increase its probability of occurrence; or (b) to avoid a bad health state or outcome, or to decrease its probability of occurrence.	HTA Glossary.net

5.5 Annex 5 – Bibliography

[1] Swedish Council on Health Technology Assessment (SBU, Sweden), French National Authority for Health (HAS, France), Institute for Economic Research (IER, Slovenia), Institute for Quality and Efficiency in Health Care (IQWiG, Germany). Methods for health economic evaluations - A guideline based on current practices in Europe. Methodological Guideline: EUnetHTA; 2015 May.

[2] Husereau D, Drummond M, Petrou S, Carswell C, Moher D, Greenberg D, et al. Consolidated Health Economic Evaluation Reporting Standards (CHEERS) statement. BMJ. 2013 Mar 25;346:f1049.

[3] Drummond MF, Jefferson TO. Guidelines for authors and peer reviewers of economic submissions to the BMJ. The BMJ Economic Evaluation Working Party. BMJ. 1996 Aug 03;313(7052):275-83.

[4] Evers S, Goossens M, de Vet H, van Tulder M, Ament A. Criteria list for assessment of methodological quality of economic evaluations: Consensus on Health Economic Criteria. Int J Technol Assess Health Care. 2005 Spring;21(2):240-5.

[5] Gomersall JS, Jadotte YT, Xue Y, Lockwood S, Riddle D, Preda A. Conducting systematic reviews of economic evaluations. Int J Evid Based Healthc. 2015 Sep;13(3):170-8.

[6] Philips Z, Ginnelly L, Sculpher M, Claxton K, Golder S, Riemsma R, et al. Review of guidelines for good practice in decision-analytic modelling in health technology assessment. Health Technol Assess. 2004 Sep;8(36):iii-iv, ix-xi, 1-158.

[7] Drummond M, Sculpher M, Claxton K, Stoddart G, Torrance G. Methods for the economic evaluation of health care programmes. 4th ed: Oxford University Press 2015.

[8] Briggs A, Claxton K, Sculpher M. Decision Modelling for Health Economic Evaluation: Oxford University Press August 2006.

[9] NIHR Evaluation, Trials and Studies Coordinating Centre (NETSCC, UK). HTA Adaptation Toolkit & Glossary: EUnetHTA; 2011 Oct.

[10] Belgian Health Care Knowledge Centre (KCE, Belgium), Haute Autorité de Santé (HAS, France), EunetHTA Joint Action Work Package 5, EUnetHTA Joint Action 2 Work Package 7. Endpoints used for Relative Effectiveness Assessment: health-related quality of life and utility measures. Methodological Guideline: EUnetHTA; 2015.

[11] Health Information and Quality Authority (HIQA), Haute Autorité de santé (HAS, France), EUnetHTA Joint Action Work Package 5, EunetHTA Joint Action 2 Work Package 7. Comparators & comparisons: direct and indirect comparisons. Methodological Guideline: EUnetHTA; 2015.

[12] National Institute for Health and Care Excellence (NICE, UK), Haute Autorité de Santé (HAS, France), EunetHTA Joint Action Work Package 5, EUnetHTA Joint Action 2 Work Package 7. Comparators & comparisons: Criteria for the choice of the most appropriate comparator(s). Methodological Guideline: EUnetHTA; 2015.

[13] Health Information and Quality Authority (HIQA), EUnetHTA Joint Action Work Package 5, EUnetHTA Joint Action 2 Work Package 7. Endpoints used for Relative Effectiveness Assessment: Clinical Endpoints. Methodological Guideline: EUnetHTA; 2015.
[14] National Health Care Institute (ZIN, Netherlands), EUnetHTA Joint Action Work Package 5, EUnetHTA Joint Action 2 Work Package 7. Levels of evidence: Applicability of evidence for the context of a relative effectiveness assessment. Methodological Guideline: EUnetHTA; 2015.
[15] Norwegian Knowledge Centre for the Health Services (NOKC, Norway), Haute Autorité de Santé (HAS, France), EUnetHTA Joint Action Work Package 5, EUnetHTA joint Action 2 Work Package 7. Endpoints used in Relative Effectiveness Assessment: Surrogate Endpoints. Methodological Guideline; 2015.
[16] Shemilt I, Mugford M, Luke V, Marsh K, Donaldson C, eds. Evidence-Based Decisions and Economics: Health Care, Social Welfare, Education and Criminal Justice. 2nd ed: Blackwell 2010.
[17] Wijnen B, Van Mastrigt G, Redekop WK, Majoie H, De Kinderen R, Evers S. How to prepare a systematic review of economic evaluations for informing evidence-based healthcare decisions: data extraction, risk of bias, and transferability (part 3/3). Expert Rev Pharmacoecon Outcomes Res. 2016 Dec;16(6):723-32.
[18] Swedish Agency for Health Technology Assessment and Assesment of Social Services (SBU). Checklist for Assessing the Quality of Trial-Based Health Economic Studies. *Assessment of methods in health care - A handbook*. Version 2017:1 ed: SBU 2018:B7:1-4.
[19] Swedish Agency for Health Technology Assessment and Assesment of Social Services (SBU). Checklist for Assessing the Quality of Health Economic Modelling Studies. *Assessment of methods in health care - A handbook*. Version 2017:1 ed: SBU 2018:B8:1-4.
[20] AOTMiT. Health Technology Assessment Guidelines. Warsaw: AOTMiT; 2016.
[21] Philips Z, Bojke L, Sculpher M, Claxton K, Golder S. Good practice guidelines for decision-analytic modelling in health technology assessment: a review and consolidation of quality assessment. Pharmacoeconomics. 2006;24(4):355-71.
[22] Institute for Quality and Efficiency in Health Care (IQWiG). Process of information retrieval for systematic reviews and health technology assessments on clinical effectiveness Methodological Guideline; 2017.
[23] CAST/SDU, Institute for Quality and Efficiency in Health Care (IQWiG). Levels of evidence - Internal validity of randomised controlled trials. Methodological Guideline; 2015.
[24] Smith GC, Pell JP. Parachute use to prevent death and major trauma related to gravitational challenge: systematic review of randomised controlled trials. BMJ. 2003 Dec 20;327(7429):1459-61.
[25] Hayes MJ, Kaestner V, Mailankody S, Prasad V. Most medical practices are not parachutes: a citation analysis of practices felt by biomedical authors to be analogous to parachutes. CMAJ Open. 2018 Jan 15;6(1):E31-E8.
[26] Institute for Quality and Efficiency in Health Care (IQWiG), Norwegian Knowledge Centre for the Health Services (NOKC), Swiss Network for Health Technology Assessment (SNHTA). Internal validity of non-randomised studies (NRS) on interventions Methodological Guideline; 2015.
[27] Phillippo DM, Ades AE, Dias S, Palmer S, Abrams KR, Welton NJ. Methods for Population-Adjusted Indirect Comparisons in Health Technology Appraisal. Med Decis Making. 2018 Feb;38(2):200-11.
[28] Wieseler B, McGauran N, Kaiser T. Finding studies on reboxetine: a tale of hide and seek. BMJ. 2010 Oct 12;341:c4942.

[29] Institute for Quality and Efficiency in Health Care (IQWiG). Bupropion, mirtazapine, and reboxetine in the treatment of depression: Executive summary of final report A05-20C. Cologne: IQWiG (Institute for Quality and Efficiency in Health Care),; 2011. Report No.: A05-20C.
[30] Jacobs A. Zombie statistics on half of all clinical trials unpublished. [Blog] 2015 2015-08-29 [cited 2018-09-27]; Available from: http://www.statsguy.co.uk/zombie-statistics-on-half-of-all-clinical-trials-unpublished/
[31] Glasziou P, Chalmers I. Can it really be true that 50% of research is unpublished? BMJ Opinion 2017 [cited 2018-08-03]; Available from: https://blogs.bmj.com/bmj/2017/06/05/paul-glasziou-and-iain-chalmers-can-it-really-be-true-that-50-of-research-is-unpublished/
[32] Rump F. Why 85% of health research is wasted with Paul Glasziou. Bio2040 - Future of Bio and Drug Discovery 2018 2018-01-12 [cited 2018-08-03]; Available from: https://bio2040.com/2018/01/12/why-85-of-health-research-is-wasted-with-paul-glasziou/
[33] The European Parliament and the Council of the European Union. REGULATION (EU) No 536/2014 OF THE EUROPEAN PARLIAMENT AND OF THE COUNCIL of 16 April 2014 on clinical trials on medicinal products for human use, and repealing Directive 2001/20/EC In: Official Journal of the European Union ed. *27052014* 2014:L158/1 - L/76.
[34] ClinicalTrials.gov. When is required clinical trial results information due? . Frequently Asked Questions 2018 2018-09 [cited 2018-09-27]; Available from: https://clinicaltrials.gov/ct2/manage-recs/faq#fr_7
[35] Prayle AP, Hurley MN, Smyth AR. Compliance with mandatory reporting of clinical trial results on ClinicalTrials.gov: cross sectional study. BMJ. 2012 Jan 3;344:d7373.
[36] Wadman M. FDA says study overestimated non-compliance with data-reporting laws. Nature News 2012 2012-05-01 [cited 2018-08-03]; Available from: https://www.nature.com/news/fda-says-study-overestimated-non-compliance-with-data-reporting-laws-1.10549
[37] Anderson ML, Chiswell K, Peterson ED, Tasneem A, Topping J, Califf RM. Compliance with results reporting at ClinicalTrials.gov. N Engl J Med. 2015 Mar 12;372(11):1031-9.
[38] Jones CW, Handler L, Crowell KE, Keil LG, Weaver MA, Platts-Mills TF. Non-publication of large randomized clinical trials: cross sectional analysis. BMJ. 2013;347:f6104
[39] Goldacre B, DeVito NJ, Heneghan C, Irving F, Bacon S, Fleminger J, et al. Compliance with requirement to report results on the EU Clinical Trials Register: cohort study and web resource. BMJ. 2018 Sep 12;362:k3218.
[40] Hwang TJ, Carpenter D, Lauffenburger JC, Wang B, Franklin JM, Kesselheim AS. Failure of Investigational Drugs in Late-Stage Clinical Development and Publication of Trial Results. JAMA Intern Med. 2016 Dec 1;176(12):1826-33.
[41] Chen R, Desai NR, Ross JS, Zhang W, Chau KH, Wayda B, et al. Publication and reporting of clinical trial results: cross sectional analysis across academic medical centers. BMJ. 2016 Feb 17;352:i637.
[42] Miller JE, Korn D, Ross JS. Clinical trial registration, reporting, publication and FDAAA compliance: a cross-sectional analysis and ranking of new drugs approved by the FDA in 2012. BMJ Open. 2015 Nov 12;5(11):e009758.
[43] Manzoli L, Flacco ME, D'Addario M, Capasso L, De Vito C, Marzuillo C, et al. Non-publication and delayed publication of randomized trials on vaccines: survey. BMJ. 2014 May 16;348:g3058.
[44] Ross JS, Tse T, Zarin DA, Xu H, Zhou L, Krumholz HM. Publication of NIH funded trials registered in ClinicalTrials.gov: cross sectional analysis. BMJ. 2012 Jan 3;344:d7292.

[45] Chang L, Dhruva SS, Chu J, Bero LA, Redberg RF. Selective reporting in trials of high risk cardiovascular devices: cross sectional comparison between premarket approval summaries and published reports. BMJ. 2015 Jun 10;350:h2613.
[46] Chalmers I, Glasziou P, Godlee F. All trials must be registered and the results published. BMJ. 2013 Jan 9;346:f105.
[47] Shah M, Avgil Tsadok M, Jackevicius CA, Essebag V, Behlouli H, Pilote L. Relation of digoxin use in atrial fibrillation and the risk of all-cause mortality in patients >/=65 years of age with versus without heart failure. Am J Cardiol. 2014 Aug 1;114(3):401-6.
[48] Turakhia MP, Santangeli P, Winkelmayer WC, Xu X, Ullal AJ, Than CT, et al. Increased mortality associated with digoxin in contemporary patients with atrial fibrillation: findings from the TREAT-AF study. J Am Coll Cardiol. 2014 Aug 19;64(7):660-8.
[49] Chan KE, Lazarus JM, Hakim RM. Digoxin associates with mortality in ESRD. J Am Soc Nephrol. 2010 Sep;21(9):1550-9.
[50] Ziff OJ, Lane DA, Samra M, Griffith M, Kirchhof P, Lip GY, et al. Safety and efficacy of digoxin: systematic review and meta-analysis of observational and controlled trial data. BMJ. 2015 Aug 30;351:h4451.
[51] Higgins J, Green S. Cochrane Handbook for Systematic Reviews of Intervention. Version 5.1.0 ed: The Cochrane Collaboration 2011.
[52] Kim SY, Park JE, Lee YJ, Seo HJ, Sheen SS, Hahn S, et al. Testing a tool for assessing the risk of bias for nonrandomized studies showed moderate reliability and promising validity. J Clin Epidemiol. 2013 Apr;66(4):408-14.
[53] Cole GD, Francis DP. Trials are best, ignore the rest: safety and efficacy of digoxin. BMJ. 2015 Aug 30;351:h4662.
[54] Grady D, Rubin SM, Petitti DB, Fox CS, Black D, Ettinger B, et al. Hormone therapy to prevent disease and prolong life in postmenopausal women. Ann Intern Med. 1992 Dec 15;117(12):1016-37.
[55] Stampfer MJ, Colditz GA. Estrogen replacement therapy and coronary heart disease: a quantitative assessment of the epidemiologic evidence. Prev Med. 1991 Jan;20(1):47-63.
[56] Pines A, Mijatovic V, van der Mooren MJ, Kenemans P. Hormone replacement therapy and cardioprotection: basic concepts and clinical considerations. Eur J Obstet Gynecol Reprod Biol. 1997 Feb;71(2):193-7.
[57] Fletcher SW, Colditz GA. Failure of estrogen plus progestin therapy for prevention. JAMA. 2002 Jul 17;288(3):366-8.
[58] Rossouw JE, Anderson GL, Prentice RL, LaCroix AZ, Kooperberg C, Stefanick ML, et al. Risks and benefits of estrogen plus progestin in healthy postmenopausal women: principal results From the Women's Health Initiative randomized controlled trial. JAMA. 2002 Jul 17;288(3):321-33.
[59] Hartz A, He T, Wallace R, Powers J. Comparing hormone therapy effects in two RCTs and two large observational studies that used similar methods for comprehensive data collection and outcome assessment. BMJ Open. 2013;3(7).
[60] Dahabreh IJ, Kent DM. Can the learning health care system be educated with observational data? JAMA. 2014 Jul;312(2):129-30.
[61] Dahabreh IJ, Sheldrick RC, Paulus JK, Chung M, Varvarigou V, Jafri H, et al. Do observational studies using propensity score methods agree with randomized trials? A systematic comparison of studies on acute coronary syndromes. Eur Heart J. 2012 Aug;33(15):1893-901.
[62] Lonjon G, Boutron I, Trinquart L, Ahmad N, Aim F, Nizard R, et al. Comparison of treatment effect estimates from prospective nonrandomized studies with propensity score analysis and randomized controlled trials of surgical procedures. Ann Surg. 2014 Jan;259(1):18-25.

[63] Tannen RL, Weiner MG, Xie D. Use of primary care electronic medical record database in drug efficacy research on cardiovascular outcomes: comparison of database and randomised controlled trial findings. BMJ. 2009 Jan 27;338:b81.
[64] Heres S, Davis J, Maino K, Jetzinger E, Kissling W, Leucht S. Why olanzapine beats risperidone, risperidone beats quetiapine, and quetiapine beats olanzapine: an exploratory analysis of head-to-head comparison studies of second-generation antipsychotics. Am J Psychiatry. 2006 Feb;163(2):185-94.
[65] Chouinard G, Jones B, Remington G, Bloom D, Addington D, MacEwan GW, et al. A Canadian multicenter placebo-controlled study of fixed doses of risperidone and haloperidol in the treatment of chronic schizophrenic patients. J Clin Psychopharmacol. 1993 Feb;13(1):25-40.
[66] Peuskens J. Risperidone in the treatment of patients with chronic schizophrenia: a multi-national, multi-centre, double-blind, parallel-group study versus haloperidol. Risperidone Study Group. Br J Psychiatry. 1995 Jun;166(6):712-26; discussion 27-33.
[67] Kane J, Honigfeld G, Singer J, Meltzer H. Clozapine for the treatment-resistant schizophrenic. A double-blind comparison with chlorpromazine. Arch Gen Psychiatry. 1988 Sep;45(9):789-96.
[68] Rosenheck R, Cramer J, Xu W, Thomas J, Henderson W, Frisman L, et al. A comparison of clozapine and haloperidol in hospitalized patients with refractory schizophrenia. Department of Veterans Affairs Cooperative Study Group on Clozapine in Refractory Schizophrenia. N Engl J Med. 1997 Sep 18;337(12):809-15.
[69] Simpson GM, Josiassen RC, Stanilla JK, de Leon J, Nair C, Abraham G, et al. Double-blind study of clozapine dose response in chronic schizophrenia. Am J Psychiatry. 1999 Nov;156(11):1744-50.
[70] Azorin JM, Spiegel R, Remington G, Vanelle JM, Pere JJ, Giguere M, et al. A double-blind comparative study of clozapine and risperidone in the management of severe chronic schizophrenia. Am J Psychiatry. 2001 Aug;158(8):1305-13.
[71] Tran PV, Hamilton SH, Kuntz AJ, Potvin JH, Andersen SW, Beasley C, Jr., et al. Double-blind comparison of olanzapine versus risperidone in the treatment of schizophrenia and other psychotic disorders. J Clin Psychopharmacol. 1997 Oct;17(5):407-18.
[72] HealthNewsReview.org. Reporting the findings: Absolute vs relative risk. 2015 [cited 20180928]; Available from: https://www.healthnewsreview.org/toolkit/tips-for-understanding-studies/absolute-vs-relative-risk/
[73] Antithrombotic Trialists C, Baigent C, Blackwell L, Collins R, Emberson J, Godwin J, et al. Aspirin in the primary and secondary prevention of vascular disease: collaborative meta-analysis of individual participant data from randomised trials. Lancet. 2009 May 30;373(9678):1849-60.
[74] Schroder FH, Hugosson J, Roobol MJ, Tammela TL, Ciatto S, Nelen V, et al. Screening and prostate-cancer mortality in a randomized European study. N Engl J Med. 2009 Mar 26;360(13):1320-8.
[75] McCartney M. The press release, relative risks, and the polypill. BMJ. 2011 Jul 27;343:d4720.
[76] Ioannidis JP, Evans SJ, Gotzsche PC, O'Neill RT, Altman DG, Schulz K, et al. Better reporting of harms in randomized trials: an extension of the CONSORT statement. Ann Intern Med. 2004 Nov 16;141(10):781-8.
[77] Chou R, Aronson N, Atkins D, Ismaila AS, Santaguida P, Smith DH, et al. AHRQ series paper 4: assessing harms when comparing medical interventions: AHRQ and the effective health-care program. J Clin Epidemiol. 2010 May;63(5):502-12.
[78] Agenzia Italiana del Farmaco (AIFA, Italy),, Haute Autorité de Santé (HAS F, Institute for Quality and Efficiency in Health Care (IQWiG), Belgian Health Care Knowledge

Centre (KCE B, EunetHTA Joint Action Work Package 5, EUnetHTA Joint Action 2 Work Package 7. Endpoints used in Relative Effectiveness Assessment: Safety. Methodological Guideline: EUnetHTA; 2015.
[79] Centre for Reviews and Dissemination (CRD). CRD's guidance for undertaking reviews in health care. York: CRD, University of York 2009.
[80] Craig D, McDaid C, Fonseca T, Stock C, Duffy S, Woolacott N. Are adverse effects incorporated in economic models? An initial review of current practice. Health Technol Assess. 2009 Dec;13(62):1-71, 97-181, iii.
[81] Caro JJ, Briggs AH, Siebert U, Kuntz KM, Force I-SMGRPT. Modeling good research practices--overview: a report of the ISPOR-SMDM Modeling Good Research Practices Task Force-1. Med Decis Making. 2012 Sep-Oct;32(5):667-77.
[82] Weinstein MC, O'Brien B, Hornberger J, Jackson J, Johannesson M, McCabe C, et al. Principles of good practice for decision analytic modeling in health-care evaluation: report of the ISPOR Task Force on Good Research Practices--Modeling Studies. Value Health. 2003 Jan-Feb;6(1):9-17.
[83] CADTH. Guidelines for the Economic Evaluation of Health Technologies: Canada. 4th Edition; March 2017.
[84] Ara R, Wailoo A. NICE DSU technical support document 12: the use of health state utility values in decision models [Internet]. Sheffield (United Kingdom): University of Sheffield; 2011.
[85] Heather EM, Payne K, Harrison M, Symmons DP. Including adverse drug events in economic evaluations of anti-tumour necrosis factor-alpha drugs for adult rheumatoid arthritis: a systematic review of economic decision analytic models. Pharmacoeconomics. 2014 Feb;32(2):109-34.
[86] National Institute for Health and Care Excellence (NICE). Guide to the methods of technology appraisal; 2013.
[87] Gabriel S, Drummond M, Maetzel A, Boers M, Coyle D, Welch V, et al. OMERACT 6 Economics Working Group report: a proposal for a reference case for economic evaluation in rheumatoid arthritis. J Rheumatol. 2003 Apr;30(4):886-90.
[88] Health Information and Quality Authority (HIQA). Guidelines for the Economic Evaluation of Health Technologies in Ireland: Health Information and Quality Authority (HIQA); 2014.
[89] Cleemput I, Neyt M, Van de Sande S, Thiry N. Belgian guidelines for economic evaluations and budget impact analyses. Health Technology Assessment (HTA). Brussels: Belgian Health Care Knowledge Centre(KCE); 2012.
[90] Goeree R, O'Brien B, Hunt R, Blackhouse G, Willan A, Watson J. Economic evaluation of long-term management strategies for erosive oesophagitis. Pharmacoeconomics. 1999 Dec;16(6):679-97.
[91] Neyt M, Van Brabandt H. The importance of the comparator in economic evaluations: working on the efficiency frontier. Pharmacoeconomics. 2011 Nov;29(11):913-6.
[92] Van Brabandt H, Camberlin C, Neyt M, De Laet C, Stroobandt S, Devriese S, et al. Cardiac resynchronisation therapy. A Health technology Assessment. Brussels: Belgian Health Care Knowledge Centre (KCE); 2010.
[93] Bristow MR, Saxon LA, Boehmer J, Krueger S, Kass DA, De Marco T, et al. Cardiac-resynchronization therapy with or without an implantable defibrillator in advanced chronic heart failure. N Engl J Med. 2004 May 20;350(21):2140-50.
[94] Dretzke J, Edlin R, Round J, Connock M, Hulme C, Czeczot J, et al. A systematic review and economic evaluation of the use of tumour necrosis factor-alpha (TNF-alpha) inhibitors, adalimumab and infliximab, for Crohn's disease. Health Technol Assess. 2011 Feb;15(6):1-244.

[95] Neyt M, Thiry N, Ramaekers D, Van Brabandt H. Cost effectiveness of implantable cardioverter-defibrillators for primary prevention in a Belgian context. Appl Health Econ Health Policy. 2008;6(1):67-80.
[96] Van Brabandt H, Thiry N, Neyt M, Van den Oever R, Galloo P, Vanoverloop J, et al. The Implantable Cardioverter Defibrillator: a Health Technology Assessment. Brussels: Belgian Health Care Knowledge Centre (KCE); 2007.
[97] Briggs AH. Handling uncertainty in cost-effectiveness models. Pharmacoeconomics. 2000 May;17(5):479-500.
[98] Eddy DM. Screening for cervical cancer. Ann Intern Med. 1990 Aug 01;113(3):214-26.
[99] Hind D, Pilgrim H, Ward S. Questions about adjuvant trastuzumab still remain. Lancet. 2007 Jan 06;369(9555):3-5.
[100] Huybrechts M, Hulstaert F, Neyt M, Vrijens F, Ramaekers D. Trastuzumab in Early Stage Breast Cancer. Brussels: Belgian Health Care Knowledge Centre (KCE); 2006.
[101] Vannieuwenhuysen C, Slegers P, Neyt M, Hulstaert F, Stordeur S, Cleemput I, et al. Towards a better managed off-label use of drugs. Brussels: Belgian Health Care Knowledge Centre (KCE); 2015.
[102] Espinoza MA, Manca A, Claxton K, Sculpher MJ. The value of heterogeneity for cost-effectiveness subgroup analysis: conceptual framework and application. Med Decis Making. 2014 Nov;34(8):951-64.
[103] Sculpher M. Subgroups and heterogeneity in cost-effectiveness analysis. Pharmacoeconomics. 2008;26(9):799-806.
[104] Randomised trial of intravenous streptokinase, oral aspirin, both, or neither among 17,187 cases of suspected acute myocardial infarction: ISIS-2. ISIS-2 (Second International Study of Infarct Survival) Collaborative Group. Lancet. 1988 Aug 13;2(8607):349-60.
[105] Fayers PM, King MT. How to guarantee finding a statistically significant difference: the use and abuse of subgroup analyses. Qual Life Res. 2009 Jun;18(5):527-30.
[106] Smeeth L, Haines A, Ebrahim S. Numbers needed to treat derived from meta-analyses--sometimes informative, usually misleading. BMJ. 1999 Jun 5;318(7197):1548-51.
[107] Sackett D, Straus S, Richardson S, Rosenberg W, Haynes R. Evidence-based medicine: how to practice and teach. London; 2000.
[108] Drummond M, Sculpher M, Torrance G, O'Brien B, Stoddart G. Methods for the economic evaluation of health care programmes 2005.
[109] Neyt M, Van Brabandt H, Devriese S, De Laet C. Cost-effectiveness analyses of drug eluting stents versus bare metal stents: a systematic review of the literature. Health Policy. 2009 Jul;91(2):107-20.
[110] Neyt M, Van Brabandt H, Devriese S, Mahieu J, De Ridder A, De Graeve D, et al. Drug Eluting Stents in Belgium: Health Technology Assessment. Brussels: Belgian Health Care Knowledge Centre (KCE); 2007.
[111] van Hout BA, Serruys PW, Lemos PA, van den Brand MJ, van Es GA, Lindeboom WK, et al. One year cost effectiveness of sirolimus eluting stents compared with bare metal stents in the treatment of single native de novo coronary lesions: an analysis from the RAVEL trial. Heart. 2005 Apr;91(4):507-12.
[112] Morice MC, Serruys PW, Sousa JE, Fajadet J, Ban Hayashi E, Perin M, et al. A randomized comparison of a sirolimus-eluting stent with a standard stent for coronary revascularization. N Engl J Med. 2002 Jun 6;346(23):1773-80.
[113] Spencer FA, Iorio A, You J, Murad MH, Schunemann HJ, Vandvik PO, et al. Uncertainties in baseline risk estimates and confidence in treatment effects. BMJ. 2012 Nov 14;345:e7401.

[114] Neyt M, Cleemput I, Thiry N, De Laet C. Calculating an intervention's (cost-)effectiveness for the real-world target population: the potential of combining strengths of both RCTs and observational data. Health Policy. 2012 Jul;106(2):207-10.
[115] Neyt M, De Laet C, De Ridder A, Van Brabandt H. Cost effectiveness of drug-eluting stents in Belgian practice: healthcare payer perspective. Pharmacoeconomics. 2009;27(4):313-27.
[116] Aronson JK. Compliance, concordance, adherence. Br J Clin Pharmacol. 2007 Apr;63(4):383-4.
[117] Cramer JA, Roy A, Burrell A, Fairchild CJ, Fuldeore MJ, Ollendorf DA, et al. Medication compliance and persistence: terminology and definitions. Value Health. 2008 Jan-Feb;11(1):44-7.
[118] King MA, Pryce RL. Evidence for compliance with long-term medication: a systematic review of randomised controlled trials. Int J Clin Pharm. 2014 Feb;36(1):128-35.
[119] Zhang Z, Peluso MJ, Gross CP, Viscoli CM, Kernan WN. Adherence reporting in randomized controlled trials. Clin Trials. 2014 Apr;11(2):195-204.
[120] Bosworth H. Causes of Medication Nonadherence. *Enhancing Medication Adherence*. Tarporley: Springer Healthcare 2012.
[121] Health Information and Quality Authority (HIQA). Health technology assessment (HTA) of extending the national immunisation schedule to include HPV vaccination of boys: Draft report for public consultation. Dublin: HIQA; 2018.
[122] Bresse X, Goergen C, Prager B, Joura E. Universal vaccination with the quadrivalent HPV vaccine in Austria: impact on virus circulation, public health and cost-effectiveness analysis. Expert Rev Pharmacoecon Outcomes Res. 2014 Apr;14(2):269-81.
[123] Haeussler K, Marcellusi A, Mennini FS, Favato G, Picardo M, Garganese G, et al. Cost-Effectiveness Analysis of Universal Human Papillomavirus Vaccination Using a Dynamic Bayesian Methodology: The BEST II Study. Value Health. 2015 Dec;18(8):956-68.
[124] Largeron N, Petry KU, Jacob J, Bianic F, Anger D, Uhart M. An estimate of the public health impact and cost-effectiveness of universal vaccination with a 9-valent HPV vaccine in Germany. Expert Rev Pharmacoecon Outcomes Res. 2017 Feb;17(1):85-98.
[125] Mennini FS, Bonanni P, Bianic F, de Waure C, Baio G, Plazzotta G, et al. Cost-effectiveness analysis of the nine-valent HPV vaccine in Italy. Cost Eff Resour Alloc. 2017;15:11.
[126] Chen G, Iezzi A, McKie J, Khan MA, Richardson J. Diabetes and quality of life: Comparing results from utility instruments and Diabetes-39. Diabetes Res Clin Pract. 2015 Aug;109(2):326-33.
[127] Hatswell AJ, Pennington B, Pericleous L, Rowen D, Lebmeier M, Lee D. Patient-reported utilities in advanced or metastatic melanoma, including analysis of utilities by time to death. Health Qual Life Outcomes. 2014 Sep 10;12:140.
[128] Siderowf A, Ravina B, Glick HA. Preference-based quality-of-life in patients with Parkinson's disease. Neurology. 2002 Jul 09;59(1):103-8.
[129] Doble B, Lorgelly P. Mapping the EORTC QLQ-C30 onto the EQ-5D-3L: assessing the external validity of existing mapping algorithms. Qual Life Res. 2016 Apr;25(4):891-911.
[130] Hess LM, Brady WE, Havrilesky LJ, Cohn DE, Monk BJ, Wenzel L, et al. Comparison of methods to estimate health state utilities for ovarian cancer using quality of life data: a Gynecologic Oncology Group study. Gynecol Oncol. 2013 Feb;128(2):175-80.
[131] Longworth L, Rowen D. Mapping to obtain EQ-5D utility values for use in NICE health technology assessments. Value in Health. 2013 Jan-Feb;16(1):202-10.
[132] Wailoo AJ, Hernandez-Alava M, Manca A, Mejia A, Ray J, Crawford B, et al. Mapping to Estimate Health-State Utility from Non-Preference-Based Outcome Measures: An ISPOR Good Practices for Outcomes Research Task Force Report. Value Health. 2017 Jan;20(1):18-27.

[133] Lewis EF, Johnson PA, Johnson W, Collins C, Griffin L, Stevenson LW. Preferences for quality of life or survival expressed by patients with heart failure. J Heart Lung Transplant. 2001 Sep;20(9):1016-24.
[134] Yao G, Freemantle N, Calvert MJ, Bryan S, Daubert JC, Cleland JG. The long-term cost-effectiveness of cardiac resynchronization therapy with or without an implantable cardioverter-defibrillator. Eur Heart J. 2007 Jan;28(1):42-51.
[135] McAlister F, Ezekowitz J, Wiebe N, Rowe B, Spooner C, Crumley E, et al. Cardiac Resynchronization Therapy for Congestive Heart Failure. Rockville MD: Agency for Healthcare Research and Quality (Prepared by the University of Alberta Evidence-based Practice Center under Contract No. 290-02-0023.); November 2004.
[136] Kirsch J, McGuire A. Establishing health state valuations for disease specific states: an example from heart disease. Health Econ. 2000 Mar;9(2):149-58.
[137] Raphael C, Briscoe C, Davies J, Ian Whinnett Z, Manisty C, Sutton R, et al. Limitations of the New York Heart Association functional classification system and self-reported walking distances in chronic heart failure. Heart. 2007 Apr;93(4):476-82.
[138] Spertus J. Assessing patients' improvement in clinical trials. BMJ. 2008 Jun 7;336(7656):1258-9.
[139] Goldman L, Cook EF, Mitchell N, Flatley M, Sherman H, Cohn PF. Pitfalls in the serial assessment of cardiac functional status. How a reduction in "ordinary" activity may reduce the apparent degree of cardiac compromise and give a misleading impression of improvement. J Chronic Dis. 1982;35(10):763-71.
[140] Caro JJ, Guo S, Ward A, Chalil S, Malik F, Leyva F. Modelling the economic and health consequences of cardiac resynchronization therapy in the UK. Curr Med Res Opin. 2006 Jun;22(6):1171-9.
[141] Biomarkers Definitions Working G. Biomarkers and surrogate endpoints: preferred definitions and conceptual framework. Clin Pharmacol Ther. 2001 Mar;69(3):89-95.
[142] Davis C, Naci H, Gurpinar E, Poplavska E, Pinto A, Aggarwal A. Availability of evidence of benefits on overall survival and quality of life of cancer drugs approved by European Medicines Agency: retrospective cohort study of drug approvals 2009-13. BMJ. 2017 Oct 4;359:j4530.
[143] Fischer A, Hernandez-Villafuerte K, Latimer N, Henshall C. Extrapolation from Progression-Free Survival to Overall Survival in Oncology; December 2016.
[144] Kemp R, Prasad V. Surrogate endpoints in oncology: when are they acceptable for regulatory and clinical decisions, and are they currently overused? BMC Med. 2017 Jul 21;15(1):134.
[145] Prasad V, Kim C, Burotto M, Vandross A. The Strength of Association Between Surrogate End Points and Survival in Oncology: A Systematic Review of Trial-Level Meta-analyses. JAMA Intern Med. 2015 Aug;175(8):1389-98.
[146] Beauchemin C, Lapierre ME, Letarte N, Yelle L, Lachaine J. Use of Intermediate Endpoints in the Economic Evaluation of New Treatments for Advanced Cancer and Methods Adopted When Suitable Overall Survival Data are Not Available. Pharmacoeconomics. 2016 Sep;34(9):889-900.
[147] Davis S, Tappenden P, Cantrell A. A review of studies examining the relationship between progression-free survival and overall survival in advanced or metastatic cancer. Sheffield: Decision Support Unit, ScHARR, University of Sheffield; 2012.
[148] Taylor RS, Elston J. The use of surrogate outcomes in model-based cost-effectiveness analyses: a survey of UK Health Technology Assessment reports. Health Technol Assess. 2009 Jan;13(8):iii, ix-xi, 1-50.
[149] Matulonis UA, Oza AM, Ho TW, Ledermann JA. Intermediate clinical endpoints: a bridge between progression-free survival and overall survival in ovarian cancer trials. Cancer. 2015 Jun 1;121(11):1737-46.

[150] Miller K, Wang M, Gralow J, Dickler M, Cobleigh M, Perez EA, et al. Paclitaxel plus bevacizumab versus paclitaxel alone for metastatic breast cancer. N Engl J Med. 2007 Dec 27;357(26):2666-76.
[151] Miles DW, Chan A, Dirix LY, Cortes J, Pivot X, Tomczak P, et al. Phase III study of bevacizumab plus docetaxel compared with placebo plus docetaxel for the first-line treatment of human epidermal growth factor receptor 2-negative metastatic breast cancer. J Clin Oncol. 2010 Jul 10;28(20):3239-47.
[152] Robert NJ, Dieras V, Glaspy J, Brufsky AM, Bondarenko I, Lipatov ON, et al. RIBBON-1: randomized, double-blind, placebo-controlled, phase III trial of chemotherapy with or without bevacizumab for first-line treatment of human epidermal growth factor receptor 2-negative, locally recurrent or metastatic breast cancer. J Clin Oncol. 2011 Apr 1;29(10):1252-60.
[153] Pouwels X, Ramaekers BLT, Joore MA. Reviewing the quality, health benefit and value for money of chemotherapy and targeted therapy for metastatic breast cancer. Breast Cancer Res Treat. 2017 Oct;165(3):485-98.
[154] Dedes KJ, Matter-Walstra K, Schwenkglenks M, Pestalozzi BC, Fink D, Brauchli P, et al. Bevacizumab in combination with paclitaxel for HER-2 negative metastatic breast cancer: an economic evaluation. Eur J Cancer. 2009 May;45(8):1397-406.
[155] Montero AJ, Avancha K, Gluck S, Lopes G. A cost-benefit analysis of bevacizumab in combination with paclitaxel in the first-line treatment of patients with metastatic breast cancer. Breast Cancer Res Treat. 2012 Apr;132(2):747-51.
[156] Refaat T, Choi M, Gaber G, Kiel K, Mehta M, Gradishar W, et al. Markov model and cost-effectiveness analysis of bevacizumab in HER2-negative metastatic breast cancer. Am J Clin Oncol. 2014 Oct;37(5):480-5.
[157] Hwang TJ, Gyawali B. Association between progression-free survival and patients' quality of life in cancer clinical trials. Int J Cancer. 2019 Apr 1;144(7):1746-51.
[158] Neyt M, Devriese S, Camberlin C, Vlayen J. Bevacizumab in the treatment of ovarian cancer. Brussels: Belgian Health Care Knowledge Centre (KCE); 2017.
[159] Helgeson VS, Tomich PL. Surviving cancer: a comparison of 5-year disease-free breast cancer survivors with healthy women. Psychooncology. 2005 Apr;14(4):307-17.
[160] Herschbach P, Keller M, Knight L, Brandl T, Huber B, Henrich G, et al. Psychological problems of cancer patients: a cancer distress screening with a cancer-specific questionnaire. Br J Cancer. 2004 Aug 2;91(3):504-11.
[161] National Institute for Health and Care Excellence (NICE). Bevacizumab in combination with carboplatin and paclitaxel for the treatment of advanced ovarian cancer [TA284] - Specification for manufacturer/sponsor submission of evidence; August 2012.
[162] Fallowfield LJ, Fleissig A. The value of progression-free survival to patients with advanced-stage cancer. Nat Rev Clin Oncol. 2011 Oct 18;9(1):41-7.
[163] Cooper K, Pickett K, Frampton G, Copley V, Bryant J. Bevacizumab in combination with carboplatin and paclitaxel for the first-line treatment of ovarian cancer. A Single Technology Appraisal: SHTAC; 2012.
[164] Food and Drug Administration (FDA), Department of health and human services. Proposal to withdraw approval for the breast cancer indication for Avastin (Bevacizumab); November 18, 2011.
[165] Neyt M, Hulstaert F. FDA withdraws accelerated approval of bevacizumab for the treatment of metastatic breast cancer: Belgian Health Care Knowledge Centre (KCE); 2012.
[166] Smith B, Cohn DE, Clements A, Tierney BJ, Straughn JM, Jr. Is the progression free survival advantage of concurrent gemcitabine plus cisplatin and radiation followed by adjuvant gemcitabine and cisplatin in patients with advanced cervical cancer worth the additional cost? A cost-effectiveness analysis. Gynecol Oncol. 2013 Sep;130(3):416-20.

[167] PBAC. Guidelines for preparing submissions to the pharmaceutical benefits advisory committee: Pharmaceutical Benefits Advisory Committee; September 2016.
[168] CADTH. Lignes directrices de l'évaluation économique des technologies de la santé au Canada: Agence canadienne des médicaments et des technologies de santé; 2017.
[169] Sanders GD, Neumann PJ, Basu A, Brock DW, Feeny D, Krahn M, et al. Recommendations for Conduct, Methodological Practices, and Reporting of Cost-effectiveness Analyses: Second Panel on Cost-Effectiveness in Health and Medicine. JAMA. 2016 Sep 13;316(10):1093-103.
[170] Gerdtham UG, Zethraeus N. Predicting survival in cost-effectiveness analyses based on clinical trials. Int J Technol Assess Health Care. 2003 Summer;19(3):507-12.
[171] Xie X, Lambrinos A, Chan B, Dhalla IA, Krings T, Casaubon LK, et al. Mechanical thrombectomy in patients with acute ischemic stroke: a cost-utility analysis. CMAJ Open. 2016 Apr-Jun;4(2):E316-25.
[172] Latimer N. Survival analysis for economic evaluations alongside clinical trials - extrapolation with patient-level data: Decision Support Unit, ScHARR, University of Sheffield; 2013.
[173] Woods B, Sideris E, Palmer S, Latimer N, Soares M. Partitioned survival analysis for decision modelling in health care: a critical review: Decision Support Unit, ScHARR, University of Sheffield; 2017.
[174] Wittenberg E, James LP, Prosser LA. Spillover Effects on Caregivers' and Family Members' Utility: A Systematic Review of the Literature. Pharmacoeconomics. 2019 Apr;37(4):475-99.
[175] Lin PJ, D'Cruz B, Leech AA, Neumann PJ, Sanon Aigbogun M, Oberdhan D, et al. Family and Caregiver Spillover Effects in Cost-Utility Analyses of Alzheimer's Disease Interventions. Pharmacoeconomics. 2019 Apr;37(4):597-608.
[176] Neumann PJ. Costing and perspective in published cost-effectiveness analysis. Med Care. 2009 Jul;47(7 Suppl 1):S28-32.
[177] Thoma A, Kaur MN, Tsoi B, Ziolkowski N, Duku E, Goldsmith CH. Cost-effectiveness analysis parallel to a randomized controlled trial comparing vertical scar reduction and inverted T-shaped reduction mammaplasty. Plast Reconstr Surg. 2014 Dec;134(6):1093-107.
[178] O'Sullivan AK, Thompson D, Drummond MF. Collection of health-economic data alongside clinical trials: is there a future for piggyback evaluations? Value Health. 2005 Jan-Feb;8(1):67-79.
[179] Busse R, Schreyogg J, Smith PC. Variability in healthcare treatment costs amongst nine EU countries - results from the HealthBASKET project. Health Econ. 2008 Jan;17(1 Suppl):S1-8.
[180] Fattore G, Torbica A. Cost and reimbursement of cataract surgery in Europe: a cross-country comparison. Health Econ. 2008 Jan;17(1 Suppl):S71-82.
[181] Briggs AH, Weinstein MC, Fenwick EA, Karnon J, Sculpher MJ, Paltiel AD, et al. Model parameter estimation and uncertainty analysis: a report of the ISPOR-SMDM Modeling Good Research Practices Task Force Working Group-6. Med Decis Making. 2012 Sep-Oct;32(5):722-32.
[182] Bojke L, Claxton K, Sculpher M, Palmer S. Characterizing structural uncertainty in decision analytic models: a review and application of methods. Value Health. 2009 Jul-Aug;12(5):739-49.
[183] Bilcke J, Beutels P, Brisson M, Jit M. Accounting for methodological, structural, and parameter uncertainty in decision-analytic models: a practical guide. Med Decis Making. 2011 Jul-Aug;31(4):675-92.

[184] Health Information and Quality Authority (HIQA). Health technology assessment of a PrEP programme for populations at substantial risk of sexual acquisition of HIV. Dublin: HIQA; 2019.
[185] Health Information and Quality Authority (HIQA). Health technology assessment of a national emergency endovascular service for mechanical thrombectomy in the management of acute ischaemic stroke. Dublin: HIQA; 2017.
[186] Mandelblatt JS, Fryback DG, Weinstein MC, et al. Assessing the effectiveness of health interventions. In: Gold MR, Siegel JE, Russell LB, et al., editors. Cost-effectiveness analysis in health and medicine. New York (NY): Oxford University Press 1996:135-64.
[187] EUnetHTA Joint Action 2, Work Package 8. HTA Core Model® version 3.0 (Pdf); 2016. Available from www.htacoremodel.info/BrowseModel.aspx.
[188] Eddy D. Technology assessment: the role of mathematical modeling. In: Masteller F. editor. Assessing medical technologies. Washington, DC: National Academy Press 1985:144-60.
[189] Vemer P, Corro Ramos I, van Voorn GA, Al MJ, Feenstra TL. AdViSHE: A Validation-Assessment Tool of Health-Economic Models for Decision Makers and Model Users. Pharmacoeconomics. 2016 Apr;34(4):349-61.
[190] Eddy DM, Hollingworth W, Caro JJ, Tsevat J, McDonald KM, Wong JB, et al. Model transparency and validation: a report of the ISPOR-SMDM Modeling Good Research Practices Task Force--7. Value Health. 2012 Sep-Oct;15(6):843-50.
[191] Haute Autorité de santé (HAS). Choices in methods for economic evaluation. 2012.
[192] Institute for Quality and Efficiency in Health Care (IQWiG). General Methods; April 2015.
[193] McCabe C, Dixon S. Testing the validity of cost-effectiveness models. Pharmacoeconomics. 2000 May;17(5):501-13.
[194] Hillston J. Model Validation and Verification. 2003.
[195] Loveman E, Jones J, Clegg AJ, Picot J, Colquitt JL, Mendes D, et al. The clinical effectiveness and cost-effectiveness of ablative therapies in the management of liver metastases: systematic review and economic evaluation. Health Technol Assess. 2014 Jan;18(7):vii-viii, 1-283.
[196] Shibata T, Niinobu T, Ogata N, Takami M. Microwave coagulation therapy for multiple hepatic metastases from colorectal carcinoma. Cancer. 2000 Jul 15;89(2):276-84.
[197] Walker D, Teerawattananon Y, Anderson R, Richardson G. Chapter 5: Generalisability, transferability, complexity and relevance. In: Shemilt I, Mugford M, Vale L, Donaldson C (editors). Evidence-based decisions and economics: health care, social welfare, education and criminal justice. Oxford: Wiley-Blackwell 2010.
[198] Drummond M, Barbieri M, Cook J, Glick HA, Lis J, Malik F, et al. Transferability of economic evaluations across jurisdictions: ISPOR Good Research Practices Task Force report. Value Health. 2009 Jun;12(4):409-18.
[199] Drummond M, Manca A, Sculpher M. Increasing the generalizability of economic evaluations: recommendations for the design, analysis, and reporting of studies. Int J Technol Assess Health Care. 2005 Spring;21(2):165-71.
[200] Sculpher MJ, Pang FS, Manca A, Drummond MF, Golder S, Urdahl H, et al. Generalisability in economic evaluation studies in healthcare: a review and case studies. Health Technol Assess. 2004 Dec;8(49):iii-iv, 1-192.
[201] Wahlster P, Brereton L, Burns J, Hofmann B, Mozygemba K, Oortwijn W, et al. An Integrated Perspective on the Assessment of Technologies: Integrate-Hta. Int J Technol Assess Health Care. 2017 Jan;33(5):544-51.
[202] Welte R, Feenstra T, Jager H, Leidl R. A decision chart for assessing and improving the transferability of economic evaluation results between countries. Pharmacoeconomics. 2004;22(13):857-76.

[203] Boulenger S, Nixon J, Drummond M, Ulmann P, Rice S, de Pouvourville G. Can economic evaluations be made more transferable? Eur J Health Econ. 2005 Dec;6(4):334-46.
[204] Goeree R, Burke N, O'Reilly D, Manca A, Blackhouse G, Tarride JE. Transferability of economic evaluations: approaches and factors to consider when using results from one geographic area for another. Curr Med Res Opin. 2007 Apr;23(4):671-82.
[205] Goeree R, He J, O'Reilly D, Tarride JE, Xie F, Lim M, et al. Transferability of health technology assessments and economic evaluations: a systematic review of approaches for assessment and application. Clinicoecon Outcomes Res. 2011;3:89-104.
[206] Antonanzas F, Rodriguez-Ibeas R, Juarez C, Hutter F, Lorente R, Pinillos M. Transferability indices for health economic evaluations: methods and applications. Health Econ. 2009 Jun;18(6):629-43.
[207] Anderson R. Systematic reviews of economic evaluations: utility or futility? Health Econ. 2010 Mar;19(3):350-64.
[208] Perrier L, Buja A, Mastrangelo G, Baron PS, Ducimetiere F, Pauwels PJ, et al. Transferability of health cost evaluation across locations in oncology: cluster and principal component analysis as an explorative tool. BMC Health Serv Res. 2014 Nov 18;14:537.
[209] Knies S, Evers SM, Candel MJ, Severens JL, Ament AJ. Utilities of the EQ-5D: transferable or not? Pharmacoeconomics. 2009;27(9):767-79.
[210] Barbieri M, Drummond M, Willke R, Chancellor J, Jolain B, Towse A. Variability of cost-effectiveness estimates for pharmaceuticals in Western Europe: lessons for inferring generalizability. Value Health. 2005 Jan-Feb;8(1):10-23.
[211] Paulden M. Recent amendments to NICE's value-based assessment of health technologies: implicitly inequitable? Expert Rev Pharmacoecon Outcomes Res. 2017 Jun;17(3):239-42.
[212] Vallejo-Torres L, Garcia-Lorenzo B, Castilla I, Valcarcel-Nazco C, Garcia-Perez L, Linertova R, et al. On the Estimation of the Cost-Effectiveness Threshold: Why, What, How? Value Health. 2016 Jul-Aug;19(5):558-66.
[213] Hutubessy R, Chisholm D, Edejer TT. Generalized cost-effectiveness analysis for national-level priority-setting in the health sector. Cost Eff Resour Alloc. 2003 Dec 19;1(1):8.
[214] Bertram MY, Lauer JA, De Joncheere K, Edejer T, Hutubessy R, Kieny MP, et al. Cost-effectiveness thresholds: pros and cons. Bull World Health Organ. 2016 Dec 1;94(12):925-30.
[215] Neyt M. Value-Based Pricing: Do Not Throw Away the Baby with the Bath Water. Pharmacoeconomics. 2018 Jan;36(1):1-3.
[216] Culyer A. Cost-Effectiveness Thresholds in Health Care: A Bookshelf Guide to their Meaning and Use. CHE Research Paper 121. December 2015.
[217] Culyer AJ. Cost-effectiveness thresholds in health care: a bookshelf guide to their meaning and use. Health Econ Policy Law. 2016 Oct;11(4):415-32.
[218] Cleemput I, Neyt M, Thiry N, De Laet C, Leys M. Threshold values for cost-effectiveness in health care. Brussels: Belgian Health Care Knowledge Centre (KCE); 2008.
[219] Roze S, de Portu S, Smith-Palmer J, Delbaere A, Valentine W, Ridderstrale M. Cost-effectiveness of sensor-augmented pump therapy versus standard insulin pump therapy in patients with type 1 diabetes in Denmark. Diabetes Res Clin Pract. 2017 Jun;128:6-14.
[220] Catala-Lopez F, Ridao M, Alonso-Arroyo A, Garcia-Altes A, Cameron C, Gonzalez-Bermejo D, et al. The quality of reporting methods and results of cost-effectiveness analyses in Spain: a methodological systematic review. Syst Rev. 2016 Jan 7;5:6.
[221] Sacristan JA, Oliva J, Del Llano J, Prieto L, Pinto JL. [What is an efficient health technology in Spain?]. Gac Sanit. 2002 Jul-Aug;16(4):334-43.

[222] Schmidt R, Majer I, Garcia Roman N, Rivas Basterra A, Grubb E, Medrano Lopez C. Palivizumab in the prevention of severe respiratory syncytial virus infection in children with congenital heart disease; a novel cost-utility modeling study reflecting evidence-based clinical pathways in Spain. Health Econ Rev. 2017 Dec 19;7(1):47.
[223] Vallejo-Torres L, Garcia-Lorenzo B, Serrano-Aguilar P. Estimating a cost-effectiveness threshold for the Spanish NHS. Health Econ. 2018 Apr;27(4):746-61.
[224] Roth JA, Ramsey SD, Carlson JJ. Cost-Effectiveness of a Biopsy-Based 8-Protein Prostate Cancer Prognostic Assay to Optimize Treatment Decision Making in Gleason 3 + 3 and 3 + 4 Early Stage Prostate Cancer. Oncologist. 2015 Dec;20(12):1355-64.
[225] Neumann PJ, Cohen JT, Weinstein MC. Updating cost-effectiveness--the curious resilience of the $50,000-per-QALY threshold. N Engl J Med. 2014 Aug 28;371(9):796-7.
[226] Nadler E, Eckert B, Neumann PJ. Do oncologists believe new cancer drugs offer good value? Oncologist. 2006 Feb;11(2):90-5.
[227] Greenberg D, Earle C, Fang CH, Eldar-Lissai A, Neumann PJ. When is cancer care cost-effective? A systematic overview of cost-utility analyses in oncology. J Natl Cancer Inst. 2010 Jan 20;102(2):82-8.
[228] Carlson JJ, Garrison LP, Ramsey SD, Veenstra DL. The potential clinical and economic outcomes of pharmacogenomic approaches to EGFR-tyrosine kinase inhibitor therapy in non-small-cell lung cancer. Value Health. 2009 Jan-Feb;12(1):20-7.
[229] Myers E, McBroom A, Shen L, Posey R, Gray R, Sanders G. Value-of-Information Analysis for Patient-Centered Outcomes Research Prioritization. Washington, DC: Patient-Centered Outcomes Research Institute; 2012.
[230] Barnett JC, Alvarez Secord A, Cohn DE, Leath CA, 3rd, Myers ER, Havrilesky LJ. Cost effectiveness of alternative strategies for incorporating bevacizumab into the primary treatment of ovarian cancer. Cancer. 2013 Oct 15;119(20):3653-61.
[231] Cohn DE, Kim KH, Resnick KE, O'Malley DM, Straughn JM, Jr. At what cost does a potential survival advantage of bevacizumab make sense for the primary treatment of ovarian cancer? A cost-effectiveness analysis. J Clin Oncol. 2011 Apr 1;29(10):1247-51.
[232] Cohn DE, Barnett JC, Wenzel L, Monk BJ, Burger RA, Straughn JM, Jr., et al. A cost-utility analysis of NRG Oncology/Gynecologic Oncology Group Protocol 218: incorporating prospectively collected quality-of-life scores in an economic model of treatment of ovarian cancer. Gynecol Oncol. 2015 Feb;136(2):293-9.
[233] Lesnock JL, Farris C, Krivak TC, Smith KJ, Markman M. Consolidation paclitaxel is more cost-effective than bevacizumab following upfront treatment of advanced epithelial ovarian cancer. Gynecol Oncol. 2011 Sep;122(3):473-8.
[234] Mehta DA, Hay JW. Cost-effectiveness of adding bevacizumab to first line therapy for patients with advanced ovarian cancer. Gynecol Oncol. 2014 Mar;132(3):677-83.
[235] Chan JK, Herzog TJ, Hu L, Monk BJ, Kiet T, Blansit K, et al. Bevacizumab in treatment of high-risk ovarian cancer--a cost-effectiveness analysis. Oncologist. 2014 May;19(5):523-7.
[236] Duong M, Wright E, Yin L, Martin-Nunez I, Ghatage P, Fung-Kee-Fung M. The cost-effectiveness of bevacizumab for the treatment of advanced ovarian cancer in Canada. Curr Oncol. 2016 Oct;23(5):e461-e7.
[237] Chappell NP, Miller CR, Fielden AD, Barnett JC. Is FDA-Approved Bevacizumab Cost-Effective When Included in the Treatment of Platinum-Resistant Recurrent Ovarian Cancer? J Oncol Pract. 2016 Jul;12(7):e775-83.
[238] National Institute for Health and Care Excellence (NICE). Bevacizumab in combination with carboplatin and paclitaxel for the treatment of advanced ovarian cancer [TA284]; August 2012.

[239] Baker CB, Johnsrud MT, Crismon ML, Rosenheck RA, Woods SW. Quantitative analysis of sponsorship bias in economic studies of antidepressants. Br J Psychiatry. 2003 Dec;183:498-506.
[240] Garattini L, Koleva D, Casadei G. Modeling in pharmacoeconomic studies: funding sources and outcomes. Int J Technol Assess Health Care. 2010 Jul;26(3):330-3.
[241] Hartmann M, Knoth H, Schulz D, Knoth S. Industry-sponsored economic studies in oncology vs studies sponsored by nonprofit organisations. Br J Cancer. 2003 Oct 20;89(8):1405-8.
[242] Ligthart S, Vlemmix F, Dendukuri N, Brophy JM. The cost-effectiveness of drug-eluting stents: a systematic review. CMAJ. 2007 Jan 16;176(2):199-205.
[243] Friedberg M, Saffran B, Stinson TJ, Nelson W, Bennett CL. Evaluation of conflict of interest in economic analyses of new drugs used in oncology. JAMA. 1999 Oct 20;282(15):1453-7.
[244] Jang S, Chae YK, Haddad T, Majhail NS. Conflict of interest in economic analyses of aromatase inhibitors in breast cancer: a systematic review. Breast Cancer Res Treat. 2010 Jun;121(2):273-9.
[245] Bell CM, Urbach DR, Ray JG, Bayoumi A, Rosen AB, Greenberg D, et al. Bias in published cost effectiveness studies: systematic review. BMJ. 2006 Mar 25;332(7543):699-703.
[246] Fleurence RL, Spackman DE, Hollenbeak C. Does the funding source influence the results in economic evaluations? A case study in bisphosphonates for the treatment of osteoporosis. Pharmacoeconomics. 2010;28(4):295-306.
[247] Lane JD, Friedberg MW, Bennett CL. Associations Between Industry Sponsorship and Results of Cost-effectiveness Analyses of Drugs Used in Breast Cancer Treatment. JAMA Oncol. 2016 Feb;2(2):274-6.
[248] Miners AH, Garau M, Fidan D, Fischer AJ. Comparing estimates of cost effectiveness submitted to the National Institute for Clinical Excellence (NICE) by different organisations: retrospective study. BMJ. 2005 Jan 8;330(7482):65.
[249] Peura PK, Martikainen JA, Purmonen TT, Turunen JH. Sponsorship-related outcome selection bias in published economic studies of triptans: systematic review. Med Decis Making. 2012 Mar-Apr;32(2):237-45.
[250] Polyzos NP, Valachis A, Mauri D, Ioannidis JP. Industry involvement and baseline assumptions of cost-effectiveness analyses: diagnostic accuracy of the Papanicolaou test. CMAJ. 2011 Apr 5;183(6):E337-43.
[251] Gilbody S, Bower P, Sutton AJ. Randomized trials with concurrent economic evaluations reported unrepresentatively large clinical effect sizes. J Clin Epidemiol. 2007 Aug;60(8):781-6.
[252] Thorn JC, Noble SM, Hollingworth W. Timely and complete publication of economic evaluations alongside randomized controlled trials. Pharmacoeconomics. 2013 Jan;31(1):77-85.
[253] Freemantle N, Mason J. Publication bias in clinical trials and economic analyses. Pharmacoeconomics. 1997 Jul;12(1):10-6.
[254] Chilcott J, Tappenden P, Rawdin A, Johnson M, Kaltenthaler E, Paisley S, et al. Avoiding and identifying errors in health technology assessment models: qualitative study and methodological review. Health Technol Assess. 2010 May;14(25):iii-iv, ix-xii, 1-107.
[255] Tappenden P, Chilcott JB. Avoiding and identifying errors and other threats to the credibility of health economic models. Pharmacoeconomics. 2014 Oct;32(10):967-79.
[256] Stinnett AA, Mullahy J. Net health benefits: a new framework for the analysis of uncertainty in cost-effectiveness analysis. Med Decis Making. 1998 Apr-Jun;18(2 Suppl):S68-80.

[257] O'Brien BJ, Briggs AH. Analysis of uncertainty in health care cost-effectiveness studies: an introduction to statistical issues and methods. Stat Methods Med Res. 2002 Dec;11(6):455-68.
[258] O'Day K, McLaughlin T, Bramley T. Is it time to eliminate the ICER? Using net benefits to report the results of deterministic cost-effectiveness analyses. *ISPOR 16th Annual International Meeting*. Baltimore, MD, USA: ISPOR 2011.
[259] Dunlop WCN, Mason N, Kenworthy J, Akehurst RL. Benefits, Challenges and Potential Strategies of Open Source Health Economic Models. Pharmacoeconomics. 2017 Jan;35(1):125-8.
[260] Sampson CJ, Wrightson T. Model Registration: A Call to Action. Pharmacoecon Open. 2017 Jun;1(2):73-7.
[261] Cohen AB. Point-Counterpoint: Cost-Effectiveness Analysis in Medical Care and the Issue of Economic Model Transparency. Med Care. 2017 Nov;55(11):907-8.
[262] Padula WV, McQueen RB, Pronovost PJ. Can Economic Model Transparency Improve Provider Interpretation of Cost-effectiveness Analysis? Evaluating Tradeoffs Presented by the Second Panel on Cost-effectiveness in Health and Medicine. Med Care. 2017 Nov;55(11):909-11.
[263] Cohen JT, Wong JB. Can Economic Model Transparency Improve Provider Interpretation of Cost-Effectiveness Analysis? A Response. Med Care. 2017 Nov;55(11):912-4.
[264] Padula WV, McQueen RB, Pronovost PJ. Finding Resolution for the Responsible Transparency of Economic Models in Health and Medicine. Med Care. 2017 Nov;55(11):915-7.
[265] Sullivan W, Hirst M, Beard S, Gladwell D, Fagnani F, Lopez Bastida J, et al. Economic evaluation in chronic pain: a systematic review and de novo flexible economic model. Eur J Health Econ. 2016 Jul;17(6):755-70.
[266] Prakash MK, Lang B, Heinrich H, Valli PV, Bauerfeind P, Sonnenberg A, et al. CMOST: an open-source framework for the microsimulation of colorectal cancer screening strategies. BMC Med Inform Decis Mak. 2017 Jun 5;17(1):80.
[267] Vataire AL, Aballea S, Antonanzas F, Roijen LH, Lam RW, McCrone P, et al. Core discrete event simulation model for the evaluation of health care technologies in major depressive disorder. Value Health. 2014 Mar;17(2):183-95.
[268] Coyle K, Coyle D, Lester-George A, West R, Nemeth B, Hiligsmann M, et al. Development and application of an economic model (EQUIPTMOD) to assess the impact of smoking cessation. Addiction. 2017 Aug 18.
[269] EQUIPT Model Technical Manual. Available at: http://www.equipt.eu/deliverables/ (Access: 19/2/2018).
[270] Govan L, Wu O, Lindsay R, Briggs A. How Do Diabetes Models Measure Up? A Review of Diabetes Economic Models and ADA Guidelines. JHEOR. 2015;3(2):132-52.
[271] Palmer AJ, Roze S, Valentine WJ, Minshall ME, Foos V, Lurati FM, et al. The CORE Diabetes Model: Projecting long-term clinical outcomes, costs and cost-effectiveness of interventions in diabetes mellitus (types 1 and 2) to support clinical and reimbursement decision-making. Curr Med Res Opin. 2004 Aug;20 Suppl 1:S5-26.
[272] Schramm W, Sailer F, Pobiruchin M, Weiss C. PROSIT Open Source Disease Models for Diabetes Mellitus. Stud Health Technol Inform. 2016;226:115-8.
[273] Schrijvers G, van Hoorn A, Huiskes N. The care pathway: concepts and theories: an introduction. Int J Integr Care. 2012 Jan;12(Spec Ed Integrated Care Pathways):e192.
[274] Vanhaecht K, De Witte K, Sermeus W. The impact of clinical pathways on the organisation of care processes. Belgium: KU Leuven; 2007.
[275] Strauss SA. 'Off-label' use of medicine: Some legal and ethical implications. SA Practice Management. 1998;19(1):12-9.

[276] Kleijnen S, George E, Goulden S, d'Andon A, Vitre P, Osinska B, et al. Relative effectiveness assessment of pharmaceuticals: similarities and differences in 29 jurisdictions. Value Health. 2012 Sep-Oct;15(6):954-60.