HTA CG | MEMBER STATE COORDINATION GROUP ON HEALTH TECHNOLOGY ASSESSMENT

Methodological Guideline for Quantitative Evidence Synthesis: Direct and Indirect Comparisons

Adopted on 8 March 2024 by the HTA CG pursuant to Article 3(7), point (d), of

Regulation (EU) 2021/2282 on Health Technology Assessment

The document is not a European Commission document and it cannot be regarded as reflecting the official position of the European Commission. Any views expressed in this document are not legally binding and only the Court of Justice of the European Union can give binding interpretations of Union law.

Contents

Table of figures

List of acronyms

CG	Coordination Group
GRADE	Grading of recommendations assessment, development and evaluation
HTA	Health technology assessment
HTAR	Health technology assessment regulation (Regulation (EU) 2021/2282)
HTD	Health technology developer
IPD	Individual patient-level data, also known as individual patient data or individual participant data
JCA	Joint clinical assessment
KH	Knapp-Hartung
MAIC	Matching-adjusted indirect comparison
ML-NMR	Multilevel network meta-regression
NMA	Network meta-analysis
PICO	Population, intervention, comparator, outcome
PRISMA	Preferred reporting items for systematic reviews and meta-analyses
RCT	Randomised controlled trial
RoB	Risk of bias
STC	Simulated treatment comparison

Summary

To assess the relative efficacy or effectiveness of a new intervention compared to one or more existing interventions (the comparators; e.g., the current standard treatment) in the presence of multiple sources of evidence, appropriate methods for evidence synthesis should be used. Randomised controlled trials (RCTs), provided they are well designed and have low risk of bias, are the gold standard for informing estimates of treatment effectiveness and should be used for evidence synthesis when possible. Thus, we assume that the evidence synthesis considered is based on adequate RCT data unless otherwise stated. The objective of this document is to describe the most commonly used methods for direct and indirect treatment comparisons, including their underlying assumptions, strengths and weaknesses. The document is not a methodological textbook and does not give a detailed description of the statistical techniques described. The guideline is aimed at assessors in the context of the EU regulation for the conduct of joint clinical assessment of health technologies (HTAR), although the relevance for other stakeholders is recognised. All methods for evidence synthesis (direct as well as indirect comparisons) are based on the fundamental assumption of exchangeability, which may be investigated by assessing similarity, homogeneity and, for indirect comparisons, consistency, of the trial data included. If this fundamental assumption is violated, the results of the corresponding evidence synthesis are unlikely to provide a meaningful estimate of treatment effectiveness. If the between-trial heterogeneity is considered to be too large to justify an overall evidence synthesis, but the heterogeneity can be explained in terms of study and patient characteristics, appropriate evidence syntheses should be performed using the corresponding groups of trials or subgroups of patients. This results in different effect estimates for the different subgroups. Meta-regression can also be a useful tool to explore heterogeneity further and to identify factors contributing to it. However, while these methods are likely to reduce heterogeneity, it is unlikely that they will eliminate it completely. Therefore, a new assessment of heterogeneity is required after the application of subgroup analysis or meta-regression.

General options for evidence synthesis involve the use of a fixed-effect or a random-effects model and application of frequentist or Bayesian methods for the effect estimation. In most practical situations, a random-effects model is the appropriate choice, although in some situations a fixed-effect model can be justified. Both frequentist and Bayesian approaches may be used. Bayesian approaches are especially useful in situations with sparse data. However, a clear justification of the prior distributions applied is required. The use of individual patient data (IPD) should be preferred as it offers the potential for exploring additional and potentially more appropriate statistical analyses compared to aggregated data.

Useful frequentist methods for direct comparisons via a fixed-effect model (fixed-effect pairwise meta-analysis) include the inverse variance method for continuous data and the Mantel-Haenszel method for binary data. The recommended frequentist approach for random-effects meta-analyses is the Knapp-Hartung (KH) method in cases involving at least five studies. In situations with fewer than five studies, and where the assumption of random effects is appropriate, alternative methods for evidence synthesis are frequently required, such as Bayesian approaches, a qualitative summary of the study results or the beta-binomial model.

In situations where no direct RCT evidence on a comparison of interest is available or where multiple treatments must be compared simultaneously, indirect comparisons may be used. In general, direct comparisons are preferable to comparisons based on indirect evidence alone because the latter are associated with greater uncertainties. If indirect comparisons are required, in general only anchored indirect comparisons respecting randomisation are appropriate, which requires that the evidence network is connected. Useful approaches for anchored indirect comparisons include the Bucher method and the frequentist and Bayesian approaches for network meta-analysis. In cases where anchored indirect comparisons are not feasible, alternative methods are available, but it is unlikely that these adequately eliminate bias due to lack of randomisation.

If the similarity assumption is not met, methods for population-adjusted indirect comparisons might be considered, provided that the network is connected and IPD are available for some of the trials included. These methods require that all effect modifiers relevant for adjustment are measured. However, this is often unverifiable and unattainable. Therefore, it is imperative that the robustness of the results derived from population-adjusted indirect comparisons are thoroughly investigated to ascertain whether the population-based adjustment leads to an estimate of the treatment effect with higher certainty of results. The model and covariate selection strategies for adjustment must be prespecified and based on transparent criteria. Owing to the greater uncertainties associated with population-adjusted methods, a sufficiently large treatment effect estimate is required, which can be formally evaluated via testing of shifted hypotheses. This means that a conclusion can be drawn regarding an effect only if the confidence interval lies completely above or below a certain pre-specified threshold shifted away from the zero effect.

In any situation with non-randomised data, such as observational evidence and single-arm trials, or in the case of disconnected networks, complete access to the IPD is required in order to apply methods that can adequately adjust for confounding. Again, these methods require that a set of all covariates (confounders, prognostic variables, or effect modifiers) relevant for adjustment are measured. However, this is often unverifiable and unattainable. Therefore, it is imperative that the model and covariate selection strategies for adjustment are prespecified and based on transparent criteria. Use of propensity scores is a method frequently applied. The assumption of conditional exchangeability needs to be met and can be assessed by investigating positivity, overlap, and balance. If conditional exchangeability es is not met, an adequate analysis is not possible and the results from the corresponding analysis are unlikely to provide a meaningful estimate of treatment effectiveness. If a propensity score approach is applied, the final target population has to be described in detail, especially if trimming or truncation methods are used. Owing to the greater uncertainty associated with non-randomised data, a sufficiently large treatment-effect estimate is required, which can be formally evaluated via testing of shifted hypotheses. Methods of sensitivity analysis that explore the potential impact of unmeasured confounders may also be applied to assess the robustness of the estimated treatment effects in this scenario.

The choice of methodology is ultimately dependent on multiple considerations (e.g., assessment scope, availability of data, underlying assumptions of methods) and should be appropriate to the data available. In many cases, the conditions will not be ideal for the use of any of the methods presented in this guideline to produce unbiased estimates of relative effectiveness. Therefore, very careful consideration of the underlying assumptions is required when making inferences. Input from a statistician with specific expertise in this area is advised for a critical assessment of the methodological approach used, assumptions potentially violated, and the corresponding uncertainty of results.

I Introduction, objective and scope

To assess the relative efficacy or effectiveness of a new intervention compared to another intervention (the comparator; e.g., the current standard treatment) in the presence of multiple sources of evidence, the best approach is given by formally combining the evidence. Broadly, we refer to this as quantitative evidence synthesis. As individual studies providing evidence with the highest certainty of results are mostly randomised controlled trials (RCTs), we assume that the evidence synthesis is based on adequate RCT data unless otherwise stated.

A systematic literature search is a prerequisite before conducting an evidence synthesis. For the purposes of this document, it is assumed that this systematic literature search has been properly conducted in accordance with good practices for the conduct of evidence synthesis and that the collection of the data contributing to the comparisons is complete, as required in the health technology assessment regulation (Regulation (EU) 2021/2282) (HTAR) (Art. 9 in [27]). Evidence must be relevant for the research question which should be formulated according to the PICO (population, intervention, comparator, outcome) framework.

Objective

The objective of this document is to describe the methods currently available for direct and indirect treatment comparisons regarding their underlying assumptions, strengths, and weaknesses. This guideline also specifies the appropriateness of methods to the data situation (e.g., the type of network and the data sources for which they can be used). The document is not a methodological textbook and does not give a detailed description of the statistical techniques described. Rather, the methods are briefly summarised, and general guidance is provided on which method(s) are appropriate in a particular situation. Specific guidance for assessors, co-assessors, and other members of the assessment team (hereafter referred to collectively as assessors) dealing with results from direct and indirect treatment comparisons submitted by health technology developers (HTDs) for performing a joint clinical assessment (JCA) is provided in the HTA CG Practical Guideline for Quantitative Evidence Synthesis: Direct and Indirect Comparisons. This methodological guideline does not cover the basic methodological principles for direct comparison of treatments using data from a single head-to-head comparative study. In addition, this guideline does not cover methods for evidence synthesis of diagnostic accuracy studies.

This methodological guideline is aimed at assessors in the context of JCA of health technologies, although the document is also relevant for other stakeholders, including those submitting evidence.

Scope and terminology

Terms used to describe types of evidence syntheses as discussed in this document are sometimes used with a slightly different understanding throughout the literature, and therefore we need to describe these terms and what they broadly describe. Pairwise meta-analysis, also known as direct comparison, refers to the synthesis of direct evidence for when exactly two interventions are compared. Network meta-analysis (NMA) is a generalisation of meta-analysis to analyse more complex evidence networks, which may include both direct and indirect evidence. We consider NMA to include other terms used in the literature to describe the synthesis of both direct and indirect evidence, such as mixed treatment comparison and indirect treatment comparison. Indirect treatment comparison is used by some authors to describe the situation in which inference about the relative efficacy or effectiveness of two treatments is made in the absence of trials comparing these treatments head-to-head. In this document, we use the term indirect comparison as the broadest term to refer to any evidence synthesis incorporating indirect evidence, which therefore includes NMA, population-adjusted methods such as matching-adjusted indirect comparison (MAIC) and simulated treatment comparison (STC), and comparisons made in disconnected evidence networks. Indirect comparisons that synthesise relative treatment effects via a shared comparator treatment (thus respecting randomisation) are referred to as anchored indirect comparisons, while those made without the use of a shared comparator are comparisons of absolute outcomes, and are called unanchored indirect comparisons.

For cases in which no data for the relevant direct comparison are available or the assessment scope requires simultaneous comparison of more than two interventions, which have not been compared directly in trials, methods for indirect comparisons can be used. However, all else being equal, results from comparisons based upon indirect evidence alone generally have greater uncertainty than results from comparisons incorporating direct RCT evidence. Therefore, direct comparisons based on adequate RCTs with low RoB should be applied whenever possible.

Throughout this document, we use the term uncertainty in its most general sense, incorporating both bias and imprecision, as well as other concerns regarding the methodological validity and applicability of the submitted evidence synthesis. This term should not be interpreted as uncertainty in the context of "certainty of evidence" used in the grading of recommendations assessment, development and evaluation (GRADE) framework, although there is overlap in some domains.

For simplicity, we use effectiveness as the common term to describe efficacy or effectiveness throughout the rest of this document. Effectiveness also includes safety within the context of this document. Furthermore, treatment, intervention and health technology are all terms used for any health technology that can be assessed.

In what follows, it is important to distinguish between prognostic variables, effect modifiers and confounders. Prognostic variables are characteristics (i.e., patient characteristics) that affect the outcome of interest irrespective of which treatment is received, while effect modifiers are characteristics that alter the relative effectiveness between two treatments. Thus, effect modifiers are specific to the pair of treatments being compared and to the scale used to measure the relative treatment effectiveness. In statistical terms, effect-modifying covariates can be considered as interactions between the treatment and the trial-level or patient-level characteristics. It is possible for a particular characteristic to be both a prognostic variable and an effect modifier, although in general not all prognostic variables will be effect modifiers. An example of a characteristic that can be both a prognostic variable and an effect modifier could be the stage of a particular disease. A patient at an earlier stage can have a better prognosis than a patient at a later stage irrespective of the treatment the patient receives, and, also, the relative effectiveness of the treatment being studied to its comparator is not the same for patients at an early stage and patients at a later stage. In the context of a comparison between two treatments, a confounder is a characteristic that affects both the likelihood of receiving the treatment and the outcome. A confounder is therefore necessarily a prognostic variable (but the inverse is not true) and can be an effect modifier. Again, an example would be the stage of a disease: as previously said it is a prognostic variable, but it can also determine which treatment a patient is more likely to receive. Confounding bias (i.e., the estimated effect cannot be causally attributed solely to the treatment) arises when there are systematic differences in the distribution of a confounder between groups.

Throughout this guideline we discuss the requirements and assumptions that must hold for the methods that we describe to produce meaningful estimates of treatment effects. Often, when using evidence synthesis methodology, some assumptions will be made, which may affect the certainty of results. This guideline aims to allow HTA assessors and HTDs to identify and evaluate potential bias and uncertainty to inform decision making as much as possible. However, we recognise that there is an element of subjectivity in the assessment of many assumptions and that decisions may vary between Member States. To answer certain PICO questions, methods of evidence synthesis will sometimes need to be applied despite uncertainty or doubt as to their validity. In these scenarios, the HTD must always submit evidence to inform the comparison of interest, together with sufficient supporting information to allow the JCA assessors to determine the extent to which the corresponding results produce meaningful estimates of relative treatment effectiveness, and to evaluate the extent of bias and uncertainty.

When treatment effect estimates are subject to uncertainty that goes beyond statistical imprecision, e.g., when there is a risk of bias arising from confounding, statistical testing against a ‘shifted null hypothesis’ is sometimes used to account for the additional systematic uncertainty in effect estimate. This method entails testing for statistical significance of this estimate against a threshold value that is shifted away from the conventional null of ‘no effect’, so that the shifted null hypothesis is rejected only if the confidence interval lies entirely to one side of this threshold. However, no general consensus exists and the determination of an appropriate threshold for decision-making is considered to be a matter for individual Member States. The results of such tests may nonetheless be presented in the JCA submission or the assessment report as they can provide useful information to assessors and national decision-makers. If an HTD presents such results, the choice of the value of the pre-specified threshold used for the test must be justified with corresponding references to the appropriate literature.

There are many useful publications on methods for evidence synthesis that advise on the theories, methods and assumptions; while we have drawn material from these texts, we advise further reference to them for completeness [7,18,37,77,88,99].

II Analysis and discussion of methodological issues

1. Types of evidence

The scope including the PICO framework for an evidence synthesis is defined elsewhere [58,59]. In order to carefully consider the analysis that is most appropriate, there needs to be a clear understanding of the types of evidence presented by HTDs. The following briefly describes the types of evidence. Further guidance is provided for assessors in the Practical Guideline for Quantitative Evidence Synthesis: Direct and Indirect Comparisons. In the case of different PICO questions, a different evidence synthesis for each PICO (e.g., pairwise meta-analysis or NMA) is generally required, but an evidence synthesis including more than one comparator (i.e., more than one PICO within a given population) could also be necessary to address Member States' evidence needs.

The gold-standard evidence is from adequate RCTs with low RoB and adequate sample size. This represents direct evidence on the benefit of a treatment over an existing comparator(s). A key feature of randomisation is that it ensures that treatment assignment is independent of all prognostic variables and effect-modifiers, i.e. there are no confounders. In this case, the underlying assumption of exchangeability holds; in other words, if patients from one group were substituted to the other, the same treatment effect would be expected. This implies that patients in each treatment group have the same average risk of presenting the outcome of interest on inclusion in the trial and therefore there is an unbiased estimation of the relative treatment effectiveness (assuming a sufficient level of internal validity for the RCT of interest). Importantly, this applies not only to known or observed patient characteristics but also to unknown characteristics for which a similar distribution in each group cannot be achieved (or even assessed) using other methods [12]. However, it is worth keeping in mind that prognostic factors and effect modifiers may be unevenly distributed between groups by chance, which could affect the outcome, particularly in small studies.

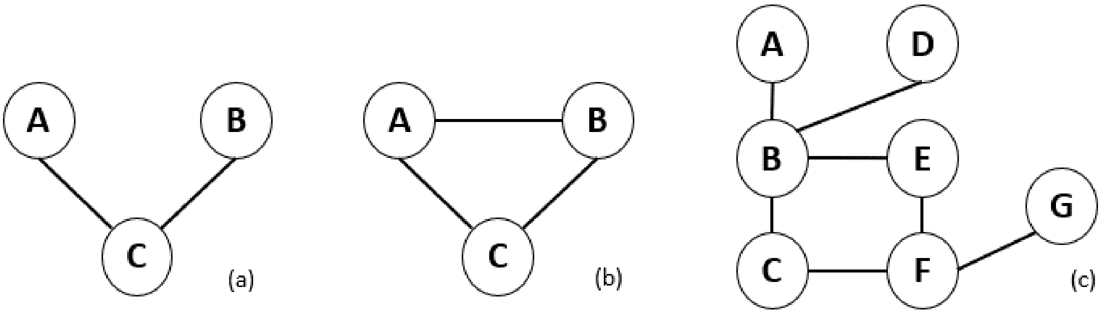
In its simplest form, "indirect evidence" for a comparison of two interventions A and B arises when there is direct RCT evidence comparing both A to C, and B to C that can be combined to indirectly estimate the benefit of treatment A versus treatment B (Figure 1a). Indirect evidence cannot ensure balance of both known and unknown effect modifiers to the same degree as direct evidence from a single RCT and, all else being equal, is more uncertain as a result. However, when direct evidence informing a comparison of interest is not available, or when simultaneous comparisons of multiple treatments is required, comparisons using indirect evidence need to be applied.

Non-randomised evidence commonly encountered are single-arm trials, cohort studies, case-control studies, other observational studies and the use of historical controls. This type of evidence also includes 'un-anchored' indirect comparisons, i.e., indirect comparisons that compare absolute outcomes between treatments across different studies (rather than relative effects along a connecting path of RCTs). Any such study has much greater potential to include material bias in the estimate of relative treatment effectiveness, especially as the underlying assumption of exchangeability is unlikely to hold, and there is a high risk of confounding bias. This high RoB is likely to carry through and can be compounded when combining evidence from these sources.

2. Network of evidence

The collection of studies relevant to the analysis forms an evidence network (Figure 1) [21]. This network consists of both direct evidence, that is, evidence from RCTs (represented by lines connecting the interventions), and indirect evidence, which exists whenever two interventions can be connected by a path of RCTs. These networks of randomised evidence allow to estimate the relative effectiveness for any pair of treatments, provided they are connected. A network is said to be connected (as in Figure 1) if any two comparators are linked by at least one path of RCTs.

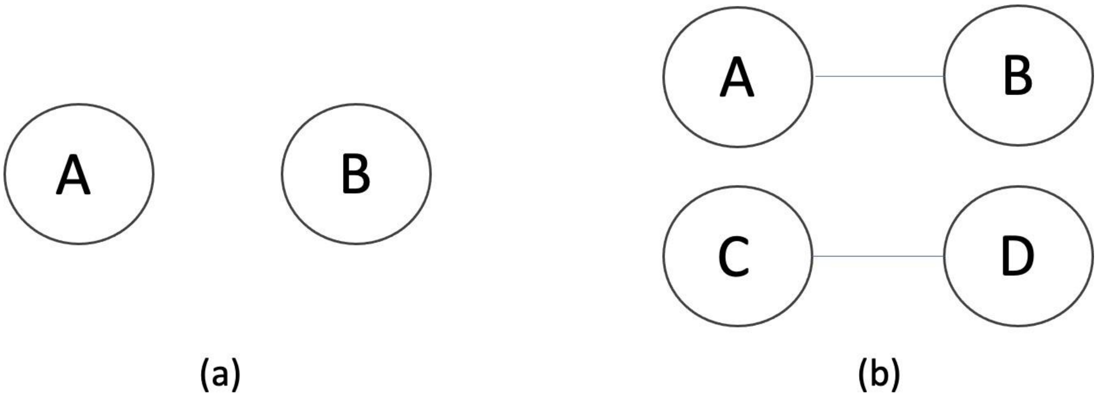
Figure 1 – Illustrative networks of evidence



(a) This is a simple star network. The solid lines represent RCTs comparing A to C and B to C (direct evidence). As A and B share a common comparator (C), an indirect comparison between A and B is allowed. (b) Direct evidence between A and B, A and C, and B and C (closed loop) allows a comparison between the direct and indirect evidence (mix of direct and indirect evidence). (c) A larger network containing evidence for many different treatments, which allow many indirect comparisons between two treatments to be performed.

When considering an evidence synthesis for the comparison of two interventions, the shape of the network (as described in Figure 1) has an impact on the type of analyses that may be carried out. Three examples are shown. There are, however, more geometric networks that can be formed [74]. Direct comparisons (i.e., standard pairwise meta-analyses) are carried out in connected evidence networks containing two interventions. For connected networks containing more than two interventions, there are a number of methods available. Bucher’s method for indirect comparison (Section 5.1) can only be used in the case of a simple star network (such as the network shown in Figure 1a). In more complex connected networks of evidence (as illustrated in Figure 1b and 1c), more complex NMA methods (Section 5.2) are needed.

Figure 2 – Disconnected networks



(a) Two single-arm studies. (b) Two RCTs (A vs B, C vs D) for which the required comparison is C versus B, so a strong assumption must be made in relation to the equivalence of the comparisons and the relative treatment effectiveness.

Disconnected networks of evidence arise when a set of RCTs does not provide sufficient information to be able to carry out an assessment of an intervention against a relevant comparator along a connected path. A disconnected network can occur in cases in which the clinical evidence stems from single-arm trials (i.e., a study carried out without a comparison group; Figure 2a) or the standard treatment is different in the studies and there are no head-to-head comparisons of the interventions being considered (Figure 2b). Disconnected networks such as those illustrated in Figure 2 are problematic since there is no way in which the comparators of interest can be compared using paths involving evidence from RCTs. Attempts to connect these networks have been proposed in the literature to deal with cases in which such evidence networks have arisen [50,75,91]. However, these approaches rely on very strong assumptions that need to be examined carefully for any specific application. Evidence from a single-arm study (i.e., non-comparative observational data) is sometimes compared to data for a group obtained elsewhere; for example, historical controls or an unrelated contemporaneous study could be used [33]. The implication of using such evidence is that the relative comparison of the results between the groups depends not only on the interventions being studied but also on all other aspects that differ between the groups and the studies (i.e., the assumption of exchangeability probably does not hold). Currently, there is no gold-standard method that addresses the issue of disconnectedness of evidence networks. The use of such evidence in JCA is highly problematic as it carries a high risk of not providing valid and reliable estimates of the relative treatment effects of interest. For this reason, the HTAR requires comparative results on the basis of adequate comparisons (PICO framework) [27].

Other types of evidence that can be problematic to incorporate in a network are comparative observational studies and registry data obtained without randomisation. Observational studies examining two or more interventions have been used to connect an otherwise disconnected network [75]. It has been proposed to use this approach to perform comparisons that otherwise would not be possible. However, relying solely on these methods to produce an unbiased estimate of the relative effectiveness of a treatment(s) of interest in practical settings remains controversial, and producing evidence using data from adequate RCTs with low RoB should always be favoured. In the context of JCA, assessors should be aware that the inclusion of evidence from non-randomised studies may lead to results that are highly uncertain and unlikely to provide a valid and reliable estimate of the relative treatment effectiveness. In some cases, it may be possible that the lack of randomisation can be compensated by rigorous adjustment for confounding. However, in general, this requires access to the full IPD information (Section 6).

3. General statistical considerations

There is not always a common understanding of the terminology in the field of evidence synthesis, in particular for indirect comparisons. In the literature, different authors use different terms for the same concept; for example, similarity and homogeneity are sometimes used interchangeably. Since there is no common terminology for the concepts that are described in the following sections, it is possible that these concepts are described with a different terminology elsewhere.

The main assumptions and other general statistical considerations applicable to all methods for evidence synthesis are outlined in this section; assumptions that are specific to a subset of methods or to one particular method will be described in the corresponding section.

3.1. Assumption of exchangeability

In a formal evidence synthesis (whether simple or more complex), exchangeability is the most fundamental assumption, that is, if individuals in one trial were substituted to another, the treatment effect observed is expected to be the same [18,40]. For practical purposes, this fundamental assumption is operationalised by assessing the properties of similarity and homogeneity and, in the case of indirect comparisons, consistency, all of which are required for exchangeability to hold [22,46,83]. Exchangeability can then be explored by first looking at observable consequences of violations in this assumption, namely by searching for dissimilarities between studies in terms of patient characteristics and study design. Moreover, there is the possibility of testing for statistical heterogeneity. If the test is statistically significant, then variations in between-study treatment effects are plausibly not due to chance (i.e., to random error) alone and thus there are systematic errors, so it is necessary to look for dissimilarities that explain this statistical heterogeneity. The three properties are not, strictly speaking, distinct assumptions, because a failure of homogeneity or consistency is often the result of a systematic difference in distribution of effect modifiers between studies (i.e., a violation of similarity).

The common requirement of exchangeability applies regardless of the complexity of the network or the methods used for conducting evidence synthesis, and should be carefully assessed before undertaking a formal evidence synthesis. When the exchangeability assumption is violated, the outputs of the analyses are affected and probably biased, and advanced statistical expertise should be sought to aid in interpretation of the outputs.

3.1.1. Similarity

A necessary condition for the exchangeability assumption to be valid is sufficient similarity of all the trials included regarding effect modifiers, which means that there are no differences in the distribution of known and unknown effect modifiers (e.g., sex or age) that modify the true difference between the treatment arms regarding the outcome of interest [46,83]. It remains impossible to assess any unknown effect modifier. However, a thorough feasibility assessment is necessary to identify differences in study design and patient characteristics that can influence similarity. If dissimilarities between studies in study design and/or patient characteristics are observed at a level that is considered substantial, it can be indicative that the fundamental assumption of exchangeability will not hold. Therefore, only if study design and the patient populations are considered similar enough can the results generate reliable estimates of the treatment effect(s) of interest. Specific guidance on assessing similarity in the context of

JCA is provided in the Practical Guideline for Quantitative Evidence Synthesis: Direct and Indirect Comparisons.

3.1.2. Homogeneity

Validity of the exchangeability assumption also requires that the relative effectiveness between each pair of treatments is sufficiently homogeneous across all studies comparing those treatments included in an evidence network (i.e., we require sufficient homogeneity of studies). If the results from the studies are very different, heterogeneity is observed and therefore combining the results may not be appropriate [26]. Heterogeneity can be clinical, methodological and statistical. It is possible to test for heterogeneity to provide evidence of whether or not the study results differ greatly [88]. However, non-significance of a statistical test for heterogeneity does not prove homogeneity because the test can be non-significant owing to lack of power. Statistical heterogeneity (i.e., when effect estimates vary more than expected by chance alone) can also be quantified via several methods [39,73]. In addition to statistical heterogeneity, clinical and methodological heterogeneity must also be examined. Clinical heterogeneity includes variability in patient inclusion criteria (e.g., age, severity of disease, line of therapy), interventions (e.g., dosage, administration route) and outcomes (e.g., different time points). It should be noted that these differences should be considered in terms of (dis)similarities if the corresponding characteristics are effect modifiers. Methodological heterogeneity includes, for example, variability in study design. To explore heterogeneity further and to identify factors contributing to it, subgroup analyses and meta-regression are useful tools [38,92]. In the context of JCA, subgroup analyses are often more useful than meta-regression, as they can help in targeting the right intervention for the right subgroup of patients. In addition, a significant statistical test for heterogeneity can be indicative of dissimilarities in study design and/or patient characteristics (Section 3.1.1) and can lead to a discussion of the plausibility of the fundamental assumption of exchangeability. Specific guidance on assessing heterogeneity in the context of JCA is provided in the Practical Guideline on Quantitative Evidence Synthesis: Direct and Indirect Comparisons.

Regardless of whether between-trial heterogeneity can be explained there must still be a decision whether or not to proceed with the comparison and whether subgroup analyses will sufficiently explore the impact of the heterogeneity on the analysis outputs. In a subgroup analysis, only studies that are considered to be sufficiently alike according to a more narrowly defined set of criteria (e.g., age range of study participants) should be included.

3.2. Robustness

The robustness of the analysis will depend on the inclusion of appropriate evidence that has been gathered in a systematic and rigorous manner and excluding any obvious bias that may occur. Further assessment of the robustness can be undertaken via sensitivity analyses of various aspects such as models and missing data, among others. Results for sensitivity analyses should be thoroughly discussed in the context of the evidence available and the results obtained.

The results of an evidence synthesis may be overly influenced by one or a small number of studies. Whether or not this is problematic should be discussed in the context of the evaluation. Similarly, some studies may be outliers in a statistical sense [55]. Outlier and influential studies are not synonymous: an outlier study is not necessarily an influential one, and vice versa. A first step for identification of outliers is to visually inspect a forest plot to identify any unusual data points or cases in which the pooled estimate appears to be driven by a single or small number of studies. In meta-analysis and NMA, visual inspection of quantile-quantile plots and other graphical tools can identify outliers and the robustness of the results of evidence syntheses [2,48,53]. Sensitivity analysis techniques can be used to determine the impact of influential studies and outliers on the results of an evidence synthesis. For example, an analysis can be conducted with and without a particular study to determine the impact of that study on the results [13]. It is also useful to characterise outliers in terms of how they might differ from other studies.

3.3. Sources of bias

In conducting treatment comparisons, bias must be minimised. Bias reflects a systematic error in the results and results in deviation of the estimated treatment effectiveness from the true treatment effectiveness. When performing evidence synthesis, some key potential sources of bias should be considered. The first is bias in the results of the individual studies included in the review. If the individual study results are biased, then a synthesised summary of the individual results will also be biased and can yield misleading conclusions. Therefore, the RoB for the results from the individual studies must be assessed [84,85].

A second potential source is bias in the result of a pairwise meta-analysis or other form of evidence synthesis. In addition to the RoB in the studies included, the result of an evidence synthesis may be affected by bias due to the absence of findings from studies that should have been included had their results been reported, known as publication bias. The issue of publication bias arises because negative results are less likely to be published [25]. The consequence of this is bias in the resulting effect estimate(s), and theevidence synthesis result may show a spurious significant effect. Publication bias may be detectable using funnel plots or regression techniques, but these methods are not without weaknesses [60]. Asymmetry in a funnel plot may indicate publication bias or it may be a reflection of how comprehensive the search strategy has been. A non-comprehensive search is a potential source of bias. Therefore, it is of critical importance that the search strategy for the systematic review is as comprehensive as possible and that clinical trial registers and any published full study protocols are searched, where possible. The presence of publication bias can impact on any evidence synthesis irrespective of the methodology used.

In the context of JCA, the HTD must provide all available evidence and data concerning the product under assessment according to the EU regulation. Therefore, the issue of publication bias typically only arises in the case of studies sponsored or conducted by other organisations (e.g., studies investigating comparator products).

3.4. Fixed-effect and random-effects approaches for evidence synthesis

Fixed-effect and random-effects approaches for evidence syntheses are available. In the fixed-effect model, also known as the common-effect model, the true treatment effectiveness is assumed to be the same in each study that compares the same treatments. Use of a fixed-effect model therefore follows from the assumption that variability between studies is entirely due to chance, which is commonly implausible [6]. In a random-effects model, the treatment effect in each study is assumed to vary around an overall average treatment effect [24]. Specifying a fixed-effect model for evidence synthesis relies on a stronger assumption than specifying a random-effects model, namely that the true effect for the same comparison is identical in all included studies. This depends on the strictness of the inclusion criteria of the studies used for pooling, the definition of outcomes (including whether they are objective or not) and how the interventions are defined (e.g., not grouping multiple doses as a single treatment), among others. For instance, in the context of a pairwise meta-analysis of two studies with identical designs, this assumption is appropriate. Without adequate justification that the assumption of a common effect holds, a random-effects model should generally be used. Incorrect use of a fixed-effect model may result in, for example, too narrow confidence intervals and consequently p-values that are too small and interpreted as that the results imply a common treatment effect [26].

Random-effects models provide an estimate of the between-study variance and the summary effect estimate. Prediction intervals provide a predicted range for the true effect size in an individual study, which incorporates the degree of heterogeneity in a random-effects evidence synthesis, together with the uncertainty surrounding the relevant average treatment effect. Therefore, the use of prediction intervals is recommended when reporting results for a random-effects evidence synthesis [95]. When the number of studies included is small, random-effects methods may not provide adequate estimates of the between-study variance and may have low statistical power. In this scenario, a fixed-effects approach should only be considered if the underlying assumptions are plausible, while a qualitative summary of the study results might also be considered as an alternative [3,76]. Bayesian methods are also an option for evidence syntheses involving sparse data and few studies (Section 4.2)[3].

3.5. Frequentist and Bayesian approaches

As with any approach to statistical inference, evidence synthesis may be performed using a frequentist or a Bayesian framework. Because of the possibility of incorporating information from existing sources of data for modelling of prior distributions, Bayesian methods are particularly useful in situations with sparse data. In the Bayesian approach, prior probability distributions for model parameters such as treatment effects and between-study heterogeneity are specified before the analysis begins. The study results are then combined with the prior distributions to derive posterior distributions for the model parameters, including the overall treatment effectiveness [93]. Prior distributions that have broad support in the parameter space are called either vague priors or non-informative priors. When non-informative prior distributions are used, results can be similar to those observed using a frequentist approach. When there is some prior knowledge (e.g., about likely between-study heterogeneity), a distribution that is narrower can be used (i.e., a distribution that will be called informative). As this has a stronger influence on the posterior distributions and hence on the estimate of relative effectiveness, informative prior distributions should generally only be used for the heterogeneity parameter and not for the treatment effect itself. The choice of prior distributions for model parameters must be accompanied by a justification and a clear description of how they were generated to maintain transparency. In addition, it is important to ensure that sensitivity analyses for the specification of the prior distribution are carried out. In particular, sensitivity analyses should explore the effect of using an informative versus a non-informative prior.

3.6. Use of IPD and aggregate data

While evidence synthesis typically combines study-level effect estimates, it is also possible to pool IPD from studies. If available, statistical analyses using raw data (i.e., IPD) should be preferred to statistical analyses that use only summary statistics (for instance, IPD can be used to investigate treatment by covariate interactions, and distributional assumptions can be assessed). The methods for evidence synthesis based on IPD can broadly be classified into two groups: a one-step analysis, in which all patients are analysed simultaneously as though in a mega-trial, but with patients clustered by trial; and a two-step analysis, in which the studies are analysed separately, but then summary statistics are combined using standard techniques. Hybrid methods are also available for combining IPD and aggregated study data [69].

Techniques to reconstruct partial IPD from published time-to-event data (Kaplan-Meier plots) could be considered also [32]. In the case of binary data, IPD are available if information on the studies' 2 × 2 tables is given. Even if only effect estimates together with confidence intervals or standard errors are given, a method for reconstructing 2 × 2 tables can be used [17]. However, these methods in general only reconstruct treatment and outcome data and will not include patient characteristics. Evidence synthesis based on full IPD (i.e. including baseline patient characteristics) have better modelling options for estimating treatment effectiveness when compared to corresponding aggregate data analyses. In particular, the availability of IPD allows valid subgroup analyses and statistical adjustment regarding patient characteristics [4,81]. However, IPD are frequently not available and cannot be reconstructed in all required detail, which limits the use of evidence synthesis based on IPD.

Key Points 1

(a) The exchangeability assumption is required to justify an evidence synthesis of the data being considered, which requires sufficient similarity and homogeneity of the included studies.

(b) If between-trial heterogeneity is too strong to justify an evidence synthesis but the heterogeneity can be explained, appropriate evidence syntheses should be performed in the corresponding groups of trials or subgroups of patients or by means of meta-regression.

(c) Fixed-effect models rely on a strong assumption that all variation observed is due to chance, which is rarely the case in practice. Therefore, the use of fixed-effects models requires rigorous justification by the HTD Incorrect use of a fixed-effect model will lead to confidence intervals that are too narrow and p-values that are too small.

(d) Application of Bayesian methods is a useful option, especially when the data are sparse and the fixed-effect assumption is not adequate.

(e) Analysis of IPD is preferred over aggregate data/summary statistics, especially for subgroup analyses regarding patient characteristics.

(f) The robustness of the results should be assessed and discussed by means of sensitivity analyses.

4. Direct comparisons

Direct comparisons are performed by means of standard pairwise meta-analyses in which results from two or more trials that all compare the treatment of interest to the same comparator are combined. In this context, the comparator is defined according to the PICO question. An investigation of whether the data considered for the meta-analysis fulfil the assumptions of similarity and homogeneity is required. If this is not the case, the results from pairwise meta-analysis of such data are unlikely to provide a meaningful and reliable estimate of treatment effectiveness and the appropriateness of the analyses should be questioned.

Pairwise meta-analysis involves computation of a summary statistic with precision for each trial followed by combination of these studies into a weighted average [24]. Outcomes can be binary, continuous or time-to-event. The summary statistic is typically an odds ratio, risk ratio, risk difference, hazard ratio, rate ratio, difference of means or standardised mean difference. The same summary statistic must be available for each study included in the pairwise meta-analysis, either extracted directly from the study publication(s) or else computed from the extracted data. The methods used for direct comparisons can be broadly split into frequentist and Bayesian approaches.

4.1. Frequentist approach

In a frequentist framework, pairwise meta-analyses can be divided into fixed-effect and random-effects methods. Fixed-effect models include inverse variance, Mantel-Haenszel and Peto methods. Inverse variance methods can be used to pool estimated summary measures with standard error and weights proportional to the inverse squared standard errors for the studies. Inverse variance methods are less reliable when data are sparse. The Mantel-Haenszel method provides more robust weighting when data are sparse and gives similar weights to inverse variance methods when data are not sparse. The Peto method is used for odds ratios and can be extended for pooling of time-to-event data. It has been shown that the Peto method fails when treatment effects are very large and when the sizes of the trial arms are very unbalanced [90]. The Peto method performs well when event rates are very low, treatment effects are small and the trial design is balanced. An undesirable feature of the Peto method is its dependence on a balanced group size ratio, which makes its interpretation difficult and limits its practical usefulness [9]. Fixed-effect methods tend to give small weights to small studies and large weights to large studies. In general, the standard approach for application of the fixed-effect model is the inverse variance method in the case of continuous data and the Mantel-Haenszel method in the case of binary data.

The most common estimation method for the random-effects model was the method of DerSimonian and Laird [15]. However, this method leads to increased type 1 errors (i.e., p-values that are too small and confidence intervals that are too narrow), especially in the case of few available studies, and is no longer recommended [3,14,95]. The recommended method for random-effects meta-analyses is the Knapp-Hartung (KH) method, also the called Hartung-Knapp-Sidik-Jonkmann method [95]. The KH approach in combination with the Paule-Mandel estimator for the heterogeneity parameter is recommended as the standard method for random-effects meta-analysis in situations with five or more studies. As a supplement to confidence intervals, use of prediction intervals is also recommended, as these reflect the potential true effect size in a future study and illustrate the uncertainty about whether an intervention is expected to work in a new study [95]. In situations with very homogeneous data, ad hoc variance correction may be required for the KH method [103].

A disadvantage of the KH method is that this approach frequently has very low power in the case of very few (i.e., <5) studies and is not recommended in these scenarios [3,76]. Alternative approaches that may be considered include a fixed-effect pairwise meta-analysis or a qualitative summary of the study results [43], and other methods, such as Bayesian pairwise meta-analysis (Section 4.2) and the beta- binomial model in the case of binary data [56]. A possible procedure for choosing a useful approach for evidence synthesis in cases involving very few studies is described by Schulz et al. [76].

For certain effect measures, such as risk ratios, a study with zero cases can be problematic for some weighting approaches such as the inverse variance method. In order to deal with this, a continuity correction can be applied to arms with zero cases. While a value of 0.5 was used historically, other non-fixed zero-cell corrections may have advantages, as has been explored by a number of authors [8,90].

For avoiding the use of zero-cell corrections, the beta-binomial model [49] can be used. Bayesian models for meta-analysis (Section 4.2) can also handle most situations involving zero cells without the need for a zero-cell correction. In general, the use of generalised linear mixed models should be considered if the available data for the evidence synthesis are sparse [5,66,82,86].

4.2. Bayesian approach

Bayesian methods for pairwise meta-analysis are analogous to frequentist methods with the primary distinction being the use of prior distributions for the model parameters [87], that is, the treatment effect and (for random-effects models) heterogeneity parameters.

Bayesian models perform well in many situations in which others do poorly, such as in analyses involving sparse data and few studies by means of the binomial-normal hierarchical model (which also avoids the need for a continuity correction in the case of zero-event studies). More generally, a hierarchical Bayesian model with weakly informative prior distributions for the heterogeneity parameter may be a better method to account for uncertainty than a non-Bayesian approach, particularly when the number of studies is small [70]. For random-effects models, selection of the prior distribution for the heterogeneity parameter is critical to any Bayesian analysis [30] and this choice should therefore be transparently justified and varied in sensitivity analyses (Section 3.4).

Key Points 2

(a) Standard frequentist methods for fixed-effect pairwise meta-analysis are the inverse variance method for continuous data and the Mantel-Haenszel method for binary data.

(b) The Knapp-Hartung method is currently the recommended frequentist approach for pairwise random- effects meta-analyses when there are five or more studies.

(c) Frequentist and Bayesian approaches to pairwise meta-analysis are both possible. The Bayesian approach allows incorporation of prior or external information for the treatment effects and heterogeneity parameters, but the choice of the prior distributions requires a clear justification.

(d) In general, non-informative prior distributions should be used for Bayesian analyses. Informative prior distributions should generally only be used for the heterogeneity parameter in pairwise random-effects meta-analyses and should be thoroughly justified.

(e) Standard approaches for random-effects meta-analysis with rare events and/or few studies often perform poorly. In this case, the use of alternative methods should be considered, such as a qualitative summary of the study results, Bayesian methods (with a weakly informative prior distribution for the heterogeneity parameter) or the beta-binomial model.

5. Indirect comparisons

When treatments have not been directly compared in RCTs, indirect comparisons can be used. Treatments should be connected in simpler or more complex networks of RCTs via one or several common comparator(s) (Figures 1 and 2).

When indirect comparisons are made, methods for anchored indirect comparisons with one or several common comparators should be used. Anchored indirect comparisons can be performed on aggregated data and hence do not require access to IPD. Population-adjusted methods (Section 5.3) are performed on a combination of aggregate data and IPD. Comparisons based on non-randomised evidence require in general access to the full IPD information (Section 6). The following sections describe methods for indirect comparisons in connected networks. The use of methods for indirect comparisons based on aggregated data is not recommended in disconnected networks (Section 6.1).

As with direct comparisons, indirect comparisons of aggregate data make the same fundamental assumption of exchangeability across studies, which requires sufficient similarity of all the trials included regarding effect modifiers, and sufficient homogeneity of the study results for all pairwise comparisons [46]. The exchangeability assumption for indirect comparisons requires the property of consistency, namely that direct pathways and indirect pathways are estimating the same treatment effect [36,46,83]. Further information on consistency is given in Section 5.2. Specific guidance on assessing similarity, homogeneity and consistency in the context of JCA is provided in the Practical Guideline for Quantitative Evidence Synthesis: Direct and Indirect Comparisons. If at least one of these assumptions are violated, the results of an anchored indirect comparison are unlikely to provide a meaningful estimate of the treatment effect.

Naive indirect comparisons combine outcome data from treatment groups across different studies as though they had come from a single large trial and thus break randomisation [23,29]. Other approaches towards unanchored indirect comparisons that attempt to compensate for the lack of randomisation are available, however these require the assumption of "conditional constancy of absolute effects" (Section 6.1), which is very unlikely to be fulfilled. Anchored indirect comparisons preserve randomisation and should always be used in preference to unanchored methods. In the absence of a connected network, methods for the analysis of non-randomised evidence should be applied when conducting unanchored indirect comparisons (Section 6).

5.1. Bucher’s method for anchored indirect comparisons

Bucher et al. [10] presented an anchored indirect method of treatment comparison for aggregate data that can estimate relative treatment effectiveness for a simple network which includes three different treatments (Figure 1a). This method is compatible with numerous measures of treatment effect including odds ratios, risk ratios, risk differences, standardised mean differences and hazard ratios [98]. The Bucher method is intended for situations in which there is no direct comparative evidence for treatments A and B and the only evidence is through comparison with treatment C. Therefore, the consistency assumptions cannot be assessed in applications of the Bucher method because only indirect evidence is available for the comparison of interest (no closed loops). In this case, a thorough assessment of similarity is even more important. For cases in which there are multiple studies for a pairwise comparison, these must be combined to obtain a summary effect estimate (e.g., using the methods discussed in Section 4) before applying the Bucher method. Certain more complex networks, including closed loops arising from multiple trials, can be analysed, but only in the form of multiple pairwise comparisons. However, this method assumes independence between the pairwise comparisons and thus it cannot be easily applied to closed loops arising from multi-arm trials, for which this assumption fails. When random-effects models have been used to synthesise treatment effects for one or more direct comparisons, use of the Bucher method, either to indirectly estimate treatment effects or to test for consistency within closed loops, is problematic and should be avoided [18,52]. More general NMA methods that appropriately incorporate random effects are available (Section 5.2) and are preferable in this scenario.

5.2. Network meta-analysis

An NMA combines direct and indirect evidence to determine the relative effectiveness of a treatment compared to two or more other treatments. The same assumption of exchangeability as for all indirect comparisons applies, which requires sufficient similarity, sufficient homogeneity and sufficient consistency. Whenever possible, all available relevant comparators should be included in the NMA [21]. In the context of JCA, this would typically include those comparators identified by the assessment scope together with additional comparators needed to form a connected network.

Measures of inconsistency are available for NMAs [20,47] for which both direct and indirect evidence is available. A statistically significant difference in the estimates of relative effectiveness between direct and indirect evidence would indicate inconsistency. A difference in the direction of relative effectiveness, even if not statistically significant, would also raise concerns about consistency. The sources of inconsistency in a complex network can be difficult to identify.

In practice, the possibility of a violation of the exchangeability assumptions for NMA may increase with increasing network complexity and greater numbers of treatments, for example due to clinical heterogeneity across study populations. This may result in inconsistency being observed. There is also a power trade-off between the number of pairwise comparisons and the number of studies included in the analysis: if there are too many comparisons with too few studies, the analysis may be underpowered for detection of true differences [96]. Therefore, the inclusion of additional comparators and studies beyond those required to connect the network should generally be avoided, unless there is strong evidence to suggest that their inclusion improves certainty of evidence. In the context of an NMA, the presence of heterogeneity may mask inconsistency. The consistency assumption cannot be assessed in cases in which corresponding direct and indirect evidence is not available. In such cases, a thorough assessment of similarity is even more important.

In what follows we briefly outline the approaches to NMA most commonly applied in HTA at the time of writing. Other methods are available, such as the original NMA method [54] and the ‘arm-based’ NMA [41]. However, these methods make different fundamental assumptions to the ones described in this document and in general unlikely to be suitable for use in JCAs [19,74,101].

5.2.1. Frequentist approaches for NMA

A method for NMA that is based on graph theory has been developed [71,77]. Methods from graph theory, which is usually applied in electrical networks, were transferred to

NMA. Using this approach, it is possible to handle multi-arm trials within a frequentist framework [72]. In general, the graph-theoretical approach produces similar results to Bayesian NMA (Section 5.2.2) [47,78]. Another frequentist method for NMA based on multivariate meta-analysis and meta-regression, has also been developed [100].

5.2.2. Bayesian NMA

The Bayesian approach for NMA is also called Bayesian mixed treatment comparison [53,74,89]. Bayesian NMA can be applied in any connected network and combines all direct and indirect evidence to obtain treatment effect estimates for all pairwise comparisons in the network. The same principles outlined in Section 3.5 and Section 4.2 are also applicable here.

5.2.3. NMA of time-to event data

In cases involving time-to-event data, evidence synthesis is often based on reported hazard ratios, which rely on the proportional hazards assumption. This assumption is often implausible; the most obvious example is when estimated survival functions intersect and can have an impact on decisions that are based on comparisons of expected survival. Although analyses based upon reported hazard ratios can be robust to minor violations in the proportional hazard assumption, more severe violations may result in bias and/or the non-interpretability of the hazard ratio as a measure of treatment effect. In these cases, NMA based on parametric survival curves [57] or fractional polynomials [45] can be applied, for which the measure of effect is multidimensional as opposed to a single hazard ratio. Other emerging methods for time-varying hazard ratios described in the literature may also be considered [28,34,102]. Whatever the method used, prerequisites and assumptions related to the method must be clearly specified and justified. Availability of IPD will improve the possibility of performing such analyses as described for time-to-event data.

Key Points 3

(a) Indirect comparisons rely on more assumptions than direct RCTs for estimating an appropriate treatment effect and may be more uncertain as a result.

(b) When indirect comparisons are carried out, only anchored indirect comparisons are appropriate, as these respect within-study randomisation.

(c) Anchored indirect comparisons of aggregate data require the assumption of exchangeability across studies. The properties of similarity, homogeneity and consistency should be assessed and reported as a means of assessing the validity of this assumption. If any of these properties do not hold, the results of an anchored indirect comparison are unlikely to provide a meaningful estimate of the treatment effect.

(d) Useful approaches for indirect comparisons include the Bucher method and the frequentist and Bayesian NMA models.

5.3. Population-adjusted methods for indirect comparisons

The methods described in Section 5.1 and Section 5.2 require the property of similarity, also known as "constancy of relative effects" [62]. When this assumption does not hold, these methods are unlikely to provide meaningful results.

In order to account for imbalances in effect modifiers between the RCTs of a connected network, several approaches have been developed to adjust for imbalance and relax the assumption of "constancy of relative effects" [63]. In these approaches, a model is specified that has to include all relevant effect modifiers, allowing anchored indirect comparisons. The new assumption is then the assumption of "conditional constancy of relative effects" (conditioned on the included effect modifiers) [62]. It is important that the relevant effect modifiers that are included are clinically justified and the strategy for selecting them was prespecified in a statistical analysis plan before analysis of the data [51]. In practice, however, one can never be sure that all the relevant effect modifiers are included. Therefore, population-adjusted methods have to be applied with the utmost care. Clear-cut decisions regarding treatment effects on the basis of population-adjusted indirect comparisons with common comparators are only possible if the size of the estimated effect is so large that this large effect could not be induced by bias due to missing effect modifiers alone. For example, this can be formally evaluated by the testing of a shifted null hypothesis. This means that a conclusion regarding an existing treatment effect can only be drawn if the confidence interval lies completely above or below a certain threshold shifted away from the zero effect. This approach accounts for the uncertainty that some relevant effect modifiers may not be included.

In order to implement the approaches, access to IPD is required for at least one study. In the case of an analysis by an HTD, this is usually limited to their own trials.

Population-adjusted methods for indirect comparisons are useful in situations in which an NMA is performed but there is some doubt regarding whether the similarity assumption is valid for some effect modifiers. This doubt can be resolved by applying a population-adjusted method that contains the corresponding effect modifiers to confirm the results of the NMA [51].

Two early approaches for population-adjusted methods for indirect comparisons were developed for situations involving two trials, one comparing treatment A versus treatment B (AB trial) and one comparing A versus C (AC trial), with IPD only available for the AB trial (Section 5.3.1 and Section 5.3.2). A third approach extended the standard NMA framework (Section 5.3.3). It is important to note that the target population for the population adjustment may differ from that of the pivotal clinical trial(s) for the intervention under assessment. The population for which the relative treatment effect is estimated must be clearly stated, while bearing in mind that this can often differ from the population of interest.

5.3.1. Simulated treatment comparison

The simulated treatment comparison (STC) method [11,44] fits an outcome regression model using IPD from the AB trial to predict the average effect of A versus B in the AC population dependent on the covariates, and finally a population- adjusted average effect of B versus C in the AC population. The method also relies on the assumption of “conditional constancy of relative effects”, which means that the model contains all relevant effect modifiers (see above). Furthermore, the validity of STC depends on the correct specification of the outcome regression model. When the outcome regression model uses a non-identity link function, the STC method as commonly applied (i.e., substitution of mean covariate values from the AC population) combines conditional and marginal treatment effects leading to bias for both estimands [65,67]. Approaches to estimating marginal treatment effects using STC have been proposed in the literature [44,68], though these do not appear to have gained widespread adoption.

5.3.2. Matching-adjusted indirect comparison

The matching-adjusted indirect comparison **(**MAIC) method [44,79,80] uses reweighting methods similar to inverse propensity score weighting (Section 6.2) to predict a population-adjusted average (marginal) effect of B versus C in the AC population. It is important to note that this method is only valid where there is sufficient overlap between the patient populations in both trials. Furthermore, MAICs are limited to providing a comparison that is adjusted to the population of the study for which only aggregate data are available, which may not match the target population for the decision. The method also requires the assumption of "conditional constancy of relative effects" to hold, which means that the model contains all relevant effect modifiers (see above). A simulation study was conducted to investigate alternative weighting approaches for MAIC in situations with a common comparator [61]. The study confirmed that none of the different weighting approaches for MAIC can estimate the true treatment effect if there are unmeasured effect modifiers. In contrast to STC, MAIC requires that effect modifiers are specified on the appropriate scale in the model for the weights in order to achieve similar distributions for the effect modifiers in the different populations after weighting.

5.3.3. Multilevel network meta-regression

The multilevel network meta-regression (ML-NMR) approach for population-adjusted indirect comparisons was proposed by extending the standard framework for NMA [64]. ML-NMR provides a formulation in a more general framework allowing for any mixture of IPD and aggregate data, for which full IPD meta-analysis, STC and aggregate NMA can be seen as specific instances. As the other population-adjusted methods, ML-NMR depends on the assumption of “conditional constancy of relative effects” and on correct specification of the outcome regression model. This approach has some conceptual advantages in facilitating inferences from larger networks with any number of treatments. ML-NMR targets conditional treatment effects, and (by contrast with STC) is compatible with non-linear link functions. Marginal (population-average) treatment effects may also be estimated using this method [65]. The population-adjusted treatment effects can be estimated for any target population with given covariate values, and not just the population of the trial for which only aggregated data are available. The application of ML-NMR generally requires either full IPD for at least one study investigating each treatment in the network, sufficiently many aggregate data studies for each treatment, invoking the "shared effect modifier" assumption [62], or else specifying informative prior distributions for treatment- covariate interactions. In practice, these requirements may be difficult to satisfy. Furthermore, ML-NMR as currently proposed cannot be applied to time-to-event data.

Key Points 4

(a) For cases in which the property of similarity does not hold, the usual methods for indirect comparisons are invalid. In this scenario, population-adjustment methods might be considered as an alternative approach, provided the network is connected and there is good evidence a priori that such an adjustment is likely to reduce bias. To this end, model and covariate selection strategies should be prespecified and based on transparent criteria.

(b) Access to IPD from at least one treatment arm in some of the studies included is required in order to adjust for imbalances between trials.

(c) Population-adjusted methods for synthesis of relative effects (i.e., in connected networks of evidence) depend on the correct specification of the weighting (MAIC) or outcome regression (STC and ML-NMR) model, and, in particular, on the assumption that all relevant effect modifiers have been included in the model.

(d) Treatment effects estimated from population-adjusted indirect comparisons are associated with additional uncertainty arising from several sources. Owing to the greater uncertainty, a large effect estimate is required. This can be formally achieved by the testing of shifted hypotheses.

(e) The target population for which the treatment effect is estimated via a population-adjusted method has to be described in detail.

6. Comparisons based on non-randomised evidence

6.1. General considerations

In cases involving an indirect comparison of two treatments observed in different studies without common comparator (i.e., disconnected networks), the data situation is similar to that for any non-randomised trial. Rather than being observed in one comparative observational study without randomisation, the data come from different trials. Naive comparisons between the treatment arms in such situations are prone to bias due to confounding and should not be performed.

When non-randomised evidence is available only at the aggregated data level, there are no adequate method available for reliable estimation of treatment effectiveness.

When a mix between IPD from a single-arm trial and aggregate statistics from another source of data is only available, unanchored STC (Section 5.3.1) and MAIC (Section 5.3.2) have been proposed and applied as a solution for adjusting for confounding bias. However, these analyses without a common comparator (i.e., use of a disconnected network) rely on the very strong assumption of "conditional constancy of absolute effects". This means that the absolute outcome in the treatment arms is assumed to be constant at any given level of the prognostic variables and effect modifiers [62]. However, in almost all practical applications this strong assumption is not justifiable. Therefore, STC and MAIC without a common comparator are highly problematic. When treatment effects are estimated from disconnected evidence networks, methods for the analysis of non-randomised data with access to full IPD from all studies should generally be used instead.

If full IPD information for all relevant co-variates (confounders; prognostic variables or effect modifiers) is available, analyses with adjustment for confounding can be performed. As for the population-adjusted methods (Section 5.3), it is important that the relevant covariates that are included are clinically justified and prespecified in a statistical analysis plan before analysis of the data [35]. Various approaches for adjusting for confounding using IPD are available, such as multiple regression, instrumental variables, g-computation and propensity scores [1,35].

In the context of estimating the relative effectiveness of treatments, methods based on propensity scores, including matching, stratification, conditional adjustment and the inverse probability of treatment weighting, are commonly used [97]. Under the propensity score framework, the most common techniques for adjustments are matching and inverse weighting. Section 6.2 provides more details regarding the application of methods based on propensity scores. The main principles for adjusting for confounding by means of propensity scores are also valid for the other approaches.

Similar to the population-adjusted methods with a common comparator that are based on IPD and aggregated data, the methods for indirect comparisons with adjustment for confounding on the basis of IPD all require that there are no unmeasured confounders. In other words, all relevant confounders and effect modifiers have to be included in the model chosen. Again, it is important that the relevant confounders and effect modifiers that are included are clinically justified and prespecified in a statistical analysis plan before analysis of the data. However, the possibility that a relevant confounder or effect modifier is not included will always remain. Therefore, clear-cut recommendations regarding treatment effects on the basis of indirect comparisons with adjustment for confounding on the basis of IPD are only possible if the size of the estimated treatment effect is so large that the effect could not be induced by bias due to missing confounders or effect modifiers alone. This can formally be evaluated by the testing of shifted null hypotheses. This means that a conclusion for an existing treatment effect can only be drawn if the confidence interval lies above or below a certain threshold shifted away from the zero effect. This approach accounts for the uncertainty that some relevant confounders or effect modifier may not be included [42]. Other methods of sensitivity analysis that explore the potential impact of unmeasured confounders are available to assess the robustness of the estimated treatment effects in this scenario [105]. An example is the use of E-value, defined as the minimum strength of association on the risk ratio scale that an unmeasured confounder would need to have with both the treatment and the outcome to fully explain away a specific treatment-outcome association, conditional on the measured covariates. The larger the E-value is, the more residual confounding would be needed to explain away a treatment effect [94].

6.2. Propensity scores

Propensity scores are mostly used to perform indirect comparisons when full IPD data of non-randomised evidence only (i.e., a disconnected network of evidence) is available. A propensity score is the conditional probability of assignment to a particular treatment given a vector of observed covariates. It can be estimated by modelling as a dependent variable the assignment to a treatment using an appropriate set of covariates. A wide range of statistical models, from logistic regression to machine learning models can be used. Parameter estimates of the model are then used to calculate the propensity score of each patient. When using propensity scores, the assumption of conditional exchangeability (i.e., conditioned on relevant confounders and effect modifiers) must be met and can be assessed by investigating the properties of positivity, overlap and balance [104]:

1. Patients in both groups must be theoretically eligible for both treatments of interest (positivity);
2. There must be sufficient overlap of the data available, as measured by the propensity score, between the populations receiving the treatments of interest;
3. The populations in the groups being compared must be sufficiently balanced after adjustment for confounding.

Relevant patient groups are specified according to the PICO question(s). To meet the positivity assumption, patients who, for instance, have a contraindication to one of the treatments investigated must not be included in the analysis. Overall, an assessment of clinical factors affecting positivity should be performed. Sufficient overlap means that the distribution of propensity score values in the different groups of patients share a common support. Lack of sufficient overlap can indicate that the positivity property is not met for some patients. After adjustment for confounding, the populations in the groups being compared must be sufficiently balanced. This means that the groups compared do not differ substantially regarding the distribution of the relevant covariates. The positivity, overlap and balance must be demonstrated before conclusions are drawn for treatment effects estimated by the use of propensity scores.

The degree of overlap and balance between the groups greatly depends on the model chosen for the propensity score. If an insufficient degree of overlap or balance is obtained by means of propensity scores, sufficient adjustment for an appropriate set of covariates cannot be achieved and consequently no robust treatment comparisons can be made [104]. In this case, switching to multiple regression is not a solution, as this would require inappropriate extrapolations in areas with no observed data [104]. The degree of overlap and balance can also be influenced by "trimming", which involves excluding patients on the basis of propensity scores without overlap [31]. If sufficient overlap and balance can be achieved by trimming, the final overlapping and balanced population of patients is ultimately the target population to whom the estimated effects apply. Therefore, if a propensity score approach is applied, the final target population must be described in detail, especially if trimming or truncation methods are applied. An investigation of whether this target population sufficiently represents the population selected for the original research question is required. If this is not the case, the estimated effects may only apply to a different population to that for the original research question [16].

Key Points 5

(a) Adjustment methods for confounding are based on full access to IPD; all methods require that there are no unmeasured confounders and no unmeasured effect modifiers.

(b) Population-adjusted methods such as STC and MAIC, when applied to disconnected networks, require stronger assumptions than the full IPD methods and are not generally sufficient to adjust for confounding.

(c) The model and covariate selection strategies to adjust for confounding should be prespecified and based on transparent criteria.

(d) Propensity score applications require sufficient positivity, sufficient overlap and sufficient balance in the populations considered. If this cannot be achieved, adequate adjustment for confounding is not possible and the results from the corresponding analysis are unlikely to provide a meaningful estimate of the treatment effect.

(e) If a propensity score approach is applied, the final target population must be described in detail, especially if trimming or truncation methods are applied.

(f) Treatment effects estimated from non-randomised data are associated with additional uncertainty arising from a number of sources. Owing to the greater uncertainty, a large effect estimate is required. This can formally be evaluated by the testing of shifted hypotheses.

III Conclusion

This guideline presents methods that are used to combine evidence to determine the relative clinical effectiveness of treatments. The guideline directs assessors towards the pathway that will ideally provide the best estimate of relative effectiveness with the least uncertainty. The most robust evidence comes from adequate RCTs with low RoB.

Pairwise meta-analyses combine data for which the treatment and comparator are the same in all the trials included. Both frequentist and Bayesian frameworks offer approaches that are suitable provided the underlying assumptions are adequately met. When direct evidence from RCTs is not available or more than two treatments are of interest and it is necessary to perform indirect treatment comparisons, uncertainty for the treatment-effect estimate increases. Indirect comparisons are commonly used in comparative effectiveness analyses for which there is a lack of trials or evidence gathered for all the comparators of interest. When conducting such analyses, the appropriate method to use is one that preserves randomisation (i.e., an anchored indirect comparison). NMA can combine both direct and indirect evidence within the network. Various frequentist and Bayesian methods have been proposed for this purpose. The more evidence that is included in a network for one treatment, the more precise the estimates may be, but as complexity increases so does the potential for violation of the assumptions, which can influence the certainty of results. Whatever the method used, prerequisites and assumptions related to that method must be clearly specified and justified.

If IPD are available, further analyses can be undertaken. Approaches that account for differences in population characteristics between studies are available. MAIC reweights the outcomes to an alternative population, which may differ from the target population of the assessment. Furthermore, this method relies on accounting for all relevant effect modifiers in the model, which is difficult to ensure, and therefore, in general not an adequate analysis. The same applies to STC, which is a regression-based approach that extrapolates the outcomes to an alternative population. ML-NMR is a recent extension of regression-based approaches that combines IPD evidence with aggregated evidence.

A number of statistical approaches has been proposed for dealing with cases in which non-randomised evidence (e.g., single-arm trials, comparative observational studies and registry data) is used to inform an estimate of relative effectiveness. Although it is possible to provide summaries of evidence syntheses generated in this way, the certainty of the results provided by these techniques remains controversial. Results from such analyses are more likely to suffer from bias and are more likely to underestimate the true uncertainty and to depend on untested assumptions in comparison to syntheses with RCT evidence alone. Therefore, any analyses extending the network using single-arm or non-randomised evidence should include sensitivity analyses and an examination of the assumptions and should provide appropriate caveats for users. The use of single-arm or non-randomised evidence usually threatens the internal validity of results. Therefore, it is incumbent on the assessor to judge whether this evidence is sufficient for adequate estimation of the relative treatment effectiveness. For some interventions, single-arm or non- randomised evidence may be the only evidence available for consideration. However, it may well be that this evidence is insufficient for estimation of the relative treatment effectiveness in the context of JCA.

In many cases the conditions will not be ideal for the use of any of the methods presented in this guideline to produce unbiased estimates of relative effectiveness. Therefore, there should be very careful consideration of the underlying assumptions when making inferences. Input from a statistician with specific expertise in this area should be sought for a critical assessment of the methodological approach used, any assumptions potentially violated and the corresponding uncertainty of the results.

References

[1] Agoritsas T, Merglen A, Shah ND et al. Adjusted analyses in studies addressing therapy and harm: Users' Guides to the Medical Literature. JAMA 2017; 317(7): 748-759. https://dx.doi.org/10.1001/jama.2016.20029.

[2] Anzures-Cabrera J, Higgins JP. Graphical displays for meta-analysis: An overview with suggestions for practice. Res Synth Methods 2010; 1(1): 66-80. https://dx.doi.org/10.1002/jrsm.6.

[3] Bender R, Friede T, Koch A et al. Methods for evidence synthesis in the case of very few studies. Res Synth Methods 2018; 9(3): 382-392. https://dx.doi.org/10.1002/jrsm.1297.

[4] Berlin JA, Santanna J, Schmid CH et al. Individual patient- versus group-level data meta-regressions for the investigation of treatment effect modifiers: Ecological bias rears its ugly head. Stat Med 2002; 21(3): 371-387. https://dx.doi.org/10.1002/sim.1023.

[5] Böhning D, Mylona K, Kimber A. Meta-analysis of clinical trials with rare events. Biom J 2015; 57(4): 633-648. https://dx.doi.org/10.1002/bimj.201400184.

[6] Borenstein M, Hedges LV, Higgins JP et al. A basic introduction to fixed-effect and random-effects models for meta-analysis. Res Synth Methods 2010; 1(2): 97-111. https://dx.doi.org/10.1002/jrsm.12.

[7] Borenstein M, Hedges LV, Higgins JPT et al. Introduction to Meta-Analysis. Chichester, UK: Wiley; 2009.

[8] Bradburn MJ, Deeks JJ, Berlin JA et al. Much ado about nothing: A comparison of the performance of meta-analytical methods with rare events. Stat Med 2007; 26(1): 53-77. https://dx.doi.org/10.1002/sim.2528.

[9] Brockhaus AC, Grouven U, Bender R. Performance of the Peto odds ratio compared to the usual odds ratio estimator in the case of rare events. Biom J 2016; 58(6): 1428-1444. https://dx.doi.org/10.1002/bimj.201600034.

[10] Bucher HC, Guyatt GH, Griffith LE et al. The results of direct and indirect treatment comparisons in meta-analysis of randomized controlled trials. J Clin Epidemiol 1997; 50(6): 683-691. https://dx.doi.org/10.1016/s0895-4356(97)00049-8.

[11] Caro JJ, Ishak KJ. No head-to-head trial? Simulate the missing arms. Pharmacoeconomics 2010; 28(10): 957-967. https://dx.doi.org/10.2165/11537420-000000000-00000.

[12] Collins R, Bowman L, Landray M et al. The magic of randomization versus the myth of real-world evidence. N Engl J Med 2020; 382(7): 674-678. https://dx.doi.org/10.1056/NEJMsb1901642.

[13] Cooper NJ, Sutton AJ, Lu G et al. Mixed comparison of stroke prevention treatments in individuals with nonrheumatic atrial fibrillation. Arch Intern Med 2006; 166(12): 1269-1275. https://dx.doi.org/10.1001/archinte.166.12.1269.

[14] Cornell JE, Mulrow CD, Localio R et al. Random-effects meta-analysis of inconsistent effects: A time for change. Ann Intern Med 2014; 160(4): 267-270. https://dx.doi.org/10.7326/M13-2886.

[15] DerSimonian R, Laird N. Meta-analysis in clinical trials. Control Clin Trials 1986; 7(3): 177-188. https://dx.doi.org/10.1016/0197-2456(86)90046-2.

[16] Desai RJ, Franklin JM. Alternative approaches for confounding adjustment in observational studies using weighting based on the propensity score: A primer for practitioners. BMJ 2019; 367: l5657. https://dx.doi.org/10.1136/bmj.l5657.

[17] Di Pietrantonj C. Four-fold table cell frequencies imputation in meta analysis. Stat Med 2006; 25(13): 2299-2322. https://dx.doi.org/10.1002/sim.2287.

[18] Dias S, Ades A, Welton N et al. Network Meta-Analysis for Decision Making. Chichester, UK: Wiley; 2018.

[19] Dias S, Ades AE. Absolute or relative effects? Arm-based synthesis of trial data. Res Synth Methods 2016; 7(1): 23-28. https://dx.doi.org/10.1002/jrsm.1184.

[20] Dias S, Welton NJ, Caldwell DM et al. Checking consistency in mixed treatment comparison meta-analysis. Stat Med 2010; 29(7-8): 932-944. https://dx.doi.org/10.1002/sim.3767.

[21] Dias S, Welton NJ, Sutton AJ et al. NICE DSU Technical Support Document 1: Introduction to Evidence Synthesis for Decision Making [online]. 2011 [Accessed: 18.04.2016]. URL: http://www.nicedsu.org.uk/TSD1%20Introduction.final.08.05.12.pdf.

[22] Donegan S, Williamson P, Gamble C et al. Indirect comparisons: A review of reporting and methodological quality. PLoS One 2010; 5(11): e11054. https://dx.doi.org/10.1371/journal.pone.0011054.

[23] Edwards SJ, Clarke MJ, Wordsworth S et al. Indirect comparisons of treatments based on systematic reviews of randomised controlled trials. Int J Clin Pract 2009; 63(6): 841-854. https://dx.doi.org/10.1111/j.1742-1241.2009.02072.x.

[24] Egger M, Davey Smith G, Altman DG. Systematic Reviews in Health Care: Meta-Analysis in Context. London: BMJ Books; 2009.

[25] Egger M, Davey Smith G, Schneider M et al. Bias in meta-analysis detected by a simple, graphical test. BMJ 1997; 315(7109): 629-634. https://dx.doi.org/10.1136/bmj.315.7109.629.

[26] Egger M, Smith GD, Phillips AN. Meta-analysis: Principles and procedures. BMJ 1997; 315(7121): 1533-1537. https://dx.doi.org/10.1136/bmj.315.7121.1533.

[27] European Parliament and the Council of the European Union. REGULATION (EU) 2021/2282 OF THE EUROPEAN PARLIAMENT AND OF THE COUNCIL of 15 December 2021 on health technology assessment and amending. Directive 2011/24/EU. Off J EU 2021; L 458/1.

[28] Freeman SC, Cooper NJ, Sutton AJ et al. Challenges of modelling approaches for network meta-analysis of time-to-event outcomes in the presence of non-proportional hazards to aid decision making: Application to a melanoma network. Stat Methods Med Res 2022; 31(5): 839-861. https://dx.doi.org/10.1177/09622802211070253.

[29] Gartlehner G, Moore CG. Direct versus indirect comparisons: A summary of the evidence. Int J Technol Assess Health Care 2008; 24(2): 170-177. https://dx.doi.org/10.1017/S0266462308080240.

[30] Gelman A. Prior distributions for variance parameters in hierarchical models (Comment on Article by Browne and Draper). Bayesian Analysis 2006; 1(3): 515-534. https://dx.doi.org/10.1214/06-BA117A.

[31] Glynn RJ, Lunt M, Rothman KJ et al. Comparison of alternative approaches to trim subjects in the tails of the propensity score distribution. Pharmacoepidemiol Drug Saf 2019; 28(10): 1290-1298. https://dx.doi.org/10.1002/pds.4846.

[32] Guyot P, Ades AE, Ouwens MJ et al. Enhanced secondary analysis of survival data: reconstructing the data from published Kaplan-Meier survival curves. BMC Med Res Methodol 2012; 12: 9. https://dx.doi.org/10.1186/1471-2288-12-9.

[33] Hatswell AJ, Sullivan WG. Creating historical controls using data from a previous line of treatment – Two non-standard approaches. Stat Methods Med Res 2020; 29(6): 1563-1572. https://dx.doi.org/10.1177/0962280219826609.

[34] Heinecke A, Tallarita M, De Iorio M. Bayesian splines versus fractional polynomials in network meta-analysis. BMC Med Res Methodol 2020; 20(1): 261. https://dx.doi.org/10.1186/s12874-020-01113-9.

[35] Hernán MA, Robins JM. Using big data to emulate a target trial when a randomized trial is not available. Am J Epidemiol 2016; 183(8): 758-764. https://dx.doi.org/10.1093/aje/kwv254.

[36] Higgins JPT, Jackson D, Barrett JK et al. Consistency and inconsistency in network meta-analysis: Concepts and models for multi-arm studies. Res Synth Methods 2012; 3(2): 98-110. https://dx.doi.org/10.1002/jrsm.1044.

[37] Higgins JPT, Thomas J, Chandler J et al. Cochrane Handbook for Systematic Reviews of Interventions, 2nd Edition. Hoboken, NJ: Wiley; 2019.

[38] Higgins JPT, Thompson SG. Controlling the risk of spurious findings from meta-regression. Stat Med 2004; 23(11): 1663-1682. https://dx.doi.org/10.1002/sim.1187.

[39] Higgins JPT, Thompson SG. Quantifying heterogeneity in a meta-analysis. Stat Med 2002; 21(11): 1539-1558. https://dx.doi.org/10.1002/sim.1186.

[40] Hoaglin DC, Hawkins N, Jansen JP et al. Conducting indirect-treatment-comparison and network-meta-analysis studies: Report of the ISPOR Task Force on Indirect Treatment Comparisons Good Research Practices – Part 2. Value Health 2011; 14(4): 429-437. https://dx.doi.org/10.1016/j.jval.2011.01.011.

[41] Hong H, Chu H, Zhang J et al. A Bayesian missing data framework for generalized multiple outcome mixed treatment comparisons. Res Synth Methods 2016; 7(1): 6-22. https://dx.doi.org/10.1002/jrsm.1153.

[42] IQWiG. Concepts for the generation of routine practice data and their analysis for the benefit assessment of drugs according to §35a Social Code Book V (SGB V), Version 1.0 [online]. 2020 [Accessed: 23.06.2022]. URL: https://www.iqwig.de/download/a19-43_routine-practice-data-for-the-benefit-assessment-of-drugs_rapid-report_v1-0.pdf.

[43] IQWiG. General Methods, Version 6.1 [online]. 2022 [Accessed: 22.04.2022]. URL: https://www.iqwig.de/en/about-us/methods/methods-paper/.

[44] Ishak KJ, Proskorovsky I, Benedict A. Simulation and matching-based approaches for indirect comparison of treatments. Pharmacoeconomics 2015; 33(6): 537-549. https://dx.doi.org/10.1007/s40273-015-0271-1.

[45] Jansen JP. Network meta-analysis of survival data with fractional polynomials. BMC Med Res Methodol 2011; 11: 61. https://dx.doi.org/10.1186/1471-2288-11-61.

[46] Kiefer C, Sturtz S, Bender R. Indirect comparisons and network meta-analyses. Dtsch Arztebl Int 2015; 112(47): 803-808. https://dx.doi.org/10.3238/arztebl.2015.0803.

[47] Kiefer C, Sturtz S, Bender R. A simulation study to compare different estimation approaches for network meta-analysis and corresponding methods to evaluate the consistency assumption. BMC Med Res Methodol 2020; 20(1): 36. https://dx.doi.org/10.1186/s12874-020-0917-3.

[48] Kossmeier M, Tran US, Voracek M. Charting the landscape of graphical displays for meta-analysis and systematic reviews: A comprehensive review, taxonomy, and feature analysis. BMC Med Res Methodol 2020; 20(1): 26. https://dx.doi.org/10.1186/s12874-020-0911-9.

[49] Kuss O. Statistical methods for meta-analyses including information from studies without any events – add nothing to nothing and succeed nevertheless. Stat Med 2015; 34(7): 1097-1116. https://dx.doi.org/10.1002/sim.6383.

[50] Leahy J, Thom H, Jansen JP et al. Incorporating single-arm evidence into a network meta-analysis using aggregate level matching: Assessing the impact. Stat Med 2019; 38(14): 2505-2523. https://dx.doi.org/10.1002/sim.8139.

[51] Leahy J, Walsh C. Assessing the impact of a matching-adjusted indirect comparison in a Bayesian network meta-analysis. Res Synth Methods 2019; 10(4): 546-568. https://dx.doi.org/10.1002/jrsm.1372.

[52] Lu G, Ades A. Modeling between-trial variance structure in mixed treatment comparisons. Biostatistics 2009; 10(4): 792-805. https://dx.doi.org/10.1093/biostatistics/kxp032.

[53] Lu G, Ades AE. Combination of direct and indirect evidence in mixed treatment comparisons. Stat Med 2004; 23(20): 3105-3124. https://dx.doi.org/10.1002/sim.1875.

[54] Lumley T. Network meta-analysis for indirect treatment comparisons. Stat Med 2002; 21(16): 2313-2324. https://dx.doi.org/10.1002/sim.1201.

[55] Madan J, Stevenson MD, Cooper KL et al. Consistency between direct and indirect trial evidence: Is direct evidence always more reliable? Value Health 2011; 14(6): 953-960. https://dx.doi.org/10.1016/j.jval.2011.05.042.

[56] Mathes T, Kuss O. A comparison of methods for meta-analysis of a small number of studies with binary outcomes. Res Synth Methods 2018; 9(3): 366-381. https://dx.doi.org/10.1002/jrsm.1296.

[57] Ouwens MJ, Philips Z, Jansen JP. Network meta-analysis of parametric survival curves. Res Synth Methods 2010; 1(3-4): 258-271. https://dx.doi.org/10.1002/jrsm.25.

[58] Page MJ, McKenzie JE, Bossuyt PM et al. The PRISMA 2020 statement: an updated guideline for reporting systematic reviews. BMJ 2021; 372: n71. https://dx.doi.org/10.1136/bmj.n71.

[59] Page MJ, Moher D, Bossuyt PM et al. PRISMA 2020 explanation and elaboration: Updated guidance and exemplars for reporting systematic reviews. BMJ 2021; 372: n160. https://dx.doi.org/10.1136/bmj.n160.

[60] Peters JL, Sutton AJ, Jones DR et al. Comparison of two methods to detect publication bias in meta-analysis. JAMA 2006; 295(6): 676-680. https://dx.doi.org/10.1001/jama.295.6.676.

[61] Petto H, Kadziola Z, Brnabic A et al. Alternative weighting approaches for anchored matching-adjusted indirect comparisons via a common comparator. Value Health 2019; 22(1): 85-91. https://dx.doi.org/10.1016/j.jval.2018.06.018.

[62] Phillippo DM, Ades AE, Dias S et al. Methods for population-adjusted indirect comparisons in health technology appraisal. Med Decis Making 2018; 38(2): 200-211. https://dx.doi.org/10.1177/0272989X17725740.

[63] Phillippo DM, Ades AE, Dias S et al. NICE DSU Technical Support Document 18: Methods for population-adjusted indirect comparisons in submission to NICE [online]. 2016. URL: http://www.nicedsu.org.uk.

[64] Phillippo DM, Dias S, Ades AE et al. Multilevel network meta-regression for population-adjusted treatment comparisons. J R Stat Soc Ser A Stat Soc 2020; 183(3): 1189-1210. https://dx.doi.org/10.1111/rssa.12579.

[65] Phillippo DM, Dias S, Ades AE et al. Target estimands for efficient decision making: Response to comments on "Assessing the performance of population adjustment methods for anchored indirect comparisons: A simulation study". Stat Med 2021; 40(11): 2759-2763. https://dx.doi.org/10.1002/sim.8965.

[66] Platt RW, Leroux BG, Breslow N. Generalized linear mixed models for meta-analysis. Stat Med 1999; 18(6): 643-654. https://dx.doi.org/10.1002/(sici)1097-0258(19990330)18:6<643::aid-sim76>3.0.co;2-m.

[67] Remiro-Azócar A, Heath A, Baio G. Methods for population adjustment with limited access to individual patient data: A review and simulation study. Res Synth Methods 2021; 12(6): 750-775. https://dx.doi.org/10.1002/jrsm.1511.

[68] Remiro-Azócar A, Heath A, Baio G. Parametric G-computation for compatible indirect treatment comparisons with limited individual patient data. Res Synth Methods 2022. https://dx.doi.org/10.1002/jrsm.1565.

[69] Riley RD, Lambert PC, Staessen JA et al. Meta-analysis of continuous outcomes combining individual patient data and aggregate data. Stat Med 2008; 27(11): 1870-1893. https://dx.doi.org/10.1002/sim.3165.

[70] Röver C, Bender R, Dias S et al. On weakly informative prior distributions for the heterogeneity parameter in Bayesian random-effects meta-analysis. Res Synth Methods 2021; 12(4): 448-474. https://dx.doi.org/10.1002/jrsm.1475.

[71] Rücker G. Network meta-analysis, electrical networks and graph theory. Res Synth Methods 2012; 3(4): 312-324. https://dx.doi.org/10.1002/jrsm.1058.

[72] Rücker G, Schwarzer G. Reduce dimension or reduce weights? Comparing two approaches to multi-arm studies in network meta-analysis. Stat Med 2014; 33(25): 4353-4369. https://dx.doi.org/10.1002/sim.6236.

[73] Rücker G, Schwarzer G, Carpenter JR et al. Undue reliance on I² in assessing heterogeneity may mislead. BMC Med Res Methodol 2008; 8: 79. https://dx.doi.org/10.1186/1471-2288-8-79.

[74] Salanti G, Higgins JP, Ades AE et al. Evaluation of networks of randomized trials. Stat Methods Med Res 2008; 17(3): 279-301. https://dx.doi.org/10.1177/0962280207080643.

[75] Schmitz S, Adams R, Walsh C. Incorporating data from various trial designs into a mixed treatment comparison model. Stat Med 2013; 32(17): 2935-2949. https://dx.doi.org/10.1002/sim.5764.

[76] Schulz A, Schürmann C, Skipka G et al. Performing meta-analyses with very few studies. In: Evangelou E, Veroniki AA (Ed). Meta-Research: Methods and Protocols. New York: Humana; 2022. S. 91-102.

[77] Schwarzer G, Carpenter J, Rücker G. Meta-Analysis with R. Basel: Springer; 2015.

[78] Shim SR, Kim SJ, Lee J et al. Network meta-analysis: Application and practice using R software. Epidemiol Health 2019; 41: e2019013. https://dx.doi.org/10.4178/epih.e2019013.

[79] Signorovitch JE, Sikirica V, Erder MH et al. Matching-adjusted indirect comparisons: A new tool for timely comparative effectiveness research. Value Health 2012; 15(6): 940-947. https://dx.doi.org/10.1016/j.jval.2012.05.004.

[80] Signorovitch JE, Wu EQ, Yu AP et al. Comparative effectiveness without head-to-head trials: A method for matching-adjusted indirect comparisons applied to psoriasis treatment with adalimumab or etanercept. Pharmacoeconomics 2010; 28(10): 935-945. https://dx.doi.org/10.2165/11538370-000000000-00000.

[81] Simmonds MC, Higgins JP. Covariate heterogeneity in meta-analysis: Criteria for deciding between meta-regression and individual patient data. Stat Med 2007; 26(15): 2982-2999. https://dx.doi.org/10.1002/sim.2768.

[82] Simmonds MC, Higgins JP. A general framework for the use of logistic regression models in meta-analysis. Stat Methods Med Res 2016; 25(6): 2858-2877. https://dx.doi.org/10.1177/0962280214534409.

[83] Song F, Loke YK, Walsh T et al. Methodological problems in the use of indirect comparisons for evaluating healthcare interventions: Survey of published systematic reviews. BMJ 2009; 338: b1147. https://dx.doi.org/10.1136/bmj.b1147.

[84] Sterne JA, Hernán MA, Reeves BC et al. ROBINS-I: A tool for assessing risk of bias in non-randomised studies of interventions. BMJ 2016; 355: i4919. https://dx.doi.org/10.1136/bmj.i4919.

[85] Sterne JAC, Savovic J, Page MJ et al. RoB 2: A revised tool for assessing risk of bias in randomised trials. BMJ 2019; 366: l4898. https://dx.doi.org/10.1136/bmj.l4898.

[86] Stijnen T, Hamza TH, Ozdemir P. Random effects meta-analysis of event outcome in the framework of the generalized linear mixed model with applications in sparse data. Stat Med 2010; 29(29): 3046-3067. https://dx.doi.org/10.1002/sim.4040.

[87] Sutton AJ, Abrams KR. Bayesian methods in meta-analysis and evidence synthesis. Stat Methods Med Res 2001; 10(4): 277-303. https://dx.doi.org/10.1177/096228020101000404.

[88] Sutton AJ, Abrams KR, Jones DR et al. Methods for Meta-Analysis in Medical Research. Chichester, UK: Wiley; 2000.

[89] Sutton AJ, Ades AE, Cooper N et al. Use of indirect and mixed treatment comparisons for technology assessment. Pharmacoeconomics 2008; 26(9): 753-767. https://dx.doi.org/10.2165/00019053-200826090-00006.

[90] Sweeting MJ, Sutton AJ, Lambert PC. What to add to nothing? Use and avoidance of continuity corrections in meta-analysis of sparse data. Stat Med 2004; 23(9): 1351-1375. https://dx.doi.org/10.1002/sim.1761.

[91] Thom H, Leahy J, Jansen JP. Network meta-analysis on disconnected evidence networks when only aggregate data are available: Modified methods to include disconnected trials and single-arm studies while minimizing bias. Med Decis Making 2022. https://dx.doi.org/10.1177/0272989X221097081.

[92] Thompson SG, Higgins JPT. How should meta-regression analyses be undertaken and interpreted? Stat Med 2002; 21(11): 1559-1573. https://dx.doi.org/10.1002/sim.1752.

[93] Vandermeer BW, Buscemi N, Liang Y et al. Comparison of meta-analytic results of indirect, direct, and combined comparisons of drugs for chronic insomnia in adults: A case study. Med Care 2007; 45(10 Supl 2): S166-172. https://dx.doi.org/10.1097/MLR.0b013e3180546867.

[94] VanderWeele TJ, Ding P. Sensitivity Analysis in Observational Research: Introducing the E-Value. Ann Intern Med 2017; 167(4): 268-274. https://dx.doi.org/10.7326/M16-2607.

[95] Veroniki AA, Jackson D, Bender R et al. Methods to calculate uncertainty in the estimated overall effect size from a random-effects meta-analysis. Res Synth Methods 2019; 10(1): 23-43. https://dx.doi.org/10.1002/jrsm.1319.

[96] Veroniki AA, Mavridis D, Higgins JP et al. Characteristics of a loop of evidence that affect detection and estimation of inconsistency: A simulation study. BMC Med Res Methodol 2014; 14: 106. https://dx.doi.org/10.1186/1471-2288-14-106.

[97] Webster-Clark M, Stürmer T, Wang T. Using propensity scores to estimate effects of treatment initiation decisions: State of the science. Stat Med 2021; 40(7): 1718-1735. https://dx.doi.org/10.1002/sim.8866.

[98] Wells GA, Sultan SA, Chen L et al. Indirect Evidence: Indirect Treatment Comparisons in Meta-Analysis. Ottawa, Canada: Canadian Agency for Drugs and Technologies in Health; 2009.

[99] Welton N, Sutton A, Cooper N et al. Evidence Synthesis for Decision Making in Healthcare. Chichester, UK: Wiley; 2012.

[100] White IR. Network meta-analysis. Stata Journal 2015; 15(4): 951-985. https://dx.doi.org/10.1177/1536867X1501500403.

[101] White IR, Turner RM, Karahalios A et al. A comparison of arm-based and contrast-based models for network meta-analysis. Stat Med 2019; 38(27): 5197-5213. https://dx.doi.org/10.1002/sim.8360.

[102] Wiksten A, Hawkins N, Piepho HP et al. Nonproportional hazards in network meta-analysis: Efficient strategies for model building and analysis. Value Health 2020; 23(7): 918-927. https://dx.doi.org/10.1016/j.jval.2020.03.010.

[103] Wiksten A, Rücker G, Schwarzer G. Hartung-Knapp method is not always conservative compared with fixed-effect meta-analysis. Stat Med 2016; 35(15): 2503-2515. https://dx.doi.org/10.1002/sim.6879.

[104] Williamson E, Morley R, Lucas A et al. Propensity scores: From naive enthusiasm to intuitive understanding. Stat Methods Med Res 2012; 21(3): 273-293. https://dx.doi.org/10.1177/0962280210394483.

[105] Zhang X, Faries DE, Li H et al. Addressing unmeasured confounding in comparative observational research. Pharmacoepidemiol Drug Saf 2018; 27(4): 373-382. https://dx.doi.org/10.1002/pds.4394.