NICE Guidelines Technical Support Unit

Meta-Analysis

Guideline Methodology Document 1

Version 1 (January 2021)

Caitlin Daly[1], Sofia Dias[2], Nicky J Welton[1], Sumayya Anwer[2], AE Ades[1]

[1] Population Health Sciences, Bristol Medical School, University of Bristol

[2] Centre for Reviews and Dissemination, University of York

About the NICE Guidelines Technical Support Unit

The NICE Guidelines Technical Support Unit (TSU) is a collaboration between the Universities of Bristol, Sheffield, York and Leicester. The TSU is commissioned by the Centre for Guidelines at the National Institute for Health and Clinical Excellence (NICE) to provide rapid-response technical support, methodology training, and methods research, in the context of guideline development. Please see this website for further information http://www.bristol.ac.uk/population-health-sciences/centres/cresyda/mpes/nice/

About the Guideline Methodology Document series

This series of Guideline Methodology Documents (GMDs) complements the Guide to the Methods of Technology Appraisal (1), the Guidelines Manual (2), and the NICE Decision Support Unit (DSU) Technical Support Documents (TSDs) (3-9).

The aim of the GMDs is to assist all those involved in guideline development, including guideline developers, guideline committee members, those commenting on draft guidelines during the consultation period, manufacturers, and stakeholders.

There is, of course, already a wealth of tutorial material on how to conduct systematic review and meta-analysis (10-12). The GMDs are in agreement with virtually all this material, although there are some significant differences in the way that meta-analytic methods are used.

The GMDs take the particular perspective of the guideline developer. They therefore go beyond standard treatments in which systematic review and meta-analysis tend to be seen as methods for producing "pooled" analyses that "summarise the literature". The decision context requires a focus on patients at specific points in their disease progression, methods that have particular properties regarding coherence and complete use of evidence, and procedures that are compatible with decision making under conditions of uncertainty.

The GMDs are aimed at a basic and introductory level: more advanced topics are indicated with an asterisk (*), and readers are referred elsewhere.

There are several areas of methodological uncertainty, controversy or rapid change. These are indicated in the GMDs. GMDs are extensively peer reviewed prior to publication (see acknowledgements). However, the responsibility for each GMD lies with the authors, who welcome any constructive feedback on the content, suggestions for updates and further guides. Readers should be aware that while the TSU is funded by NICE, these documents do not constitute formal NICE guidance or policy.

Acknowledgements

The TSU thanks the NICE Centre for Guidelines Methods and Economic Team, their NMA Working Group and Guidelines Methodology Group for their substantial contribution to this document. The joint editors for the GMD series are Nicky Welton (University of Bristol) and Sofia Dias (University of York). The production of this document was funded by the National Institute for Health and Clinical Excellence (NICE) through the NICE Guidelines Technical Support Unit. We are especially grateful to the external reviewers: Julian Higgins (University of Bristol), Alex Sutton (University of Leicester), Tom Trikalinos (Brown University). The views, and any errors or omissions, expressed in the Guideline Methodology Documents are those of the authors only. NICE and NICE Guideline Developers may take account of any part of this document, but they are not bound to do so.

Contents

* indicates advanced material that may be skipped

1. INTRODUCTION: WHAT THIS GUIDELINE METHODOLOGY DOCUMENT COVERS

This guideline methodology document (GMD) focuses on pairwise meta-analysis (PMA) but also includes references to general evidence synthesis issues that come up in guideline development, including network meta-analysis (NMA), and construction of a "baseline" natural history model of disease progression for use in economic assessment. Where appropriate, references to other material on these topics are provided.

Synthesis of data on diagnostic tests, medical devices, and prognostic markers is not covered.

This document begins by describing the motivation for conducting a meta-analysis and then gives an overview of the issues that will require attention when carrying out evidence synthesis in the context of guideline development: selecting the evidence (Section 3), defining the interventions (Section 4), multiple outcomes or follow-up times (Section 5), fixed and random effects (Section 6), the baseline model (Section 7), effect-modifiers (Section 8), bias-adjustment (Section 9), sensitivity analysis (Section 10), what should be reported (Section 11), software (Section 12), and embedding the model in a probabilistic cost effectiveness analysis (Section 13).

The aim is to provide a general orientation and to outline a recommended approach. It is not possible to cover every situation in a single document, so when non-standard analyses are mentioned we refer readers to published examples in NICE guidelines, where possible.

Technical details on ways of implementing these recommendations can be found in companion GMDs on continuous outcomes and event outcomes. Further technical guidance is available in the NICE Decision Support Unit documents (3-9), which are abridged in *Medical Decision Making* (13-19), and elsewhere (20). Finally, the document should be considered alongside the NICE Guidelines Manual (2).

2. WHEN TO CONDUCT A META-ANALYSIS IN THE COURSE OF GUIDELINE DEVELOPMENT

In the context of guideline development, the purpose of a meta-analysis, like the purpose of a randomised trial, is to inform a decision about how to treat a specified group of patients. This group of patients is usually defined by the specific point they have reached on their disease pathway. We refer to this as the target population and assume that this has been defined in advance.

A meta-analysis pools results on specified outcomes from multiple studies comparing treatment options for a defined target patient population. In this document it is assumed that information on relative treatment effects is exclusively sourced from randomised controlled trials (RCTs). Meta-analysis of RCTs pools *relative* treatment effects, which are measures of how treatments compare, for example mean differences or odds ratios. Like every statistical analysis, a meta-analysis embodies a *model* of the evidence, delivering pooled relative treatment effect estimates which can be used as a basis for treatment recommendations. Often treatment recommendations are based on cost-effectiveness evaluated in an economic model, where the key efficacy estimates are taken from a meta-analysis. For treatment recommendations to be transparent and robust, they should reflect all the relevant available evidence. A systematic review to identify the relevant available evidence, followed by a meta-analysis to pool the quantitative results of those studies achieves this.

A meta-analysis should therefore be undertaken whenever a quantitative assessment of a relative treatment effect and its uncertainty is needed, as in a cost-effectiveness or decision analysis, or whenever a quantitative assessment would be helpful in making treatment recommendations.

The task of making treatment recommendations is not necessarily be straightforward. To support decision making, the model of the trial evidence would have to be extended to consider the long-term effects of the treatment, and perhaps side effects. This would be necessary even if there was just a single trial conducted in the target population. If the trial was conducted in a somewhat different group of patients, its relevance might also be questioned. When the "evidence base" consists of two or more trials, the task facing decision makers can only become more difficult. Variations in trial quality, follow-up times, methods of outcome assessment, patient population, health-care settings, even different doses and co-therapies, may differ across studies and will need to be taken into account (see Section 9 on bias adjustment). Because meta-analysis pools *relative effects*, factors that need to be taken into account are those that lead to different relative effects (called *effect modifiers*). Prognostic factors which change the absolute outcome, but not the relative effects (i.e., those that are not effect modifiers), do not need to be accounted for in meta-analysis of RCTs.

3. SELECTING TRIALS TO INCLUDE IN THE META-ANALYSIS

The criteria for including trials in the meta-analysis will be constructed to match the target population. The extent to which inclusion criteria might be broadened to include trials on patients sampled from different, but similar, populations, is a matter for guideline developers, and this will often be determined before literature searches are conducted. In particular, if trial populations only differ in prognostic factors, even those these may influence the absolute outcome, and the relative effects from those trials are expected to be unaffected, it would be reasonable to include them in a meta-analysis.

One strategy would be to restrict inclusion to trials with the exact same target population; alternatively, inclusion can be broadened to studies on patients who are "similar", but not identical to the target population. As the inclusion criteria are broadened, which may be necessary if evidence is especially sparse, the relevance ("applicability") of the trial becomes more questionable, increasing the potential need for regression or bias adjustment (Sections 8 and 9) to take account of possible effect modifiers that would alter the relative treatment effect.

3.1. STUDIES WITH DIFFERENT DESIGNS

It is widely accepted that studies with different designs, such as cluster randomised trials, or cross-over trials (where these are appropriate), estimate the same treatment effects as the standard parallel RCT design where individuals are randomised. Therefore, these should be included in meta-analyses. However, care should be taken to assess that they have been analysed correctly, and that the estimators are similar across trial designs (10, 11, 21) (see Appendix A).

4. SELECTION AND DEFINITION OF TREATMENTS

Decision makers will want to consider all the eligible treatments for their target population. If there are more than two treatments under consideration for the recommendations, a network meta-analysis (NMA) will be needed; if there are only two treatments, a pairwise meta-analysis (PMA) will suffice. NMA and PMA are identical in every way, except for the number of treatments being considered. Both rely on the assumption that the included trials are similar in terms of effect modifiers, so that relative treatment effects are similar across all the included trials (22). This means that the considerations discussed in this GMD are equally relevant to NMAs and PMAs. Similarly, existing guidance on NMA methods and software are also relevant to PMA (3-9, 13-20).

We define "treatment" as a specific intervention or a pharmaceutical product at a specific dose and regime. Different products – for example different selective serotonin reuptake inhibitors (SSRIs), or different statins, or different tissue plasminogen activators – should generally speaking not be "lumped" together as if they are the same treatment, nor should different doses or regimens of the same product. Although there are some exceptions (see below), the practice of "lumping" treatments together, in order to have enough trials to put into a meta-analysis (23), should be avoided. Further, different treatments, doses and regimens should be not be studied using subgroup analysis, but should be treated as distinct treatments using NMA (4, 14, 24). This is more efficient and will provide the best set of relative treatment effects for decision making.

The definition of control groups requires a similar approach: no treatment, waitlist control, treatment as usual, attention placebo, and pill placebo should be considered different "treatments" and analysed in NMAs. This allows for the different placebo effects to be accounted for (25), and has been done in several NICE guidelines on: social anxiety disorder (26, 27), post-traumatic stress disorder (PTSD) (28) and post-operative management of Crohn's disease (29).

Nevertheless, in some circumstances, Guideline Developers may feel that it is reasonable to assume that particular treatments, doses or co-therapies have virtually the same effect. If this assumption is made, it must be stated explicitly and justified based on clinical grounds, and checked statistically in preliminary analyses (see Section 11). Note that treatments should never be lumped together solely on the basis that no difference could be detected in preliminary statistical checks.

4.1. MULTI-ARM TRIALS

Treatment arms that are not of interest can be ignored. If the effects of two treatments are considered to be virtually identical, the two arms can be pooled as a weighted average of the aggregate data and treated as a single arm. As noted above, this assumption should be based on clinical judgement, stated explicitly, and statistically checked. "Splitting" the control arm into two parts, to pair them with two active treatments of interest should never be done: in this scenario an NMA is required to properly account for the multi-arm structure.

4.1.1. Multi-arm trials in a network meta-analysis*

If more than two treatments are of interest, methods for NMA should be used. In an NMA, if it is assumed that two arms in a multi-arm trial have virtually the same effect, it is preferable to keep them as separate arms and index them as being the same treatment. In theory, this will contribute to the estimation of the between-study standard deviation and examples of this can be found in the PTSD guideline (28) and in the Bipolar disorder guideline (30).

4.2. * TREATMENT EFFECT MODELS

In some contexts it is useful to express the notion that an entire class of treatments – such as SSRIs, or statins – have "similar" treatment effects, by assuming a "class effect". Such models were used in NICE guidelines for Social Anxiety (26, 27) and Chronic Obstructive Pulmonary Disease (COPD) (31), which was informed by a Cochrane review of COPD treatments (32), and elsewhere for migraine (33), treatments for pressure ulceration (34), and for over-active bladder (35). Other examples of treatment that might be considered to fall into classes are: all cognitive therapies, all group therapies, all combination cognitive and SSRI therapies. Class models can take two forms: either assuming that all treatments in the class have identical effects, or allowing for differences between members of the class, as a compromise between identical effects and the entirely unrelated effects that would be assumed in a standard NMA. It can be useful to run all three models to investigate whether or not

there is evidence for differences in treatment effects among a set of related treatments (see the NICE guideline on tocolytic treatments for preterm labour (36) and management of COPD (31)).

A second kind of treatment effect model has been proposed for dose-response relationship (20, 37). In these models a functional relationship for the dose-response relationship is fitted, such as linear, log-linear or Emax relationships (37). This approach can improve precision of estimates compared with modelling doses as different treatments nodes in a network, but relies on an appropriate functional form for the dose-response relationship.

Finally, when treatments have multiple components, such as complex psychological interventions (38, 39), or double or triple therapies for chronic obstructive pulmonary disease (1, 40), it is possible to model the combined effect on the a priori assumption that the components act independently.

All these ways of modelling treatment effects require clinician input and preliminary analysis should be undertaken to check that the evidence available is consistent with the assumptions being made. Examples can be found in the references cited.

5. MULTIPLE OUTCOMES INCLUDING MULTIPLE FOLLOW-UP TIMES

It is assumed that guideline developers will have identified which trial outcomes are relevant to making a treatment decision in the specified target population. If more than one outcome is of interest, or if there are outcomes at multiple time points, it is important to consider how they are related. Each outcome or follow-up time can be examined in separate meta-analyses. However, this is an inefficient analysis that fails to capture the relationships between the multiple outcomes and does not help formulate treatment recommendations based on all the available evidence. This section summarises the main issues; specific methods and recommendations are made in the relevant sections of GMD2 and GMD3.

5.1. REDUCING MULTIPLE OUTCOMES TO A SINGLE OBSERVATION

If the selected trials report outcomes at different time points, guideline developers might determine the optimal follow-up time that is most relevant to their decision, and a range of follow-up times, on either side of the optimal time, within which they are prepared to assume that the relative treatment effects are virtually indistinguishable. Then, for a basic standard analysis, a single result can be selected from each trial, within this prescribed range and with a follow-up time as close as possible to the optimum. Trials not reporting the outcome within the prescribed range are excluded from the analysis.

The same approach can be adopted with trials reporting more than one outcome, out of a set of similar outcomes, for example HAMD, BDI, MADRS scales of depression. It is common to select a single outcome from each trial, based on a preference hierarchy. Rather than conduct separate meta-analyses of similar outcomes reported on different scales (for example HAMD, BDI, MADRS scales of depression, or, for social anxiety, the Leibovitz Social Anxiety Scale and the Brief Social Phobia Scale), it is preferable to adopt a strategy that allows a single unified analysis. Several methods of conducting a basic standard analysis are available, depending on how trial results are reported and what assumptions can be made. One is to analyse outcomes such as “proportion showing improvement”, another is “standardisation” using external or internal standards, and a third is the Ratio of Means method. These topics are discussed in more detail in GMD2.

5.2. SYNTHESIS OF MULTIPLE OUTCOMES

5.2.1.Within-trial pooling of multiple continuous outcomes

If it can be assumed that the relative treatment effect is the same regardless of follow-up time, pooling multiple observations into a single composite observation, taking account of the correlation between them, is a simple and transparent approach (see Section 5.4 in GMD2 for more details on within-trial synthesis). The pooled result, along with its variance, can then be used in conventional meta-analyses. Note that the benefit of pooling outcomes within-trials will depend on the correlations between the outcomes. If the correlations are close to 1, then the benefit will be minimal and data at the timepoint with the most precise estimate or the most commonly reported outcome could be used instead.

One difficulty in the case of continuous outcomes reported at different timepoints or on different scales is that the within-study correlations between outcomes are seldom reported. Guideline developers adopting this approach would need to source suitable values from external data. In this case, given the uncertainty involved, it is best to err towards a higher correlation, to avoid over-stating the precision of the pooled estimate. Standardised mean differences from multiple scales reported in the same trial were combined in this way in the NICE guideline of social anxiety (26, 27).

A quite different approach with correlated outcomes relies on the creation of a composite score: an example commonly given is a composite outcome combining continuous results of verbal and maths reasoning tests (10, 41). This is statistically valid, but only useful in a decision-making context if the composite score has a meaningful interpretation.

5.2.2.Within-trial pooling of multiple event outcomes

If a trial reports event probabilities at, say, 3 follow-up times, the data can be converted to 3, independent 2-by-2 tables representing the numbers at risk at the start of each follow-up period and the numbers reaching the endpoint during the period, by treatment. Odds ratios (or hazards ratios, see Example 1 in GMD3) can then be calculated and pooled within the study using a fixed effect procedure. These results are then pooled across trials either using fixed or random effects models as appropriate. Note this process assumes that the relative treatment effects remain constant over the trials' follow-up periods.

Alternatively, composite scores are often used with event data when decisions must be based on rare events, such as major adverse cardiac events (MACE) or major adverse neurological events (MANE).

5.2.3. Between-trial pooling of correlated outcomes: Multivariate synthesis*

If two or more outcomes are correlated, and if they all serve as inputs into a decision model, it is important that the correlations are taken into account, in order for the decision uncertainty to be correctly represented. Multivariate normal random effects models are now being proposed with increasing frequency, and they can be applied to both continuous or event outcomes. The purpose of these methods is to "borrow strength", and thus increase precision, by taking account of within- and between-trial correlations between outcomes. However, these models are sometimes difficult to fit, and may achieve little benefit in improved precision unless the within-study correlations are strong *and* only a few trials report all outcomes (42, 43). Indeed, if all trials report all the outcomes, there is no difference in precision between univariate and multivariate approaches.

5.3. * MULTIPLE TIME POINTS: EVENT DATA

Rather than reducing repeat observations over time to a single observation, the alternative is to include all the follow-up times, taking account of both the within-study dependencies between results at different follow-up times, and of possible ways in which time may affect event rates. Although introducing further complexity, these analyses have many advantages. First, they allow guideline developers to investigate whether event rates change over time. This is likely to be of critical importance if recommendations are to be based on a cost effectiveness analysis (CEA) or other forms of decision analysis. A second advantage of carrying out this more complex modelling is that it avoids arbitrary selection of which data to use, and in using all the data available it can generate more precise and robust results than the standard analysis. Piecewise constant hazards models offer a way of handling this and have been applied to gastro-esophageal reflux disease (44) and stents for cardiovascular disease (45).

5.4. * MULTIPLE STRUCTURALLY RELATED OUTCOMES

In many cases outcomes are *structurally related*. It is always possible to analyse them separately, but a single, combined analysis, that correctly reflects the structural relationships between outcomes, has many advantages. As with synthesis of data at multiple time points (Section 4.2), it represents all the available data in a coherent way, incorporates more data, providing more precise estimates, more robust decision-making, while avoiding arbitrary selection of evidence. In GMD-3 we refer to ordered categorical outcomes, such as those seen with the PASI score for psoriasis and ACR for rheumatoid arthritis; also to competing risk analysis of multiple outcomes. These are examples where the *within-trial* relationships between outcomes are taken into account.

Some more complex examples with structurally-related outcomes are:

- Combining data on median time to an endpoint, mean time to an endpoint, and proportion reaching an endpoint at a given time: influenza (46, 47)
- Chain of evidence: intravenous antibacterial prophylaxis for early onset neonatal Group B strep (20, 48)
- Combining data on time to tumour progression, survival and probability of response: advanced meta-static breast cancer (49)
- Combining data on time to tumour progression and survival in a partitioned survival model: NICE guideline on lung cancer (50)
- Synthesis of intermediate outcomes: coronary patency and mortality following ischaemic heart attacks (51)
- Simultaneous within- and between-trial synthesis and mapping between similar measurement scales: ankylosing spondylitis (52), social anxiety disorder (53), depression (54)
- Partially observed transitions in a Markov model with identification of the transition at which the treatment effect occurs: treatments for asthma (55)

Clinical input is essential to confirm that relationships between outcomes are being modelled in a realistic way, and assumptions should be checked wherever possible in preliminary analyses. The relationships between outcomes are typically discussed during the development of the economic model; however, these discussions should start as early as possible, with the intention of capturing these relationships in the meta-analyses.

6. HETEROGENEITY AND CHOICE OF FIXED OR RANDOM EFFECT MODELS

Heterogeneity is considered to be present when the differences between the estimated treatment effects from different trials are greater than what would be expected from sampling variation alone. Setting aside biased selection of studies due to mechanisms like publication bias, heterogeneity is a result of differences in effect modifiers between included studies.

6.1. CHOICE OF RANDOM OR FIXED EFFECT MODELS

In a fixed effect model, it is assumed that every trial estimates the same treatment effect for a specific comparison. If all trials were infinitely large, we would observe identical treatment effect estimates across trials. The random effects model assumes each trial estimates a different treatment effect drawn from a common distribution. The study-specific treatment effect estimates generated by the random effects model are not the same as the observed treatment effects, but are "shrunken" towards the mean effect (see Appendix B, Table B.3 for an example) (56).

Statistically significant tests of heterogeneity, such as the chi-square tests (57, 58), or meaningful improvement in the goodness of fit of the random effects model compared to the fixed effect model, should generally be interpreted as ruling out fixed effect models. However, failure to clearly establish that there is heterogeneity should not necessarily be interpreted as meaning that a fixed effect model is appropriate if there are insufficient studies to estimate the between studies variance.

Guideline developers should seek clinician input on whether they believe the true treatment effects differ across trials. (The "true" treatment effects being those that would be observed if the trials were infinitely large). This would be enough to justify the use of random effects models, regardless of how much data are available to inform the extent of random variation. However, imposing a random effects model when there is not enough data to estimate between-study heterogeneity will generally require that Bayesian methods are adopted with informative priors on the between-trials variation (see Section 6.3.1).

6.2. MEASURES OF BETWEEN TRIAL VARIATION

The best measure to present is the between-trials standard deviation (SD), because this is measured in the same units as the modelled treatment effect, whether that is in units of continuous measurements such as kilograms, units of blood pressure, log odds ratio, or log hazard ratio. The degree of heterogeneity, reflected by the magnitude of the between-study SD, should be interpreted in context of the scale of the treatment effect. Some software reports the between-trial variance, which is the square of the between-trials SD. The I^2 measure (57) does not measure the amount of between-trial variation and should not be used for this purpose (59, 60).

Alongside an estimate of the between-trials SD, it is also important that there is a realistic assessment of the *uncertainty* in the estimate, usually in the form of confidence interval, or credible interval. For this reason, we prefer the *metafor* R package to RevMan, as *metafor* provides confidence intervals for both between-trials variance and between-trials SD. In the *metafor* package, the Paule-Mandel estimator for the between-study variance should be specified, as recommended by the authors in (61).

Between-trials variation impacts the uncertainty in the mean "pooled" effect, and has a major effect on the *predictive uncertainty,* which represents uncertainty in the true treatment effect we might expect to see in the target population assuming that it is similar to the populations in the included studies (59, 62). As a result it has been recommended that the "predictive treatment effect" is reported instead of the mean treatment effect, as its wider confidence interval better represents the

range of plausible treatment effects we might expect to observe in a "new" population similar to those in the included trials (63) (see Appendix B).

6.3. HETEROGENEITY AND SPARSE DATA

Sparse data may arise when dealing with rare events or when there are simply a small number of studies making a treatment comparison. Both situations have implications in meta-analysis. Zero events in either or both arms impact the calculation of the relative effect in a trial and methods for dealing with this are discussed in GMD3.

Where data are sparse, there can be considerable uncertainty about the degree of between-trial variation. With some frequentist software, sparse data may lead to an estimate of zero between-study variance, implying a fixed effect model. In other situations, depending on the meta-analytic methods used (64) sparse data may lead to wide confidence or credible intervals not only on the between-study SD, but also the treatment effect itself. Occasionally this will result in upper limits on the relative treatment effect that are far beyond the bounds of clinical plausibility. In these situations, borrowing external information may help estimation, as described below. Alternatively, placing a model on the baseline effect offers another way of dealing with sparse data (34, 65).

6.3.1. Informative priors in a Bayesian framework*

Excessively wide confidence or credible intervals occur in both Bayesian analyses assuming vague priors for the between-study standard deviation, or -equally - in frequentist analyses. This is because frequentist analyses implicitly assume an infinitely vague prior.

In these circumstances, guideline developers should consider using Bayesian methods with informative prior distributions on the between-trials variance. Meta-epidemiological data can be used to provide an "evidence based" prior for the extent of between-trials variation (66, 67), or one might use the findings from previous Cochrane reviews of similar outcomes in similar trials. This approach has been adopted in NICE guidelines on: bronchopulmonary dysplasia as an outcome in Specialist Neonatal Care for Babies born Preterm (68); post-traumatic stress disorder symptom scale scores in treatments for children and young people (28). However, priors derived from Cochrane reviews tend to be rather weakly informative as the studies on which they are based are quite heterogeneous.

Another approach is to derive informative priors from clinical opinion: for example, it is possible to create a prior that puts limits on how much treatment effects drawn from a random effect distribution can depart from their mean value: for example, one can specify that 95% of ORs will be within a factor 2 above and below their median value (20). Whichever approach is taken, if data are sparse and credible/confidence intervals unreasonably wide, use of suitable informative prior distributions is likely to yield results that are closer to "the truth" than frequentist analyses or a Bayesian analyses with vague priors.

6.4. SOURCES OF HETEROGENEITY

There are many sources of heterogeneity, including:

1. Clinical heterogeneity, i.e. variation across trials in the distribution of patient-level, or trial-level, effect modifiers ("external" biases), including baseline severity. (An external bias means that the trial does not estimate the treatment effect in the guideline developer's target population (69)).
2. Random variation in "internal" biases. (An internal bias prevents a trial from estimating the true effect in *the trial's* target population (69)).

3. Variation in how the outcome is reported, for example different definitions of what constitutes an event (e.g. response), different measurement scales or follow-up times.

Note that reported variations in treatments, such as different doses or different co-treatments, should not be seen as a source of heterogeneity, as we assume this will usually be handled by treating them as separate treatments in an NMA (Section 3). Alternatively, they might be explicitly assumed to have no impact on efficacy (Section 4), and "lumped" together, which would also mean they were not a source of heterogeneity.

Factors contributing to clinical heterogeneity will have been assessed during guideline development when considering trial inclusion criteria. If it is believed that heterogeneity within the included trials can be related to specific factors – *effect modifiers* – this raises a number of further issues (see Section 8).

Internal biases will be documented by application of risk of bias (RoB) tools, such as the one proposed by the Cochrane Collaboration (70). If it is considered that there are random internal biases which impact on the treatment effect, and which vary from trial to trial, various methods are available which have the effect of simultaneously (a) estimating the size and direction of the bias, (b) adjusting the treatment effects for bias, thus recovering the "true" treatment effect, and (c) down-weighting evidence from trials vulnerable to potential bias (see Section 9).

7. META-ANALYSIS TO ESTIMATE THE BASELINE MODEL

Guideline developers may wish to have an evidence-based estimate of the *absolute* outcomes (eg probability of an event) on each treatment, as well as an estimate of the relative treatment effect (eg log-odds ratio). This can be achieved by applying the pooled relative effect from the meta-analysis to the estimated absolute outcome(s) on the control/reference treatment. For example,

$$\text{log-odds}(pB)=\text{log-odds}(pA)+\text{logOR}$$

where pA and pB are the absolute probabilities of outcome on treatments A and B respectively, and logOR is the log odds-ratio for treatment B relative to treatment A.

The absolute outcome (or outcomes for multi-outcome meta-analysis) on the reference treatment (e.g. pA) is sometimes referred to as the *baseline model.* This is needed whenever there is a need to offset the benefits of treatment against side effects and other harms, or when the Number Needed to Treat is to be estimated.

When a CEA or other formal decision analysis is to be performed, the requirement is not simply for an estimate of the absolute outcome in the time horizon of the trials, but for a full natural history model to inform estimates of lifetime costs and expected quality of life. Methods for extrapolation to a lifetime model are not covered here, and we refer to general text (71) and the NICE TSDs on economic modelling (72-77) which cover these issues.

The construction of the baseline model is an entirely separate exercise to the construction of a model for the relative treatment effect. One option, which is not recommended here, is to estimate the baseline effect by carrying out a meta-analysis of all the arms on the reference treatment – usually placebo, or the most common standard treatment. Instead, guideline developers should ask themselves the question "which trial(s), or which observational data source(s), best represent the outcomes that would be observed on the reference treatment if it was to be rolled out now, given

contemporary standards of care?". The answer might be the control arms in a subset of the trials informing the relative treatment effect, a single trial, or one or more observational databases. Cost-effectiveness may be sensitive to outcomes on the reference treatment, and so the evidence sources for the baseline model should be reported and justified in detail.

7.1. ESTIMATION OF THE BASELINE MODEL FROM ONE OR MORE CONTROL ARMS

The simplest solution is to carry out a meta-analysis of a subset of the control arms, selected to be representative of absolute outcomes that would be observed on the control treatment. The meta-analysis should be on the same scale as the meta-analysis of the relative effects. Thus, if the relative effect is in continuous units, a log OR, or a log HR, the baseline model will be in the same units, a log-odds, or a log hazard rate. An example can be found in the NICE guideline on attention deficit hyperactivity disorder (ADHD) (78), and Appendix B illustrates how to do this in various software.

There have been situations where there is some heterogeneity between the selected studies' baseline effects, but there is not enough evidence to estimate the between-study SD. Pooling these studies in a fixed effect meta-analysis is not appropriate, as the uncertainty due to the variability of the baseline effects will not be fully captured. In this case, one option would be to conduct sensitivity analyses, where a single study is selected to inform a base-case analysis, and the robustness of conclusions is assessed through a sensitivity analysis using another study/ies to inform the baseline effect. This was the case for the acute coronary syndrome guideline (79).

7.2. ESTIMATES FROM COHORT STUDIES AND MODELS

If guideline developers believe that there are cohort studies that provide the best estimate of the baseline model for the target population, then these can be synthesised in the same way. Cohort studies may be pooled with control arms from RCTs. Cohort studies may also provide information about longer-term extrapolation of outcomes on the reference treatment. An example from NICE Guidelines is in treatments for post-surgical maintenance of remission from Crohn's disease (29). In the guideline on tocolytic treatment for preterm labour (36), the absolute event probabilities were dependant on gestational age, based on a regression of baseline effect against gestational age.

7.3. INFERRING A BASELINE FROM A STUDY OF NATURAL HISTORY UNDER TREATMENT

It is also possible to use cohort studies which have followed up patients not on the reference treatment A, but on one or more of the other treatments under consideration. The absolute effects on reference treatment A can then be found by subtracting the relative treatment effect d_{AB} from the absolute effect on treatment B. The calculations should be carried out on the same scale as the meta-analysis. See the NICE guideline on surgical site infections for an example (80).

7.4. MODELLING THE BASELINE EVENT RATE OVER TIME

A particular issue to consider when conducting a CEA is whether the baseline event rate changes over time, and indeed whether the relative treatment effect changes over time as well. It should be emphasised that a reduction of event rates over time is highly likely, just because individuals in trials have different levels of baseline risk. Inevitably, those at higher risk reach endpoints earlier, so that the hazard rate must diminish over time. This effect is called "depletion of susceptibles". Models allowing for changes in baseline event rates over time have appeared in NICE guidelines (29). Models in which both baseline event rates and relative treatment effects (hazard ratios) change over time have been applied to gastro-esophageal reflux disease (44), and to stents for cardiovascular disease (45).

8. EFFECT MODIFIERS: META-REGRESSION AND SUBGROUPS

8.1. WHAT IS AN EFFECT MODIFIER?

Effect modifiers are variables which change the relative treatment effects. Their presence has a profound effect on the way a meta-analysis is used when making a treatment recommendation. Among the most frequently considered effect modifiers are: age, gender, disease severity, number of years since diagnosis and previous lines of therapy. Disease severity may be measured as severity at "baseline", for example at recruitment, or represented by the event rate in the control arm. Note that it is not possible to randomly allocate patients to effect modifiers, which means that analyses of effect modification are effectively observational, even if the data come from randomised trials.

8.2. MODIFICATIONS OF THE TREATMENT

We noted in Section 4 that the treatment, whether a pharmaceutical or other intervention, could be modified. Pharmaceutical treatments can be delivered in different doses, with different co-treatments, and in different regimes. Similarly, psychological treatments can be delivered by staff of different grades, trained in different ways, in groups or individually, and for different lengths of time. These are *not* effect-modifiers, but modifications of the treatment. The distinction can be easily recognised from the fact that, unlike effect modifiers like disease severity, it is possible to randomise patients to the different varieties of treatment.

8.3. IDENTIFYING EFFECT MODIFIERS

8.3.1. Categorical covariates and aggregate data

At the simplest level it is possible to check for effect modifiers via "sub-group analysis". This is quite inefficient as a separate meta-analysis is carried out for each group, with different between-study variances in the case of random effects models. A more efficient way to implement a sub-group analysis is by a meta-regression in which a categorical variable represents the groups. Asingle heterogeneity parameter can be assumed for both (or all) groups. This is available in *metafor*.

Both methods are, however, vulnerable to confounding: it must not be forgotten that meta-regression is quite unlike meta-analysis, as there is no randomisation to the different covariate values. A positive association with, say, gender, might reflect other factors which vary between trials which happen to be associated with gender.

8.3.2. Analysis of aggregate "within-trial" subgroups*

A somewhat more powerful analysis is possible when trials report results for different subgroups separately. This avoids cross-trial comparisons which are highly vulnerable to confounding. However, there is still a vulnerability to confounders if, within trials, the covariate – say gender – is correlated with – for example – disease severity.

8.3.3. Continuous covariatesand aggregate vs. IPD meta-regression

If the variable is continuous (age, duration of illness), a meta-regression of the treatment effect against the *average* value of the covariate in each trial can be undertaken. However, this is unlikely to find clear evidence of an interaction, unless the effect is very strong, and trials differ markedly in average

covariate values. With any individual-level covariate, aggregate data are highly inefficient compared to individual patient data (IPD) (81), as well as being vulnerable to ecological bias (82).

A more thorough analysis of whether or not a continuous covariate is an effect modifier, *must* therefore include a literature search of the trial evidence, looking for reports where interactions were tested and reported. It is possible to carry out a meta-analysis of the interaction terms (83), but the possibility of selective reporting of "significant" results must also be borne in mind.

8.3.4. Mixed IPD and aggregate data*

There is literature on the combination of IPD and AgD in a pairwise and network meta-analysis context (84-88). These methods typically assume common regression coefficients at both the individual and aggregate level, which leads to aggregation bias (a form of ecological bias) when the model is non-linear (89). One solution is to estimate *both* an IPD-level *and* an AgD level interaction effect. This can be conceptualized as having a between-study regression model in which the study mean covariate values are the covariates, and a within-study regression model in which the covariates are the patient covariate values *minus* the study means.

A second solution similarly estimates regressions at both the within- and between-trial levels, but accounts for the mathematical relation between the two sets of coefficients (90, 91). This is known as population-adjusted evidence synthesis (92).

8.3.5. Summary: Deciding whether a covariate is an effect modifier

Because of the impact of effect modifiers on the nature of the decision problem, it is essential that the process of identifying potential effect modifiers is carried out diligently. While all the above analyses are available to assist this modelling decision, guideline developers should be very cautious in concluding that a variable is an effect modifier based on statistical evidence alone. It is important to pre-specify which variables are going to be considered as potential effect modifiers in advance, and variables should only be accepted as effect modifiers if this is clinically reasonable *a priori, as specified in the protocol* (1).

8.4. Different recommendations for different patients, or a single recommendation in a mixed population

Once it is accepted that a factor – say disease severity – *is* an effect modifier, two courses of action are open. One option is to make separate recommendations for patients with and without severe disease. For a continuous variable such as disease severity, or baseline risk, this requires an explicit threshold to separate the groups. Separate guidance was issued for mild-to-moderate and moderate-to-severe depression (93).

The second option is to make a single recommendation for the (combined) target population; in this case, if a formal analysis of efficacy or cost-effectiveness in the combined population is required, it is necessary that the proportions of "severe" and "non-severe" patients are known. Then it is possible to produce an estimate of the relative efficacy, or the cost-effectiveness, of each of several treatments in any specified population (94).

Note that if the meta-analysis is of a continuous outcome, or on a risk difference, then the treatment effect in the combined population can be calculated as a simple weighted average of the treatment effects in the two populations. However, log hazard ratios, log risk ratios, and log odds ratios *cannot* be averaged in this way, and require a different computational approach to account for non-linearity, such as numerical integration via Monte Carlo (MC) simulation.

It is important to appreciate that the issue of whether there are different decisions for different patients, or a single decision for a combined population comes up regularly when *absolute treatment effects vary,* even though relative effects may not vary by subgroup. For example, with statins an individual risk equation is routinely applied, with different decisions for those above and below a specified threshold (95).

9. * BIAS ADJUSTMENT

In this section we consider steps that can be taken when it is believed that the trial evidence is biased. There is a useful distinction to be made between "internal" biases, which result in trials failing to estimate their intended target parameter, and "external" biases, in which a trial correctly estimates its target parameter, but this is not the target parameter for the decision (69). Among the causes of internal bias are: lack of blinding, failure to conceal treatment allocation and selective loss to follow-up. The potential for external bias arises when the patient population is not the same as the decision population, and they differ in a variable that is an effect modifier. It is possible for a trial to be vulnerable to *both* kinds of bias.

As well as specific shortcomings in the conduct of trials, there is also literature referring to "novelty bias", "sponsor bias", "optimism bias", "outcome reporting bias", and of course "publication bias", which may be related to study size, . The latter is a form of bias which are not related to the trial itself, but which attaches to the ensemble of trials, due to the mechanism by which they were identified.

Application of the Cochrane Risk of Bias tool will only identify and record internal biases. Bias-adjustment methods are more pro-active: recognising that bias may be present, they aim to correct for it, and, because there is uncertainty about the degree of bias, the evidence is down-weighted. Three methods have been proposed which are summarised below. Adjustment for small study effects through regression may also be carried out, which may address publication bias (96, 97).

9.1. ADJUSTMENT OF TRIAL TREATMENT EFFECTS BASED ON EXPERT OPINION

This method relies on having a panel of experts who are able to provide a quantitative assessment of the extent of bias in the form of a probability distribution (69). Separate elicitation exercises are used for internal and external biases. The distributions from the different assessors are combined, and then used to modify and down-weight the trial data before a standard meta-analysis takes place.

9.2. ADJUSTMENT USING META-EPIDEMIOLOGY DATA

Studies have shown that trials which lack blinding, or in which allocation was not properly concealed tend to have larger treatment effects relative to the control treatment, than trials without these markers of poor quality (98-101). Empirically, the extent of the mean bias observed in meta-epidemiological studies has been higher with "subjective" outcomes, and lowest with mortality outcomes (99, 100).

Using meta-epidemiological data, it is possible to estimate both the mean "bias" associated with these markers of risk of bias, and also the between-study variation in risk of bias (102). Once these statistics have been estimated, they can be applied in a particular meta-analysis to adjust out the bias in the studies with risk of bias markers, and also down-weight them. This approach avoids having to choose between using high quality evidence alone, or ignoring the potential bias in low quality evidence.

9.3. ADJUSTMENT USING NMA

In a network of more than two treatments, it is often possible to simultaneously estimate and adjust for bias in trials with risk of bias markers. This has been done for markers related to trial quality, such as blinding or allocation concealment (25), so-called "sponsor bias" (103), "novelty bias" (104), or small study bias/ publication bias (97, 105). A worked example and further references can be found in (20); the method has been used in the NICE guideline for eating disorders (106).

10. SENSITIVITY ANALYSES (SAs)

In a decision-making context involving a formal CEA or decision analysis, studies of sensitivity of recommendations to assumptions are routinely carried out. This lies within the scope of CEA. However, in this section we mention specifically sensitivity analyses that are focused on the treatment efficacy parameters. We assume throughout that the recommendations are based on a probabilistic model, so that uncertainties and non-linearity are taken into account (107).

Given that, we would recommend the following SAs:

1. If there are doubts about the quality or relevance of specific trials, or sets of trials, they can be omitted in SAs.
2. A more formal approach can be adopted which asks the question: how different would the results of this trial, or this set of trials, have to be before it changed the treatment recommendation (108). This is a form of threshold analysis, which can be applied to recommendations based on treatment efficacy whether from PMA or NMA, and to some recommendations based on CEA. They have been applied in a NMA of social anxiety (109) and in NICE guidelines on specialist neonatal respiratory care (68). Application to the full range of CEAs is under development.

All sensitivity analyses should be reported and commented on.

11. REPORTING RESULTS OF AN EVIDENCE SYNTHESIS

Although it is important that both results and modelling decisions are reported accurately and in full, at the same time it is unhelpful to burden readers with information which has not contributed to the guideline development. It is useful to distinguish between three kinds of analysis that are used to support the guideline development process:

- Results of meta-analytic models that are used to drive treatment recommendations, whether these are based on cost-effectiveness, or efficacy alone
- Results of preliminary or exploratory analyses whose purpose is to check model assumptions
- Other analyses

We distinguish between the first two types of analyses, which must be explained in full, and the third, which may be included out of completeness, but should be kept together in a separate appendix, which is specifically marked as containing material that is not used in the guideline development, for reference purposes. Among items in this category are: meta-analyses relating to each individual outcome measure at each individual follow-up time and GRADE confidence assessments. Risk of Bias tables could be regarded in the first two categories if they have impacted the treatment recommendations, otherwise they should be included under the third category.

11.1.PMA & NMA: IF QUANTITATIVE ESTIMATES ARE USED DIRECTLY IN RECOMMENDATIONS

- Meta-analytic method (by name), with citations

- Software, including version number and any packages used
- Code from statistical software (e.g., WinBUGS, R, STATA)
- Data (in appendix or on web)
- Reasons for selecting FE or RE model
- Between trial SD & 95%CI
- Goodness of fit statistics with comments
- Relative treatment effects of each treatment relative to a placebo, or a commonly used "Control" treatment, on the log scale or natural scale (ie LogOR, or OR).
- Predictive relative effects, if these are used in recommendations
- If CEA or other formal decision analysis is used, the absolute effects on each treatment should be tabulated
- Any formal checks for effect modifiers
- Key trial level covariates: we suggest a table with a row for each trial, and columns for each key covariate, showing for example: %male; Mean, SD, range of age; etc. (Characteristics of included studies table).

11.2. RESULTS SPECIFIC TO PMAS

- Forest plots

11.3. RESULTS SPECIFIC TO NMAS

- the relative treatment effects of each treatment relative to every other treatment (upper or lower triangle table) - optional
- "forest"-like plots presenting direct, indirect, and NMA relative effect estimates on each contrast
- Rankograms or posterior rank statistics including 95% credible intervals (110) (optional)
- SuCRA plots (Surface under the cumulative Ranking curve) (110) (optional)
- Inconsistency analyses. See Dias (2018) (20) for suitable methods and worked examples. Examples using the Bucher method (111), *inconsistency models (6, 16), and *node splitting (6, 16) can be found in NICE guidelines

11.4. MODELS WITH MULTIPLE OUTCOMES

- Describe method, with citations, or include algebra, and software code and data
- Summaries of relationships between outcomes in tables or graphs

11.5. PRELIMINARY ANALYSES, CHECKING OF ASSUMPTIONS

- Summary narratives of the evidence base: for example, tables of the number of trials reporting each outcome at each time point
- Analyses used to check assumptions in a model on which recommendations are based, for example.
    - relative treatment effects against FU time, with line drawn between points from same study
    - equivalence of slightly different treatments
    - class-effect assumptions / dose effect assumptions / component model assumptions etc
- Comparisons of separate analyses by outcome with output of multi-outcome analyses

11.6. SENSITIVITY ANALYSES, INCLUDING THRESHOLD ANALYSES

- Sensitivity analysis: Forest plot or table of the relative effects resulting from the base-case and sensitivity analyses.
- Threshold analysis: Forest plot displaying the study or contrast-level estimates, together with the invariant intervals.

12. SOFTWARE

The following sections summarizes available software for conducting pairwise or network meta-analysis. Bayesian software are marked with ** and it is important to assess convergence when using such software (112). A summary table is provided (Table 1); note this list of software is not exhaustive.

12.1. R

There are several packages available in R to conduct a pairwise meta-analysis. These include, *bayesmeta***, *bmeta***, *dmetar*, *meta*, *metafor*, *rmeta*. The package *metafor* (113) is widely used and can carry out a variety of analyses (e.g., Mantel-Haenszel method for pooling event data, inverse-variance weighted approach, meta-regression which is referred to as moderator analysis in the package documentation). We note here that for event data, the escalc() command's default option for handling zero-cells is to impose a continuity correction, where 0.5 is added to all events and non-events in a trial. This is not recommended for the Mantel-Haenszel method and this is accounted for in the rma.mh() command of *metafor* version 2.4-0. By default, a continuity correction is applied to trials with zero cells in order to include them in forest plots, but the Mantel-Haenszel estimates do not include a continuity correction. Pairwise meta-analysis may also be carried out using network meta-analysis software. To conduct a network meta-analysis in R, the main packages are: *gemtc*** (114), *BUGSnet*** (115), and *netmeta* (116). We also note that published WinBUGS code for pairwise or network meta-analysis may be implemented in R using the *R2WinBUGS*** or *R2OpenBUGS*** packages (117). The *gemtc* package can handle the different types of outcomes covered in (4, 14) apart from the multinomial likelihood, and includes the inconsistency model (6, 16), and has automated the node-splitting process to assess inconsistency (6, 16, 118, 119). One thing to note is that the *gemtc* package does not accept non-integers for event data, as a continuity correction is not usually required in a Bayesian framework to estimate relative effects in studies containing a zero cell. Nevertheless, the continuity correction might help stabilise results, in which case data will have be inputted as relative effects (e.g., log-odds ratios) in order to be accepted by the *gemtc* package. *BUGSnet* is a fairly new package for Bayesian NMA that can model the same type of outcomes as *gemtc* (115). It was motivated by the need to generate the statistical outputs recommended by reporting guidelines, including the NICE DSU reviewer's checklist (9). In *netmeta*, the network meta-analysis function requires trial data to be inputted on a contrast-level (e.g., mean differences or log-odds ratios), which may be calculated within the package using the pairwise() function. In addition to traditional network meta-analysis models, models assuming additive effects of components within complex interventions may be fitted in the *netmeta* package.

Note: Avoid the *pcnetmeta* package for network meta-analysis, which provides functions for arm-based network meta-analysis and is not recommended.

12.2. REVIEW MANAGER 5.3 (REVMAN)

Fixed- and random-effects pairwise meta-analysis may be conducted in RevMan (120). Binary and continuous data may be inputted at arm- or contrast-level, while rate data can only be inputted at

contrast-level (i.e., log-HRs). The Mantel-Haenszel method may be used to pool binary data. However, when a study contains a zero cell, the software computes the odds ratio with a continuity correction imposed on the trial data, even though this is not required. Thus, we recommend the *metafor* package in R when dealing with this situation. Sub-group analyses may be performed in RevMan, however meta-regression and network meta-analysis cannot be conducted in this software.

12.3.** WinBUGS or OpenBUGS

Code for fitting pairwise meta-analysis and network meta-analysis may be found in (4); meta-regression models in (5); models that help assess consistency in (6); and models for estimating baseline effects in (7).

12.4.(**) MetaInsight

A more user-friendly interface for conducting NMAs is offered by MetaInsight (121), where binary (probability) or continuous outcomes may be modelled. This software platform calls upon the *gemtc* or *netmeta* packages in R conduct an NMA in a Bayesian or frequentist framework, respectively. Inconsistency may be assessed through node-splitting. Currently, only arm-level summaries are accepted (i.e., not relative effects), and a maximum of 6 arms if permitted in each trial.

Table 1: Summary of key software features for (network) meta-analysis

Software Package	Code/Function/Analysis			
	Name or reference	Description	Outcome	Notes
R: *metafor*	rma()	Fixed or random effects meta-analysis, including meta-regression	Generalised for most outcomes. Outcome measure can be specified using the 'measure' argument.	Meta-regression may be conducted by specifying covariate under 'mods' argument.
	rma.mh()	Fixed effect meta-analysis using Mantel-Haenzsel method	Dichotomous (2 x 2 tables) and person-time data.	Mantel-Haenzsel estimate does not include continuity correction by default in version 2.4-0.
R: *gemtc*	mtc.model()	Fixed or random effects network meta-analysis	Generalised for most outcomes described in TSD2 (4)	Continuity correction cannot be applied to arm-level data. Duplicate treatment arms within a study are not accepted.
R: *BUGSnet*	nma.model()	Fixed or random effects network meta-analysis	Generalised for most outcomes described in TSD2 (4)	Continuity correction cannot be applied to arm-level data. Duplicate treatment arms within a study are not accepted.
R: *netmeta*	netmeta()	Fixed or random effects network meta-analysis	Generalised for most outcomes, provided they are inputted on contrast-level.	Ratios should be inputted on log-scale. Duplicate treatment arms within a study are not accepted.
	netcomb()	Fixed or random effects network meta-analysis, assuming additive effects of components within complex interventions		Network meta-analysis model must be fitted first using netmeta().

Table 1 (Continued): Summary of key software features for (network) meta-analysis

Software Package	Code/Function/Analysis			
	Name or reference	Description	Outcome	Notes
RevMan	Inverse-variance	Fixed or random effects meta-analysis	Dichotomous and continuous data may be inputted at arm-level. All outcome data may be pooled as contrasts.	Subgroup analysis is possible, but not meta-regression.
	Mantel-Haenszel	Fixed or random effects meta-analysis	Dichotomous data (2 x 2 tables)	Continuity correction applied in Mantel-Haenszel method.
WinBUGS	TSD2	Fixed and random effects meta-analysis	Generalised for most outcomes, refer to Table A1 in Appendix to find appropriate code for data type.	All programs may be used for pairwise MA by specifying na[]=2 in data.
	TSD3	Meta-regression	Can be generalised for most outcomes, by adding appropriate code to programs set out in TSD2.	Subgroup analysis may be conducted using this code, where a common between-study SD is assumed in random effects models.
MetaInsight		Fixed or random effects network meta-analysis	Dichotomous and continuous data may be inputted at arm-level.	Sensitivity analysis may be conducted, where trials may be removed.

13. EMBEDDING EVIDENCE SYNTHESIS IN A PROBABILISTIC CEA

Absolute outcomes for each of the treatments under consideration are key inputs for economic models assessing cost-effectiveness. As described in Section 7, absolute outcomes are obtained by applying the relative treatment effect from the meta-analysis model to the absolute effect on the reference treatment from the baseline model, for example:

$$\text{log-odds}(pB)=\text{log-odds}(pA)+\text{logOR}$$

where pA and pB are the absolute probabilities of outcome on treatments A and B respectively, and logOR is the log odds-ratio for treatment B relative to treatment A.

However, the baseline and meta-analysis models are estimated with uncertainty and it is important that this uncertainty is reflected into the economic model (71). This is typically achieved by simulating each model parameter from a probability distribution that reflects the uncertainty in the estimation. Two approaches are commonly used (8, 18).

The 2-stage approach first approximates the uncertainty in the estimation with a probability distribution. For example, Beta distributions are often used for probability parameters, and Normal distributions for log-ORs, log-HRs, and continuous outcome scales. In the second stage, the parameters are simulated from this distribution and the economic model is evaluated at each simulation step, giving a set of simulations reflecting uncertainty in the cost-effectiveness estimates.

The 1-stage approach (122) can be used when a Bayesian approach has been taken for the meta-analysis using Monte Carlo Markov chain (MCMC) simulation. The MCMC simulations can be used directly as the simulated parameter values with which to evaluate the economic model. Note if there are multiple outcomes estimated in the meta-analysis, then the correlations in these parameter estimates must be preserved in the economic model.

13.1.* SUMMARISING RANDOM EFFECTS MODELS

When we have fitted a fixed effect meta-analysis model, it is clear that the relevant summary to use in an economic model is the pooled relative effect (eg logOR). However, when we fit a random effects model, we have estimated a distribution of relative effects that has a mean relative effect and a between-study standard deviation. A variety of different summary measures have been proposed, and are appropriate in different situations (56, 94, 123, 124). The mean of the random effects distribution is appropriate if we consider that the relative effects we expect to see in our target population is most likely to lie in middle of the effects seen previously. However, if our target population is similar to the trial population in a particular RCT , then we might consider the shrunken estimate for that study may be the most appropriate estimate to use (see Appendix B, Table B3 for example). If a meta-regression has been used to explain some, but not all of the residual heterogeneity, then the study shrunken estimate obtained at a specified value of the covariate in the meta-regression model may be used (see Appendix C for example). Finally, we may not know where in the random effects distribution our target population may lie, in which case the effect we expect to see may be described by a randomly selected trial effect drawn from the random effects distribution, known as the *predictive distribution*. The predictive distribution is centred on the mean of the random effects distribution, but it has a wider confidence/credible interval than the random effects mean because it represents both the uncertainty in the mean of the distribution and also the uncertainty as to where in the distribution the relative effect for the target population might lie.

If it is possible to fully or partially explain heterogeneity with covariates (Section 8.4), then either sub-group specific recommendations may be made, or a recommendation for all patients. If recommendation is made for all patients, then the subgroup specific estimates need to be properly averaged over in the economic model (94).

APPENDICES

APPENDIX A

Inclusion of non-standard trial designs

Results from non-standard trial designs can be included in meta-analyses, unless there are special reasons for dis-allowing this – so long as the reported effect estimates and their standard errors have been correctly derived or can be inferred from what is reported. Aside from the potential biases that may arise in these trials due to their design, guideline developers should be attentive to the statistical analysis employed. If the statistical analysis within the trial does not account for its design, then the data should be adjusted. Where there is any ambiguity or hesitation, a statistician should be consulted to evaluate the statistical methods employed in such trials.

Cluster randomised trials

Cluster trials are common when the setting of the treatment involves natural groups (e.g., hospitals, families, schools). The groups or clusters are randomised, and observations from individuals within clusters are not independent (11). Data from a cluster randomised controlled trial (cRCT) must be analysed using a method that accounts for the clusters, such as: multilevel, hierarchical, or mixed effects models, variance components analysis, or generalized estimating equations (GEE) (125). Results will typically be reported as an effect size with its corresponding standard error, which can be extracted and input into the MA.

If it is clear that clustering has *not* been accounted for, then the extracted (naïve) standard error,

SE, must be adjusted using the design effect, DE, which is calculated based on the average cluster size, M, and the intraclass correlation coefficient, ICC, in the trial (126)

$$DE = 1 + (M - 1)ICC$$
$$SE^{Adj} = SE\sqrt{DE}$$

If an ICC is not reported in a trial, it may be imputed by taking the average of the ICCs reported by other cRCTS in the meta-analysis or by cRCTs in the same area of medicine. A sensitivity analysis is recommended to assess the robustness of the results to any imputations.

Cross-over trials

In a typical cross-over trial, a participant receives two treatments and the order in which they receive the treatments is randomised (e.g., first receive treatment A and then cross-over to treatment B, or vice versa) (125). In a cross-over trial, the effect estimate and its standard error should result from an analysis which accounts for the correlation of the repeated measurements made on the same individual (127). While the Mantel-Haenszel OR for paired outcomes and conditional logistic regression are appropriate methods to analyse binary data in a cross-over trial, the magnitude of the resulting OR depends on the within-patient correlation, which may vary across trials (127-129). The Becker-Balagtas method for cross-over binary data does not depend on the within-patient correlation, and thus odds ratios calculated using , this approach should be included in the meta-analysis (130).

For continuous outcomes, the effect estimate is the mean of the within-patient differences (128). The corresponding standard error based on the results of a paired t-test or analysis of variance where the participant is included among the factors may be derived using the appropriate formula(s) found in the Appendix of GMD2 (127). Alternatively, if a cross-over trial reports continuous arm-

based data by treatment group, the effect estimate may be computed using standard methods, and its corresponding standard error, SE, may be computed as

$$SE = \sqrt{S_1^2 + S_2^2 - 2rS_1S_2}$$

where S_1 and S_2 are the standard deviations in treatment arms 1 and 2, and r is the within-patient correlation (128). The key issue here is to ensure that the correlation of within-patient measurements across periods is accounted for when computing the standard error. If this is not reported in a trial, it may be imputed by taking the average of correlations reported by the trials in the meta-analysis. Correlations may be derived from trials reporting the SDs in each treatment arm, S_1 and S_2, and the SD of the within-patient differences, $S_{Diff_{12}}$ (128):

$$r = \frac{S_1^2 + S_2^2 - S_{Diff_{12}}^2}{2S_1S_2}$$

A sensitivity analysis is recommended to assess the robustness of the results to any imputations.

Note that change-from-baseline is not a recommended outcome measure in cross-over designs, as it is unlikely that the carryover effect of a treatment in a particular period has the same influence on the baseline and post-treatment measurements in a subsequent period (131). However, it is acceptable to adjust for the baseline scores in the first period only using ANCOVA.

Split-body trials

Split-body (e.g., split-face, split-mouth) trials randomise treatments to different sections of the body (e.g., left or right side of face, quadrants in mouth) (125). Similar to cross-over trials, correlation arises in such within-person trials since individuals will receive more than one treatment, whether it be concurrently or sequentially (132). Appropriate methods for analysis of split-body trials are similar to those listed for cross-over trials, so long as the correlation between sites is accounted for.

Factorial trials

Factorial trials are designed to compare multiple treatments that are assumed to have independent effects (125). When considering a factorial trial for inclusion in a meta-analysis, clinical judgement is required to determine if the assumption of no interaction is reasonable. That is, the effect of a treatment is independent of the "levels" (e.g., presence or absence; dose) of the other treatment(s). In the simplest design, there are two active treatments of interest (e.g., A and B) and a patient will be randomly allocated to one of four combinations: A and B, A only, B only, or neither A nor B. This is referred to a 2 x 2 factorial design.

If the effects of the treatments are indeed independent of each other, their effects are essentially estimated by merging the outcome data across the levels of the other treatment (133, 134). The outcome data may be extracted and inputted into the analyses as if they were derived from independent trials, since independent effects implies zero correlation. Data can be analysed within the trial using a generalised linear model (e.g., linear regression in the case of continuous outcomes, logistic regression in the case of binary outcomes), where each treatment is included as a covariate so that the effect of each treatment is adjusted for the effects of other treatments, along with other covariates such as baseline measurements (134). Again, these analyses rely on the assumption that there is no interaction between the effects of the treatments. If there is evidence of an interaction, then a subset of the outcome data (e.g., A vs. neither A nor B) could be extracted, provided this level of detail is reported in the trial.

Combination of N-of-1 studies

An N-of-1 study typically involves a single patient taking two interventions in a random order across multiple periods (135).

Single N-of-1 studies reporting a mean effect and its standard error can be included in a meta-analysis. It is also possible to include results from meta-analytic studies which have brought together several N-of-1 studies on patients in the target population, and have reported an overall mean and its standard error. Methods for combining N-of-1 studies go beyond the scope of this document, but might include: linear mixed models, which make use of the individual patient data (IPD) in each trial, or Bayesian hierarchical models, which use the same hierarchical structure as linear mixed models (135).

APPENDIX B

Estimating a baseline effect in a meta-analysis

Fixed effect model

To illustrate how to estimate a baseline effect using available software, consider the following event data from two trials comparing the effectiveness of treatments for acute coronary syndrome (Table B.1). The outcome is all-cause mortality after 30-days and we are interested in estimating the baseline effect of clopidogrel. Since the relative effects for this outcome were pooled as log-odds ratios, we require an estimate of the baseline effect in the form of a log-odds. The log-odds and corresponding variance for each study were calculated using the log-odds worksheet in GMD3 Data Conversion Workbook (Table B.1).

Table B.1: 30 day all-cause mortality among patients receiving clopidogrel

Study	Number of Deaths	Total Randomised	log-odds	Var(log-odds)
1	212	9186	-3.7455	0.0048
2	45	1765	-3.6434	0.0228

Similar estimates were obtained across three different software (Table B.2). Assuming the normality of the baseline effect (in terms of log odds), values may be simulated from a Normal distribution, where the mean is the pooled log-odds, and the variance is the standard error squared. Note that to perform the correct analysis in RevMan, the data must be inputted in terms of log-odds, rather than numerators and denominators, and pooled using the Generic Inverse Variance approach. In addition, RevMan does not directly provide an estimate of the standard error, however this may be derived from the confidence interval.

Table B.2: Pooled baseline effect estimates using a FE model based on data in Table B1

Software	Accepted data entry formats	mean (pooled) log-odds	standard error	95% confidence or credible interval
metafor package in R	1) numerator & denominator 2) log-odds & SE	-3.73	0.063	(-3.85, -3.60)
WinBUGS	1) numerator & denominator 2) log-odds & SE	-3.73	0.063	(-3.85, -3.61)
RevMan	1) log-odds & SE	-3.73	N/A	(-3.85, -3.60)

N/A = not available

Random effects model

When there is some heterogeneity in the study-specific baseline effects, but the inclusion of multiple studies is justifiable, then a random effects model may be fitted. We illustrate this using the following data from nine trials comparing the effectiveness treatments for depression (Table B.3). The outcome is remission and we are interested in estimating the baseline effect of placebo. The observed and shrunken estimates are provided in Table B.3. The observed log-odds and corresponding variance for each study were calculated using the log-odds worksheet in *GMD3 Data Conversion Workbook*, while the shrunken estimates were produced by a baseline model in WinBUGS.

Table B.3: Remission status among patients receiving placebo

Study	Number of Remitters	Total Randomised	Observed Estimates		Shrunken Estimates	
			log-odds	Var(log-odds)	log-odds	Var(log-odds)
1	2	37	-2.8622	0.5286	-2.225	0.3299
2	9	21	-0.2877	0.1944	-0.4406	0.1671
3	38	97	-0.4400	0.0433	-0.4753	0.0422
4	43	270	-1.6637	0.0277	-1.637	0.0268
5	10	31	-0.7419	0.1476	-0.7884	0.1276
6	10	30	-0.6931	0.1500	-0.7461	0.1295
7	8	12	0.6931	0.3750	0.167	0.3137
8	13	42	-0.8023	0.1114	-0.83	0.0984
9	10	34	-0.8755	0.1417	-0.896	0.1219

The following estimates were obtained using various software options (Table B.4). The pooled log-odds estimates are similar across studies; however, there is more uncertainty in the estimate produced by a model in WinBUGS, compared to the estimates produced by the other software. This may be explained by 1) the different estimators of the between-study SD employed across the software and 2) the model in WinBUGS also accounts for the uncertainty in the between-study SD, whereas the models run in the other software do not. To simulate values of the baseline effect from a Normal distribution, the pooled log-odds may be inputted as the mean and variance. In terms of the variance, the variance of the predicted effect (i.e., standard error squared) should be used. This is not produced by the *metafor* package or RevMan. However, the standard error may be derived from the outputted 95% predictive interval, or alternatively can be estimated as $\sqrt{V(\theta)+\tau^2}$, where $V(\theta)$ is the variance of the pooled log-odds and τ is the between-study SD (59).

Table B.4: Pooled baseline effect estimates using a RE model based on data in Table B.3

Software	Accepted data entry formats	mean (pooled) log-odds	standard error	95% confidence or credible interval	standard error of predicted effect	95% predictive interval	Between-study SD (95% confidence interval)
metafor package in R (PM estimator)	1) numerator & denominator 2) log-odds & SE	-0.83	0.25	(-1.39, -0.27)	N/A	(-2.42, 0.76)	0.76 (0.35, 1.76)
WinBUGS*	1) numerator & denominator 2) log-odds & SE	-0.88	0.36	(-1.60, -0.14)	1.07	(-3.04, 1.28)	0.85 (0.41, 1.89)
RevMan	1) log-odds & SE	-0.83	N/A	(-1.33, -0.33)	N/A	N/A	0.66** (N/A)

* Prior on between-study standard deviation: Uniform(0, 5)

** Obtained by taking the square root of the between-study variance

N/A = not available

APPENDIX C

Obtaining shrunken estimates from a study in a random effects meta-regression model

In a meta-regression model with a continuous covariate, the covariate should be centred to improve convergence. This is done in the BCG vaccine example in Technical Support Document (TSD) 3, where the efficacy of a BCG vaccine for preventing tuberculosis is evaluated (5). Program 4(a) in TSD3 provides the WinBUGS code for a random effects meta-regression model, where the covariate of interest, the absolute latitude in degrees, is centred. In this case, the study-specific effect estimates are shrunken towards the overall mean effect at the average of the absolute study latitudes, 33.46°.

To inform an economic model (Section 13.1), we may wish to use a study-specific effect shrunken towards an overall mean at a specific absolute latitude, e.g., 50°. To obtain this, the following code would have to be added to the code provided in TSD3:

```
# Study-specific effects shrunken toward overall mean at |latitude| = 50 degrees
for(i in 1:ns) {
for(k in 2:na[i]){
delta.adj[i,k] <- delta[i,k] + (beta[t[i,k]]-beta[t[i,1]]) * (50-mx)
}
}
```

The observed and shrunken estimates are provided in Table C.1. The observed log-odds and corresponding variance for each study were calculated using the log-odds worksheet in *GMD3 Data Conversion Workbook,* while the shrunken estimates were produced in WinBUGS. The covariate coefficient estimated is -0.03 (95% CrI: -0.05, -0.01), suggesting that the efficacy of the vaccine increases with absolute latitude (log odds ratios (LOR) < 0 favours vaccine). In Study 1, which has an absolute latitude of 44°, the observed LOR = -0.9387. The shrunken estimate at a smaller latitude, 33.46°, increases, while the shrunken estimate at a larger latitude, 50°, decreases. This makes sense given the values of the latitude corresponding to each estimate.

Table C.1: Remission status among patients receiving placebo

Study	Absolute latitude (degrees)	Observed Estimates		Shrunken Estimates (latitude = 33.46°)		Shrunken Estimates (latitude = 50°)	
		LOR	Var(LOR)	LOR	Var(LOR)	LOR	Var(LOR)
1	44	-0.9387	0.5976	-0.7348	0.0812	-1.2600	0.0942
2	55	-1.6662	0.4562	-0.8320	0.0733	-1.3570	0.0705
3	42	-1.3863	0.6583	-0.8298	0.0871	-1.3550	0.0993
4	52	-1.4564	0.1425	-0.8346	0.0318	-1.3600	0.0166
5	13	-0.2191	0.2279	-0.8081	0.0455	-1.3340	0.0936
6	44	-0.9581	0.0995	-0.6532	0.0140	-1.1790	0.0127
7	19	-1.6338	0.4765	-1.0990	0.1449	-1.6250	0.1696
8	13	0.0120	0.0633	-0.6493	0.0284	-1.1750	0.0868
9	27	-0.4717	0.2387	-0.7125	0.0327	-1.2380	0.0613
10	42	-1.4012	0.2746	-0.9348	0.0504	-1.4600	0.0548
11	18	-0.3408	0.1119	-0.8057	0.0227	-1.3310	0.0682
12	33	0.4466	0.7309	-0.5751	0.1127	-1.1010	0.1378
13	33	-0.0173	0.2676	-0.4228	0.0678	-0.9483	0.0924

REFERENCES

1.National Institute for Health and Care Excellence. Guide to the methods of technology appraisal. London; 2013.

2.National Institute for Health and Care Excellence. Developing NICE guidelines: the manual. National Institute for Health and Care Excellence; 2014.

3.Dias S, Welton NJ, Sutton AJ, Ades AE. NICE DSU Technical Support Document 1: Introduction to evidence synthesis for decision making. 2011.

4.Dias S, Welton NJ, Sutton AJ, Ades AE. NICE DSU Technical Support Document 2: A generalised linear modelling framework for pair-wise and network meta-analysis of randomised controlled trials. 2011.

5.Dias S, Sutton AJ, Welton NJ, Ades AE. NICE DSU Technical Support Document 3: Heterogeneity: subgroups, meta-regression, bias and bias-adjustment. 2011.

6.Dias S, Welton NJ, Sutton AJ, Caldwell DM, Lu G, Ades AE. NICE DSU Technical Support Document 4: Inconsistency in networks of evidence based on randomised controlled trials. 2011.

7.Dias S, Welton NJ, Sutton AJ, Ades AE. NICE DSU Technical Support Document 5: Evidence synthesis in the baseline natural history model. 2011.

8.Dias S, Sutton AJ, Welton NJ, Ades AE. NICE DSU Technical Support Document 6: Embedding evidence synthesis in probabilistic cost-effectiveness analysis: Software choices. 2011.

9.Ades AE, Caldwell DM, Reken S, Welton NJ, Sutton AJ, Dias S. NICE DSU Technical Support Document 7: Evidence synthesis of treatment efficacy in decision making: a reviewer's checklist. 2012.

10.Borenstein M, Hedges LV, Higgins JPT, Rothstein HR. Introduction to meta-analysis. Chichester: Wiley; 2009.

11.Higgins JPT, Thomas J, Chandler J, Cumpston M, Li T, Page MJ, et al. Cochrane handbook for systematic reviews of interventions. 2nd ed. Chichester: John Wiley & Sons; 2019.

12.Sutton AJ, Abrams KR, Jones DR, Sheldon TA, Song F. Methods for meta-analysis in medical research. London: Wiley; 2000.

13.Dias S, Welton NJ, Sutton AJ, Ades AE. Evidence Synthesis for Decision Making 1: Introduction. Medical Decision Making. 2013;33:597-606.

14.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. Medical Decision Making. 2013;33:607-17.

15.Dias S, Sutton AJ, Welton NJ, Ades AE. Evidence Synthesis for Decision Making 3: Heterogeneity - subgroups, meta-regression, bias and bias-adjustment. Medical Decision Making. 2013;33:618-40.

16.Dias S, Welton NJ, Sutton AJ, Caldwell DM, Lu G, Ades AE. Evidence Synthesis for Decision Making 4: Inconsistency in networks of evidence based on randomized controlled trials. Medical Decision Making. 2013;33:641-56.

17.Dias S, Welton NJ, Sutton AJ, Ades AE. Evidence Synthesis for Decision Making 5: The baseline natural history model. Medical Decision Making. 2013;33:657-70.

18.Dias S, Sutton AJ, Welton NJ, Ades AE. Evidence Synthesis for Decision Making 6: Embedding evidence synthesis in probabilistic cost-effectiveness analysis. Medical Decision Making. 2013;33:671-8.

19.Ades AE, Caldwell DM, Reken S, Welton NJ, Sutton AJ, Dias S. Evidence Synthesis for Decision Making 7: A reviewer's checklist. Medical Decision Making. 2013;33:679-91.

20.Dias S, Ades AE, Welton NJ, Jansen JP, Sutton AJ. Network meta-analysis for decision making: Wiley; 2018. 488 p.

21.Borenstein M. Effect sizes for continuous data. In: Cooper H, Hedges LV, Valentine JC, editors. The handbook of research synthesis and meta-analysis. 2nd ed. New York, NY, US: Russell Sage Foundation; 2009. p. 221-35.

22.Lu G, Ades AE. Modelling between-trial variance structure in mixed treatment comparisons. Biostatistics. 2009;10:792-805.

23.Gotzsche PC. Why we need a broad perspective on meta-analysis. BMJ. 2000;321:585-6.

24.Caldwell DM, Ades AE, Higgins JPT. Simultaneous comparison of multiple treatments: combining direct and indirect evidence. BMJ. 2005;331:897-900.

25.Dias S, Welton NJ, Marinho VCC, Salanti G, Higgins JPT, Ades AE. Estimation and adjustment of bias in randomised evidence by using mixed treatment comparison meta-analysis. Journal of the Royal Statistical Society (A). 2010;173(3):613-29.

26.Mayo-Wilson E, Dias S, Mavranezouli I, Kew K, Clark DM, Ades AE, et al. Psychological and pharmacological interventions for social anxiety disorder in adults: a systematic review and network meta-analysis. Lancet Psychiatry. 2014;1:368-76.

27.National Collaborating Centre for Mental Health. Social anxiety disorder: Recognition, assessment and treatment. National Clinical Guideline Number 159. 2013.

28.National Institute for Health and Care Excellence. Post-traumatic stress disorder. National Institute for Health and Care Excellence; 2018.

29.National Institute for Health and Care Excellence. Crohn's disease: management. National Institute for Health and Care Excellence; 2019.

30.National Institute for Health and Care Excellence. Bipolar disorder: assessment and management. National Institute for Health and Care Excellence; 2014.

31.National Institute for Health and Care Excellence. Chronic obstructive pulmonary disease in over 16s: diagnosis and management. National Institute for Health and Care Excellence; 2018.

32.Oba Y, Keeney E, Ghatehorde N, Dias S. Dual combination therapy versus long-acting bronchodilators alone for chronic obstructive pulmonary disease (COPD): A systematic review and network meta-analysis. Cochrane Database of Systematic Reviews. 2018(12).

33.Dominici F, Parmigiani G, Wolpert RL, Hasselblad V. Meta-analysis of migraine headache treatments: combining information from heterogenous designs. Journal of the American Statistical Association. 1999;94:16-28.

34.Soares MO, Dumville J, Ades AE, Welton NJ. Treatment comparisons for decision making: facing the problems of sparse and few data. Journal of the Royal Statistical Society (A). 2014;177:259-79

35.Owen RK, Tincello DG, Keith RA. Network meta-analysis: Development of a three-level hierarchical modeling approach incorporating dose-related constraints. Value in Health. 2015;18(1):116-26.

36.National Collaborating Centre for Women's and Children's Health. Preterm labour and birth [NG25]. London; 2015.

37.Mawdsley D, Bennetts M, Dias S, Boucher M, Welton NJ. Model-based network meta-analysis: A framework for evidence synthesis of clinical trial data. CPT Pharmacometrics & Systems Pharmacology. 2016;5(8):393-401.

38.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.

39.Caldwell DM, Welton NJ. Approaches for synthesising complex mental health interventions in meta-analysis. Evidence Based Mental Health. 2016;19(1):16.

40.Mills E, Druyts E, Ghement I, Puhan MA. Pharmacotherapies for chronic obstructive pulmonary disease: a multiple treatment comparison meta-analysis. Clinical Epidemiology. 2011;3:107-29.

41.Gleser LJ, Olkin I. Stochastically dependent effect sizes. In: Cooper H, Hedges LV, Valentine JC, editors. The handbook of research synthesis and meta-analysis. 2nd ed: Russell Sage Foundation; 2009. p. 357-76.

42.Jackson D, Riley R, White IR. Multivariate meta-analysis: potential and promise. Statistics In Medicine. 2011;30:2481-598.

43.Price MJ, Blake HA, Kenyon S, White IR, Jackson D, Kirkham JJ, et al. Empirical comparison of univariate and multivariate meta-analyses in Cochrane Pregnancy and Childbirth reviews with multiple binary outcomes. Research Synthesis Methods. 2019;10(3):440-51.

44.Lu G, Ades AE, Sutton AJ, Cooper NJ, Briggs AH, Caldwell DM. Meta-analysis of mixed treatment comparisons at multiple follow-up times. Statistics In Medicine. 2007;26(20):3681-99.

45.Stettler C, Allemann S, Wandel S, Kastrati A, Morice MC, Schomig A, et al. Drug eluting and bare metal stents in people with and without diabetes: collaborative network meta-analysis. BMJ. 2008;337(7671):1331.

46.Welton NJ, Cooper NJ, Ades AE, Lu G, Sutton AJ. Mixed treatment comparison with multiple outcomes reported inconsistently across trials: evaluation of antivirals for treatment of influenza A and B. Statistics In Medicine. 2008;27:5620-39.

47.Burch J, Paulden M, Conti S, Stock C, Corbette M, Welton NJ, et al. Antiviral drugs for the treatment of influenza: a systematic review and economic evaluation. Health Technology Assesment. 2010;13(58):1-290.

48.Anwer S, Ades AE, Dias S. Joint synthesis of conditionally related multiple outcomes makes better use of data than separate meta-analyses. Research Synthesis Methods. 2020;11(4):496-506.

49.Welton NJ, Willis SR, Ades AE. Synthesis of survival and disease progression outcomes for health technology assessment of cancer therapies. Research Synthesis Methods. 2010;1:239-57.

50.National Institute for Health and Care Excellence. Lung cancer: diagnosis and management. National Institute for Health and Care Excellence; 2019.

51.Ades AE. A chain of evidence with mixed comparisons: models for multi-parameter evidence synthesis and consistency of evidence. Statistics in Medicine. 2003;22:2995-3016.

52.Lu G, Kounali D, Ades AE. Simultaneous multi-outcome synthesis and mapping of treatment effects to a common scale. Value in Health. 2014;17:280-7.

53.Ades AE, Lu G, Dias S, Mayo-Wilson E, Kounali D. Simultaneous synthesis of treatment effects and mapping to a common scale: an alternative to standardisation. Research Synthesis Methods. 2015;6:96-107.

54.Kounali DZ, Button KS, Lewis G, Ades AE. The relative responsiveness of test instruments can be estimated using a meta-analytic approach: an illustration with treatments for depression. Journal of Clinical Epidemiology. 2016;77:68-77.

55.Price MJ, Welton NJ, Ades AE. Synthesis of Markov data from RCTs: presentation of a structured approach to modelling treatments of asthma. Society for Medical Decicion Making 10th Bienniel European Meeting; Birmingham, England 2006.

56.Ades AE, Lu G, Higgins JPT. The interpretation of random effects meta-analysis in decision models. Medical Decision Making. 2005;25(6):646-54.

57.Higgins JPT, Thompson SG. Quantifying heterogeneity in a meta-analysis. Statistics in Medicine. 2002;21:1539-58.

58.Higgins JPT, Thompson SG, Deeks JJ, Altman DG. Measuring inconsistency in meta-analyses. BMJ. 2003;327:557-60.

59.Borenstein M, Higgins JPT, Hedges LV, Rothstein HR. Basics of meta-analysis: I2 is not an absolute measure of heterogeneity. Research Synthesis Methods. 2017;8(1):5-18.

60.Higgins JPT. Commentary: Heterogeneity in meta-analysis should be expected and appropriately quantified. Internation Journal of Epidemiology. 2008;37:1158-60.

61.Veroniki AA, Jackson D, Viechtbauer W, Bender R, Bowden J, Knapp G, et al. Methods to estimate the between-study variance and its uncertainty in meta-analysis. Research Synthesis Methods. 2016;7(1):55-79.

62.Higgins JPT, Thompson SG, Spiegelhalter DJ. A re-evaluation of random-effects meta-analysis. Journal of the Royal Statistical Society (A). 2009;172:137-59.

63.Spiegelhalter D, Best N, Carlin B, Linde AVD. Bayesian measures of model complexity and fit. Journal of the Royal Statistical Society (B). 2002;64(4):583-639.

64.Jackson D, Law M, Stijnen T, Viechtbauer W, White IR. A comparison of seven random-effects models for meta-analyses that estimate the summary odds ratio. Statistics in Medicine. 2018;37(7):1059-85.

65.Senn S, Gavini DM, Scheen A. Issues in performing a network meta-analysis. Statistical Methods in Medical Research. 2013;22:169-89.

66.Turner RM, Jackson D, Wei Y, Thompson SG, Higgins JPT. Predictive distributions for between-study heterogeneity and simple methods for their application in Bayesian meta-analysis. Statistics in Medicine. 2015;34(6):984-98.

67.Rhodes KM, Turner RM, Higgins JPT. Predictive distributions were developed for the extent of heterogeneity in meta-analyses of continuous outcome data. Journal of Clinical Epidemiology. 2015;68:52-60.

68.National Institute for Health and Care Excellence. Specialist neonatal respiratory care for babies born preterm. National Institute for Health and Care Excellence; 2019.

69.Turner RM, Spiegelhalter DJ, Smith GCS, Thompson SG. Bias modelling in evidence synthesis. Journal of the Royal Statistical Society (A). 2009;172:21-47.

70.Sterne JAC, Savović J, Page MJ, Elbers RG, Blencowe NS, Boutron I, et al. RoB 2: a revised tool for assessing risk of bias in randomised trials. BMJ. 2019;366:l4898.

71.Briggs A, Sculpher M, Claxton K. Decision modelling for health economic evaluation. Oxford: Oxford University Press; 2006 2006.

72.Longworth L, Rowen D. NICE DSU Technical Support Document 10: The use of mapping methods to estimate health state utility values. 2011.

73.Kaltenthaler E, Tappenden P, Paisley S, Squires H. NICE DSU Technical Support Document 13: Identifying and reviewing evidence to inform the conceptualisation and population of cost-effectiveness models. 2011.

74.Latimer N. NICE DSU Technical Support Document 14: Survival analysis for economic evaluations alongside clinical trials - extrapolation with patient-level data. 2013.

75.Latimer NR, Abrams KR. NICE DSU Technical Support Document 16: Adjusting survival time estimates in the presence of treatment switching. 2014.

76.Faria R, Alava MH, Manca A, Wailoo AJ. NICE DSU Technical Support Document 17: The use of observational data to inform estimates of treatment effectiveness in technology appraisal: Methods for comparative individual patient data. 2015.

77.Woods B, Sideris E, Palmer S, Latimer N, Soares M. NICE DSU Technical Support Document 19: Partitioned survival analysis for decision modelling in health care: A critical review. 2017.

78.National Institute for Health and Care Excellence. Attention deficit hyperactivity disorder: diagnosis and management. National Institute for Health and Care Excellence; 2018.

79.National Institute for Health and Care Excellence. Acute coronary syndromes. National Institute for Health and Care Excellence; 2020.

80.National Institute for Health and Care Excellence. Surgical site infections: prevention and treatment. National Institute for Health and Care Excellence; 2019.

81.Lambert PC, Sutton AJ, Abrams KR, Jones DR. A comparison of summary patient-level covariates in meta-regression with individual patient data meta-analysis. Journal of Clinical Epidemiology. 2002;55:86-94.

82.Berlin JA, Begg CB, Louis TA. An assessment of publication bias using a sample of published clinical trials. Journal of the American Statistical Association. 1989;84:381-92.

83.Fisher DJ, Copas AJ, Tierney JF, Parmar MKB. A critical review of methods for the assessment of patient-level interactions in individual participant data meta-analysis of randomized trials, and guidance for practitioners. Journal of Clinical Epidemiology. 2011;64(9):949-67.

84.Sutton AJ, Kendrick D, Coupland CAC. Meta-analysis of individual- and aggregate-level data. Statistics in Medicine. 2008;27:651-69.

85.Donegan S, Williamson P, D'Alessandro U, Garner P, Tudor Smith C. Combining individual patient data and aggregate data in mixed treatment comparison meta-analysis: Individual patient data may be beneficial if only for a subset of trials. Statistics in Medicine. 2013;32:914-30.

86.Jansen JP. Network meta-analysis of individual and aggregate level data. Research Synthesis Methods. 2012;3(2):177-90.

87.Saramago P, Sutton AJ, Cooper NJ, Manca A. Mixed treatment comparisons using aggregate and individual participant level data. Statistics in Medicine. 2012;31:3516-36.

88.Thom HH, Capkun G, Cerulli A, Nixon RM, Howard LS. Network meta-analysis combining individual patient and aggregate data from a mixture of study designs with an application to pulmonary arterial hypertension. BMC Medical Research Methodology. 2015;15:34.

89.Rothman KJ, Greenland S, Lash TL. Modern epidemiology. 3rd ed. Philadelphia: Lippincott, Williams & Wilkins; 2012.

90.Jansen JP, Cope S. Meta-regression models to address heterogeneity and inconsistency in network meta-analysis of survival outcomes. BMC Medical Research Methodology. 2012;12:152.

91.Phillippo DM, Dias S, Ades AE, Belger M, Brnabic A, Schacht A, et al. Multilevel network meta-regression for population-adjusted treatment comparisons. Journal of the Royal Statistical Society: Series A. 2020;183(3):1189-210.

92.Phillippo DM, Ades AE, Dias S, Palmer S, Abrams KR, Welton NJ. NICE DSU Technical Support Document 18: Methods for population-adjusted indirect comparisons in submissions to NICE. 2016.

93.National Collaborating Centre for Mental Health. The treatment and management of depression in adults. National Clinical Practice Guideline CG90. London: National Instiutute for Health and Care Excellence; 2010.

94.Welton NJ, Soares MO, Palmer S, Ades AE, Harrison D, Shankar-Hari M, et al. Accounting for heterogeneity in relative treatment effects for use in cost-effectiveness models and value-of-information analyses. Medical Decision Making. 2015;35:608-21.

95.National Clinical Guideline Centre. Lipid modification: Cardiovascular risk assessment and the modification of blood lipids for the primary and secondary prevention of cardiovascular disease. London; 2014.

96.Moreno SG, Sutton AJ, Ades AE, Stanley TD, Abrams KR, Peters JL, et al. Assessment of regression-based methods to adjust for publication bias through a comprehensive simulation study. BMC Medical Research Methodology. 2009;9(2).

97.Moreno SG, Sutton AJ, Turner EH, Abrams KR, Cooper NJ, Palmer TM, et al. Novel methods to deal with publication biases: Secondary analysis of antidepressant trials in the FDA trial registry database and related journal publications. BMJ. 2009;339:b2981.

98.Schulz KF, Chalmers I, Hayes RJ, Altman DG. Empirical evidence of bias: Dimensions of methodological quality associated with estimates of treatment effects in controlled trials. JAMA. 1995;273(5):408-12.

99.Wood L, Egger M, Gluud LL, Schulz K, Juni P, Altman D, et al. Empirical evidence of bias in treatment effect estimates in controlled trials with different interventions and outcomes: Meta-epidemiological study. BMJ. 2008;336:601-5.

100.Savovic J, Jones H, Altman D, Harris R, Juni P, Pildal J, et al. Influence of study design characteristics on intervention effect estimates from randomised controlled trials: Combined analysis of meta-epidemiological studies. Health Technology Assessment. 2012;16(35).

101.Savovic J, Jones HE, Altman DG, Harris RJ, Juni P, Pildal J, et al. Influence of Reported Study Design Characteristics on Intervention Effect Estimates From Randomized, Controlled Trials. Annals of Internal Medicine. 2012;157:429-38.

102.Welton NJ, Ades AE, Carlin JB, Altman DG, Sterne JAC. Models for potentially biased evidence in meta-analysis using empirically based priors. Journal of the Royal Statistical Society (A). 2009;172(1):119-36.

103.Naci H, Dias S, Ades AE. Industry sponsorship bias in research findings: A network meta-analytic exploration of LDL cholesterol reduction in the randomised trials of statins. BMJ. 2014;349:g5741.

104.Salanti G, Dias S, Welton NJ, Ades AE, Golfinopoulos V, Kyrgiou M, et al. Evaluating novel agent effects in multiple treatments meta-regression. Statistics in Medicine. 2010;29:2369-83.

105.Moreno SG, Sutton AJ, Ades AE, Cooper NJ, Abrams KR. Adjusting for publication biases across similar interventions performed well when compared with gold standard data. Journal of Clinical Epidemiology. 2011;64(11):1230-41.

106.National Institute for Health and Care Excellence. Eating disorders: recognition and treatment. National Institute for Health and Care Excellence; 2017.

107.Ades AE, Claxton K, Sculpher M. Evidence synthesis, parameter correlation and probabilistic sensitivity analysis. Health Economics. 2005;14:373-81.

108.Phillippo DM, Dias S, Ades AE, Didelez V, Welton NJ. Sensitivity of treatment recommendations to bias in network meta-analysis. Journal of the Royal Statistical Society (A). 2018;181(3):843-67.

109.Phillippo DM, Dias S, Welton NJ, Caldwell DM, Taske N, Ades AE. Threshold analysis as an alternative to GRADE for assessing confidence in guideline recommendations based on network meta-analyses. Annals of Internal Medicine. 2019;170(8):538-46.

110.Salanti G, Ades AE, Ioannidis JPA. Graphical methods and numerical summaries for presenting results from multiple-treatment meta-analysis: an overview and tutorial. Journal of Clinical Epidemiology. 2011;64:163-71.

111.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.

112.Lunn D, Jackson C, Best N, Thomas A, Spiegelhalter D. The BUGS book. Boca Raton, FL: CRC Press; 2013.

113.Viechtbauer W. Conducting meta-analyses in R with the metafor package. Journal of Statistical Software. 2010;36(3).

114.van Valkenhoef G, Kuiper J. gemtc: Network meta-analysis using Bayesian methods. R package version 0.8-2. 2016.

115.Béliveau A, Boyne DJ, Slater J, Brenner D, Arora P. BUGSnet: an R package to facilitate the conduct and reporting of Bayesian network meta-analyses. BMC Medical Research Methodology. 2019;19(1):196.

116.Rücker G, Krahn U, König J, Efthimiou O, Schwarzer G. netmeta: Network meta-analysis using frequentist methods. R package version 1.2-1. 2020.

117.Sturtz S, Ligges U, Gelman A. R2WinBUGS: A package for running WinBUGS from R. Journal of Statistical Software. 2005;12(3):1-16.

118.van Valkenhoef G, Dias S, Ades AE, Welton NJ. Automated generation of node-splitting models for assessment of inconsistency in network meta-analysis. Research Synthesis Methods. 2016;7:80-93.

119.Dias S, Welton NJ, Caldwell DM, Ades AE. Checking consistency in mixed treatment comparison meta-analysis. Statistics in Medicine. 2010;29:932-44.

120.The Nordic Cochrane Centre. Review Manager (RevMan) Version 5.3. Copenhagen: The Cochrane Collaboration; 2014.

121.Owen RK, Bradbury N, Xin Y, Cooper N, Sutton A. MetaInsight: An interactive web-based tool for analyzing, interrogating, and visualizing network meta-analyses using R-shiny and netmeta. Research Synthesis Methods. 2019;10(4):569-81.

122.Cooper NJ, Sutton AJ, Abrams KR, Turner D, Wailoo A. Comprehensive decision analytical modelling in economic evaluation: a Bayesian approach. Health Economics. 2003;13:203-26.

123.Jones HE, Ades AE, Sutton AJ, Welton NJ. Use of a random effects meta-analysis in the design and analysis of a new clinical trial. Statistics in Medicine. 2018;37(30):4665-79.

124.Welton NJ, White I, Lu G, Higgins JPT, Ades AE, Hilden J. Correction: Interpretation of random effects meta-analysis in decision models. Medical Decision Making. 2007;27:212-4.

125.Higgins JPT, Eldridge S, Li T (editors). Chapter 23: Including variants on randomized trials. In: Higgins JPT, Thomas J, Chandler J, Cumpston M, Li T, Page MJ, et al., editors. Cochrane handbook for systematic reviews of interventions. 2nd ed. Chichester: John Wiley & Sons; 2019. p. 569–94.

126.Donner A, Klar N. Issues in the meta-analysis of cluster randomized trials. Statistics in Medicine. 2002;21(19):2971-80.

127.Dwan K, Li T, Altman DG, Elbourne D. CONSORT 2010 statement: Extension to randomised crossover trials. BMJ. 2019;366:l4378.

128.Elbourne DR, Altman DG, Higgins JPT, Curtin F, Worthington HV, Vail A. Meta-analyses involving cross-over trials: Methodological issues. Internation Journal of Epidemiology. 2002;31:140-9.

129.Curtin F, Elbourne D, Altman DG. Meta-analysis combining parallel and cross-over clinical trials. II: Binary outcomes. Statistics in Medicine. 2002;21:2145-59.

130.Becker NG, Marschner IC. A method for estimating the age-specific relative risk of HIV infection from AIDS incidence data. Biometrika. 1993;80:165-78.

131.Fleiss JL. A critique of recent research on the two-treatment crossover design. Controlled clinical trials. 1989;10(3):237-43.

132.Moher D, Hopewell S, Schulz KF, Montori V, Gøtzsche PC, Devereaux PJ, et al. CONSORT 2010 explanation and elaboration: Updated guidelines for reporting parallel group randomised trials. Journal of Clinical Epidemiology. 2010;63(8):e1-e37.

133.McAlister FA, Straus SE, Sackett DL, Altman DG. Analysis and reporting of factorial trials: A systematic review. JAMA. 2003;289(19):2545-53.

134.Montgomery AA, Peters TJ, Little P. Design, analysis and presentation of factorial randomised controlled trials. BMC Medical Research Methodology. 2003;3(1):26.

135.Zucker DR, Ruthazer R, Schmid CH. Individual (N-of-1) trials can be combined to give population comparative treatment effect estimates: Methodologic considerations. Journal of Clinical Epidemiology. 2010;63(12):1312-23.