



**Health  
Information  
and Quality  
Authority**

An tÚdarás Um Fhaisnéis  
agus Cáilíocht Sláinte

# Guidelines for Evaluating the Clinical Effectiveness of Health Technologies in Ireland

December 2018

*Safer Better Care*



## About the Health Information and Quality Authority

The Health Information and Quality Authority (HIQA) is an independent authority established to drive high-quality and safe care for people using our health and social care services in Ireland. HIQA's role is to develop standards, inspect and review health and social care services and support informed decisions on how services are delivered.

HIQA aims to safeguard people and improve the safety and quality of health and social care services across its full range of functions.

HIQA's mandate to date extends across a specified range of public, private and voluntary sector services. Reporting to the Minister for Health and engaging with the Minister for Children and Youth Affairs, HIQA has statutory responsibility for:

- **Setting Standards for Health and Social Services** – Developing person-centred standards, based on evidence and best international practice, for health and social care services in Ireland.
- **Regulation** – Registering and inspecting designated centres.
- **Monitoring Children's Services** – Monitoring and inspecting children's social services.
- **Monitoring Healthcare Safety and Quality** – Monitoring the safety and quality of health services and investigating as necessary serious concerns about the health and welfare of people who use these services.
- **Health Technology Assessment** – Providing advice that enables the best outcome for people who use our health service and the best use of resources by evaluating the clinical effectiveness and cost-effectiveness of drugs, equipment, diagnostic techniques and health promotion and protection activities.
- **Health Information** – Advising on the efficient and secure collection and sharing of health information, setting standards, evaluating information resources and publishing information about the delivery and performance of Ireland's health and social care services.

## **Table of contents**

<b>About the Health Information and Quality Authority.....</b>	<b>3</b>
<b>Table of contents .....</b>	<b>4</b>
<b>Foreword .....</b>	<b>6</b>
<b>Process and acknowledgements.....</b>	<b>8</b>
<b>Record of updates .....</b>	<b>10</b>
<b>List of abbreviations .....</b>	<b>11</b>
<b>1 Introduction .....</b>	<b>12</b>
<b>1.1 Clinical effectiveness guidelines .....</b>	<b>13</b>
<b>1.2 Document layout .....</b>	<b>13</b>
<b>1.3 Explanation of terms .....</b>	<b>13</b>
<b>1.4 Summary of guideline statements for measures of effect.....</b>	<b>14</b>
<b>1.5 Summary of guideline statements for methods of meta-analysis. ....</b>	<b>16</b>
<b>2 Measures of effect .....</b>	<b>19</b>
<b>2.1 Common considerations when assessing endpoints .....</b>	<b>19</b>
2.1.1 Endpoint data .....	19
2.1.2 Relative and absolute endpoints.....	20
2.1.3 Efficacy and effectiveness.....	21
2.1.4 Endpoint reliability and validity.....	23
2.1.5 Internal and external validity of a study.....	24
2.1.6 Survival data.....	26
2.1.7 Multiple endpoints.....	28
2.1.8 Subgroup analysis .....	28
<b>2.2 Types of endpoint .....</b>	<b>29</b>
2.2.1 Patient-reported outcomes (PROs) .....	29
2.2.2 Clinical endpoints .....	32
2.2.3 Surrogate endpoints.....	33
2.2.4 Composite endpoints.....	35
2.2.5 Adverse events .....	36
2.2.6 Sensitivity and specificity .....	39
<b>3 Methods of comparison .....</b>	<b>41</b>
<b>3.1 Common considerations when undertaking or evaluating meta-analysis .....</b>	<b>41</b>

3.1.1	Gathering evidence .....	41
3.1.2	Individual patient data .....	43
3.1.3	Types of study .....	44
3.1.4	Data and study quality .....	46
3.1.5	Heterogeneity .....	47
3.1.6	Meta-regression .....	48
3.1.7	Fixed and random effects .....	48
3.1.8	Sources of bias .....	49
3.1.9	Frequentist and Bayesian approaches.....	51
3.1.10	Outliers and influential studies.....	51
3.1.11	Sensitivity analysis .....	52
3.2	Networks of evidence .....	53
3.3	Selecting the method of comparison.....	56
3.4	Methods of meta-analysis .....	58
3.4.1	Direct meta-analysis.....	58
3.4.2	Unadjusted indirect comparison .....	60
3.4.3	Adjusted indirect comparison.....	61
3.4.4	Network meta-analysis .....	62
3.4.5	Meta-analysis of diagnostic test accuracy studies .....	65
3.4.6	Generalised linear mixed models.....	67
3.4.7	Confidence profile method.....	68
<b>4</b>	<b>References .....</b>	<b>70</b>
<b>5</b>	<b>Glossary of terms and abbreviations.....</b>	<b>78</b>
	<b>Appendix A — Network meta-analysis example .....</b>	<b>84</b>
	<b>Appendix B — Further reading .....</b>	<b>95</b>

## Foreword

The Health Information and Quality Authority (HIQA) has a statutory remit to evaluate the clinical and cost-effectiveness of health technologies and provide advice to the Minister for Health and the Health Service Executive (HSE). It is recognised that the findings of a health technology assessment (HTA) may also have implications for other key stakeholders in the Irish healthcare system, such as patient groups, the general public, clinicians, other healthcare providers, academic groups, and the manufacturing industry.

HTA guidelines provide an overview of the principles and methods used in assessing health technologies. They are intended as a guide for all those who are involved in the conduct or use of HTA in Ireland. The purpose of the national guidelines is to promote the production of assessments that are timely, reliable, consistent and relevant to the needs of decision makers and key stakeholders in Ireland.

The HTA guidelines include documents on economic evaluation, budget impact analysis, assessment of clinical effectiveness, stakeholder involvement in HTA and recommended reporting formats. Each of these areas is important. For ease of use, the guidelines have been developed as stand-alone documents.

These guidelines are intended to inform health technology assessments conducted by, or on behalf of the Health Information and Quality Authority (HIQA), the National Centre for Pharmacoeconomics, the Department of Health and the Health Service Executive (HSE), to include health technology suppliers preparing applications for reimbursement. The guidelines are intended to be applicable to all healthcare technologies, including pharmaceuticals, procedures, medical devices, broader public health interventions and service delivery models.

This document, *Guidelines for Evaluating the Clinical Effectiveness of Health Technologies in Ireland*, is part of the series of guidelines, and is limited to methodological guidance on the evaluation of clinical effectiveness. The guidelines will be reviewed and revised as necessary. For ease of use, guideline statements that summarise key points are included prior to each section in italics.

HIQA would like to thank the members of the Scientific Advisory Group and its Chairperson, Dr Michael Barry from the National Centre for Pharmacoeconomics, and all who have contributed to the production of these guidelines.

A handwritten signature in black ink, appearing to read 'Máirín Ryan', enclosed within a thin black rectangular border.

**Dr Máirín Ryan,**  
Deputy CEO and Director of Health Technology Assessment  
Health Information and Quality Authority

## Process and acknowledgements

This document, *Guidelines for Evaluating the Clinical Effectiveness of Health Technologies in Ireland* is a complementary document to *Guidelines for Economic Evaluation of Health Technologies in Ireland* (2018), *Guidelines for Budget Impact Analysis of Health Technologies in Ireland* (2018), *Guidelines for Stakeholder Engagement in Health Technology Assessment in Ireland* (2014), and *Guidelines for the Retrieval and Interpretation of Economic Evaluations of Health Technologies in Ireland* (2014). These guidelines are limited to methodological guidance on the evaluation of clinical effectiveness and are intended to promote best practice in this area. They will be reviewed and revised as necessary, with updates provided through HIQA's website ([www.hiqa.ie](http://www.hiqa.ie)). The above documents form part of a series of national guidelines for health technology assessment (HTA) in Ireland that the HIQA will develop and continuously review in the coming years.

The guidelines have been developed by HIQA with technical input from the National Centre for Pharmacoeconomics and in consultation with its Scientific Advisory Group. Providing broad representation from key stakeholders in Irish healthcare, this group includes methodological experts from the field of HTA. The group provides ongoing advice and support to HIQA in its development of national HTA guidelines. The terms of reference for this group are to:

- contribute fully to the work, debate and decision-making processes of the group by providing expert technical and scientific guidance at Scientific Advisory Group meetings as appropriate
- be prepared to occasionally provide expert advice on relevant issues outside of Scientific Advisory Group meetings, as requested
- support HIQA in the generation of guidelines to establish quality standards for the conduct of HTA in Ireland
- support HIQA in the development of methodologies for effective HTA in Ireland
- advise HIQA on its proposed HTA Guidelines Work Plan and on priorities as required
- support HIQA in achieving its objectives outlined in the HTA Guidelines Work Plan
- review draft guidelines and other HTA documents developed by HIQA and recommend amendments as appropriate
- contribute to HIQA's development of its approach to HTA by participating in an evaluation of the process as required.

Minor updates to the guidelines have been made based on feedback received in the four years since the publication of the second edition of these guidelines. Following public consultation, these guidelines were revised where necessary and, subsequently, approved by the HIQA Board.

**The membership of the HTA Scientific Advisory Group is as follows:**

Dr Michael Barry (Chair)	National Centre for Pharmacoeconomics
Orlaith Brennan	Irish Pharmaceutical Healthcare Association
Professor Kerri Clough	National Cancer Registry
Dr Kathleen MacLellan	Department of Health
Dr Anne Dee	Health Service Executive
Professor Mike Drummond	University of York
Shaun Flanagan	Health Service Executive
Dr Patricia Harrington	Health Information and Quality Authority
Ciara Finlay	Irish Medtech Association
Dr Teresa Maguire	Department of Health
Dr Brendan McElroy	University College Cork
Stephen McMahon	Irish Patients Association
Dr Peter Kiely	Health Products Regulatory Authority
Dr Derek Mitchell	Irish Platform for Patients' Organisations Science & Industry
Dr Mairead O'Driscoll	Health Research Board
Professor Ciarán O'Neill	National University of Ireland, Galway
Mark Chapman	Irish Medical & Surgical Trade Association
Dr Máirín Ryan	Health Information and Quality Authority
Professor Mark Sculpher	University of York
Professor Susan Smith	Royal College of Surgeons in Ireland
Dr Conor Teljeur	Health Information and Quality Authority
Dr Lesley Tilson	National Centre for Pharmacoeconomics
Dr Valerie Walshe	Health Service Executive
Professor Cathal Walsh	Trinity College Dublin

**Contributors**

HIOA would like to thank Finbarr Leacy of the Health Products Regulatory Authority for detailed comments.

We would also like to thank the evaluation team at the National Centre for Pharmacoeconomics, in particular Joy Leahy, for their feedback.

## Record of updates

Date	Title / Version	Summary of changes
November 2011	Guidelines for Evaluating the Clinical Effectiveness of Health Technologies in Ireland	First national clinical effectiveness guidelines
September 2014	Guidelines for Evaluating the Clinical Effectiveness of Health Technologies in Ireland	Minor revisions and reorganisation of text. Addition of text on extrapolating survival from censored trial data and investigator assessed endpoints.
December 2018	Guidelines for Evaluating the Clinical Effectiveness of Health Technologies in Ireland	Minor revisions and reorganisation of text. Addition of appendix with illustrative network meta-analysis.

Guidelines for Evaluating the Clinical Effectiveness of Health Technologies in Ireland

Issued: December 2018

This document is one of a set that describes the methods and processes for conducting health technology assessment in Ireland.

The document is available from the HIQA website ([www.hiqa.ie](http://www.hiqa.ie)).

Suggested citation:

Health Information and Quality Authority. *Guidelines for Evaluating the Clinical Effectiveness of Health Technologies in Ireland*. Dublin, Ireland: HIQA, 2018.

## List of abbreviations

EUnetHTA	European Network for Health Technology Assessment
HIQA	Health Information and Quality Authority
HSE	Health Service Executive
HSROC	hierarchical summary receiver operating characteristic
HRQoL	health-related quality of life
HTA	health technology assessment
IPD	individual patient data
ITT	intention-to-treat
MTC	mixed treatment comparison
QALY	quality-adjusted life years
PRO	patient-reported outcome
RCT	randomised controlled trials
ROC	receiver operating characteristic
sROC	summary receiver operating characteristic

## 1 Introduction

Health technology assessment (HTA) has been described as ‘a multidisciplinary process that summarises information about the medical, social, economic and ethical issues related to the use of a health technology in a systematic, transparent, unbiased, robust manner’.<sup>(1)</sup> The scope of the assessment depends on the technology being assessed, but may include any, or all of these issues. The purpose of HTA is to inform health policy decisions that promote safe, effective, efficient and patient-focussed healthcare.

The primary audience for HTAs in Ireland is decision makers within the publicly-funded health and social care system. It is recognised that the findings of a HTA may also have implications for other stakeholders in the system. Stakeholders include patient groups, the general public, clinicians and other healthcare professionals, other healthcare providers, academic groups and the manufacturing industry.

HIQA continues to develop a series of methodological guidelines that are intended to assist those that conduct HTA for or on behalf of the Health Information and Quality Authority, the National Centre for Pharmacoeconomics, the Department of Health and the Health Service Executive. They underpin assessments of health technologies carried out within the framework agreement between the Irish Pharmaceutical Healthcare Association and the Department of Health and the Health Service Executive on the supply terms, conditions and prices of medicines in Ireland. Their purpose is to promote the production of assessments that are timely, reliable, consistent and relevant to the needs of decision makers and other stakeholders.

The series of guidelines are intended to be applicable to all healthcare technologies, including pharmaceuticals, procedures, medical devices, broader public health interventions and service delivery models. They are, therefore, broad in scope and some aspects may be more relevant to particular technologies than others.

The Clinical Effectiveness Guidelines represent one component of the overall series. They are limited to the methodological guidance on the evaluation of the clinical effectiveness of technologies in HTA. These guidelines are intended to be viewed as a complementary document to the *Guidelines on Economic Evaluation of Health Technologies in Ireland* and the *Guidelines for Budget Impact Analysis of Health Technologies in Ireland*.

The content of this document was partly derived from text prepared by HIQA for inclusion in an overarching HTA guideline being prepared by the European Network for HTA (EUnetHTA) collaboration. These guidelines have drawn on published

research and will be reviewed and revised as necessary following consultation with the various stakeholders, including those in the Scientific Advisory Group.

## **1.1 Clinical effectiveness guidelines**

Clinical effectiveness describes the ability of a technology to achieve a clinically significant impact on a patient's health status. The evaluation of clinical effectiveness is considered in this document under two headings: measures of effect and methods of comparison or meta-analysis. The former are used to determine the impact of a technology on a patient's health status. The effect of a technology may be assessed in isolation or it may be considered relative to one or more other technologies in terms of comparative effectiveness. In these guidelines, clinical effectiveness is considered relative to another treatment, whether that is usual care, placebo or some other comparator. There are numerous methods available to measure and report treatment effects and many associated methodological issues. Measures of effect are discussed in Section 2 of this document.

To compare two or more technologies, the measured effects of those technologies are often combined across a number of studies to maximise the evidence base. Data from multiple studies are typically combined in a meta-analysis. There are a variety of meta-analysis methodologies available that are appropriate in different contexts. Section 3 of this document outlines the methods of meta-analysis available, the main issues and considerations associated with meta-analysis, and it provides guidance on selecting the most appropriate method for a given analysis.

In this document, the descriptions of type of effect measure and method of meta-analysis follow a standard format using the following set headings: description, examples, usage, strengths, limitations, and critical questions.

## **1.2 Document layout**

For ease of use, a list of the guideline statements that summarise the key points of the guidance is included at the end of this chapter. These guideline statements are also included in italics at the beginning of each section for the individual elements described in Chapters 2 and 3.

## **1.3 Explanation of terms**

A number of terms used in the guidelines may be interpreted more broadly elsewhere or may have synonymous terms that could be considered to be interchangeable. The following outlines the specific meanings that may be inferred for these terms within the context of these guidelines and identifies the term that will be used throughout the guidelines for the purpose of consistency.

'Technology' includes any intervention that may be used to promote health, to prevent, diagnose or treat disease, or that is used in rehabilitation or long-term care. This includes pharmaceuticals, devices, medical equipment, medical and surgical procedures, and the organisational and supportive systems within which healthcare is provided. Within the context of these guidelines, the terms 'intervention' and 'technology' should be considered to be interchangeable, with the term 'technology' used throughout for the purpose of consistency.

Efficacy is the extent to which a treatment has the ability to achieve the intended effect under ideal circumstances. Effectiveness is the extent to which a treatment achieves the intended effect in the typical clinical setting. Efficacy studies usually precede effectiveness studies. Both efficacy and effectiveness studies provide valuable information about the treatment effect in a specified patient group.

## **1.4 Summary of guideline statements for measures of effect**

**Endpoint data (Section 2.1.1)** Endpoints can be expressed as continuous, categorical or count data. When continuous data are expressed as categorical, the selection of cut-points must be clearly described and justified.

**Relative and absolute endpoints (Section 2.1.2)** Absolute measures are presented as a difference and are dependent on the baseline risk in the study population. Relative measures are presented as a ratio and are variationally independent of the baseline risk. Endpoints should be expressed in both absolute and relative terms where possible.

**Efficacy and effectiveness (Section 2.1.3)** Efficacy is the extent to which a treatment has the ability to achieve the intended effect under ideal circumstances. Effectiveness is the extent to which a treatment achieves the intended effect in the typical clinical setting. Both efficacy and effectiveness studies provide valuable information about the treatment effect for a specified group of patients. Where available, both efficacy and effectiveness must be reported. Statistical methods for handling missing data and underlying assumptions regarding the missing data mechanism should be clearly stated.

**Endpoint reliability and validity (Section 2.1.4)** A reliable endpoint returns the same value with repeated measurements on the same individual. A valid endpoint accurately measures the endpoint it was intended to measure. Endpoints used in an assessment must have demonstrated reliability and validity.

**Internal and external validity of a study (Section 2.1.5)** Internal validity is the extent to which bias is minimised in a particular trial. External validity is the extent to which the findings of a particular trial can be generalised to other settings or

populations. Treatment effect should be measured in trials that have both internal and external validity.

**Survival data (Section 2.1.6)** In survival analysis, overall survival should be considered the gold standard for demonstrating clinical benefit. In assessing progression-free, relapse-free and disease-free survival, patients must be evaluated on a regular basis to ensure that the time of progression is measured accurately. The length of follow up must be clearly defined and relevant to the disease in question. It should be clear whether all or only the first post-treatment event was recorded. When extrapolating longer-term survival from trial data, alternative models should be tested and reported with goodness-of-fit measures.

**Multiple endpoints (Section 2.1.7)** All relevant endpoints used in the literature should be reported in an assessment.

**Subgroup analysis (Section 2.1.8)** Consideration should be given to the inclusion of relevant subgroups that have been clearly defined and identified based on an a priori expectation of differences, supported by a plausible biological or clinical rationale for the subgroup effect.

**Types of endpoint (Section 2.2)** An endpoint must be clearly defined and measurable. It must be reliable and valid. An endpoint should be relevant to the condition being treated and sensitive to change.

**Patient-reported outcomes (PROs) (Section 2.2.1)** PROs should be used to measure changes in health and functional status that are of direct relevance to the patient. A PRO should be sensitive to changes in health status. If a multi-dimensional measure is used, it should be clearly stated whether all or some of the dimensions were evaluated. The PRO should encompass domains relevant to the illness being treated. The use of mapping from one PRO to another must be clearly stated and justified. Only a validated mapping function based on empirical data should be used. The statistical properties of the mapping function should be clearly described. All PROs collected in a study should be reported.

**Clinical endpoints (Section 2.2.2)** The choice of clinical endpoint must be justified on the basis of a clear link between the disease process, technology and endpoint.

**Surrogate endpoints (Section 2.2.3)** A surrogate endpoint must have a clear biological or medical rationale or have a strong and validated link to a final endpoint of interest.

**Composite endpoints (Section 2.2.4)** A change in a composite endpoint should be clinically meaningful. All of the individual components of a composite must be

reliable and valid endpoints. The components that drive the composite result should be identified.

**Adverse outcomes (Section 2.2.5)** All adverse effects that are of clinical or economic importance must be reported. Both the severity and frequency of harms should be reported. It should be clear whether harms are short term or of lasting effect.

**Sensitivity and specificity (Section 2.2.6)** The sensitivity and specificity of a diagnostic or screening test should be measured in relation to a recognised reference test. The threshold for a positive test result should be clearly defined.

## **1.5 Summary of guideline statements for methods of meta-analysis**

**Gathering evidence (Section 3.1.1)** The methods used to gather evidence for a meta-analysis must be clearly described. Evidence is typically gathered using a systematic review.

**Individual patient data (Section 3.1.2)** Individual patient data can be analysed in a meta-analysis. Individual patient data meta-analysis should not be used to the exclusion of other relevant data. Results should be compared to a study-level analysis.

**Types of study (Section 3.1.3)** Evidence to support the effectiveness of a technology should be derived by clearly defined methods. Where available, evidence from high quality RCTs should be used to quantify efficacy.

**Data and study quality (Section 3.1.4)** Studies included in a meta-analysis should be graded for quality of evidence. The quality of evidence should be clearly stated. The results of a meta-analysis should be reported according to relevant standards.

**Heterogeneity (Section 3.1.5)** Heterogeneity of treatment effect between studies must be assessed. Where significant heterogeneity is observed, attempts should be made to identify its causes. Substantial heterogeneity must be dealt with appropriately and may preclude a meta-analysis.

**Meta-regression (Section 3.1.6)** When there is significant between-study heterogeneity, meta-regression is a useful tool for identifying study-level covariates that modify the treatment effect.

**Fixed and random effects (Section 3.1.7)** The choice between a fixed and random effects analysis is context specific. Heterogeneity should be assessed using

standard methods. Significant heterogeneity suggests the use of a random effects model. Justification must be given for the choice of fixed or random effects model.

**Sources of bias (Section 3.1.8)** Attempts should be made to identify possible sources of bias such as publication bias, sponsorship bias and bias arising from the inclusion of poor quality studies. Potential sources of bias must be reported along with steps taken to minimise the impact of bias.

**Frequentist and Bayesian approaches (Section 3.1.9)** Both frequentist and Bayesian approaches are acceptable in meta-analysis. The approach taken must be clearly stated.

**Outliers and influential studies (Section 3.1.10)** Influential studies and those that are statistical outliers should be identified and reported. The methods used for identifying outliers must be clearly stated. Studies that are outliers should be characterised to determine if they are comparable to the other included studies.

**Sensitivity analysis (Section 3.1.11)** If potential outliers have been identified, or if plausible subgroups of patients or studies have been identified, a comprehensive sensitivity analysis should be conducted. In a Bayesian analysis, the choice of priors should be tested using a sensitivity analysis.

**Networks of evidence (Section 3.2)** The network of available evidence should be described and used to guide the selection of the method of meta-analysis. The selection of direct and indirect evidence must be clearly defined. The exclusion of relevant evidence, either direct or indirect, should be highlighted and justified. Where direct and indirect evidence are combined, inconsistencies between the direct and indirect evidence must be assessed and reported.

**Selecting the method of comparison (Section 3.3)** The choice of method of comparison depends on the quality, quantity and consistency of direct and indirect evidence. The available evidence must be clearly described along with a justification for the choice of method.

**Methods of meta-analysis (Section 3.4)** For any method of meta-analysis, all included trials must be sufficiently comparable and measuring the same treatment effect.

**Direct meta-analysis (Section 3.4.1)** Direct meta-analysis should be used when there are sufficient comparable head-to-head studies available. If indirect evidence is available, then consideration should also be given to a multiple treatment comparison.

**Unadjusted indirect comparison (Section 3.4.2)** The method of unadjusted indirect comparisons should not be used.

**Adjusted indirect comparison (Section 3.4.3)** Adjusted indirect comparison is appropriate for comparing two technologies using a common comparator.

**Network meta-analysis (Section 3.4.4)** A network meta-analysis can be appropriate for comparing multiple treatments when both direct and indirect evidence are available.

**Meta-analysis of diagnostic test accuracy studies (Section 3.4.5)** The bivariate random effects and hierarchical summary receiver operating characteristic models (HSROC) should be used for pooling sensitivity and specificity from diagnostic and screening test accuracy studies. The correlation between sensitivity and specificity should be reported.

**Generalised linear mixed models (Section 3.4.6)** Generalised linear mixed models can be appropriate when analysing individual patient data from trials.

**Confidence profile method (Section 3.4.7)** The confidence profile method can be used to combine direct and indirect evidence. Network meta-analysis or Bayesian mixed treatment comparison should be used in preference to the confidence profile method. The use of this method over other available methods should be justified.

## 2 Measures of effect

Measures of effect are used to determine treatment impact in terms of changes in health status. That impact is usually in the form of improved health status (for instance survival, cure, remission), but it can also be worsening health status (such as adverse reactions, hospitalisations, deaths). Measures of effect should be clearly relevant to the disease, condition, complaint or process of interest. It should be possible to diagnose and interpret them, and they should be sensitive to treatment differences. Effects may be observed for any technology such as pharmaceutical, surgical, or therapeutic.

In these guidelines, measures of effect are referred to as endpoints. This chapter will describe issues that are relevant to a variety of endpoint types before describing types of endpoint.

### 2.1 Common considerations when assessing endpoints

A number of common important considerations must be taken into account when assessing endpoints.

#### 2.1.1 Endpoint data

*Endpoints can be expressed as continuous, categorical, count or time-to-event data. When continuous data are expressed as categorical, the selection of cut-points must be clearly described and justified.*

Endpoint data can be expressed in a variety of ways<sup>(2)</sup>:

- Continuous — a continuous variable has numeric values (for example 1, 2, 3). The relative magnitude of the values is significant (for instance, a value of 2 indicates twice the magnitude of 1). Examples include blood pressure and prostate specific antigen.
- Categorical — a categorical variable classifies a subject into one of 2 or more unique categories (such as disease status — remission, mild, moderate, or severe relapse). A binary variable is a categorical variable with only two levels (such as mortality, stroke). An ordinal variable is a categorical variable that can be rank-ordered (for example, self-reported health status or Clinical Global Impression).
- Count data — variables in which observations can only have non-negative integer values (such as number of hospitalisations).

Many endpoints are reported as proportions or rates and hence are continuous data. The techniques available for summarising and analysing the endpoint data are

affected by how the data are expressed. The conversion of a variable from continuous to categorical results in the loss of information. The quality of the conversion depends on how homogeneous the observations are within each category. While a categorical variable, particularly if binary, is often simpler to interpret, categorisation of a continuous variable should be avoided where possible.

Categorical endpoints, particularly when expressed in binary form, can be open to manipulation when derived from a continuous measure. For example, the threshold for distinguishing between healthy and ill in an endpoint can be set to show a treatment in the best light if there is no commonly agreed cut-off. Dichotomising does not typically introduce significant bias if the split is made at the median or some other pre-specified percentile. However, if the cut-point is chosen based on analysis of the data, in particular by splitting at the value which produces the largest difference in endpoints between categories, then severe bias will be introduced. Endpoints that are binary by nature, such as myocardial infarction, may still vary considerably in clinical interpretation.<sup>(3)</sup>

### 2.1.2 Relative and absolute endpoints

*Absolute measures are presented as a difference and are dependent on the baseline risk in the study population. Relative measures are presented as a ratio and are variationally independent of the baseline risk. Endpoints should be expressed in both absolute and relative terms where possible.*

The endpoints of a trial can typically be expressed in absolute or relative terms. Absolute measures are presented as a difference and are dependent on the baseline risk in the study population. Relative measures are presented as a ratio and are independent of the baseline risk.

Absolute risk measures:

- are generally useful to clinicians as they provide a more realistic quantification of treatment effect than relative measures<sup>(4)</sup>
- have limited generalisability due to their dependence on baseline risk
- should not be pooled in a meta-analysis due to fact that the variation in baseline risk is not accounted for<sup>(5)</sup>
- cannot be applied to different subgroups unless they have been explicitly calculated for those subgroups
- examples include absolute risk reduction and number needed to treat.

Relative risk measures:

- are usually stable across populations with different baseline risks and are useful when combining the results of different trials in a meta-analysis
- do not take into account a patient's risk of achieving the intended endpoint without the treatment and thus do not give a true reflection of how much benefit the patient would derive from the treatment<sup>(6)</sup>
- can be applied to different subgroups with the understanding that baseline risk will vary by subgroup and ignoring that subgroup characteristics may modify the treatment effect
- examples include relative risk, odds ratio and hazard ratio.

Further detail on the properties, advantages and disadvantages of the various effect measures are available in the Cochrane Handbook.<sup>(7)</sup>

The choice between absolute and relative is sometimes made to maximise the perceived effect. In some instances the absolute risk difference will be small whereas the relative risk might be large.<sup>(8)</sup> Wherever possible, both relative and absolute measures should be presented as together they provide the magnitude and context of an effect.<sup>(9)</sup> If both are not included, then justification of the presented measure must be included.

### 2.1.3 Efficacy and effectiveness

*Efficacy is the extent to which a treatment has the ability to achieve the intended effect under ideal circumstances. Effectiveness is the extent to which a treatment achieves the intended effect in the typical clinical setting. Both efficacy and effectiveness studies provide valuable information about the treatment effect for a specified group of patients. Where available, both efficacy and effectiveness must be reported. Statistical methods for handling missing data and underlying assumptions regarding the missing data mechanism should be clearly stated.*

Efficacy is the extent to which a treatment has the ability to achieve the intended effect under ideal circumstances. Effectiveness is the extent to which a treatment achieves the intended effect in the typical clinical setting. Efficacy studies usually precede effectiveness studies.

Efficacy studies tend to utilise condition-specific endpoints with strong links to the mechanism of action. Such endpoints also tend to be collected in a short-term time horizon. Efficacy studies sometimes have stringent exclusion criteria to prevent the enrolment of patients who are less likely to observe a significant treatment effect

(such as those with comorbidities, lower risk patients). Efficacy is generally measured in randomised controlled trials (RCTs).

Effectiveness studies tend to collect more comprehensive endpoint measures that reflect the range of benefits expected from the treatment that are relevant to the patient and to the payer, including improvement in ability to function and quality of life. These measures often have a weaker link to the mechanism of action. Both short- and longer-term horizons are typically considered in effectiveness studies. Effectiveness studies can be useful in identifying the true benefit of a technology in a real world or community setting. Examples of effectiveness studies include pragmatic RCTs, observational cohorts or registry data. Pragmatic RCTs, such as those nested in population-based screening programmes, can generate high quality effectiveness data.

Efficacy does not necessarily correlate with effectiveness. The distinction between efficacy and effectiveness may be more pronounced for some endpoints, particularly endpoints that are sensitive to individual-level factors (for example, comorbidities, smoking status). Factors that may impact on the effectiveness of a treatment, such as adherence, should be documented.<sup>(10)</sup>

In clinical trials, the analysis for efficacy and effectiveness are referred to as intention-to-treat (ITT) and per-protocol, respectively. Intention-to-treat analysis patients are analysed according to the group into which they were randomised irrespective of whether or not they received that treatment. Per-protocol analysis, on the other hand, only considers the patients who fully adhered to the clinical trial instructions as specified in the protocol. Intention-to-treat analyses can provide a more realistic view of how technologies work in practice whereas effect estimates from a per-protocol analysis, as typically implemented, may be biased by non-random loss of patients. While ITT estimates are affected by the adherence patterns observed in the trial and may not reflect the adherence patterns that would be observed in clinical practice, ITT may indicate acceptability for a technology not captured in the per-protocol analysis.

All available data should be used where possible, even if some data are missing for a specific patient. Missing data can pose problems where outcomes are not recorded due to loss to follow up. There are a number of techniques for handling missing data. The appropriateness of a specific technique depends on the underlying missing data mechanism (for example, missing at random). Ad-hoc methods such as last observation carried forward (LOCF) and the related methods baseline observation carried forward (BCOF) and worst observation carried forward (WOCF) impute a single value for each missing data point using previous observations for a patient. While convenient, these methods are rarely appropriate and resulting estimates should be interpreted with caution. For dichotomous outcomes, non-responder

imputation (NRI), which assumes that all dropouts are non-responders, can be appropriate in some contexts when the proportion of missing values is low. Alternative methods such as multiple imputation (MI) and mixed-effect model for repeated measures (MRMM) are becoming increasingly popular but rely on the assumption that the data are missing at random. LOCF methods tend to result in inflated rates of Type I errors compared to MI and MRMM methods.<sup>(11, 12)</sup> Preference is for the use of MI and MRMM methods. Where missing data have been imputed, the approach for handling missing data should be clearly stated and, where possible, the impact of departures from the underlying assumption on the missing data mechanism should be explored in a sensitivity analysis.

#### 2.1.4 Endpoint reliability and validity

*A reliable endpoint returns the same value with repeated measurements on the same individual. A valid endpoint accurately measures the endpoint it was intended to measure. Endpoints used in an assessment must have demonstrated reliability and validity.*

Endpoints should be both reliable and valid. Reliability refers to whether repeated measurements return the same value. Differences can arise due to the individual who takes the measurement (inter-rater reliability), the instruments used to make the measurements or the context in which the measurement is made.<sup>(13)</sup> The reliability of the instruments used can be checked using test-retest reliability, whereby the measurement is repeated and differences compared. The measures should be recorded at an interval over which no change is expected. Particularly for subjective measures, the inter-rater reliability should be investigated. Depending on the subjectivity of the measure, substantial variability may occur across clinical or patient raters. Where possible, all raters should be blinded to treatment assignment.

Validity refers to how accurately an instrument measures the endpoint it was intended to measure. There are a number of forms of validity:

- Construct validity — how well the endpoint represents reality in terms of cause and effect
- Content validity — how well the endpoint measures what it is intended to measure
- Criterion validity — how well the endpoint compares to a reference or gold-standard measure
- Face validity — if the endpoint appears to be valid to the clinician, patient or assessor.

Direct measures of objective endpoints are presumed to have validity. As shown in independent empirical studies, any subjectively measured endpoint must have established validity as shown in independent empirical studies.

Reliability and validity are not independent as an endpoint cannot be valid if it is not reliable, but could be reliable without being valid.<sup>(14)</sup> If an endpoint is unreliable, then it will not return the same value on repeated measures and, hence, will not have criterion validity. However, a reliable endpoint would not be valid if, for example, it is not related to the effect of the technology being assessed.

### 2.1.5 Internal and external validity of a study

*Internal validity is the extent to which bias is minimised in a study. External validity is the extent to which the findings of the trial can be generalised to other settings or populations. Treatment effect should be measured in trials that have both internal and external validity.*

Internal validity is concerned with the extent to which bias is minimised in a study. A number of types of bias can impact on internal validity<sup>(15)</sup>:

- selection bias due to biased allocation to study arms
- performance bias due to unequal provision of care
- detection bias due to biased endpoint assessment
- attrition bias due to loss to follow up.

Internal validity can be maximised by a combination of careful study design, conduct and analysis. Proper randomisation prevents allocation bias. Endpoint measurement is prone to detection bias if adequate blinding has not been used in a study.<sup>(16)</sup> Proper blinding of patients, clinicians and independent raters can reduce or eliminate performance and detection bias. Blinding is particularly important for more subjective and or self-reported endpoint measures. Independently-assessed endpoints are considered more reliable and less prone to bias than investigator-assessed endpoints, and are preferred for measuring clinical effectiveness. Bias can also be introduced by systematic withdrawals or exclusions from the study for patients receiving the technology. Maximising response rates in all study arms will reduce attrition bias. From an analytical point of view, it is important to know how a study has dealt with drop-outs and missing data when computing summary effect measures (see Section 2.1.3). Failure to properly account for poor response rates or missing data will introduce further bias into an analysis. If the internal validity of a study is doubtful then the measured treatment effect must be questioned.

The external validity of a study impacts on the extent to which the findings of the study are generalisable to other settings or populations. The main factors that impact on external validity are:

- patient profile (such as age-sex distribution, ethnicity, disease severity, risk factors, comorbidity)
- treatment regimen, including dosage, frequency and comparator treatment
- setting (for example, primary, secondary or tertiary care)
- factors specific to the country or jurisdiction where the study took place (such as eligibility for follow-up care)
- endpoints (such as definition of endpoints, length of follow up)
- participation rate, as a poor participation rate may mean that the study population is not representative of the target population.

To achieve a given statistical power for a study, the number of patients required is a function of both the risk in the control group and of the hypothesised reduction in the risk due to treatment. For rarer endpoints the required sample size is larger. By enrolling high-risk patients, trials can be run with a smaller sample size. It is also noted that many studies do not report the power and sample size calculations, or whether they are testing for superiority, inferiority or equivalence which impacts on the ability to detect a statistically significant difference. The power and sample size should be appropriate for the type of test being carried out.

External validity can be maximised by ensuring that the study characteristics closely match those found in routine clinical practice. The patients should be typical of those who would generally be eligible for the type of treatment being assessed. The treatment regimen should reflect what would realistically be achieved in routine practice in terms of dose, frequency, adherence and compliance. The technology should be applied in a similar setting to routine practice and the measured endpoints should be those that are commonly accepted as relevant to the disease being treated. A lack of external validity does not imply that the measured treatment effect is incorrect but may prevent the effect estimate from being generalised to other populations.

Issues of validity affect both experimental trials and observational studies. In general, randomised controlled trials can achieve internal validity through careful study design, but they may lack external validity for a variety of reasons such as patient selection.<sup>(17)</sup> Observational studies, on the other hand, may have sufficient external validity but often lack internal validity due to confounding factors that may or may not be challenging to identify.<sup>(18)</sup>

### 2.1.6 Survival data

*In survival analysis, overall survival should be considered the gold standard for demonstrating clinical benefit. In assessing progression-free, relapse-free and disease-free survival, patients must be evaluated on a regular basis to ensure that the time of progression is measured accurately. The length of follow up must be clearly defined and relevant to the disease in question. It should be clear whether all or only the first post-treatment event was recorded. When extrapolating longer-term survival from trial data, alternative models should be tested and reported with goodness-of-fit measures.*

Survival analysis measures when the endpoint occurred as well as whether the endpoint occurred. Common survival endpoints include overall survival, disease-free survival, relapse-free survival and progression-free survival. When the primary aim of the technology is to extend life, overall survival is the gold standard for demonstrating clinical benefit. Defined as the time from randomisation to death, this endpoint is unambiguous and is not subject to investigator interpretation. Where overall survival is not measurable in a practical study time horizon, the alternatives of progression-free, relapse-free and disease-free survival could be considered. Both the overall survival rate and the intermediate data (for example, progression-free, relapse-free and disease-free survival) should be reported, if available. When intermediate data are reported, they must be clearly defined. It cannot be automatically assumed that an intermediate outcome is a suitable surrogate for overall survival, and the relationship should be demonstrated by evidence. In some cases, overall and progression-free survival may be modelled separately: in these cases the two models should be consistent with each other and generate mutually plausible results.

In assessing progression-free, relapse-free and disease-free survival, patients must be evaluated on a regular basis to ensure that the time at which a change in health status occurs is measured accurately. When combining time-to-event data from multiple studies, it is critical that appropriate methods are used to account for differences in the intervals at which endpoints were measured.

Survival can also be expressed in terms of a hazard ratio which is a widely used metric to compare survival in two groups.<sup>(19)</sup> The hazard ratio gives the relative risk of an endpoint at any given time with a value of 1 corresponding to equal hazards. The ratio is based on the entire study period and assumes that the ratio does not change through the study period. With a large enough sample size, it is possible to calculate the hazard ratio for smaller time periods during a study. The validity of the assumptions underpinning the hazard ratio should be demonstrated. In the event that the analysis is not valid, alternate methods of expressing survival must be considered.

A key issue in survival analysis is censoring, that is, ceasing observation at the end of the study period.<sup>(20)</sup> The length of follow up should be explicitly stated and justification provided as to its relevance to the disease in question. Different studies may use quite different follow-up periods, rendering their findings incompatible. Data analysis cut-off dates and schedule of assessment have an impact on the probability of observing events related to the time frame. Incomplete reporting has been shown to be common, affecting the definition of survival terms and the numbers of patients at risk. It should be clear whether only the first post-treatment event was recorded or if all non-fatal events were recorded in the follow-up period. A key assumption of many time-to-event models and analytical methods is that censoring is non-informative, meaning that censoring is independent of the disease process and occurs at random. Where censoring is informative (such as a patient is withdrawn from a trial because of deteriorating health), particular care needs to be taken when analysing the data and appropriate adjustments need to be made to the data.

Sometimes studies report results from interim analyses. Such analyses are often conducted to protect the interests of trial participants by allowing for early stopping of the trial due to demonstrated efficacy or futility based on pre-specified criteria. In the case of rare diseases with limited available treatment options, results from pre-specified interim analyses may form the basis of a conditional marketing authorisation pending final approval based on analysis of the pre-specified endpoints. Interim analyses are often based on an intermediate outcome rather than the pre-specified endpoint and may be biased with respect to the true treatment effect as measured when trial follow up is complete. If interim analysis results are incorporated into an economic model, they are likely to bias the estimate of cost-effectiveness. Wherever possible, treatment effect should be based on an analysis of pre-specified endpoints when trial follow up is complete. Results from ad-hoc interim analyses that were not pre-specified in the study protocol should be interpreted with significant caution.

Due to the relatively short time-horizon of clinical trials and consequent censoring, full survival benefit is often extrapolated using a variety of methods.<sup>(20, 21)</sup> The different extrapolation methods can produce substantially different estimates of long-term survival that can have a significant impact on the estimated effectiveness of the technology. Alternative extrapolation models should be tested and the relative fit to the observed trial data should be reported.

Time-to-event data may be reported for outcomes other than survival (for example, cessation of breast feeding), and the methodology used for survival analysis can be applied to other endpoints that are subject to censoring.<sup>(22)</sup>

### 2.1.7 Multiple endpoints

*All relevant endpoints used in the literature should be reported in an assessment.*

The use of multiple endpoints can give rise to Type 1 error whereby the probability of false-positive findings by chance is increased. A single primary endpoint and multiple secondary endpoints should be defined in the study protocol and consideration should be given to adjusting for multiple testing.<sup>(19)</sup> In reality, there may be multiple primary endpoints which may include safety endpoints. If multiple endpoints are included, they should be justified. Reported endpoints may not be per-protocol — in some instances, studies may report results for the endpoint(s) where the most significant effect was observed.<sup>(23)</sup> There is debate as to whether or not secondary endpoints should even be reported if the effect on the primary endpoint is not significant.<sup>(24)</sup> To reduce reporting bias, all relevant endpoints used in the literature should be reported in an assessment.<sup>(25)</sup> Where multiple endpoints are used, they should be specified in advance and not selected on a post-hoc basis.

### 2.1.8 Subgroup analysis

*Consideration should be given to the inclusion of relevant subgroups that have been clearly defined and identified based on an a priori expectation of differences, supported by a plausible biological or clinical rationale for the subgroup effect.*

Subgroup analysis should be considered where there are potentially important differences in patient characteristics or treatment benefit that may be observed between groups. Subgroups should have been defined a priori and pre-specified in the study protocol and statistical analysis plan, with plausible reasons for expecting different treatment effects across subgroups. Where subgroups were not pre-specified in trials, the results are less likely to be valid and should be treated as exploratory.<sup>(26)</sup> The use of subgroups increases the number of statistical tests undertaken and hence increases the chance of generating false-positive results. Trials should be suitably powered for subgroup analysis, with appropriate adjustment for multiple testing, and this should be considered when evaluating trial results. A test for interaction should be used to provide statistical evidence of a differential treatment effect.<sup>(27)</sup>

A subgroup analysis may be required if the licensed indication is narrower than the indications included for the entire study population. In this instance, it may be possible to restrict the analysis to the subgroup of patients treated for the licensed indication.

## 2.2 Types of endpoint

*An endpoint must be clearly defined and measurable. It must be reliable and valid. An endpoint should be relevant to the condition being treated and sensitive to change.*

The choice of endpoints used in a study or comparison will be influenced by the purpose for which they are measured.<sup>(28)</sup> For example, if the primary purpose of a technology is to improve survival, then mortality will be the relevant endpoint. If, however, a technology is designed to improve mobility, then functional status may be a more appropriate endpoint. When selecting endpoints for inclusion in an evaluation, it may be useful to include patients and clinicians in the process to ensure the relevance of the selected endpoints.<sup>(29)</sup>

This section looks at different endpoints types that have distinct modes of collection or purpose. For each type of endpoint there is a brief description, some typical examples, a brief note on usage in the literature, the strengths and limitations of that type of endpoint and then some critical questions that should be asked when assessing such an endpoint.

### 2.2.1 Patient-reported outcomes (PROs)

*PROs should be used to measure changes in health and functional status that are of direct relevance to the patient. A PRO should be sensitive to changes in health status. If a multi-dimensional measure is used, it should be clearly stated whether all or some of the dimensions were evaluated. The PRO should encompass domains relevant to the illness being treated. The use of mapping from one PRO to another must be clearly stated and justified. Only a validated mapping function based on empirical data should be used. The statistical properties of the mapping function should be clearly described. All PROs collected in a study should be reported.*

#### Description

The term patient-reported outcome (PRO) covers a whole range of measurement types, but usually refers to self-reported patient health status focussing on how the patient functions or feels in relation to a health condition and its treatment. PROs encompass simple symptom measures (for example, pain measured by Likert scale), more complex measures (such as activities of daily living or function), multidimensional measures (for example, health-related quality of life) and satisfaction with treatment. PROs can be generic or disease specific. Generic PROs can be used for any condition, but they can be less responsive to changes in health

status than disease specific measures. When using a multidimensional PRO, it is important to ensure that the PRO covers all the domains relevant to the illness and technology being assessed, including adverse events. The choice of PRO should therefore be justified based on coverage of relevant domains for the indication of interest.

Health-related quality of life (HRQoL) measures can be susceptible to change due to a variety of external factors (such as life circumstances unrelated to the illness being treated) with the exception of HRQoL questionnaires that have been specifically developed to capture the impact of a specific disease process. It is possible to map one HRQoL measure onto another, such as EQ-5D onto SF-36. This may be done for comparability to present results using a different HRQoL measure to the one used in a study. Mapping may over- or underestimate the effectiveness of a technology.<sup>(30)</sup> When mapping from another HRQoL measure, only a validated mapping function based on empirical data should be used. The statistical properties of the mapping function should be clearly described.

Utility measures are used to generate quality-adjusted life years (QALYs) which can be used in economic analyses. QALY data are often collected, but not reported in a study. QALYs are PROs and provide useful endpoint data for assessing the clinical effectiveness of a technology.

A detailed guidance on the use of HRQoL measures is beyond the scope of this document. References for further reading on HRQoL measures are provided in Appendix 1.

## **Examples**

- EQ-5D – a general self-administered questionnaire used to rate health-related quality of life across five dimensions (mobility, self-care, usual activities, pain/discomfort, anxiety/depression)<sup>(31)</sup>
- SF-36 – a multi-purpose, short-form health questionnaire consisting of eight scaled scores relating to aspects of physical and mental health<sup>(32)</sup>
- WOMAC – a self-administered questionnaire used to evaluate the condition of patients with osteoarthritis of the knee and hip.<sup>(33)</sup>

## **Usage**

PROs are often used as primary endpoints for technologies that do not have a clear impact on final clinical endpoints, but do improve a patient's well-being or functional status. PROs are often collected as secondary endpoints for a trial, but may not be reported. Rather than using objective measures of a patient's health status, PROs use subjective self-assessment.

It is possible to detect improvements in a PRO in the absence of a change in a clinical endpoint and vice versa. Where such an apparent inconsistency arises, it is valuable to consider whether the difference is plausible and which outcome might be more relevant in the context of the assessment.

## **Strengths**

PROs measure changes in health and functional status that are of direct relevance to the patient. The use of PROs therefore gives a patient-centred perspective on the effect of a technology. PROs can also highlight where clinicians and patients have divergent views on what endpoints are considered important to patients. A PRO can encompass both the positive and negative effects of a technology in a single summary measure.

PROs can be used to detect endpoints, such as pain, that may be difficult or unfeasible to measure in clinical tests.

PROs are often collected in the form of self-administered questionnaires which do not have to be filled out in a clinical setting. There are no requirements for biological samples and they can be assessed by non-clinicians.

## **Limitations**

Some generic PRO measures have been shown to be unresponsive to modest changes in status. If a PRO is not sensitive to change then it may not be able to adequately capture the effect of a technology.

The clinical relevance of PROs can be difficult to determine, except in cases where a PRO is the main efficacy endpoint of the treatment, for example pain used to assess the efficacy of a pain-killer drug.

PRO results are often non-specific to a particular condition and are often susceptible to general changes in a patient's circumstances making it difficult to directly associate a change in score with the health technology under assessment.

As PROs frequently return a score, it can be difficult to translate the change in score into a marker of clinical improvement. Concepts such as 'minimal perceptible clinical improvement' and 'responders' are used to define clinically significant improvements. The definitions of a clinically significant change and a 'responder' are open to question and must be clearly justified if used.

PROs can be time consuming to complete. If the patient cohort has literacy issues then response rates may be low or the answers may not accurately reflect the true perceptions of the patients. PRO data collected amongst patients who are not

blinded to their treatment assignment can be biased. Where possible, study participants should, therefore, be blinded to treatment assignment while completing PRO measures to prevent bias.

### Critical questions

- Is the PRO a reliable and valid measure of effect?
- Is the PRO sensitive to change?
- Is a change in the PRO clinically significant?
- Is the PRO condition-specific or general?
- Were the participants blinded to their treatment assignment when completing the PRO?

### 2.2.2 Clinical endpoints

*The choice of clinical endpoint must be justified on the basis of a clear link between the disease process, technology and endpoint.*

### Description

A clinical endpoint is an aspect of a patient's clinical or health status that is measured to assess the efficacy or harm of a treatment relative to the best available alternative. A clinical endpoint should be a valid measure of clinical benefit due to treatment – it is clinically relevant, sensitive (responsive to change) and is both recognised and used by physicians. Clinical endpoints are based on the presence or absence of measurable clinical events. Clinical endpoints tend to be unambiguous, impartially measured events to minimise potential bias. The choice of clinical endpoint must be justified on the basis of a clear link between the disease process, technology and endpoint.

### Examples

- mortality
- stroke
- lower limb amputation.

### Usage

Clinical endpoints are perhaps the most common type of endpoints used in clinical trials. They are used in trials where clear clinical events are achieved or avoided due to the treatments being studied. All-cause mortality is considered to be the most unbiased clinical endpoint as it is final and its measurement is unambiguous.

## Strengths

Clinical endpoints tend to be both valid and reliable. Clinical endpoints are typically objectively measured, reducing the occurrence of assessment bias. Clinical endpoints are usually generalisable across settings and can therefore improve the external validity of a trial.

## Limitations

Clinical endpoints can be poorly defined in studies (for example, does non-fatal myocardial infarction include silent events?). Differences in definition can lead to different results. A clinical endpoint may be a rare event which raises issues of statistical power and the need for large sample sizes. Some clinical endpoints may be clinically important, but of little direct relevance to the patient.

## Critical questions

- Is the clinical endpoint clearly defined?
- Is there a clear mechanism of action between the technology and the clinical endpoint?
- Is the clinical endpoint objectively or subjectively measured?

### 2.2.3 Surrogate endpoints

*A surrogate endpoint must have a clear biological or medical rationale or have a strong and validated link to a final endpoint of interest.*

## Description

A surrogate endpoint, also called an intermediate endpoint, is an objectively measured endpoint that is expected to predict clinical benefit or harm based on epidemiologic, pathophysiologic, therapeutic and other scientific evidence. They are typically physiological or biochemical markers that can be relatively quickly and easily measured. The effect of the technology on the surrogate endpoint must predict the effect on the clinical endpoint.<sup>(34)</sup> The effect on the surrogate should be of a similar magnitude to the effect on a final endpoint.

If surrogate endpoints are assessed, caution must be exercised in directly extrapolating from these to final endpoints unless underpinned by a clear biological or medical rationale or have a strong or validated link. Although a surrogate endpoint may have a strong link to an endpoint of interest, it may not itself represent a meaningful endpoint to the patient.

## **Examples**

- blood pressure as a surrogate endpoint for cardiovascular mortality
- bone mineral density as a surrogate for bone fracture
- HIV1-RNA viral load as an indicator of viral suppression.

## **Usage**

Surrogate endpoints are common where final endpoints might require a long follow-up period. Surrogate markers are often used when the primary endpoint is either undesired (such as death) or when the number of events is very small, thus making it impractical to conduct a clinical trial to gather a statistically significant number of endpoints.

## **Strengths**

When there is a clear and strong link to a final endpoint, a surrogate can enable a shorter follow-up period and greatly reduce the cost of a trial.

## **Limitations**

If the mechanism of action of the technology is not fully understood, it is possible that the surrogate endpoints will fail to accurately predict the true clinical effect of the technology. Furthermore, if multiple causal pathways exist between the technology and the clinical endpoint then the surrogate may also fail to accurately predict the clinical effect.

The magnitude of the effect on the surrogate may be substantially different to that on the final endpoint. Thus the use of a surrogate may under- or over-estimate the effect of the technology. Depending on the treatment pathway, a surrogate endpoint may fail to capture the effects of subsequent technologies on the final endpoint. This may be particularly relevant for complex interventions or cross-over studies where the sequencing of interventions could have an important influence on outcomes.

Statistical methods used to predict final outcomes from surrogate endpoints are imperfect and do not in any way negate the need to gather data on final endpoints. An analysis predicting final endpoints from surrogate endpoints should be supported by extensive sensitivity analyses.

## **Critical questions**

- What is the reason for using a surrogate endpoint?
- Does the surrogate have a clear biological or medical rationale or have a strong or validated link to a final endpoint of interest?
- Can the biomarker be reliably detected?

- Is the magnitude of the effect on the surrogate similar to that on the final endpoint?

#### 2.2.4 Composite endpoints

*A change in a composite endpoint should be clinically meaningful. All of the individual components of a composite must be reliable and valid endpoints. The components that drive the composite result should be identified and highlighted in the analysis.*

#### Description

Composite endpoints combine multiple single events into one endpoint that attempts to capture an overall and clinically relevant treatment effect. They are often used to increase event rates and decrease the sample size required where statistical power is poor and to avoid the issue of multiple testing. Each of the endpoints included in the composite must meet the requirements of validity, reliability, relevance and accurate measurement. The composite may include a mixture of direct clinical and surrogate endpoints. It is important that patients are followed up after the first non-fatal event as they may subsequently experience further events, including a fatal event.<sup>(35)</sup> If non-fatal events are included in a composite endpoint, it is important to state whether all non-fatal events were evaluated or just the first event to occur.

Although trials are often underpowered to report disaggregated endpoints, they should be reported individually where possible.

#### Examples

- mortality, hospitalisation and cardiac arrest in patients with chronic heart failure
- mortality, myocardial infarction and stroke in patients with hypertension
- mortality and new-onset diabetes.

#### Usage

Composite endpoints are most commonly used in studies of cardiovascular technologies. On average, composites include three endpoints but can range from two to nine or more.<sup>(36)</sup>

#### Strengths

Composite endpoints can make it possible to estimate the net benefit of a treatment. A composite avoids the problem of selecting a single endpoint where there may be a

number of endpoints of equal importance. The use of composites can avoid the need to adjust for multiple comparisons.

## Limitations

Interpretation of composites can cause problems particularly if the combination consists of endpoints with very different clinical importance or a combination of objective and subjective measures.<sup>(37)</sup> Identifying what could be considered a clinically significant change may be difficult. Interpretation of the results will be complicated if the effect on the composite is primarily driven by the effect on one of the components. Although it may be tempting to conclude that the treatment has a significant impact on the component, it is likely that the data are underpowered to draw such a conclusion. The components that drive the composite result should be identified and highlighted in any presentation of the analysis.

If the composite endpoint is not given in disaggregated form it may not be viable to combine the results of several studies due to differences in definition (such as the use of different components). Varying definitions of composite endpoints can lead to substantially different results and conclusions.<sup>(38)</sup>

As a composite requires a number of endpoints, there is an increased risk of missing data. Inappropriate adjustment for missing data can result in biased estimates of the proportions of successes in composite endpoints.<sup>(39)</sup>

## Critical questions

- Does the composite endpoint really measure treatment effect for a disease?
- Does the use of a composite endpoint solve a medical problem or is it just for statistical convenience?
- Are the individual components of the composite endpoint valid, biologically plausible, and of importance for patients?
- Are the results clear and clinically meaningful? Do they provide a basis for therapeutic decisions? Does each single endpoint support the overall result?
- Is the statistical analysis adequate?

### 2.2.5 Adverse events

*All adverse events that are of clinical or economic importance must be reported. Both the severity and frequency of harms should be reported. It should be clear whether harms are short-term or of lasting effect.*

## **Description**

Many technologies have side-effects - these are unintended effects that may be harmful. It is generally anticipated that the benefits of a technology will exceed the potential harms. Endpoints can include adverse events that reflect the safety of a technology. Harms caused by a technology provide an important counterbalance to benefits and can include harm to the patient or to the clinician providing the technology (for instance, radiation exposure during diagnostic imaging). Harms can be broadly classified as effects and reactions. Effects are caused by a technology, while patients experience a reaction.

For many adverse events it may be difficult to definitively ascribe them to a technology. Adverse events are often collected as secondary endpoints and there is likely to be variation across studies in how these are reported in terms of both detail and terminology.

Any differences between the trial population and the intended target population should be reported, as the adverse event profile may differ between the two populations.

As serious adverse events are usually anticipated to be relatively rare, studies are typically underpowered to detect differences in their occurrence. To overcome the problem of statistical power, studies often aggregate the adverse events even though they may be of varying importance or severity. Furthermore, relatively minor events, such as low grade fever, will not be of much importance in studies where the primary objective is a reduction in mortality. Sufficient follow up is required to capture important adverse events such as mortality.

## **Examples**

- hospitalisations due to an adverse drug reaction
- postoperative complications
- toxicity-related side effects due to external beam radiotherapy.

## **Usage**

Trials of technologies are generally designed to evaluate benefits rather than harms. Trials are generally run over relatively short time horizons with small numbers of patients. Such trials are therefore at most able to detect and quantify frequent adverse events that occur early during treatment. In addition, to be recorded systematically in a trial it must be known beforehand or anticipated that there will be adverse events.<sup>(40)</sup>

Due to the difficulties in making a causal link between a technology and a harm, adverse events are often distinguished from potential adverse events which may have arisen despite the technology. The distinction between preventable and unavoidable adverse events is also used. Preventable events stem from errors such as incorrect drug, dose or frequency.

### **Strengths**

Harms are relevant to patients and may influence whether or not a treatment is acceptable to patients.

Adverse events can have a major impact on cost-effectiveness as they may generate substantial additional treatment costs.

### **Limitations**

Most clinical studies are underpowered to detect statistically significant differences for adverse events, especially for rare events. If a difference in aggregated harms is observed then it may be difficult to conclude whether the difference is due to harms of greater or lesser severity.

Adverse events can be quite different to the endpoints collected to determine treatment effect making it difficult to carry out a direct comparison of benefit and harm other than through the impact on a quality of life measure.

Adverse events may be recorded by a variety of means (for example, by a patient, nurse or doctor) leading to variable quality of reporting and coverage. Furthermore, the events reported may be very different depending on whether they were reported by a clinician or a patient, although reported adverse events usually undergo assessment of causality to determine whether they are due to medications or are related to disease.

Drug therapies often fail due to interactions with concomitant medications taken by a patient. If a study excludes patients with comorbidities or older patients, then there will be less opportunity for serious drug interactions to arise, even though they may occur frequently in routine practice.

### **Critical questions**

- How have the safety endpoints been collected and reported?
- Are both the severity and frequency of harms quantified?
- Do the harms have lasting effect or are they short-term only?
- Has sufficient follow up been used to capture all important adverse events?

## 2.2.6 Sensitivity and specificity

*The sensitivity and specificity of a diagnostic or screening test should be measured in relation to a recognised reference test. The threshold for a positive test result should be clearly defined.*

### Description

Sensitivity and specificity are standard measures of diagnostic and screening test accuracy. Although they are not a direct measure of clinical effect, diagnostic tests are used to identify and monitor the existence, onset, severity or risk of disease. As such, they are used as a means to evaluate clinical effects.

Sensitivity and specificity are calculated by comparing the index test to a gold standard test. Sensitivity shows positive index test results as a proportion of those that are genuinely positive based on a gold standard diagnostic or screening test. The specificity indicates negative index test results as a proportion of those that are genuinely negative based on the same gold standard. A perfect test would have a sensitivity and specificity both equal to 100. A test with high sensitivity helps rule out the disease when the result is negative, whereas a test with high specificity helps rule in the disease when the result is positive.

As the calculation of sensitivity and specificity require a dichotomous outcome (that is to say, positive or negative), a threshold value must be used to convert continuous or categorical parameters into dichotomous values. Varying the threshold will impact on the sensitivity and specificity of the test.

Different thresholds result in different sensitivities and specificities and the resulting pairs can be illustrated on a receiver operator characteristic (ROC) plot. Sensitivity and specificity are typically negatively correlated, so the choice of threshold is a trade-off between high sensitivity at the expense of low specificity or *vice versa*.

### Examples

- magnetic resonance imaging for detection of acute vascular lesions
- computed tomography in the diagnosis of lymph node metastases in patients with cancer
- electrocardiography for the diagnosis of left ventricular hypertrophy.

### Usage

Sensitivity and specificity have no clinical value of themselves. However, they are used to calculate other useful characteristics such as the positive and negative

likelihood ratios and the diagnostic odds ratio. The likelihood ratios are combined with pre-test odds to calculate the post-test odds of disease. Hence a clinician can determine the probability of presence or absence of disease on foot of a positive or negative test result.

If the sum of sensitivity and specificity is equal to 100, then the test provides no diagnostic evidence.

### **Strengths**

Sensitivity and specificity provide a combined measure of diagnostic test accuracy.

### **Limitations**

Diagnostic test accuracy studies are common, but the reporting is often of poor quality and subject to numerous forms of bias.<sup>(41)</sup> In particular it is vital that those assessing the results of the gold standard test are blinded to the results of the index test. The same gold standard should be applied throughout and the index test should not form part of the gold standard.

Both sensitivity and specificity need to be reported together. Subject to threshold effects, the pre-test odds must be known in order to calculate the post-test probability of a given test result.

### **Critical questions**

- What gold standard is the diagnostic test being compared to?
- Does the test have diagnostic value?
- Should the test be used to rule in or rule out disease?
- Does the test make use of a threshold value and how has that been defined?
- How high was the disease prevalence in the study sample?

## 3 Methods of comparison

The clinical effectiveness of a technology is generally measured in a randomised controlled trial (RCT) setting. Multiple trials may attempt to estimate the clinical effectiveness of the same technology and will provide different estimates of the effectiveness. Often a single trial may fail to detect a modest, but clinically and statistically significant difference between two technologies, mainly due to inadequate numbers of patients. To maximise the evidence base and improve precision it is common to combine results from several trials in a meta-analysis. The process of combining trials involves a weighted average typically related to the precision of each trial estimate.

For the purposes of these guidelines, it is presumed that sufficient data of acceptable quality are available to justify a meta-analysis. It is assumed that the collected measures of effect comply with Chapter 2 of the guidelines. It is also assumed that the collection of the data contributing to the comparisons involves an exhaustive search of published and unpublished trials, and a rigorous selection process based on the methodological quality of the trials.

The purpose of this chapter is to give guidance on appropriate methods of combining measures of effect from multiple trials and to outline some of the common issues associated with those methods. The chapter is structured as follows:

- Section 3.1 discusses common considerations to be taken into account when conducting or assessing a meta-analysis.
- Section 3.2 describes networks of evidence.
- Section 3.3 provides guidance on how to select the most appropriate method for a given meta-analysis.
- Section 3.4 outlines the various methods of meta-analysis.

### 3.1 Common considerations when undertaking or evaluating meta-analysis

There are a number of common important considerations that must be taken into account when undertaking or evaluating the results of a meta-analysis.

#### 3.1.1 Gathering evidence

*The methods used to gather evidence for a meta-analysis must be clearly described. Evidence is typically gathered using a systematic review.*

Data from trials are used as evidence of treatment effect. In combining data from multiple trials a first step is to identify the relevant studies. The methods used to gather evidence for a meta-analysis must be clearly described. Evidence is typically gathered using a systematic review.

A systematic review of a clinical technology is a review of the evidence regarding that technology prepared using a systematic approach.<sup>(42)</sup> The study question to be addressed should be defined in advance along with clear inclusion and exclusion criteria. A clear protocol should be prepared outlining the steps of the review. The use of a systematic approach will reduce the likelihood of bias.

The typical steps in a systematic review are as follows<sup>(43)</sup>:

- formulate the review question
- define inclusion and exclusion criteria
- identify studies
- select studies for inclusion
- assess study quality
- extract data
- analyse and present results
- interpret results.

When conducting a systematic review, specialist expertise on bibliographic searching is required. Where available, it may be beneficial to seek the support of a librarian or information specialist to assist with literature searches, devise a search strategy for key resources, or for training on literature search techniques for systematic reviews.

A systematic review may include a meta-analysis of the evidence, but it is not a prerequisite. However, a meta-analysis should preferably be undertaken as part of a systematic review to minimise bias in study selection. In some instances, such as where the researchers have conducted all relevant trials on a particular technology, a systematic review would be unnecessary prior to a meta-analysis. Attempts should be made to identify any relevant unpublished trials or studies as this could reduce the bias in the findings of the review. It is critical that a systematic review is current at the point of informing an economic evaluation and that all relevant evidence is captured. Ideally a systematic review would be no more than six months old when the cost-effectiveness analysis it informs is carried out. There are a number of texts available that provide clear guidance on how to carry out a systematic review. A list of appropriate guidance texts is provided in Appendix 1.

The pooling of data is sometimes used as an alternative to formal meta-analysis. Data pooling can be used to combine data from across multiple trial locations or when using individual patient data. Although data pooling does not require a

systematic review, any such analysis is at risk of bias by potentially not including all relevant evidence.

The choice of comparators should reflect the policy question being addressed. In the context of HTA, the choice of comparators should at a minimum encompass the recommended standard of care and those that are used in routine clinical practice in Ireland.<sup>(44)</sup> In some contexts, it may be appropriate to include potential comparators that are not yet reimbursed but may reasonably be expected to become the standard of care in the short to medium term.

### 3.1.2 Individual patient data

*Individual patient data can be analysed in a meta-analysis. Individual patient data meta-analysis should not be used to the exclusion of other relevant data. Results should be compared to a study-level analysis.*

While meta-analyses typically combine study-level effect estimates, it is also possible to combine individual patient data (IPD) from studies. Use of individual data can improve the ability to include comparable subgroups or common endpoints which may not be reported in published studies. Analysis of patient data also enables more detailed time-to-event data to be combined.

The methods of IPD meta-analysis can be broadly classified into two groups: a one-step analysis, in which all patients are analysed simultaneously as though in a mega-trial, but with patients clustered by trial; or a two-step analysis in which the studies are analysed separately, but then summary statistics are combined using standard meta-analysis techniques.<sup>(45)</sup> Although one-step analysis can be undertaken without clustering at the study level thereby ignoring the distinction between studies, this approach is not recommended.

A number of advantages of IPD meta-analysis over aggregate data analysis have been cited, including<sup>(46)</sup>:

- original study data are used which gives access to all endpoints recorded
- consistent inclusion and exclusion criteria and subgroups can be defined across studies
- potentially longer follow-up data may be available than in published results
- results of unpublished studies can be included, reducing potential publication bias
- uniform analytical methods can be applied across all studies
- better handling of covariates and prognostic factors.

By modelling the individual risk across hundreds or thousands of patients, IPD meta-analyses generally have much higher power to detect differences in treatment effect

than the equivalent aggregate data analyses that may have fewer than 10 studies.<sup>(46)</sup> An IPD analysis can also be used to determine the potential treatment effect for individual patients rather than at a group level, which may be more relevant to patients. Methods are also available to combine IPD and aggregate data in a single model.<sup>(47)</sup>

The main disadvantage of IPD meta-analysis is that data collection is both expensive and time-consuming, and it may not be possible to acquire data from all relevant studies. When the number of patients involved is large, of the order of tens of thousands, the analysis becomes computationally intensive. Using data from a limited number of studies may distort the results and the estimates from an IPD analysis should be compared to the equivalent study-level analysis.

IPD from a single or small number of trials may also be used as a basis for developing a micro-simulation model. Patient characteristics are used to populate the model and simulate the impact of introducing a treatment in terms of endpoints and costs. Such an exercise should not be considered as either evidence synthesis or meta-analysis, but rather a form of subgroup analysis. The use of IPD for micro-simulation is beyond the scope of these Guidelines.

### 3.1.3 Types of study

*Evidence to support the effectiveness of a technology should be derived by clearly defined methods. Where available, evidence from high quality RCTs should be used to quantify efficacy.*

Controlled trials and observational studies are generally used to evaluate the effectiveness of technologies.

RCTs demonstrate the effect of the technology and randomisation to minimise bias between cases and controls. Patients enrolled in an RCT are often carefully selected, may have few if any comorbidities and may not be receiving any concurrent treatment, thereby making it difficult to generalise the results. RCTs are often limited to non-rare diseases and short durations, and ethical considerations can impact on the choice of comparator technology.

Observational studies follow patients in the real world where treatment may be less carefully monitored and the patients will often have comorbid conditions for which they are also being treated. Observational studies can have large sample sizes and longer follow-up periods compared with RCTs and can, therefore, provide useful data with respect to rare events and outcomes (for example, adverse events). They can be based on routinely collected administrative data, which greatly reduces the cost of data collection. Due to the numerous possible confounders, it can be difficult to

obtain unbiased estimates of treatment effect in observational studies. They are also open to numerous sources of bias.

The spectrum of study types is broad, including RCT, non-randomised controlled trial, controlled before-and-after, interrupted time series, historically controlled, cohort, case-control, cross-sectional and case series. The different study types have distinct risks of bias or limitations associated with them, and careful consideration must be given to how those risks might affect the interpretation of results. Non-randomised studies can detect associations between an technology and a health outcome but cannot rule out that the association arose due to a confounding factor linked to both the technology and outcome.<sup>(48)</sup>

For the purposes of a comparison, efficacy should be quantified using high-quality RCT data where available. Given the difficulties in assessing bias, observational studies do not always offer the best level of evidence, but they do provide valuable evidence on the impact a treatment will have in routine care. In the context of public health interventions, evidence of effectiveness is often only available through non-randomised studies. Where an analysis combines data from different study designs, it is essential that careful consideration is given to whether the studies can be viewed to be estimating the same treatment effect. Of particular importance are the presence of systematic differences in the study populations, the manner in which the technology is delivered (for example, the setting), the comparator (for example, differences in the definition of usual care) and how outcomes were assessed. For example, non-randomised studies may include patients with a different spectrum of disease to those included in randomised studies or may allow patient self-selection, thus biasing the potential for benefit from the technology. The same considerations apply to an analysis combining data from phase I, phase II and phase III studies.

In the event that RCT evidence is not available and it is not possible to estimate treatment effect through indirect methods, then data from single arm studies may be available. Where such data are used in a comparison, the data and analytical approach must be presented with complete clarity and transparency. The assumptions underpinning the analysis must be clearly stated along with estimates of potential error associated with the analytical approach. Where single-arm study data are included, it should be on the basis of an anchored comparison.<sup>(49)</sup> Where weights are used to adjust the data, the methods and data used to derive the weights must be clearly described. If weights are developed through a modelling approach including covariates, clinical and statistical justification should be provided for the selection of covariates and the model fit statistics should be reported. The choice of model should be based on highest predictive power. Where alternative approaches can be used for derivation of weights or incorporation of single-arm studies into the analysis, a sensitivity analysis should be presented to determine the impact of methodology on the estimate of treatment effect.

### 3.1.4 Data and study quality

*Studies included in a meta-analysis should be graded for quality of evidence. The quality of evidence should be clearly stated. The results of a meta-analysis should be reported according to relevant standards.*

When combining a number of trials, it is essential to assess the quality of the data that are to be combined. A meta-analysis of low-quality data will not yield a high-quality effect estimate. Study quality must be evaluated using a recognised tool and reported, and it should be explicitly taken into consideration in the interpretation of the findings of the analysis. The tool used must be appropriate to the design of the study being assessed.

A trial may be of genuinely poor quality due to inadequate study design, or it may be poorly reported irrespective of the actual study quality. It can be anticipated that a poor quality study will generate a biased estimate of effect. A poorly reported study may be of good quality, but there is insufficient information to safely draw that conclusion. However, a well-designed study will typically adhere to good reporting guidelines.

The CONSORT statement was developed to give guidance on best practice for reporting RCTs.<sup>(50)</sup> CONSORT includes a 25 item checklist of key characteristics (such as trial design, endpoints, technologies, blinding) that must be included in the reporting of an RCT. A trial that reports according to the checklist provides sufficient information to accurately gauge study quality. CONSORT is based on a standard two-group parallel study design, but variations are available for other randomised study designs. Standards are also available for the reporting of observational studies (STROBE) and the meta-analysis of observational studies (MOOSE).<sup>(51, 52)</sup>

While guidelines for reporting strive to improve standards, they do not provide an explicit means of assessing study quality. There are a number of systems available for grading the quality of evidence presented in a study, including GRADE<sup>(53)</sup> and the NHMRC Designation of Levels of Evidence.<sup>(54)</sup> The level of evidence is primarily driven by the study design with an RCT providing the best evidence. The GRADE system can also be applied to studies of diagnostic test accuracy where different considerations apply.<sup>(55)</sup>

There are also guidelines and standards available for the reporting of meta-analyses. The QUORUM statement lists the key features of a meta-analysis that should be clearly reported (for example, study selection, data abstraction, heterogeneity assessment).<sup>(56)</sup> The QUORUM statement has been revised to encompass advances in systematic reviews and is now known as PRISMA.<sup>(57)</sup> An equivalent standard, the

QUADAS statement, outlines the critical elements when reporting a meta-analysis of diagnostic test accuracy studies.<sup>(58)</sup> Network meta-analyses represent an additional level of complexity over and above head-to-head analyses, and a framework is available for evaluating the confidence in the results of a network meta-analysis.<sup>(59)</sup> Critical appraisal of meta-analyses is very important as a poorly conducted analysis can provide a very inaccurate estimate of treatment effect, irrespective of the quality of the data that has been used.

### 3.1.5 Heterogeneity

*Heterogeneity of treatment effect between studies must be assessed. Where significant heterogeneity is observed, attempts should be made to identify its causes. Substantial heterogeneity must be dealt with appropriately and may preclude a meta-analysis.*

It is assumed that the relative effectiveness of a treatment is the same across all studies included in a meta-analysis, that is, similarity of studies is assumed. If the results of the studies are very different then heterogeneity is observed and combining the results may not be appropriate.<sup>(60)</sup> Three broad forms of heterogeneity have been identified: statistical, where effect estimates vary more than expected by chance alone; clinical, due to differences in patient populations or study settings; and methodological, arising from differences in study design and analysis.<sup>(61)</sup> These three forms of heterogeneity are not mutually exclusive and will sometimes overlap.

It is possible to test for heterogeneity to provide evidence on whether or not the study results differ greatly or whether or not they are all measuring the same treatment effect. If studies agree on the direction of treatment effect, but disagree on the scale then it may still be possible to draw conclusions from a meta-analysis. However, if significant differences are observed in both the direction and scale of effect then it is unlikely that conclusions can be drawn from a meta-analysis.

Examples of common heterogeneity measures include  $I^2$  and Q statistics although the interpretation of these is subjective. These measures do not provide an optimal way to assess heterogeneity and, where significant heterogeneity is observed, it is critical to closely examine the studies being combined. Such an examination is typically based on a qualitative assessment of the studies in terms of study populations, endpoint measures and other study characteristics.

There can be many causes of heterogeneity such as variations in study design, study subjects, setting, geographic location, and endpoint measures. In some instances it will be possible to partially explain heterogeneity between studies by differences such as those listed above. Even if the variability can be explained, there must still be a decision as to whether or not to proceed with the meta-analysis and whether to

consider subgroup analyses. A subgroup analysis can involve including studies that are considered equivalent according to a more narrowly defined set of criteria (for instance, age range of study participants). It may also be possible to analyse a common subgroup of patients across studies based on a characteristic, for example age group, gender or disease risk.

The presentation of the results of a meta-analysis is frequently accompanied by a forest plot, also called a blobbogram, showing the treatment effect estimate of each individual study along with the pooled average. A forest plot provides a relatively simple means of visually assessing heterogeneity and study precision.

### 3.1.6 Meta-regression

*When there is significant between-study heterogeneity, meta-regression is a useful tool for identifying study-level covariates that modify the treatment effect.*

The interpretation of the results of a meta-analysis can become complicated when there is significant between-study heterogeneity. While it is possible to allow for the between-study variation by using a random effects meta-analysis, it can be useful to try and understand the sources of heterogeneity by using a method called meta-regression. This technique enables the incorporation of study characteristics that may help explain some of the observed heterogeneity into the meta-analysis. Meta-regression may lack power to detect significant study differences when there are few studies.

Meta-regression would generally be considered as part of a random effects model in that it is understood that covariates are required to explain differences between the studies, whereas a fixed effect meta-analysis typically presumes equivalence of the studies.

### 3.1.7 Fixed and random effects

*The choice between a fixed and random effects analysis is context specific. Heterogeneity should be assessed using standard methods. Significant heterogeneity suggests the use of a random effects model. Justification must be given for the choice of a fixed or random effects model.*

In fixed effect meta-analyses, the true effect of treatment is typically assumed to be the same in each study. Use of a fixed effect model therefore follows from the assumption that variability between studies is entirely due to chance. In a random effects meta-analysis, the treatment effect in each study is assumed to vary around

an overall average treatment effect.<sup>(62)</sup> As the random effects model assumes a different underlying effect for each study, it tends to result in wider confidence intervals than the fixed effect model.<sup>(60)</sup> When the reported effect sizes are homogeneous the fixed and random effects approaches yield very similar results. The choice between fixed and random effects models is context specific and the decision is often made following an assessment of heterogeneity. Although choosing between fixed and random effects models on the basis of heterogeneity is common practice, the scientific basis for this approach is unclear. Substantial heterogeneity suggests the use of a random effects model but also raises the question of whether the studies are actually comparable, sometimes referred to as comparing apples, oranges and pears. In analyses of sparse event data, as is common for adverse outcomes, it is common to use a fixed effect analysis possibly due to a lack of evidence of heterogeneity. The use of random effects has implications for the interpretation of results and the distribution of effect estimates should be discussed.<sup>(63)</sup> When random effects are used, it is strongly recommended that prediction intervals are reported in addition to the confidence bounds.<sup>(64)</sup> Confidence bounds indicate the precision of the estimate of average effect, whereas prediction intervals give bounds to the potential effect in an individual study setting.

A measure of heterogeneity should be reported to support the choice between a fixed and random effects model. Where there are few studies with small sample sizes, there may be limited ability to detect statistical heterogeneity. It should be noted that the absence of statistical heterogeneity does not imply a lack of clinical heterogeneity. A fixed effect model assumes that if all studies had infinitely large sample sizes then they would report precisely the same treatment effect. Due to differences in study populations and implementations of technologies, this assumption is unlikely to hold in most cases. As such, a random effects approach is usually justified. Where heterogeneity is present and a meta-analysis is justified, then use of a fixed effect model is not recommended. In such instances a fixed effect model should only be presented in special situations, such as few studies and strongly differing sample sizes. Where a fixed effect approach is used, the random effects analysis should also be presented.

Ideally, a review of clinical effectiveness is based on a pre-defined protocol, and the protocol should clearly state what steps will be taken to evaluate and address heterogeneity in the identified evidence. The protocol should define the conditions under which a fixed effect analysis would be considered appropriate.

### **3.1.8 Sources of bias**

*Attempts should be made to identify possible sources of bias such as publication bias, sponsorship bias and bias arising from the inclusion of poor*

*quality studies. Potential sources of bias must be reported along with steps taken to minimise the impact of bias.*

The issue of publication bias arises due to journals being more likely to publish studies showing beneficial effects of treatments, while equally valid studies showing no significant effect remain unpublished.<sup>(65)</sup> The consequence of this bias is that a meta-analysis may show a spurious significant effect. Publication bias may be detectable using funnel plots or regression methods, but these are not particularly powerful techniques.<sup>(66)</sup> Asymmetry in a funnel plot may indicate publication bias or it may be a reflection of how comprehensive the search strategy has been. The trim and fill technique can be used to adjust for observed publication bias.<sup>(67)</sup>

English language bias and citation bias are forms of publication bias in which studies with negative findings are more likely to appear in non-English language publications and are less likely to be cited, respectively. It is of critical importance that the search strategy element of the systematic review is as comprehensive as possible and that clinical trial registers are searched, where relevant. The presence of publication bias can affect any meta-analysis irrespective of the methodology used (that is to say, direct, indirect or network meta-analysis).

Bias may also be introduced where some studies are sponsored by the technology manufacturer. In such trials, there is a risk that the comparator technology may be applied in a sub-optimal manner to show the sponsor's treatment in a more favourable light. Published studies should be examined for stated conflict of interest or study funding that might indicate potential sponsorship bias.

Studies of diagnostic test accuracy are also subject to a variety of biases relating to the patients, the index test and the reference standard.<sup>(68)</sup> Spectrum bias, for example, relates to the observed spectrum of severity for the target condition. The study populations should be representative of the types of people who would normally be subject to the diagnostic test. Disease progression bias arises when the condition of patients changes between application of the index and reference tests. Depending on the type of bias and how it has arisen it will lead to over- or under-estimation of the diagnostic test accuracy. The inclusion of healthy control participants and the differential use of reference standards have both been shown to lead to over-estimation of the diagnostic test accuracy.<sup>(68)</sup>

A recognised structured risk of bias assessment tool should be used and reported as part of evidence synthesis. The tool should be appropriate to the design of the included studies. The interpretation of evidence synthesis should explicitly refer to risk of bias and the potential impact on the estimated treatment effect. In the event that the analysis included a subset of studies considered at low risk of bias, a

sensitivity analysis should be considered reporting treatment effect based only on studies at low risk of bias.

### 3.1.9 Frequentist and Bayesian approaches

*Both frequentist and Bayesian approaches are acceptable in meta-analysis. The approach taken must be clearly stated.*

There are two broad approaches to statistical inference: frequentist and Bayesian. Frequentists state that data are a repeatable random sample, and that parameters are constant during this repeatable process. Bayesians state that data are observed from the realised sample and that parameters are unknown and described by a probability distribution. In essence, for frequentists the parameters are fixed, whereas for Bayesians the data are fixed.<sup>(69)</sup>

A Bayesian approach incorporates prior information about the parameters of interest. The prior information is combined with the observed data to generate a posterior distribution for the parameters of interest. Prior information can come from a variety of sources such as previous studies or expert opinion. If there is no useful prior information then non-informative or vague priors are used. In the event of non-informative priors a Bayesian analysis typically generates results that are comparable to those from an equivalent frequentist analysis.

A key distinction between the two approaches is evident from the interpretation of their associated statistical interval estimates. In the frequentist approach, a 95% confidence interval means that in repeated samples the confidence interval will include the true parameter value 95% of the time. In the Bayesian approach, the 95% credible interval means that given the realised sample there is a 95% probability that the parameter value is in the interval.

Frequentist methods are common and have been widely implemented and applied. Bayesian methods have gained ground in recent years due to increased computing power and readily available software.

### 3.1.10 Outliers and influential studies

*Influential studies and those that are statistical outliers should be identified and reported. The methods used for identifying outliers must be clearly stated. Studies that are outliers should be characterised to determine if they are comparable to the other included studies.*

The results of a meta-analysis may be overly influenced or distorted by one or a small number of studies. Similarly, some studies may be outliers in a statistical sense. Outliers and influential studies are not synonymous: an outlier need not necessarily be influential and an influential study need not be an outlier. A first step is to visually inspect a forest plot to identify any unusual data points or where the pooled estimate appears to be driven by a single or small number of studies.

A variety of techniques are available to identify influential studies and potential outliers. These include metrics such as standardised residuals, Cook's distance, DFFITS and DFBETAS.<sup>(70)</sup> Sensitivity analysis techniques based on leave-one-out can be used to determine the impact of influential studies and outliers on the results of a meta-analysis. It is also useful to characterise outliers and gain an understanding of why they might be different from other studies.

### **3.1.11 Sensitivity analysis**

*If potential outliers have been identified, or if plausible subgroups of patients or studies have been identified, a comprehensive sensitivity analysis should be conducted. In a Bayesian analysis the choice of priors should be tested using a sensitivity analysis.*

The results of a meta-analysis can be sensitive to a variety of factors, but the choice of included studies is clearly critical. To test the effects of decisions about which studies to include or exclude, it is advisable to use sensitivity analyses. If potential outliers have been identified, then it is pertinent to examine the effect of excluding those studies from the analysis. Similarly it is useful to determine the impact of influential studies on the results.

If plausible subgroups have been identified, then it may be possible to carry out a separate meta-analysis for each subgroup. Subgroups can sometimes be identified based on patient characteristics such as age bands or disease risk. Alternatively, subgroups of trials may be identified according to study characteristics. For example, geographic region or study quality if measured using a recognised scale.

In many cases there may be a limited number of studies available for a meta-analysis. Clearly if there are limited data available then removing studies may not be feasible and it may not be possible to carry out a full or comprehensive sensitivity analysis.

In a Bayesian analysis, there are decisions relating to the choice of priors which may be informative or non-informative. Where there are informative priors it is important to test how the results compare to those using non-informative priors. However, in

the case of non-informative priors a variety of distributions are often available and the choice of distribution may impact on results.

### 3.2 Networks of evidence

*The network of available evidence should be described and used to guide the selection of the method of meta-analysis. The selection of direct and indirect evidence must be clearly defined. The exclusion of relevant evidence, either direct or indirect, should be highlighted and justified. Where direct and indirect evidence are combined, inconsistencies between the direct and indirect evidence must be assessed and reported.*

The studies available for a meta-analysis form a network of evidence. The most common comparison is between two technologies based on a number of head-to-head trials. Such a comparison is called a direct comparison. In cases where two treatments are compared, there is sometimes insufficient data available to reliably estimate the relative effectiveness of the two treatments in which case it may be possible to estimate the relative effectiveness using an indirect comparison.<sup>(71)</sup>

When there are no head-to-head trials, but two technologies can be compared based on a common comparator, it is possible to use indirect methods of meta-analysis. There are also approaches that allow direct and indirect evidence to be combined.

Depending on the method of comparison used, there may be restrictions on the type of networks that can be analysed. For direct comparisons only a standard pair-wise meta-analysis can be used.

Alternative networks can include: a star pattern in which two or more treatments have a common comparator (such as A-B, C-B, D-B); a ladder where treatment comparisons appear in a sequence (such as A-B, B-C, C-D); a closed loop in which there is both direct and indirect evidence (such as A-B, A-C, C-B); or a complex combination of patterns such as a closed loop with a star (see Figure 1).<sup>(72)</sup>

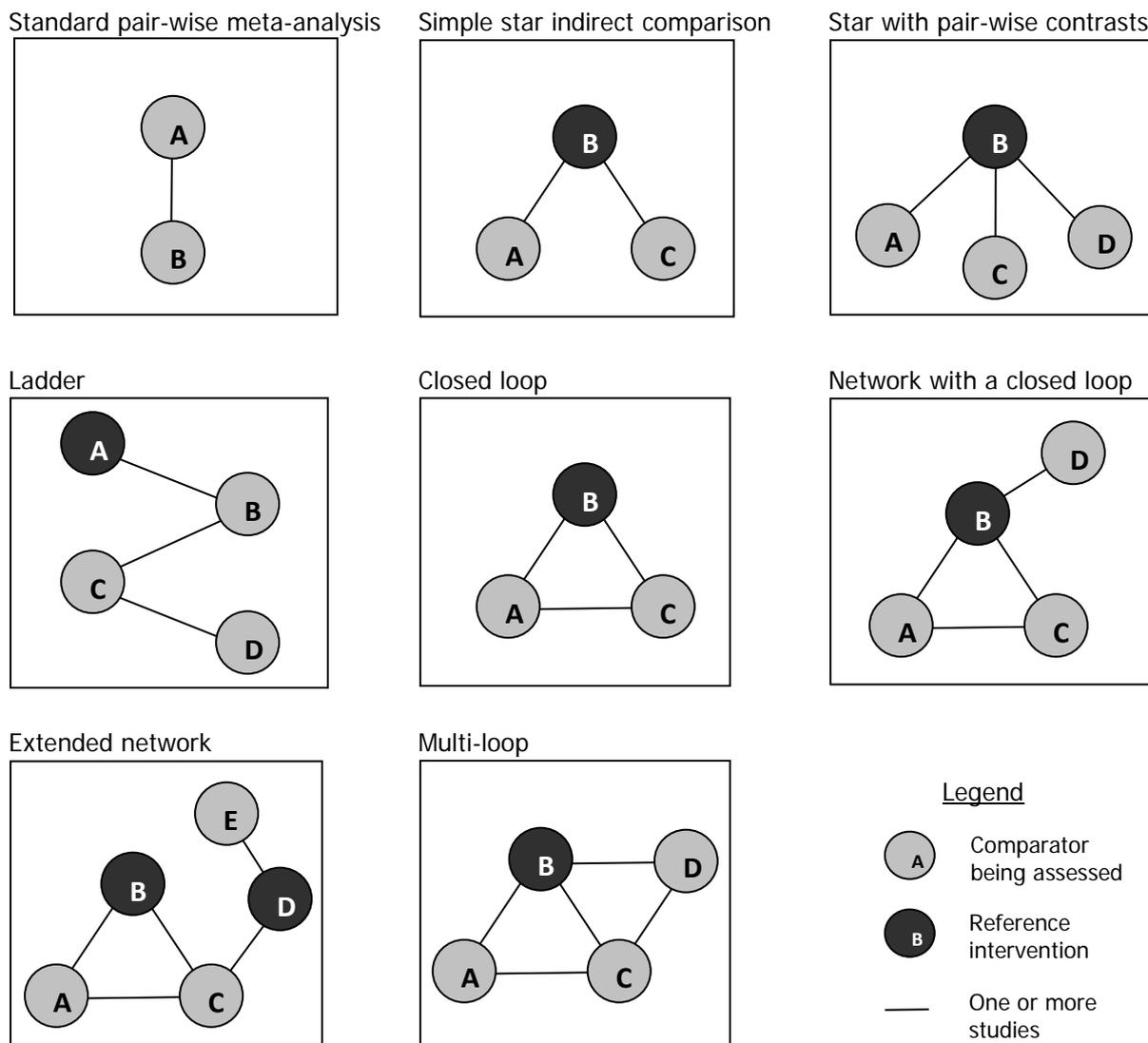
Direct comparisons involve a meta-analysis combining the results of multiple trials that all compare the treatment of interest to the same comparator (such as placebo). Standard meta-analytic techniques are applied for direct comparisons. The primary decision in direct comparisons relates to the choice between fixed and random effects meta-analysis.

Approaches to direct comparisons meta-analysis can be sub-divided into two methodologies: frequentist and Bayesian. The former are standard for direct comparisons primarily due to the ease of application and the variety of software packages available to apply them.

The need for indirect comparisons arises when treatments A and B are being compared, but only studies comparing A to C and B to C are available. By using a common comparator, in this case treatment C, an indirect comparison of treatments A and B can be carried out.

Placebo-controlled trials are commonly conducted in preference to head-to-head trials giving rise to the need for indirect comparisons when comparing two active treatments.<sup>(72)</sup> Depending on the amount of evidence available, indirect comparisons can sometimes make comparisons via two or more different paths. In comparing treatments A and B, the relative effectiveness should be similar whether derived via common comparator C or D. A statistically significant difference in the estimate of effectiveness would indicate inconsistency.

**Figure 1. Networks of evidence**



A network meta-analysis combines direct and indirect evidence to compare a technology to two or more other treatments. As a combination of direct and indirect evidence is used, these methods generally provide a measure of inconsistency or incoherence between the direct and indirect evidence. In network meta-analyses, consistency between direct and indirect evidence is assumed. That is, if direct evidence suggests that treatment A is better than treatment B, then that evidence should not be contradicted by the indirect evidence.

In the context of a multiple treatment comparison through network meta-analysis, there is an implicit assumption of transitivity, that is, it was equally likely that any patient in the network of studies could have been given any of the treatments in the network. In reality, there may be systematic differences between the trials in terms of patient characteristics such that the assumption of transitivity does not hold. In a network meta-analysis, transitivity should be investigated.

In a network meta-analysis involving both direct and indirect evidence, the evidence network can become very complex with many comparisons based on only one or two studies. With increasing complexity and greater numbers of treatments, the prospect of inconsistency increases. There is also a power trade-off between the number of pair-wise comparisons and the number of studies included in the analysis — too many comparisons with too few studies and the analysis may be underpowered to detect true differences.<sup>(73)</sup> Inconsistency should be measured within closed loops and reported as part of an analysis.

Where a combination of direct and indirect evidence is used, there may be options to increase the size of the evidence network to include comparators that are not strictly necessary to address the policy question. In these cases, consideration should be given to sensitivity analyses in which the evidence network is extended to include the additional comparators and to test the impact on the estimated treatment effect.

### 3.3 Selecting the method of comparison

*The choice of method of comparison depends on the quality, quantity and consistency of direct and indirect evidence. The available evidence must be clearly described along with a justification for the choice of method.*

When undertaking a meta-analysis a decision must be made regarding which method of comparison to use. A first question is to ask whether or not there is sufficient evidence to warrant combining data. There are no hard and fast rules to define 'sufficient evidence' in this context, but it is based on an evaluation of the quality, quantity and agreement of evidence. Substantial heterogeneity highlights where trials may not be measuring the same effect or where there may be systematic effect moderators.

The choice of method of comparison must take into account the network of evidence and the number of technologies being compared. When there is only direct evidence then the only questions relate to whether the studies should be combined and, if so, whether to use a frequentist or Bayesian approach. When indirect evidence is available then one must evaluate whether or not to include it and, if so, by which method. An important aspect in evaluating indirect evidence is whether or not it is in agreement with direct evidence. Disagreement between direct and indirect evidence must be fully investigated and it may preclude pooling data if the disagreement cannot be adequately explained. Certain networks of evidence limit the number of methods available but the researcher will often have some discretion as to how many comparisons to incorporate.

The network can be restricted to include the minimum number of comparisons required to enable an indirect comparison between the technologies of interest. Alternatively, it can be expanded to include as many relevant comparators as possible. It is important to ensure that the evidence network is comprehensive enough to accurately estimate the relative treatment effect of all relevant comparators. For example, by excluding a comparator that is not considered relevant to the policy question, a substantial amount of evidence may be excluded on the relevant comparators. Extending the evidence base can have a substantial impact on the precision of effect estimates.<sup>(74)</sup> When comparative evidence is available, either directly or indirectly, then non-comparative evidence should not be used.

Some questions that will assist in selecting the appropriate method of comparison include:

- Are there sufficient head-to-head trials for a direct comparison?
- Is there reliable indirect evidence available?
- If direct evidence has been excluded, why?
- If indirect evidence has been excluded, why?
- If indirect evidence is used, were all or a subset of available indirect comparisons used?
- For the indirect evidence, is there a single or multiple common comparators?

If more than one method is potentially appropriate in a given context, then the choice of method should be justified, with consideration being given to the possible impact of that choice on the outputs of the analysis.

In some cases it will not be advisable to carry out a meta-analysis. For example, a meta-analysis would not be recommended if there is unexplained heterogeneity across studies which may render an average treatment effect difficult to interpret or potentially misleading.<sup>(75)</sup> Omitting a meta-analysis does not negate the need to carry out a quality appraisal of the studies or to carry out data extraction. In the event that a meta-analysis is not recommended, an alternative is to summarise the available studies with a focus on the best quality evidence that is applicable in the context of the policy question being addressed. That is, there should be a focus on studies that are most applicable to the policy question in terms of target population, method of delivery of the technology, the comparator and the outcomes of interest.

Where the treatment effect is not pooled across studies, providing counts of studies with positive and negative results is not recommended.<sup>(75)</sup> Summary of findings tables should be used for key outcomes to describe and summarise treatment effect. If the studies are sufficiently similar in design and context, it may be possible to present a numerical summary using the range or inter-quartile range of treatment effects.

### 3.4 Methods of meta-analysis

*For any method of meta-analysis, all included trials must be sufficiently comparable and measuring the same treatment effect.*

A variety of meta-analysis methods are available depending on the type of evidence network being analysed. In any method of meta-analysis it is assumed that the relative effectiveness of a technology is the same across all trials used in the comparison. The assumption of constant efficacy requires all trials included in the analysis to be equivalent and attempting to measure the same treatment effect, that is, the results of one set of trials (A vs. B) should be generalisable to the other set of trials (A vs. C). Determining whether the assumption of generalisability holds is a subjective assessment based on a detailed review of the included studies in both comparisons. It should be considered whether the sets of studies treating the same indications were in comparable populations and if they were applying the common treatment in the same manner (for instance, dosing and frequency).

This section looks at different methods of meta-analysis. For each method of meta-analysis there is a brief description, some examples of published meta-analyses using that method, a brief note on usage in the literature, the strengths and limitations of the methodology and then some critical questions that should be asked when considering that method of meta-analysis.

Methods of direct, unadjusted and adjusted indirect meta-analysis are treated as distinct methodologies. The structure is partly influenced by the chronology of new methodologies to address the issue of indirect comparisons. It should be recognised that analyses that incorporate any degree of indirect evidence or consider multiple treatments simultaneously are now usually analysed using one of a number of approaches referred to as network meta-analysis.

#### 3.4.1 Direct meta-analysis

*Direct meta-analysis should be used when there are sufficient comparable head-to-head studies available. If indirect evidence is available then consideration should also be given to a multiple treatment comparison.*

#### Description

Direct meta-analysis is used for combining head-to-head trials. The methods available for direct comparison meta-analysis are divided into fixed and random effects methods. The confidence intervals around the pooled random effects estimate tend to be wider than would be observed in the fixed effect meta-analysis.

Bayesian methods for direct comparisons meta-analysis are analogous to frequentist methods with the primary distinction being the use of prior distributions for the mean of the overall estimate, the means of the individual estimates of each study, and the between-study variance (for random effects models).<sup>(76)</sup> The use of non-informative priors will generally result in effect estimates that are comparable to those in a frequentist approach. However, in some instances it may be appropriate to form informative priors by way of other data, such as expert opinion, which are likely to generate results that may differ to those from a frequentist approach.

For certain endpoints, such as rate ratios, a study with zero cases can be problematic. The common solution to this problem is to apply a continuity correction by adding a constant (typically 0.5) to the number of cases. The use of a continuity correction can impact on the significance and interpretation of results.<sup>(16)</sup>

### **Examples**

- the safety and efficacy of carotid endarterectomy versus carotid artery stenting in the treatment of carotid artery stenosis<sup>(77)</sup>
- aprotinin compared to tranexamic acid in cardiac surgery<sup>(78)</sup>
- the impact of omega-3 fatty acids on mortality and restenosis in high risk cardiovascular patients<sup>(79)</sup>
- adjunctive thrombectomy for acute myocardial infarction<sup>(80)</sup>
- double versus single stenting for coronary bifurcation lesions.<sup>(81)</sup>

### **Usage**

The application of Bayesian methods for direct meta-analysis is uncommon primarily because of the greater complexity in computing the models, and the fact that the results tend to be quite similar to those obtained using standard frequentist methods.

### **Strengths**

The methods for direct meta-analysis are well described and can be implemented in a wide variety of software packages. Analyses can be easily reproduced if the underlying data are available.

The strength of Bayesian approaches in this context is that they can incorporate data from a wide variety of sources and can, for example, use expert opinion to elicit useful information. Rather than computing confidence intervals, a Bayesian meta-analysis computes a credible interval which has a different interpretation. A Bayesian approach allows the computation of the probability that one treatment is better than another, which is useful to decision makers.

## Limitations

Direct meta-analysis requires head-to-head trials to compare two technologies. For some treatments it is becoming increasingly difficult to obtain sufficient studies to enable a direct comparison.

A common criticism of Bayesian techniques rests on the use of priors for key parameters. Critics of the Bayesian approach suggest that priors are subjectively chosen. In reality, most Bayesian analyses employ vague or non-informative priors. However, even with a non-informative prior, assumptions are made about the distribution of that prior and often there are alternative formulations available so the use of sensitivity analysis is important.<sup>(82)</sup>

## Critical questions

- Are the studies comparable?
- Has heterogeneity been assessed?
- Was the choice between fixed and random effects clearly justified?
- In a Bayesian analysis, how were the priors defined and were alternatives tested?

### 3.4.2 Unadjusted indirect comparison

*The method of unadjusted indirect comparisons should not be used.*

## Description

Unadjusted indirect comparisons combine study data as though they had come from a single large trial.<sup>(71)</sup> A weighted summary effect is computed for all study arms involving treatment A and is compared to a weighted summary effect for all study arms including treatment B. The relative effectiveness of treatment A is compared to treatment B using the two summary effects. This method is called an 'unadjusted indirect comparison' because the indirect comparison does not adjust for events in the control group.<sup>(83)</sup>

## Examples

- rectal corticosteroids versus alternative treatments in ulcerative colitis<sup>(84)</sup>
- effectiveness of anticoagulant or platelet anti-aggregant treatment for stroke prevention in patients at elevated risk for stroke<sup>(85)</sup>
- the effects of non-steroidal anti-inflammatory drugs on blood pressure.<sup>(86)</sup>

## Usage

The application of unadjusted indirect comparisons is very unusual.<sup>(87)</sup> Given the shortcomings of the method it would be difficult to publish an analysis using this methodology.

## Limitations

Although unadjusted indirect methods provide a simple and easily implemented method of calculating relative effectiveness in the absence of head-to-head evidence, the primary flaw of this approach is that it ignores the randomised nature of individual trials. When compared to direct estimates, unadjusted direct comparisons result in a large number of discrepancies in the significance and direction of relative effectiveness.<sup>(87)</sup> Although unbiased, this method yields unpredictable results and is flawed by not acknowledging randomisation. As such this method of indirect comparison should not be used.

## Critical questions

- Why was an unadjusted indirect comparison used rather than an adjusted indirect method?
- How would the results have differed if an adjusted indirect comparison had been applied?

### 3.4.3 Adjusted indirect comparison

*Adjusted indirect comparison is appropriate for comparing two technologies using a common comparator.*

## Description

Bucher *et al.* presented an adjusted indirect method of treatment comparison that can estimate relative treatment effects for star pattern networks.<sup>(88)</sup> This method is based on the odds ratio as the measure of treatment effect, although it can be trivially extended for other measures.<sup>(72)</sup> This method is intended for situations where there is no direct evidence (such as comparing treatments A and B, but the only evidence is through comparison with C). Certain more complex networks including closed loops can be analysed, but only in the form of pair-wise comparisons.

As the method assumes independence between the pair-wise comparisons, it cannot readily be applied to multi-arm trials where this assumption fails. In a multi-armed trial it is expected that the treatment effect will be correlated between arms.

## Examples

- Effectiveness of gemcitabine-based combinations compared to single agent gemcitabine in patients with advanced pancreatic cancer.<sup>(89)</sup>
- Effectiveness of nifedipine compared to atosiban for tocolysis in preterm labour.<sup>(90)</sup>
- Comparison of pravastatin, simvastatin, and atorvastatin for cardiovascular disease prevention.<sup>(91)</sup>

## Usage

Although initially popular, Bucher's adjusted indirect comparison method is gradually being replaced by other methods, particularly network meta-analysis.<sup>(83, 92)</sup>

## Strengths

This method is relatively simple to implement and superior to an unadjusted indirect comparison. It is possible to combine pooled estimates of direct and indirect evidence using inverse variance weights as in a standard meta-analysis.<sup>(76)</sup>

## Limitations

This method is applied in the absence of any direct evidence and can only be used in more complex networks of evidence in the form of pair-wise comparisons.

This method is not appropriate when using data derived from multi-arm trials.

## Critical questions

- Does the analysis include or exclude multi-arm trials?
- Is direct evidence available that could be incorporated into a network meta-analysis?

### 3.4.4 Network meta-analysis

*A network meta-analysis can be appropriate for comparing multiple treatments when both direct and indirect evidence are available.*

## Description

The method of network meta-analysis (NMA) first proposed by Lumley allows the combination of both direct and indirect evidence.<sup>(93)</sup> This methodology requires the data to contain a closed loop structure. Depending on the complexity of the closed loop design, it is generally possible to compute relative effectiveness by a number of routes. It is possible to compute the amount of agreement between the results

obtained when different linking treatments are used. This agreement forms the basis of an incoherence measure which is used to estimate the consistency of the network paths. Incoherence is used to compute the 95% confidence interval for the indirect comparison. It is assumed that the comparison between two treatments will occur through a closed loop. The measure of incoherence, which is an integral part of the calculation, requires a closed loop.

Network meta-analysis can also be undertaken within a Bayesian framework, sometimes also referred to as Bayesian mixed treatment comparison (MTC). This form of NMA is a generalisation of standard pair-wise meta-analysis for A vs. B trials to more complex networks of evidence.<sup>(94)</sup> Any combination of studies can be combined as long as every study is connected to at least one other study. Both direct and indirect evidence can be combined and there is no restriction to the number of arms in any given trial. Bayesian NMA facilitates simultaneous inference about all of the treatments included in the analysis, allowing estimation of effect estimates for all pair-wise comparisons and for treatments to be ranked according to relative effectiveness. Bayesian NMA can incorporate meta-regression enabling the addition of study-level covariates as a means to reduce inconsistency although this adaptation has implications for compromised power.<sup>(63, 73)</sup> Being a Bayesian approach, there is scope for defining informative priors. While priors may be legitimately generated, it is critical that they are credible and clearly justified.

## **Examples**

- effects of glucosamine, chondroitin, or placebo in patients with osteoarthritis of hip or knee<sup>(95)</sup>
- efficacy and tolerability of second-generation antidepressants in social anxiety disorder<sup>(96)</sup>
- comparison of common antiplatelet regimens after transient ischaemic attack or stroke<sup>(97)</sup>
- the relative efficacy of existing treatments and combinations to reduce the risk for COPD exacerbations<sup>(98)</sup>
- the efficacy and safety of bronchodilators and steroids, alone or combined, for the acute management of bronchiolitis in children under two years<sup>(99)</sup>
- the effectiveness of psychological interventions compared to usual care in coronary heart disease.<sup>(100)</sup>

A worked example of the output of a network meta-analysis is included in Appendix A.

## Usage

This method can be implemented in a variety of software packages and can be estimated using both frequentist and Bayesian approaches. Network meta-analysis has gained popularity over the last decade and is widely used.

## Strengths

Network meta-analysis generates an adjusted indirect treatment comparison that partially preserves the randomisation of study groups in the included trials.

This method simultaneously combines direct and indirect evidence, and provides an estimate of the agreement between different results. Direct evidence is not required for this methodology.

Bayesian approaches are becoming increasingly popular due to its versatility and the greater availability of indirect compared to direct evidence. Although forest plots are often presented, the studies have to be grouped by comparator given that multiple comparisons are shown. As this method can be applied to very complex networks it allows more evidence to be incorporated into an analysis. All of the technologies included can be ranked according to the probability that they are the best treatment. While Bayesian approaches are complex, the outputs are directly interpretable and lend themselves to incorporation into a cost-effectiveness model.

Network meta-analysis pools the effect estimates across trials rather than individual treatment groups. Multi-arm trials can be included and the correlations between arms are taken into account.

## Limitations

Network meta-analysis does not automatically account for correlations that may exist between different effect estimates when they are obtained from a single multi-armed trial. In trials with more than two treatment arms, estimates of relative treatment effects will be correlated due to the structure of the network of evidence. For example, in a multi-arm placebo-controlled trial the comparison between any two treatments will be correlated with the comparison of each of those treatments with placebo. Accounting for this correlation is possible using a random effects model but this is not considered to be an optimal solution.<sup>(72)</sup> A commonly implemented approach to dealing with correlation was published by the NICE Decision Support Unit in the UK.<sup>(101)</sup>

Bayesian approaches can be complex and do not lend themselves to easy application or interpretation. The Bayesian framework requires an in-depth understanding, particularly with regard to model checking and the definition of priors.

The key strength that more evidence can be utilised can also represent a weakness as it may be difficult to define limits for the network of evidence, particularly for an indication that has a wide range of treatment options available. Also, in a complex network there may be very little evidence for many of the comparisons.

### Critical questions

- How was it decided which comparisons should be included?
- What model was used and is it appropriate for the data?
- How were the priors defined?
- Has the assumption of transitivity been investigated and reported?
- Is there evidence of inconsistency and, if so, has it been explained?
- Does the analysis include multi-arm trials?
- Were multiple paths available to compare two treatments and were they all used to test the consistency of results?

#### 3.4.5 Meta-analysis of diagnostic test accuracy studies

*The bivariate random effects and hierarchical summary receiver operating characteristic models (HSROC) should be used for pooling sensitivity and specificity from diagnostic and screening test accuracy studies. The correlation between sensitivity and specificity should be reported.*

### Description

Diagnostic accuracy studies measure the level of agreement between the results of the test under evaluation and that of the reference standard. The primary endpoint is binary (that is, a positive or negative test result) and is recorded for both the diagnostic test being assessed and the reference standard. Diagnostic test accuracy is most often represented by sensitivity and specificity.

Meta-analysis of diagnostic test accuracy studies have traditionally used the summary Receiver Operating Characteristic (sROC) curve approach whereby sensitivity and specificity are converted into a single measure called the diagnostic odds ratio.<sup>(102)</sup> Pooled estimates of sensitivity and specificity can be derived from the sROC curve. However, such an approach ignores the fact that sensitivity and specificity are often correlated.

Bivariate random effects models have come to the fore more recently and take into account any observed correlation between sensitivity and specificity.<sup>(103)</sup> Another method, the hierarchical sROC (HSROC), generates equivalent results in the absence of covariates. By analysing these two parameters and producing a summary estimate

of each, it is possible to determine whether the test is better for ruling in or ruling out a particular diagnosis.

Likelihood ratios can also be computed which are of more use to clinicians as they quantify the extent to which a test result changes the probability of disease. If there is no correlation between sensitivity and specificity then it may be more appropriate to carry out separate univariate analyses to pool sensitivities and specificities.<sup>(104)</sup>

It is possible to carry out a meta-analysis of diagnostic test accuracy adjusting for different test thresholds used within and across studies.<sup>(105)</sup> Methods of network meta-analysis have been extended to enable an NMA of diagnostic test accuracy using the bivariate random effects approach, and these also facilitate the inclusion of data on multiple test thresholds.<sup>(106)</sup>

### **Examples**

- diagnostic accuracy of natriuretic peptides and ECG in the diagnosis of left ventricular systolic dysfunction<sup>(107)</sup>
- diagnostic accuracy of rectal bleeding in combination with other symptoms, signs and tests in relation to colorectal cancer<sup>(108)</sup>
- diagnostic accuracy of FDG PET for the characterization of adrenal masses.<sup>(109)</sup>

### **Usage**

The Moses-Littenberg approach is still common for pooling diagnostic test accuracy studies. The bivariate random effects and HSROC methods have gained popularity. It is recommended that the bivariate random effects and HSROC models should be used although they often generate similar results to the traditional techniques.<sup>(110, 111)</sup>

### **Strengths**

By combining the results of several diagnostic test accuracy studies, it is possible to determine the typical performance of the test. If the threshold for a positive test can be varied, it is possible to determine the performance under different thresholds.

The bivariate random effects and HSROC model the distribution of pairs of sensitivity and specificity from each study. These models give valid estimates of the average sensitivity and specificity and can be extended to include covariates that may explain between-study heterogeneity.

## Limitations

The reference standard test itself may not be an accurate measure of disease. This can arise due to a poor choice or because of inconsistent application of the reference standard.

Test accuracy can vary between patient subgroups, disease spectrum, clinical setting, or with the test interpreters and may depend on the results of previous testing.<sup>(68)</sup> Failure to ensure that the included studies are fully equivalent will lead to a biased estimate of test accuracy. The Moses-Littenberg approach fails to consider the precision of the study estimates and does not estimate between-study heterogeneity.

## Critical questions

- Do the various studies being pooled use the same threshold for a positive test result?
- Do all the studies use the same reference standard?
- Has the correlation between sensitivity and specificity been reported?

### 3.4.6 Generalised linear mixed models

*Generalised linear mixed models can be appropriate when analysing individual patient data from trials.*

## Description

Regression techniques can be used to combine trial data to evaluate relative effectiveness. Where the primary endpoint is binary and data are available in the form of 2×2 frequency tables for each study, logistic regression can be used.

Generalised linear mixed models (GLMMs) have also been proposed as a method of combining trial data in a regression framework.<sup>(87)</sup> The application of GLMMs to continuous or time-to-event endpoints requires individual level patient data. The advantage of regression techniques is the potential for including study level covariates that may be used to explain heterogeneity in the measured effects. Although not restricted to meta-analysis of individual patient data, it is the application where GLMMs offer the greatest advantage over other techniques.

GLMMs can offer benefits for meta-analysis as they do not have to be restricted to the within-study normal distribution assumption. They can be extended to other distributions which may be more appropriate particularly for rare event data.<sup>(112)</sup>

## Examples

- comparison of low-molecular-weight heparin to unfractionated heparin for the treatment of pulmonary embolism and deep vein thrombosis.<sup>(113)</sup>

## Usage

The application of GLMMs to meta-analysis is relatively rare as less complex techniques for direct meta-analysis and meta-regression are sufficient in most applications. Given the difficulties in obtaining individual patient data, it is unlikely that the advantages of GLMMs will be realised.

## Strengths

GLMMs can be applied through most of the leading statistical software packages. It is possible to use exact rather than approximate likelihood approaches by using GLMMs.<sup>(112)</sup> They are versatile and can be applied to network meta-analysis (see Section 3.4.4).

## Limitations

Individual level patient data can be very difficult if not impossible to obtain.

In many cases GLMMs may not offer substantial advantages over other methods that can be applied more easily.

## Critical questions

- If individual patient data are used, have data from an adequate number of studies been included?

### 3.4.7 Confidence profile method

*The confidence profile method can be used to combine direct and indirect evidence. Network meta-analysis or Bayesian mixed treatment comparison should be used in preference to the confidence profile method. The use of this method over other available methods should be justified.*

## Description

The confidence profile method provides a general framework for undertaking multi-parameter meta-analysis.<sup>(114)</sup> As well as incorporating trials with different treatment comparisons, it can encompass different designs, outcomes and measures of effect. The confidence profile method also allows explicit modelling of biases. Although this method is typically implemented as a fully Bayesian model, it can be formulated

without prior distributions and fitted using maximum likelihood methods.<sup>(87)</sup> Where there is direct and indirect evidence available, cross-validators predictive checking can be used to determine evidence consistency.<sup>(72)</sup> If different doses of the same drug treatment were studied, looking at dose-response relationships can also provide cross-validators information, provided the trial populations are comparable. The models available for this methodology are based on fixed-treatment effects although both fixed and random study-effects are possible.

## **Examples**

- the efficacy of the ketogenic diet in reducing seizure frequency for children with refractory epilepsy<sup>(115)</sup>
- the efficacy of antibiotics for patients undergoing tube thoracostomy<sup>(116)</sup>
- the efficacy and complications of cervical spine manipulation and mobilisation for the treatment of neck pain and headache.<sup>(117)</sup>

## **Usage**

The confidence profile method of meta-analysis never entered common usage and has been replaced by other methods of indirect and multiple treatment comparison.<sup>(114)</sup>

## **Strengths**

This method preserves the randomised nature of RCT data. The appropriateness of combining direct and indirect evidence can be assessed using various model-checking statistics.

## **Limitations**

The models for the confidence profile method are relatively complex which has partly restricted their diffusion into general use. Although improvements in computing power and software now make these models more feasible, other methodological developments have come to the fore (such as Bayesian mixed treatment comparison).

When there is no direct evidence the cross-validators predictive checking cannot be carried out to determine whether or not the selected studies can be validly combined in an indirect comparison.

## **Critical questions**

- Does the analysis combine direct and indirect evidence?
- Has cross-validators predictive checking been carried out?
- If there is direct evidence, is it consistent with the indirect evidence?

## 4 References

1. European network for Health Technology Assessment. Online [Internet]. 2014 7/9/2014. Available from: <http://www.eunetha.eu/faq/Category%201-0#t287n73>.
2. Pharmaceutical Benefits Advisory Committee. Guidelines for preparing submissions to the Pharmaceutical Benefits Advisory Committee (Version 4.3). Canberra: 2008 2008. Report No.
3. Shaw LJ, Iskandrian AE, Hachamovitch R, Germano G, Lewin HC, Bateman TM, et al. Evidence-Based Risk Assessment in Noninvasive Imaging. *The Journal of Nuclear Medicine*. 2001;42(9):1424-36.
4. Replogle WH, Johnson WD. Interpretation of absolute measures of disease risk in comparative research. *Fam Med*. 2007;39(6):432-5.
5. Smeeth L, Haines A, Ebrahim S. Numbers needed to treat derived from meta-analyses--sometimes informative, usually misleading. *BMJ*. 1999;318(7197):1548-51.
6. Akobeng AK. Understanding measures of treatment effect in clinical trials. *Arch Dis Child*. 2005;90(1):54-6.
7. Higgins JPT, Green S, (editors). *Cochrane handbook for systematic reviews of interventions*. Version 5.1.0 ed: The Cochrane Collaboration; 2011 2011.
8. Jaeschke R, Guyatt G, Shannon H, Walter S, Cook D, Heddle N. Basic statistics for clinicians: 3. Assessing the effects of treatment: measures of association. *CMAJ*. 1995;152(3):351-7.
9. Moher D, Hopewell S, Schulz KF, Montori V, Gotzsche PC, Devereaux PJ, et al. CONSORT 2010 Explanation and Elaboration: updated guidelines for reporting parallel group randomised trials. *J Clin Epidemiol*. 2010.
10. Berger ML, Mamdani M, Atkins D, Johnson ML. Good Research Practices for Comparative Effectiveness Research: Defining, Reporting and Interpreting Nonrandomized Studies of Treatment Effects Using Secondary Data Sources: The ISPOR Good Research Practices for Retrospective Database Analysis Task Force Report - Part I. *Value in Health*. 2009;12(8):1044-52.
11. Barnes SA, Mallinckrodt CH, Lindborg SR, Carter MK. The impact of missing data and how it is handled on the rate of false-positive results in drug development. *Pharmaceutical Statistics*. 2008;7(3):215-25.
12. Siddiqui O, Hung HMJ, O'Neill R. MMRM vs. LOCF: A Comprehensive Comparison Based on Simulation Study and 25 NDA Datasets. *Journal of Biopharmaceutical Statistics*. 2009;19(2):227-46.
13. Knottnerus JA, Muris JW. Assessment of the accuracy of diagnostic tests: the cross-sectional study. In: Knottnerus JA, editor. *The evidence base of clinical diagnosis*. London: BMJ Books; 2002. p. 39-59.
14. Goetz CG, Poewe W, Rascol O, Sampaio C, Stebbins GT, Fahn S, et al. The Unified Parkinson's Disease Rating Scale (UPDRS): Status and Recommendations. *Movement Disorders*. 2003;18(7):738-50.
15. Jüni P, Altman DG, Egger M. Assessing the quality of randomised controlled trials. In: Egger M, Smith GD, Altman DG, editors. *Systematic reviews in*

- health care: meta-analysis in context. London: BMJ Publishing Group; 2011. p. 87-108.
16. Diamond GA, Bax L, Kaul S. Uncertain Effects of Rosiglitazone on the Risk for Myocardial Infarction and Cardiovascular Death. *Annals of Internal Medicine*. 2007;147(8):578-W162.
  17. Rothwell PM. Factors that can affect the external validity of randomised controlled trials. *PLoS clinical trials*. 2006;1(1):e9.
  18. Lu CY. Observational studies: a review of study designs, challenges and strategies to reduce confounding. *Int J Clin Pract*. 2009;63(5):691-7.
  19. Altman DG. *Practical statistics for medical research*. London: Chapman & Hall; 1991 1991.
  20. Latimer NR. Survival Analysis for Economic Evaluations Alongside Clinical Trials - Extrapolation with Patient-Level Data: Inconsistencies, Limitations, and a Practical Guide. *Medical Decision Making*. 2013;33(6):743-54.
  21. Davies C, Briggs A, Lorgelly P, Garellick G, Malchau H. The "hazards" of extrapolating survival curves. *Med Decis Making*. 2013;33(3):369-80.
  22. Altman DG, Bland JM. Time to event (survival) data. *Bmj*. 1998;317(7156):468-9.
  23. Chan AW, Hrobjartsson A, Haahr MT, Gotzsche PC, Altman DG. Empirical evidence for selective reporting of outcomes in randomized trials: comparison of protocols to published articles. *JAMA*. 2004;291(20):2457-65.
  24. O'Neill RT. Secondary endpoints cannot be validly analyzed if the primary endpoint does not demonstrate clear statistical significance. *Controlled Clinical Trials*. 1997;18(6):550-6.
  25. Bekkering GE, Kleijnen J. Procedures and methods of benefit assessments for medicines in Germany. *The European Journal of Health Economics*. 2008;9(Supplement 1):5-29.
  26. Sun X, Ioannidis JP, Agoritsas T, Alba AC, Guyatt G. How to use a subgroup analysis: users' guide to the medical literature. *Jama*. 2014;311(4):405-11.
  27. Sun X, Briel M, Busse JW, You JJ, Akl EA, Mejza F, et al. The influence of study characteristics on reporting of subgroup analyses in randomised controlled trials: systematic review. *BMJ*. 2011;342.
  28. Mancía G, Grassi G. Efficacy of antihypertensive treatment: which endpoints should be considered? *Nephrology Dialysis Transplantation*. 2005;20(11):2301-3.
  29. Health Information and Quality Authority. *Guidelines for Stakeholder Engagement in Health Technology Assessment in Ireland*. Dublin: 2014 2014. Report No.
  30. Rowen D, Brazier J, Roberts J. Mapping SF-36 onto the EQ-5D index: how reliable is the relationship? *Health Qual Life Outcomes*. 2009;7:27.
  31. Rabin R, de CF. EQ-5D: a measure of health status from the EuroQol Group. *Ann Med*. 2001;33(5):337-43.
  32. McHorney CA, Ware JE, Jr., Raczek AE. The MOS 36-Item Short-Form Health Survey (SF-36): II. Psychometric and clinical tests of validity in measuring physical and mental health constructs. *Med Care*. 1993;31(3):247-63.
  33. Bellamy N, Buchanan WW, Goldsmith CH, Campbell J, Stitt LW. Validation study of WOMAC: a health status instrument for measuring clinically important

- patient relevant outcomes to antirheumatic drug therapy in patients with osteoarthritis of the hip or knee. *J Rheumatol*. 1988;15(12):1833-40.
34. Fleming TR, DeMets DL. Surrogate end points in clinical trials: are we being misled? *Annals of Internal Medicine*. 1996;125(7):605-13.
  35. Chi GYH. Some issues with composite endpoints in clinical trials. *Fundamental & Clinical Pharmacology*. 2005;19:609-19.
  36. Cordoba G, Schwartz L, Woloshin S, Bae H, Gøtzsche PC. Definition, reporting, and interpretation of composite outcomes in clinical trials: systematic review. *BMJ*. 2010;341.
  37. Kleist P. Composite Endpoints for Clinical Trials: Current Perspectives. *International Journal of Pharmaceutical Medicine*.21(3).
  38. Lim E, Brown A, Helmy A, Mussa S, Altman DG. Composite Outcomes in Cardiovascular Research: A Survey of Randomized Trials. *Annals of Internal Medicine*. 2008;149(9):612-7.
  39. Li X, Caffo BS. Comparison of Proportions for Composite Endpoints with Missing Components. *Journal of Biopharmaceutical Statistics*. 2011;21(2):271-81.
  40. Vandenbroucke JP, Psaty BM. Benefits and Risks of Drug Treatments. *JAMA: The Journal of the American Medical Association*. 2008;300(20):2417-9.
  41. Whiting P, Rutjes AWS, Reitsma JB, Glas AS, Bossuyt PMM, Kleijnen J. Sources of Variation and Bias in Studies of Diagnostic Accuracy. *Annals of Internal Medicine*. 2004;140(3):189-202.
  42. Egger M, Smith GD, O'Rourke K. Rationale, potentials, and promise of systematic reviews. In: Egger M, Smith GD, Altman DG, editors. *Systematic reviews in health care: meta-analysis in context*. London: BMJ Publishing Group; 2001. p. 3-19.
  43. Egger M, Smith GD. Principles of and procedures for systematic reviews. In: Egger M, Smith GD, Altman DG, editors. *Systematic reviews in health care: meta-analysis in context*. London: BMJ Publishing Group; 2001. p. 23-42.
  44. Health Information and Quality Authority. *Guidelines for the Economic Evaluation of Health Technologies in Ireland*. Dublin, Ireland: HIQA, 2018.
  45. Simmonds MC, Higgins JPT, Stewart LA, Tierney JF, Clarke MJ, Thompson SG. Meta-analysis of individual patient data from randomized trials: a review of methods used in practice. *Clinical Trials*. 2005;2(3):209-17.
  46. Riley RD, Lambert PC, Abo-Zaid G. Meta-analysis of individual participant data: rationale, conduct, and reporting. *BMJ*. 2010;340.
  47. Riley RD, Lambert PC, Staessen JA, Wang J, Gueyffier F, Thijs L, et al. Meta-analysis of continuous outcomes combining individual patient data and aggregate data. *Statistics in Medicine*. 2008;27(11):1870-93.
  48. Sibbald B, Roland M. Understanding controlled trials. Why are randomised controlled trials important? *Bmj*. 1998;316(7126):201.
  49. 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;38(2):200-11.
  50. Begg C, Cho M, Eastwood S, Horton R, Moher D, Olkin I, et al. Improving the Quality of Reporting of Randomized Controlled Trials. *JAMA: The Journal of the American Medical Association*. 1996;276(8):637-9.

51. Stroup DF, Berlin JA, Morton SC, Olkin I, Williamson GD, Rennie D, et al. Meta-analysis of Observational Studies in Epidemiology. *JAMA: The Journal of the American Medical Association*. 2000;283(15):2008-12.
52. Vandembroucke JP, Elm Ev, Altman DG, Gøtzsche PC, Mulrow CD, Pocock SJ, et al. Strengthening the Reporting of Observational Studies in Epidemiology (STROBE): Explanation and Elaboration. *Annals of Internal Medicine*. 2007;147(8):W-163.
53. GRADE Working Group. Grading quality of evidence and strength of recommendations. *BMJ*. 2004;328(7454):1490.
54. Merlin T, Weston A, Tooher R. Extending an evidence hierarchy to include topics other than treatment: revising the Australian 'levels of evidence'. *BMC Medical Research Methodology*. 2009;9(1):34.
55. Schünemann HJ, Oxman AD, Brozek J, Glasziou P, Jaeschke R, Vist GE, et al. Grading quality of evidence and strength of recommendations for diagnostic tests and strategies. *BMJ*. 2008;336(7653):1106-10.
56. Moher D, Cook DJ, Eastwood S, Olkin I, Rennie D, Stroup DF. Improving the quality of reports of meta-analyses of randomised controlled trials: the QUOROM statement. *The Lancet*. 1999;354(9193):1896-900.
57. Moher D, Liberati A, Tetzlaff J, Altman DG, Group TP. Preferred Reporting Items for Systematic Reviews and Meta-Analyses: The PRISMA Statement. *PLoS Med*. 2009;6(7):e1000097.
58. Whiting P, Rutjes A, Reitsma J, Bossuyt P, Kleijnen J. The development of QUADAS: a tool for the quality assessment of studies of diagnostic accuracy included in systematic reviews. *BMC Medical Research Methodology*. 2003;3(1):25.
59. Salanti G, Del Giovane C, Chaimani A, Caldwell DM, Higgins JP. Evaluating the quality of evidence from a network meta-analysis. *PloS one*. 2014;9(7):e99682.
60. Egger M, Smith GD, Phillips AN. Meta-analysis: Principles and procedures. *BMJ*. 1997;315:1533-7.
61. Ades AE, Sculpher M, Sutton A, Abrams K, Cooper N, Welton N, et al. Bayesian Methods for Evidence Synthesis in Cost-Effectiveness Analysis. *PharmacoEconomics*. 2006;24(1):1-19.
62. Deeks JJ, Altman DG, Bradburn MJ. Statistical methods for examining heterogeneity and combining results from several studies in meta-analysis. In: Egger M, Smith GD, Altman DG, editors. *Systematic reviews in health care: meta-analysis in context*. London: BMJ Publishing Group; 2001. p. 285-312.
63. Sutton A, Ades AE, Cooper N, Abrams K. Use of Indirect and Mixed Treatment Comparisons for Technology Assessment. *PharmacoEconomics*. 2008;26(9):753.
64. Higgins JPT, Thompson SG, Spiegelhalter DJ. A re-evaluation of random-effects meta-analysis. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*. 2009;172(1):137-59.
65. Egger M, Smith GD, Schneider M, Minder C. Bias in meta-analysis detected by a simple, graphical test. *BMJ*. 1997;315:629-34.
66. Peters JL, Sutton AJ, Jones DR, Abrams KR, Rushton L. Comparison of Two Methods to Detect Publication Bias in Meta-analysis. *JAMA*. 2006;295:676-80.

67. Duval S, Tweedie R. Trim and Fill: A Simple Funnel-Plot-Based Method of Testing and Adjusting for Publication Bias in Meta-Analysis. *Biometrics*. 2000;56(2):455-63.
68. Leeflang MMG, Deeks JJ, Gatsonis C, Bossuyt PMM, on behalf of the Cochrane Diagnostic Test Accuracy Working G. Systematic Reviews of Diagnostic Test Accuracy. *Ann Intern Med*. 2008;149:889-97.
69. Bland JM, Altman DG. Bayesians and frequentists. *BMJ*. 1998;317(7166):1151-60.
70. Viechtbauer W, Cheung MW-L. Outlier and influence diagnostics for meta-analysis. *Research Synthesis Methods*. 2010;1(2):112-25.
71. Gartlehner G, Moore CG. Direct versus indirect comparisons: A summary of the evidence. *International Journal of Technology Assessment in Health Care*. 2008;24(02):170-7.
72. Wells GA, Sultan SA, Chen L, Khan M, Coyle D. Indirect Evidence: Indirect Treatment Comparisons in Meta-Analysis. Ottawa: 2009 2009. Report No.
73. Cooper NJ, Sutton AJ, Morris D, Ades AE, Welton NJ. Addressing between-study heterogeneity and inconsistency in mixed treatment comparisons: Application to stroke prevention treatments in individuals with non-rheumatic atrial fibrillation. *Statistics in Medicine*. 2009;28(14):1861-81.
74. Caldwell DM, Dias S, Welton NJ. Extending Treatment Networks in Health Technology Assessment: How Far Should We Go? *Value Health*. 2015;18(5):673-81.
75. Cochrane Effective Practice and Organisation of Care (EPOC). Synthesising results when it does not make sense to do a meta-analysis. EPOC Resources for review authors, 2017.
76. Vandermeer BW, Buscemi N, Liang Y, Witmans M. Comparison of meta-analytic results of indirect, direct, and combined comparisons of drugs for chronic insomnia in adults: a case study. *Medical Care*. 2007;45(10 Suppl 2):S166-S72.
77. Yavin D, Roberts DJ, Tso M, Sutherland GR, Eliasziw M, Wong JH. Carotid endarterectomy versus stenting: a meta-analysis of randomized trials. *Can J Neurol Sci*. 2011;38(2):230-5.
78. Takagi H, Manabe H, Kawai N, Goto Sn, Umemoto T. Aprotinin increases mortality as compared with tranexamic acid in cardiac surgery: a meta-analysis of randomized head-to-head trials. *Interactive Cardiovascular and Thoracic Surgery*. 2009;9(1):98-101.
79. Filion K, El Khoury F, Bielinski M, Schiller I, Dendukuri N, Brophy J. Omega-3 fatty acids in high-risk cardiovascular patients: a meta-analysis of randomized controlled trials. *BMC Cardiovascular Disorders*. 2010;10(1):24.
80. Mongeon FP, Bélisle P, Joseph L, Eisenberg MJ, Rinfret S. Adjunctive Thrombectomy for Acute Myocardial Infarction. *Circulation: Cardiovascular Interventions*. 2010;3(1):6-16.
81. Katritsis DG, Siontis GCM, Ioannidis JPA. Double Versus Single Stenting for Coronary Bifurcation Lesions. *Circulation: Cardiovascular Interventions*. 2009;2(5):409-15.
82. Jansen JP, Crawford B, Bergman G, Stam W. Bayesian meta-analysis of multiple treatment comparisons: an introduction to mixed treatment

- comparisons. *Value In Health: The Journal Of The International Society For Pharmacoeconomics And Outcomes Research*. 2008;11(5):956-64.
83. Schöttker B, Lühmann D, Boukhemair D, Raspe H. Indirekte Vergleiche von Therapieverfahren. (German). *GMS Health Technology Assessment*. 2009;5:1-13.
84. Marshall JK, Irvine EJ. Rectal corticosteroids versus alternative treatments in ulcerative colitis: a meta-analysis. *Gut*. 1997;40(6):775-81.
85. Matchar DB, McCrory DC, Barnett HJM, Feussner JR. Medical Treatment for Stroke Prevention. *Annals of Internal Medicine*. 1994;121(1):41-53.
86. Pope JE, Anderson JJ, Felson DT. A Meta-analysis of the Effects of Nonsteroidal Anti-inflammatory Drugs on Blood Pressure. *Archives of Internal Medicine*. 1993;153(4):477-84.
87. Glenny AM, Altman DG, Song F, Sakarovitch C, Deeks JJ, D'Amico R, et al. Indirect comparisons of competing interventions. *Health Technology Assessment*. 2005;9(26).
88. Bucher HC, Guyatt GH, Griffith LE, Walter SD. The results of direct and indirect treatment comparisons in meta-analysis of randomized controlled trials. *Journal of Clinical Epidemiology*. 1997;50(6):683-91.
89. Sultana A, Ghaneh P, Cunningham D, Starling N, Neoptolemos J, Smith C. Gemcitabine based combination chemotherapy in advanced pancreatic cancer-indirect comparison. *BMC Cancer*. 2008;8(1):192.
90. Coomarasamy A, Knox EM, Gee H, Song F, Khan KS. Effectiveness of nifedipine versus atosiban for tocolysis in preterm labour: a meta-analysis with an indirect comparison of randomised trials. *BJOG: An International Journal of Obstetrics & Gynaecology*. 2003;110(12):1045-9.
91. Zhou Z, Rahme E, Pilote L. Are statins created equal? Evidence from randomized trials of pravastatin, simvastatin, and atorvastatin for cardiovascular disease prevention. *American Heart Journal*. 2006;151(2):273-81.
92. Song F, Loke YK, Walsh T, Glenny AM, Eastwood AJ, Altman DG. Methodological problems in the use of indirect comparisons for evaluating healthcare interventions: survey of published systematic reviews. *BMJ*. 2009;338(apr03\_1):b1147.
93. Lumley T. Network meta-analysis for indirect treatment comparisons. *Statistics in Medicine*. 2002;21(16):2313-24.
94. Lu G, Ades AE. Combination of direct and indirect evidence in mixed treatment comparisons. *Statistics in Medicine*. 2004;23(20):3105-24.
95. Wandel S, Jüni P, Tendal B, Nuesch E, Villiger PM, Welton NJ, et al. Effects of glucosamine, chondroitin, or placebo in patients with osteoarthritis of hip or knee: network meta-analysis. *BMJ*. 2010;341.
96. Hansen RA, Gaynes BN, Gartlehner G, Moore CG, Tiwari R, Lohr KN. Efficacy and tolerability of second-generation antidepressants in social anxiety disorder. *International Clinical Psychopharmacology*. 2008;23(3).
97. Thijs V, Lemmens R, Fieuws S. Network meta-analysis: simultaneous meta-analysis of common antiplatelet regimens after transient ischaemic attack or stroke. *European Heart Journal*. 2008;29(9):1086-92.

98. Mills EJ, Druyts E, Ghement I, Puhan MA. Pharmacotherapies for chronic obstructive pulmonary disease: a multiple treatment comparison meta-analysis. *Clin Epidemiol*. 2011;3:107-29.
99. Hartling L, Fernandes RM, Bialy L, Milne A, Johnson D, Plint A, et al. Steroids and bronchodilators for acute bronchiolitis in the first two years of life: systematic review and meta-analysis. *BMJ*. 2011;342:d1714.
100. Welton NJ, Caldwell DM, Adamopoulos E, Vedhara K. Mixed Treatment Comparison Meta-Analysis of Complex Interventions: Psychological Interventions in Coronary Heart Disease. *American Journal of Epidemiology*. 2009;169(9):1158-65.
101. Dias S, Welton NJ, Sutton AJ, Ades AE. NICE DSU Technical Support Document 2: a generalised linear modelling framework for pairwise and network meta-analysis of randomised controlled trials. Sheffield, UK: Decision Support Unit, SchARR, University of Sheffield, 2014.
102. Moses LE, Shapiro D, Littenberg B. Combining independent studies of a diagnostic test into a summary roc curve: Data-analytic approaches and some additional considerations. *Statistics in Medicine*. 1993;12:1293-316.
103. Reitsma JB, Glas AS, Rutjes AWS, Scholten RJPM, Bossuyt PM, Zwinderman AH. Bivariate analysis of sensitivity and specificity produces informative summary measures in diagnostic reviews. *Journal of Clinical Epidemiology*. 2005;58:982-90.
104. Chappell FM, Raab GM, Wardlaw JM. When are summary ROC curves appropriate for diagnostic meta-analyses? *Statistics in Medicine*. 2009;28(21):2653-68.
105. Steinhäuser S, Schumacher M, Rucker G. Modelling multiple thresholds in meta-analysis of diagnostic test accuracy studies. *BMC Med Res Methodol*. 2016;16(1):97.
106. Owen RK, Cooper NJ, Quinn TJ, Lees R, Sutton AJ. Network meta-analysis of diagnostic test accuracy studies identifies and ranks the optimal diagnostic tests and thresholds for health care policy and decision-making. *J Clin Epidemiol*. 2018;99:64-74.
107. Davenport C, Cheng EY, Kwok YT, Lai AH, Wakabayashi T, Hyde C, et al. Assessing the diagnostic test accuracy of natriuretic peptides and ECG in the diagnosis of left ventricular systolic dysfunction: a systematic review and meta-analysis. *Br J Gen Pract*. 2006;56(522):48-56.
108. Olde Bekkink M, McCowan C, Falk GA, Teljeur C, Van de Laar FA, Fahey T. Diagnostic accuracy systematic review of rectal bleeding in combination with other symptoms, signs and tests in relation to colorectal cancer. *Br J Cancer*. 2009;102(1):48-58.
109. Boland GWL, Dwamena BA, Jagtiani Sangwaiya M, Goehler AG, Blake MA, Hahn PF, et al. Characterization of Adrenal Masses by Using FDG PET: A Systematic Review and Meta-Analysis of Diagnostic Test Performance. *Radiology*. 2011;259(1):117-26.
110. Harbord RM, Whiting P, Sterne JAC, Egger M, Deeks JJ, Shang A, et al. An empirical comparison of methods for meta-analysis of diagnostic accuracy showed hierarchical models are necessary. *Journal of Clinical Epidemiology*. 2008;61:1095-103.

111. Simel DL, Bossuyt PMM. Differences between univariate and bivariate models for summarizing diagnostic accuracy may not be large. *Journal of Clinical Epidemiology*. 2009;62(12):1292-300.
112. Stijnen T, Hamza TH, Özdemir P. Random effects meta-analysis of event outcome in the framework of the generalized linear mixed model with applications in sparse data. *Statistics in Medicine*. 2010;29(29):3046-67.
113. Morris TA, Castrejon S, Devendra G, Gamst AC. No Difference in Risk for Thrombocytopenia During Treatment of Pulmonary Embolism and Deep Venous Thrombosis With Either Low-Molecular-Weight Heparin or Unfractionated Heparin\*. *Chest*. 2007;132(4):1131-9.
114. Sutton AJ, Higgins JPT. Recent developments in meta-analysis. *Statistics in Medicine*. 2008;27:625-50.
115. Lefevre F, Aronson N. Ketogenic diet for the treatment of refractory epilepsy in children: A systematic review of efficacy. *Pediatrics*. 2000;105(4):E46.
116. Evans JT, Green JD, Carlin PE, Barrett LO. Meta-analysis of antibiotics in tube thoracostomy. *Am Surg*. 1995;61(3):215-9.
117. Hurwitz EL, Aker PD, Adams AH, Meeker WC, Shekelle PG. Manipulation and mobilization of the cervical spine. A systematic review of the literature. *Spine (Phila Pa 1976)*. 1996;21(15):1746-59.
118. Hutton B, Salanti G, Caldwell DM, Chaimani A, Schmid CH, Cameron C, et al. The PRISMA extension statement for reporting of systematic reviews incorporating network meta-analyses of health care interventions: checklist and explanations. *Ann Intern Med*. 2015;162(11):777-84.
119. Health Information and Quality Authority. Health technology assessment (HTA) of smoking cessation interventions. Dublin: HIQA, 2017.
120. Dias S, Sutton AJ, Ades AE, Welton NJ. Evidence synthesis for decision making 2: a generalized linear modeling framework for pairwise and network meta-analysis of randomized controlled trials. *Med Decis Making*. 2013;33(5):607-17.
121. Dias S, Sutton AJ, Welton NJ, Ades AE. Evidence synthesis for decision making 3: heterogeneity--subgroups, meta-regression, bias, and bias-adjustment. *Med Decis Making*. 2013;33(5):618-40.
122. Dias S, Welton NJ, Caldwell DM, Ades AE. Checking consistency in mixed treatment comparison meta-analysis. *Stat Med*. 2010;29(7-8):932-44.
123. van Valkenhoef G, Dias S, Ades AE, Welton NJ. Automated generation of node-splitting models for assessment of inconsistency in network meta-analysis. *Res Synth Methods*. 2016;7(1):80-93.
124. Salanti G, Ades AE, Ioannidis JP. Graphical methods and numerical summaries for presenting results from multiple-treatment meta-analysis: an overview and tutorial. *J Clin Epidemiol*. 2011;64(2):163-71.

## 5 Glossary of terms and abbreviations

Some of the terms in this glossary will not be found within the body of these guidelines. They have been included here to make the glossary a more complete resource for users.

**Adverse event:** an undesirable effect of a health technology.

**Bayesian:** a form of statistical inference in which data are observed from a realised sample and the underlying parameters (such as mean) are unknown and described by probability distributions. Prior knowledge about the parameters is updated using observed data to generate posterior distributions for the unknown parameters (See also **Frequentist**).

**Bias:** systematic (as opposed to random) deviation of the results of a study from the 'true' results.

**Biomarker:** a substance used as an indicator of a response to a therapeutic intervention. An example of a biomarker is the presence of an antibody that may indicate infection.

**Comorbidity:** the coexistence of a disease, or more than one disease, in a person in addition to the disease being studied or treated.

**Comparator:** the alternative against which the technology is compared.

**Confidence interval:** the computed interval with a specified probability (by convention, 95%) that the true value of a variable such as mean, proportion, or rate is contained within the interval over repeated sampling.

**Cost-effectiveness:** a comparison of both the costs and health effects of a technology to assess whether the technology provides value for money.

**Covariate:** a variable that may be predictive of the endpoint being analysed. Covariates can be specified for individual patients (such as age, sex, disease risk) and for studies (for example, mean patient age, proportion males).

**Critical appraisal:** a strict process to assess the validity, results and relevance of evidence.

**Direct comparison:** a meta-analysis combining multiple head-to-head trials comparing the technology of interest to the same comparator (See also **Indirect comparison** and **Multiple treatment comparison**).

**Effectiveness:** the extent to which a technology produces an overall health benefit (taking into account adverse and beneficial effects for a specified group of patients) in routine clinical practice (contrast with **Efficacy**).

**Efficacy:** the extent to which a technology produces an overall health benefit (taking into account adverse and beneficial effects for a specified group of patients) when studied under controlled research conditions (contrast with **Effectiveness**).

**EQ-5D:** the EQ-5D is a standardised instrument (questionnaire) used to measure health outcomes. The instrument is applicable to a wide range of health conditions and treatments and can be used to generate a single index value for health status. The EQ-5D questionnaire describes five attributes (mobility, self-care, usual activity, pain/discomfort, and anxiety/depression) each of which has three levels (no problems, some problems, and major problems). This combination defines 243 possible health states which added to the health states 'unconscious' and 'dead', allow for 245 possible health states. Each EQ-5D health state (or profile) provides a set of observations about a person by way of a five digit code number. This EQ-5D health state is then converted to a single summary index by applying a formula that attaches weights to each of these levels in each dimension and subtracting these values from 1.0. Additional weights that are applied are a constant (for any deviation from perfect health) and a weight if any of the dimensions are at level three (major problems). The scores fall on a value scale that ranges from 0.0 (dead) to 1.0 (perfect health). For further information on EQ-5D see: [www.euroqol.org](http://www.euroqol.org).

**Final outcome:** a health outcome that is directly related to the length of life, e.g. life-years gained or quality-adjusted life years.

**Fixed effect analysis:** the true effect of the treatment is typically assumed to be the same in each study and that the variability between studies is entirely due to chance (See also **Random effects analysis**).

**Frequentist:** a form of statistical inference in which data are considered a repeatable random sample whereas the underlying parameters (e.g. mean) are fixed. If a trial is repeated enough times the sample mean will approach the true mean (See also **Bayesian**).

**Generalisability:** the problem of whether one can apply or extrapolate results obtained in one setting or population to another. Term may also be referred to as 'transferability', 'transportability', 'external validity', 'relevance', or 'applicability'.

**Health outcome:** a change (or lack of change) in health status caused by a therapy or factor when compared with a previously documented health status using disease-specific measures, general quality of life measures or utility measures.

**Health technology:** the application of scientific or other organised knowledge – including any tool, technique, product, process, method, organisation or system – in healthcare and prevention. In healthcare, technology includes drugs, diagnostics, indicators and reagents, devices, equipment, and supplies, medical and surgical procedures, support systems and organisational and managerial systems used in prevention, screening diagnosis, treatment and rehabilitation.

**Health technology assessment (HTA):** this is a multidisciplinary process that summarises information about the medical, social, economic and ethical issues related to the use of a health technology in a systematic, transparent, unbiased, and robust manner. Its aim is to inform the formulation of safe, effective health policies that are patient focused and seek to achieve best value.

**Heterogeneity:** in the context of meta-analysis, clinical heterogeneity means dissimilarity between studies. It can be because of the use of different statistical methods (statistical heterogeneity), or evaluation of people with different characteristics, treatments or outcomes (clinical heterogeneity). Heterogeneity may render pooling of data in meta-analysis unreliable or inappropriate. Finding no significant evidence of heterogeneity is not the same as finding evidence of no heterogeneity. If there are a small number of studies, heterogeneity may affect results but not be statistically significant.

**Incidence:** the number of new cases of a disease or condition that develop within a specific time frame in a defined population at risk. It is usually expressed as a ratio of the number of affected people to the total population.

**Indication:** a clinical symptom or circumstance indicating that the use of a particular technology would be appropriate.

**Indirect comparison:** a meta-analysis in which the technology of interest is compared to the comparator technology via a third technology. This method is used in the absence of any head-to-head trials comparing the technology of interest to the comparator technology (See also **Direct comparison** and **Multiple treatment comparison**).

**Meta-analysis:** systematic methods that use statistical techniques for combining results from different studies to obtain a quantitative estimate of the overall effect of a particular technology or variable on a defined outcome. This combination may produce a stronger conclusion than can be provided by any individual study. (Also known as data synthesis or quantitative overview).

**Multiple treatment comparison:** a meta-analysis using a combination of direct and indirect comparisons to determine the relative effectiveness of three or more technologies (See also **Direct comparison** and **Indirect comparison**).

**Multi-arm trial:** a trial evaluating more than two treatments with a patient group for each treatment.

**Outcome:** consequence of condition or intervention; in Economic Guidelines, outcomes most often refer to health outcomes, such as surrogate outcomes or patient outcomes.

**Prevalence:** the number of people in a population with a specific disease or condition at a given time and is usually expressed as a ratio of the number of affected people to the total population.

**Prior:** the probability distribution that would express one's beliefs about an uncertain quantity or parameter before some evidence is taken into account. Priors are used in Bayesian analysis.

**Probability:** expression of degree of certainty that an event will occur, on scale from zero (certainty that event will not occur) to one (certainty that event will occur).

**Probability distribution:** portrays the relative likelihood that a range of values is the true value of a treatment effect. This distribution often appears in the form of a bell-shaped curve. An estimate of the most likely true value of the treatment effect is the value at the highest point of the distribution. The area under the curve between any two points along the range gives the probability that the true value of the treatment effect lies between those two points. Thus, a probability distribution can be used to determine an interval that has a designated probability (e.g. 95%) of including the true value of the treatment effect.

**Quality-adjusted life year (QALY):** a unit of healthcare outcomes that adjusts gains (or losses) in years of life subsequent to a healthcare intervention by the quality of life during those years. QALYs can provide a common unit for comparing cost-utility across different technologies and health problems. Analogous units include Disability-Adjusted Life Years (DALYs) and Healthy-Years Equivalents (HYEs).

**Random effects analysis:** the treatment effects in each study is assumed to vary around an overall average treatment effect. A random effects analysis therefore assumes a different underlying true effect for each study (See also **Fixed effect analysis**).

**Receiver operating characteristic (ROC) curve:** a graphical plot of the true positive rate against the false positive rate. The ROC curve is used as a fundamental tool for evaluating diagnostic tests.

**Reliability:** the extent to which repeated measures of the same endpoint on the same individual patient return the same value (See also **Validity**).

**Sensitivity analysis:** a means to determine the robustness of a mathematical model or analysis by examining the extent to which results are affected by changes in methods, parameters or assumptions.

**SF-36:** the SF-36 is a standardised instrument (questionnaire) used to measure health outcomes. It is a multi-purpose, short-form health survey with 36 questions. It yields an 8-scale profile of functional health and well-being scores as well as psychometrically-based physical and mental health summary measures and a preference-based health utility index. It is a generic measure, as opposed to one that targets a specific age, disease, or treatment group. Accordingly, the SF-36 has proven useful in surveys of general and specific populations, comparing the relative burden of diseases, and in differentiating the health benefits produced by a wide range of different treatments. For further information on SF-36 see: [www.sf-36.org](http://www.sf-36.org).

**Statistical significance:** a conclusion that a technology has a true effect, based upon observed differences in outcomes between the treatment and control groups that are sufficiently large so that these differences are unlikely to have occurred due to chance, as determined by a statistical test. Statistical significance is related to the probability of observing a difference between treatment and control groups at least as large as the observed value if the null hypothesis of no true treatment effect is true. A cut-off value of 0.05 is commonly used, with statistical significance declared if the calculated probability is less than or equal to 0.05. It does not provide information about the magnitude of a treatment effect. (Statistical significance is necessary but not sufficient for clinical significance.)

**Stratified analysis:** a process of analysing smaller, more homogeneous subgroups according to specified criteria such as age groups, socioeconomic status, where there is variability (heterogeneity) in a population.

**Subgroup:** a subset of individuals in a population group or of participants in a study that share one or more common characteristics (for example, sex, age, risk status).

**Subgroup analysis:** an analysis in which the technology effect is evaluated in a subgroup of a trial, including the analysis of its complementary subgroup. Subgroup analyses can be pre-specified, in which case they are easier to interpret. If not pre-specified, they are difficult to interpret because they tend to uncover false positive results.

**Surrogate endpoint:** a measure that is used in place of a primary endpoint (outcome). Examples are decrease in blood pressure as a predictor of decrease in strokes and heart attacks in hypertensive patients, and increase in T-cell (a type of white blood cell) counts as an indicator of improved survival of patients with HIV or

AIDS. Use of a surrogate endpoint assumes that it is a reliable predictor of the primary endpoint(s) of interest.

**Target population:** in the context of a budget impact analysis the individuals with a given condition or disease who might avail of the technology being assessed within the defined time horizon.

**Technology:** the application of scientific or other organised knowledge – including any tool, technique, product, process, method, organisation or system – to practical tasks. In healthcare, technology includes drugs, diagnostics, indicators and reagents, devices, equipment and supplies, medical and surgical procedures, support systems, and organisational and managerial systems used in prevention, screening, diagnosis, treatment and rehabilitation.

**Time horizon:** in the context of a clinical trial it is the time span over which patients are monitored for treatment effect.

**Type I error:** occurs when the null hypothesis is incorrectly rejected, also known as a false positive finding (See also **Type II error**).

**Type II error:** occurs when a false null hypothesis fails to be rejected, also known as a false negative finding (See also **Type I error**).

**Uncertainty:** where the true value of a parameter or the structure of a process is unknown.

**Validity:** the extent to which an endpoint measures what it is intended to measure (See also **Reliability**).

**Variability:** this reflects known differences in parameter values arising out of inherent differences in circumstances or conditions. It may arise due to differences in patient population (e.g. patient heterogeneity – baseline risk, age, gender), differences in clinical practice by treatment setting or geographical location.

**WOMAC:** the Western Ontario and McMaster Universities Osteoarthritis Index (WOMAC) is a proprietary set of standardized questionnaires used to evaluate the condition of patients with osteoarthritis of the knee and hip. The questionnaire consists of 24 items divided into three subscales, and can be self-administered. The instrument measures for pain, stiffness, and physical functioning of the joints.

## **Appendix A — Network meta-analysis example**

A network meta-analysis is more complex to present than a standard direct comparison as there are a variety of outputs that can assist in interpreting the results of the analysis. Reporting guidelines specific to network meta-analysis, such as those produced by PRISMA,<sup>(118)</sup> are available and should be used.

A brief example is provided here to show some of the outputs that might be included as part of the presentation of an analysis. The example is of a network meta-analysis of pharmacological smoking cessation interventions (including electronic cigarettes) in unselected adults.<sup>(119)</sup> The primary outcome was long-term smoking cessation rates as indicated by quit rates at greater than or equal to six months. All data are derived from published randomised control trials.

### **A.1 Methodology for indirect comparisons**

The methodology used should be described in sufficient detail that the choice of model is clear.

In this example, a Bayesian approach was used. Where there was sufficient indirect and direct evidence and the assumption of transitivity was justified, a network meta-analysis approach was considered. In the case of a network meta-analysis, the consistency model was used.<sup>(120)</sup> An unrelated mean effects model, also referred to as an inconsistency model, was also applied.<sup>(121)</sup> A random effects model was used. The node splitting approach was used to compare direct and indirect evidence, and an examination of deviance statistics was used to identify studies that were providing potentially inconsistent estimates.<sup>(122, 123)</sup> Node splitting generates separate models for direct and indirect evidence, and the network evidence is not a mathematical combination of the two. Some multi-arm trials may be excluded from the node splitting analysis if they provide both direct and indirect evidence for a given comparison. Node splitting has only been applied to comparisons for which there is both direct and indirect evidence.

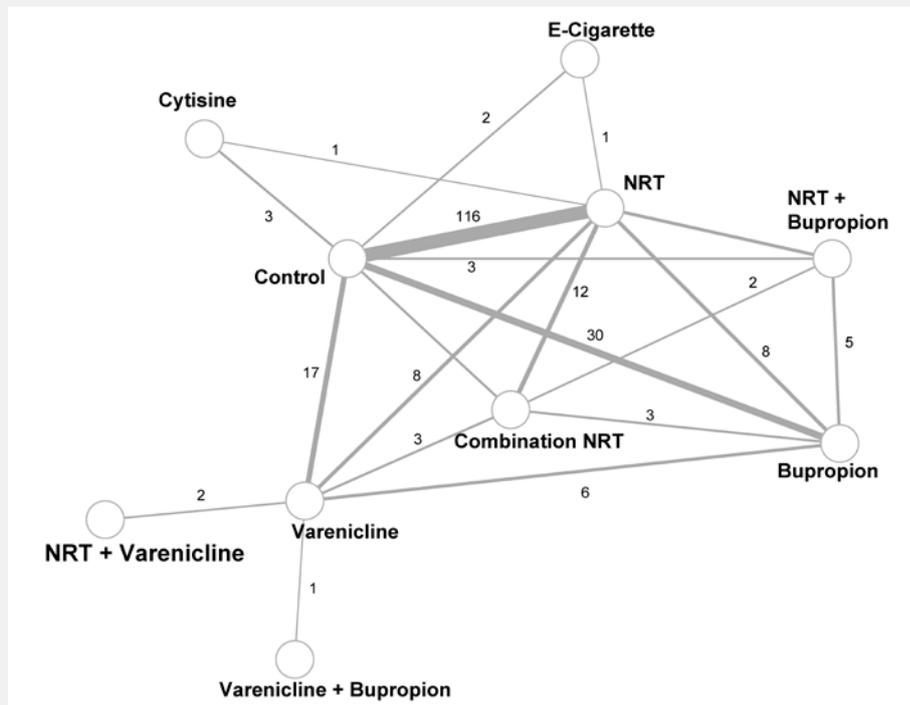
Meta-regression approaches were also applied to determine if study-level covariates could explain some of the observed variance. Models were compared using the deviance information criterion (DIC). Models were run with a burn-in of 20,000 iterations followed by 50,000 iterations on four chains. Model convergence in the adaption phase was checked using the Gelman and Rubin convergence diagnostic.

### **A.2 Network of evidence**

The network of studies and interventions needs to be clearly described. A network diagram allows the reader to quickly understand the comparisons available. The

diagram should be clearly labelled in terms of the interventions (Figure A.1). The diagram can also be formatted to help the reader see how many studies underpin each comparison. There should be a commentary to draw attention to any important features of the evidence network, such as important comparisons with limited direct evidence. In the example below, an intervention of particular interest, e-cigarettes, appeared in only three trials, only one of which included an active comparator.

**Figure A.1 Network of evidence**



In this example there were 232 comparisons available across 176 pharmacotherapy trials (Figure A.1). Of those comparisons, 174 were between intervention and control. The largest quantity of evidence was for nicotine replacement therapy (NRT), with 152 comparisons. There were 20 head-to-head comparisons between interventions in total. For the purposes of the analysis, combinations of interventions are treated as distinct interventions. For example, NRT plus varenicline is a distinct combination therapy. Most of the interventions appear in numerous different comparisons. Others, such as NRT plus varenicline and varenicline plus bupropion, each appear in a single comparison, although there may be multiple studies providing evidence for those comparisons.

### A.3 Analysis results

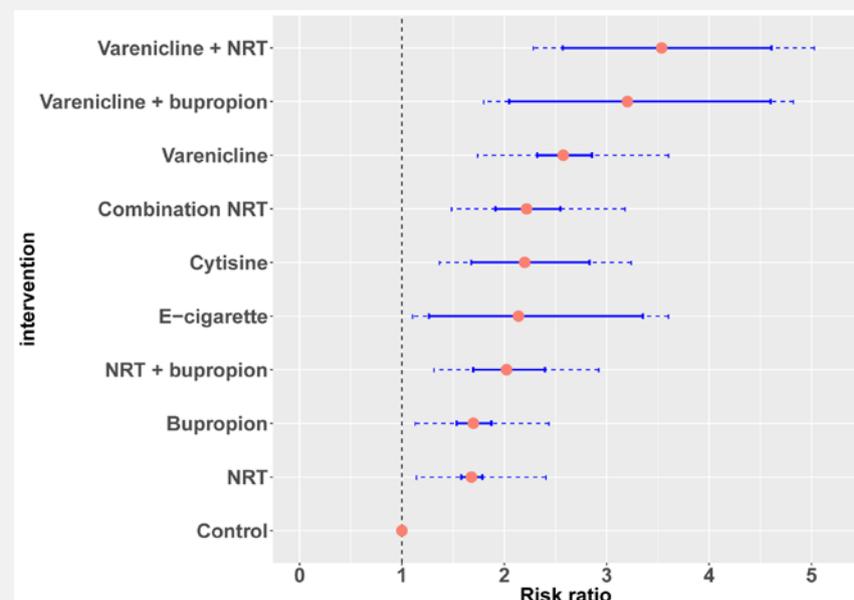
Ideally, both tabular and graphical methods are used to describe the treatment effect. A description of the results should also describe the outputs of the model in terms of heterogeneity and inconsistency between direct and indirect evidence.

Treatment effects were estimated using both consistency and inconsistency models to determine whether the assumption of consistency has a substantial impact on estimates of treatment effect. The consistency and inconsistency models produced very similar estimates of treatment effect, agreeing in terms of direction and magnitude of effect. All of the estimates from the consistency model were well within the confidence bounds for the corresponding inconsistency model estimates. The DIC was marginally lower for the consistency model (639.2 versus 643.9), although the difference ( $<5$ ) was not large enough to be considered important. The random effects standard deviation was 0.236 (95% credible interval (CI): 0.173 to 0.303) for the consistency model, and 0.239 (95% CI: 0.215 to 0.238) for the inconsistency model. As such, the consistency model was considered appropriate.

An analysis of heterogeneity estimated a global  $I^2$  of 29%. Based on an analysis of heterogeneity, potential issues were identified for two comparisons: varenicline versus control ( $p=0.077$ ) and varenicline versus NRT monotherapy ( $p=0.054$ ). A node-splitting analysis was used to investigate the contribution of direct and indirect evidence to treatment effect estimates. There was no statistically significant difference in the direct and indirect evidence for any of the comparisons. For almost all comparisons, the direct and indirect treatment effects were in agreement in terms of direction and, for the most part, in terms of magnitude, although there were some differences. For example, the direct evidence showed combination NRT to be better than NRT monotherapy. The indirect evidence showed a non-significant treatment benefit associated with the monotherapy. The pooled estimate was driven by the direct evidence.

A forest plot can be used to present the treatment effect of each intervention relative to a common comparator (Figure A.2). In the context of an NMA, a forest is primarily useful to understand the treatment effect for each intervention relative to the same common comparator. The treatment effects for each pair of interventions can be presented in a table (Table A.1). It is important to provide the full set of comparisons as the common reference comparator may be of limited interest and does not provide information about the uncertainty regarding any of the other comparisons. By viewing all pair-wise comparisons, it is possible to see how some comparisons are subject to greater uncertainty than others. In this example, the estimated risk ratio for NRT versus placebo control is 1.68 (95% CI: 1.58 to 1.78) while the risk ratio for NRT versus e-cigarettes is 0.76 (95% CI: 0.41 to 1.34). The much greater imprecision in the estimate relative to e-cigarettes reflects the limited evidence in relation to e-cigarettes for smoking cessation.

**Figure A.2 Forest plot of treatment effect of smoking cessation interventions relative to control**



Note: solid lines refer to the confidence bounds, dashed lines represent the prediction intervals.

Aside from the effect estimates, a critical element of a network meta-analysis is the consistency (or lack of) between the direct and indirect evidence. By including the direct, indirect and combined direct and indirect estimates in a table, one can see comparisons where inconsistencies may exist between the direct and indirect evidence (Table A.2). Statistical tests can be used to determine if the direct and indirect evidence are markedly different or inconsistent. Such inconsistencies can be very important and should be fully explored. In the example presented here, there were no comparisons for which the difference between direct and indirect evidence was statistically significant. However, there were two comparisons for which the p-value was less than or equal to 0.1 (Varenicline versus control and Varenicline versus NRT).

Another output of network meta-analysis that can assist in interpreting the findings is the ranking of interventions. Depending on the method of network meta-analysis used, it is possible to extract the probability of an intervention being ranked best, second best and so on. The uncertainty of ranking may be intuitively interpreted and facilitate an understanding of which is the best treatment of those assessed (Figure A.3). These plots are sometimes called rankograms. From the analysis presented here, control has a 99.7% chance of being the lowest ranked intervention. The combination of nicotine replacement therapy and varenicline has a 64.1% chance of being the highest ranked intervention. Rankograms can also be presented as cumulative probability of having a particular ranking or higher (for example, a 70% probability of being ranked sixth or better out of ten interventions).

**Table A.1 Network meta-analysis treatment effect estimates**

	Risk ratio (95% credible interval)								
	Control	Bupropion	Cytisine	E-cigarette	NRT	NRT + bupropion	NRT + varenicline	Combination NRT	Varenicline
Bupropion	1.70 (1.53 - 1.87)								
Cytisine	2.20 (1.68 - 2.83)	1.33 (0.97 - 1.81)							
E-cigarette	2.14 (1.26 - 3.35)	1.29 (0.72 - 2.20)	0.97 (0.49 - 1.80)						
NRT	1.68 (1.58 - 1.78)	0.99 (0.88 - 1.11)	0.73 (0.53 - 1.00)	0.76 (0.41 - 1.34)					
NRT + bupropion	2.02 (1.70 - 2.40)	1.21 (0.99 - 1.48)	0.91 (0.62 - 1.30)	0.94 (0.50 - 1.68)	1.23 (1.01 - 1.48)				
NRT + varenicline	3.54 (2.57 - 4.61)	2.33 (1.58 - 3.27)	1.80 (1.08 - 2.81)	1.86 (0.93 - 3.30)	2.35 (1.61 - 3.28)	1.96 (1.27 - 2.89)			
Combination NRT	2.22 (1.91 - 2.55)	1.35 (1.12 - 1.60)	1.01 (0.70 - 1.41)	1.04 (0.57 - 1.84)	1.36 (1.16 - 1.58)	1.11 (0.88 - 1.40)	0.53 (0.33 - 0.86)		
Varenicline	2.57 (2.32 - 2.85)	1.60 (1.39 - 1.84)	1.21 (0.87 - 1.65)	1.25 (0.69 - 2.13)	1.61 (1.43 - 1.83)	1.33 (1.06 - 1.65)	0.65 (0.42 - 0.99)	1.20 (0.99 - 1.44)	
Varenicline + bupropion	3.20 (2.05 - 4.60)	2.07 (1.22 - 3.25)	1.58 (0.85 - 2.75)	1.64 (0.75 - 3.20)	2.08 (1.24 - 3.27)	1.73 (0.98 - 2.86)	0.87 (0.43 - 1.69)	1.57 (0.90 - 2.61)	1.32 (0.77 - 2.18)

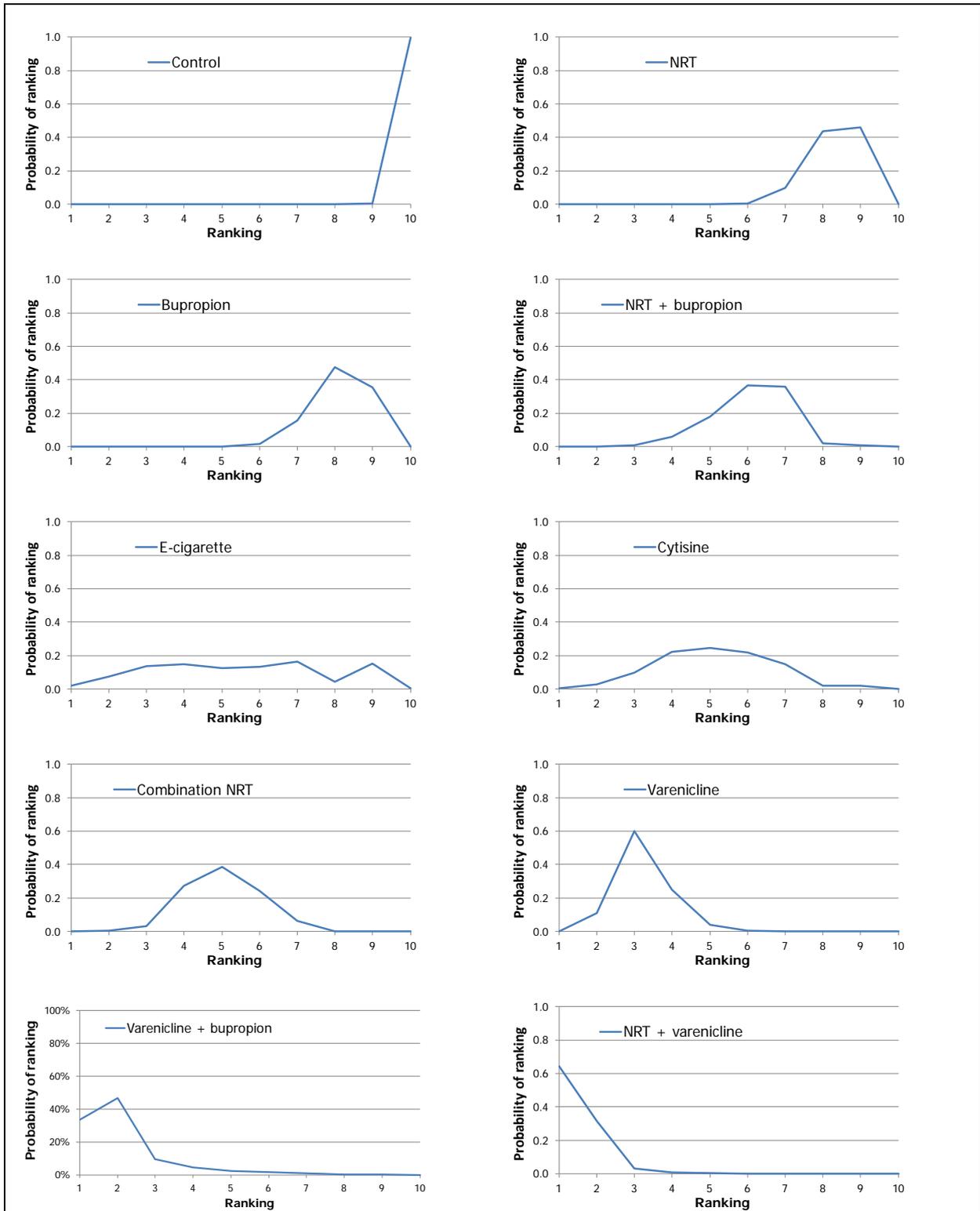
Note: NRT, nicotine replacement therapy. Shaded cells indicate statistically significant treatment effect.

**Table A.2 Comparison of direct and indirect treatment effect estimates: pharmacological interventions**

Comparison	Risk ratio						p-value
	Direct estimate (95% CI)		Indirect estimate (95% CI)		Network estimate (95% CI)		
Bupropion vs Control	1.67	(1.49 to 2.18)	1.67	(1.27 to 2.53)	1.70	(1.54 to 2.16)	0.98
Cytisine vs Control	2.10	(1.49 to 3.55)	2.28	(1.50 to 4.64)	2.20	(1.68 to 3.55)	0.66
NRT vs Control	1.67	(1.56 to 2.05)	1.78	(1.47 to 2.51)	1.68	(1.58 to 2.05)	0.53
NRT + bupropion vs Control	1.85	(1.34 to 2.97)	2.07	(1.60 to 3.25)	2.02	(1.70 to 2.97)	0.60
Combination NRT vs Control	1.82	(1.19 to 3.25)	2.28	(1.96 to 3.25)	2.21	(1.93 to 2.97)	0.28
Varenicline vs Control	2.83	(2.45 to 3.88)	2.23	(1.78 to 3.55)	2.64	(2.28 to 3.55)	0.10
NRT vs Bupropion	0.97	(0.78 to 1.33)	1.00	(0.86 to 1.30)	0.99	(0.88 to 1.25)	0.80
NRT + bupropion vs Bupropion	1.18	(0.92 to 1.67)	1.17	(0.66 to 2.28)	1.21	(0.99 to 1.67)	0.98
Combination NRT vs Bupropion	1.36	(0.97 to 2.12)	1.38	(1.12 to 1.96)	1.34	(1.13 to 1.82)	0.90
Varenicline vs Bupropion	1.58	(1.24 to 2.30)	1.60	(1.33 to 2.26)	1.60	(1.38 to 2.12)	0.91
NRT vs Cytisine	0.67	(0.40 to 1.24)	0.78	(0.51 to 1.32)	0.74	(0.53 to 1.13)	0.65
NRT vs E-cigarette	0.79	(0.36 to 1.87)	0.60	(0.17 to 1.87)	0.76	(0.41 to 1.53)	0.68
NRT + bupropion vs NRT	1.33	(1.06 to 1.89)	1.02	(0.58 to 1.92)	1.22	(1.01 to 1.67)	0.35
Combination NRT vs NRT	1.37	(1.16 to 1.86)	0.89	(0.47 to 1.84)	1.37	(1.17 to 1.79)	0.18
Varenicline vs NRT	1.38	(1.09 to 1.94)	1.72	(1.49 to 2.36)	1.62	(1.43 to 2.10)	0.09
Combination NRT vs NRT + bupropion	1.07	(0.74 to 1.75)	1.15	(0.84 to 1.76)	1.11	(0.88 to 1.58)	0.77
Varenicline vs Combination NRT	1.16	(0.83 to 1.82)	1.20	(0.95 to 1.70)	1.19	(0.99 to 1.63)	0.88

Notes: NRT refers to NRT monotherapy (only one pharmaceutical form used); Combination NRT refers to use of more than one formulation (for example, transdermal patch plus gum or spray). CI, credible interval. The direct, indirect and network evidence are created from three different models, and the network estimate was not a weighted average of the indirect and direct studies. Results are presented for studies where there was both direct and indirect evidence not limited to data from a single multi-arm study).

**Figure A.3 Rankograms for 10 pharmacological interventions for smoking cessation**



Another method for considering ranks of treatments is the surface under the cumulative ranking curve (SUCRA).<sup>(124)</sup> The SUCRA presents a single number that incorporates the uncertainty associated with the ranking of each treatment. SUCRA values range from 0 to 100%: the closer to 100% the higher the likelihood that a therapy is in the top rank or one of the top ranks; closer to 0 the more likely that a therapy is at or near the bottom rank. For the example of smoking cessation interventions, the SUCRA values reinforce the notion that placebo control is the lowest ranked treatment and the combination of nicotine replacement therapy and varenicline is the highest ranked intervention (Table A.3).

**Table A.3 Surface under the cumulative ranking curve (SUCRA) for pharmacological smoking cessation interventions**

Treatment	SUCRA
Control	0.000
NRT	0.183
Bupropion	0.204
NRT + bupropion	0.432
E-cigarette	0.500
Cytisine	0.540
Combination NRT	0.553
Varenicline	0.753
Varenicline + bupropion	0.881
NRT + varenicline	0.953

While ranks are useful for determining the hierarchy of the interventions included in the analysis, they are not used as an input for economic evaluation.

#### **A.4 Sensitivity analysis**

As with a meta-analysis of direct comparisons, the methodology for a network meta-analysis is underpinned by a number of assumptions. The extent of the network, the quality of the included studies and other study-level characteristics may all impact on the estimates of treatment effect. As with direct comparisons, the use of sensitivity analysis can aid interpretation of findings.

In a network meta-analysis, there is a risk that there may be systematic differences across studies that may invalidate the assumption of transitivity. One approach is to consider the inclusion of study-level covariates through a meta-regression.

In the smoking cessation example, the network meta-analysis was also run as a network meta-regression to determine if certain study-level characteristics might be acting as effect modifiers.

Six different covariates were considered in network meta-regression: two continuous variables (study year, length of follow up) and four dichotomous variables (high quality, biochemical verification of abstinence, continuous abstinence and no provision of supplementary care). The meta-regression assumed a shared effect across treatments. Longer follow up was associated with a reduced effect size, while measuring continuous abstinence (rather than point prevalence) was associated with a larger effect size (Table A.4). The other covariates were not associated with statistically significant effects. Inclusion of covariates did not impact on the DIC.

**Table A.4 Network meta-regression results**

Covariate	Coefficient	(95% CI)	DIC	Random effects SD
No covariates	-	-	639.2	0.24 (0.17 to 0.30)
Study year	0.12	(-0.03 to 0.27)	640.7	0.22 (0.16 to 0.29)
Follow up	-0.12	(-0.25 to 0.00)	639.8	0.22 (0.16 to 0.29)
High quality	0.02	(-0.13 to 0.16)	640.4	0.24 (0.17 to 0.30)
Biochemically verified	0.11	(-0.08 to 0.31)	640.1	0.23 (0.17 to 0.30)
Continuous abstinence	0.17	(0.03 to 0.31)	638.3	0.22 (0.16 to 0.29)
No supplementary care	0.06	(-0.17 to 0.29)	640.2	0.24 (0.17 to 0.30)

Notes: CI, credible interval; DIC, Deviance Information Criterion; SD, standard deviation.

When considering the impact of covariates it is important to take model fit into account. If the addition of covariates does not meaningfully improve the model fit then their inclusion may reduce rather than increase clarity.

When investigating the impact of covariates, it is also useful to look at the impact on estimated treatment effect. In the present example, a study-level covariate was included that distinguished between trials on the basis of whether continuous abstinence was required. Some trials recorded smoking cessation on the date of follow-up (for example, at six months), not taking into account that the individual may have relapsed between the start of the trial and follow up. Other trials required continuous abstinence, which is considered a better marker of cessation. The coefficient for continuous abstinence was statistically significant, suggesting that trials using continuous abstinence for outcome measurement estimated a different effect to those that did not. Length of follow up also had an impact on treatment effect. Although the change to model fit did not justify including either covariate, it is useful to look at the impact on the estimated clinical effectiveness (Table A.5).

The impact on treatment effects (relative to control) of including covariates in the model is shown in Table A.5. Including length of follow up reduces the effect size for all interventions apart from cytisine. Including continuous abstinence increases the treatment effect for all interventions. The addition of covariates has a negligible impact on DIC and random effects standard deviation, indicating that inclusion of covariates has a small impact on reducing heterogeneity.

**Table A.5 Impact on treatment effect (relative to control) of including covariates in analysis**

Intervention	No covariate		Follow up = 12 months		Continuous abstinence	
	RR	(95% CI)	RR	(95% CI)	RR	(95% CI)
Bupropion	1.70	(1.53 - 1.87)	1.65	(1.49 - 1.82)	1.76	(1.59 - 1.94)
Cytisine	2.20	(1.68 - 2.83)	2.31	(1.78 - 2.95)	2.34	(1.80 - 3.00)
E-cigarette	2.14	(1.26 - 3.35)	2.09	(1.25 - 3.28)	2.19	(1.32 - 3.42)
NRT	1.68	(1.58 - 1.78)	1.64	(1.54 - 1.75)	1.76	(1.64 - 1.88)
NRT + bupropion	2.02	(1.70 - 2.40)	1.97	(1.65 - 2.33)	2.12	(1.78 - 2.50)
NRT + varenicline	3.54	(2.57 - 4.61)	3.44	(2.51 - 4.49)	3.60	(2.66 - 4.65)
Combination NRT	2.22	(1.91 - 2.55)	2.16	(1.87 - 2.49)	2.30	(2.00 - 2.64)
Varenicline	2.57	(2.32 - 2.85)	2.49	(2.24 - 2.78)	2.63	(2.37 - 2.91)
Varenicline + bupropion	3.20	(2.05 - 4.60)	3.11	(2.01 - 4.46)	3.26	(2.13 - 4.61)

Notes: RR, risk ratio; CI, credible interval.

The impact of length of follow up suggests that if all trials were followed up to 12 months, the treatment effects observed would be lower. This is plausible if the rate of failure (that is to say, recommencing smoking) is different in the control and intervention arms after six months. Given that the pharmacological treatments typically last for up to 12 weeks, it is possible that those in the intervention arm reach the point of no nicotine or active treatment three months after the participants in the control arm, and the failure curve may, therefore, be different.

The influence of continuous abstinence implies that studies that record cessation on the basis of continuous abstinence observe a greater treatment effect than those using a point prevalence estimate. Continuous abstinence is considered a better measure of smoking cessation, as point prevalence does not account for those who have short-term relapses. People with short relapses may be less likely to succeed in long-term quitting. How the choice of abstinence measure might lead to a consistent bias is unclear.

The meta-regression results should be interpreted with caution. The inclusion of covariates has a negligible impact on model fit, and the estimated impact may be influenced more by certain comparisons than others. For example, the use of continuous abstinence is least common in NRT trials, which also contribute the most evidence to the network. The potentially counter-intuitive findings, particularly with regard to continuous abstinence, may be an artefact or proxy for some other study feature.

## **Appendix B – Further reading**

Throughout the guidelines, key publications are cited as appropriate. However, a number of informative texts are available for more detailed treatments of some of the topics covered in these guidelines.

### **Evaluating interventions:**

Medical Research Council. Developing and evaluating complex interventions: new guidance. 2008. MRC ([www.mrc.ac.uk/documents/pdf/complex-interventions-guidance/](http://www.mrc.ac.uk/documents/pdf/complex-interventions-guidance/))

### **Patient-relevant outcomes (PROs):**

Brazier JE, Ratcliffe J, Saloman JA, Tsuchiya A. Measuring and valuing health for economic evaluation (2<sup>nd</sup> edition). 2017. Oxford: Oxford University Press.

Krabbe PFM. The measurement of health and health status. 2017. London: Academic Press.

### **Systematic reviews:**

Higgins JPT, Green S (editors). Cochrane Handbook for Systematic Reviews of Interventions (Version 5.1.0). 2011. The Cochrane Collaboration ([https://handbook-5-1.cochrane.org/front\\_page.htm](https://handbook-5-1.cochrane.org/front_page.htm))

### **Meta-analysis techniques:**

Borenstein M, Hedges LV, Higgins JPT, Rothstein HR. Introduction to Meta-Analysis. 2009. Chichester, UK: John Wiley & Sons.

Egger M, Smith GD, Altman DG (editors). Systematic reviews in health care: meta-analysis in context. 2001. London, BMJ Publishing Group.

Wells GA, Sultan SA, Chen L, Khan M, Coyle D. Indirect Evidence: Indirect Treatment Comparisons in Meta-Analysis. 2009. Ottawa, Canadian Agency for Drugs and Technologies in Health.

**Published by the Health Information and Quality Authority**

**For further information please contact:**

**Health Information and Quality Authority  
Dublin Regional Office  
George's Court  
George's Lane  
Smithfield  
Dublin 7**

**Phone: +353 (0) 1 814 7400**

**Email: [info@hiqa.ie](mailto:info@hiqa.ie)**

**URL: [www.hiqa.ie](http://www.hiqa.ie)**

**© Health Information and Quality Authority 2018**