eunethta

EUROPEAN NETWORK FOR HEALTH TECHNOLOGY ASSESSMENT

EUnetHTA21 - Individual Practical Guideline Document

D4.3.1: DIRECT AND INDIRECT COMPARISONS

Version 0.2, 26.07.2022

Template version 04 February 2022

Document history and contributors

Version	Date	Description
V0.1	27/04/2022	First draft
V0.2	26/07/2022	Second draft

Disclaimer

This Guideline Document was produced under the Third EU Health Programme through a service contract with the European Health and Digital Executive Agency (HaDEA) acting under the mandate from the European Commission. The information and views set out in this Project Plan are those of the author(s) and do not necessarily reflect the official opinion of the Commission/Executive Agency. The Commission/Executive Agency do not guarantee the accuracy of the data included in this study. Neither the Commission/Executive Agency nor any person acting on the Commission's/Executive Agency's behalf may be held responsible for the use which may be made of the information contained therein.

Participants

Hands-on group	Gemeinsamer Bundesausschuss (G-BA), Germany Haute Autorité de Santé (HAS), France Institut für Qualität und Wirtschaftlichkeit im Gesundheitswesen (IQWIG), Germany National Centre for Pharmacoeconomics, St James Hospital (NCPE), Ireland Norwegian Medicines Agency (NOMA), Norway
Project Management	ZIN, The Netherlands
CSCQ CEB	Agencia Española de Medicamentos y Productos Sanitarios [AEMPS), Spain Austrian Institute for Health Technology Assessment (AIHTA), Austria Belgian Health Care Knowledge Centre (KCE), Belgium G-BA, Germany HAS, France Institut für Qualität und Wirtschaftlichkeit im Gesundheitswesen (IQWIG), Germany Italian Medicines Agency (AIFA), Italy National Authority of Medicines and Health Products, I.P. (INFARMED), Portugal NCPE, Ireland National Institute of Pharmacy and Nutrition (NIPN), Hungary NOMA, Norway The Dental and Pharmaceutical Benefits Agency (TLV), Sweden ZIN, The Netherlands

The work in European Network for Health Technology Assessment (EUnetHTA) 21 is a collaborative effort. While the agencies in the Hands-on Group will be actively writing the deliverable, the entire EUnetHTA 21 consortium is involved in its production throughout various stages. This means that the Committee for Scientific Consistency and Quality (CSCQ) will review and discuss several drafts of the deliverable prior to validation. Afterwards the Consortium Executive Board (CEB) will endorse the final deliverable prior to publication.

Conflict of interest

The authors declare that there is no conflict of interest.

Copyright

TABLE OF CONTENTS

1 INTRODUCTION 5
1.1 Definitions 5
1.2 Relevant articles in Regulation (EU) 2021/2282 5
2 SCOPE AND OBJECTIVE OF THE GUIDELINE 6
3 GENERAL CONSIDERATIONS 7
3.1 Initial feasibility questions 7
3.2 Assessment of exchangeability 7
3.2.1 Assessment of similarity 7
3.2.2 Assessment of homogeneity 9
3.2.3 Assessment of consistency 10
3.3 Possible approaches when the assumptions are violated 12
4 METHODS APPLICABLE TO DIRECT OR INDIRECT COMPARISONS 14
4.1 Methods for direct comparisons 14
4.1.1 Standard approaches 14
4.1.2 Application of the Knapp–Hartung method 14
4.1.3 Direct comparisons with very few studies 14
4.2 Indirect comparisons 15
4.3 Evidence synthesis of time-to-event data 17
4.3.1 Assessment of the proportional hazards assumption 17
4.3.2 (Network) meta-analysis of restricted mean survival time 17
4.3.3 (Network) meta-analysis with flexible survival time models 18
5 Assessment of population-adjusted methods 20
5.1 General considerations: is population adjustment for indirect comparisons appropriate? 20
5.2 Assessing covariate selection (all population-adjusted methods) 20
5.3 Additional considerations for outcome regression approaches 21
5.4 Additional considerations for matching-adjusted indirect comparisons 23
5.5 Dealing with unanchored MAICs and STCs: additional challenges 24
5.6 Interpretation and use of population-adjusted results 24
6 ASSESSMENT OF COMPARISONS BASED UPON NON-RANDOMISED EVIDENCE 26
6.1 General considerations 26
6.2 Propensity scores 26
6.2.1 Checking that the assumptions of propensity score matching and/or weighting are valid 26
6.2.2 Interpreting results of propensity score 27
7 FURTHER RELEVANT DOCUMENTS (UNDER DEVELOPMENT) 29
8 FUTURE RECOMMENDATION 29
9 References 30

61 # LIST OF ABBREVIATIONS

Abbreviation	Meaning
AgD	Aggregate data
AIC	Akaike information criterion
ATE	Average treatment effect
ATT	Average treatment effect among treated
BIC	Bayesian information criterion
CA	Collaborative assessment
DIC	Deviance information criterion
DSL	DerSimonian-Laird
EMA	European Medicines Agency
EU	European Union
EUnetHTA	European Network for Health Technology Assessment
FP	Fractional polynomials
HR	Hazard ratio
HTD	Health technology developer
IPD	Individual patient-level data, also known as individual patient data or individual participant data
IQWiG	Institut für Qualität und Wirtschaftlichkeit im Gesundheitswesen
ITC	Indirect treatment comparison
JCA	Joint clinical assessment
KH	Knapp-Hartung
MAIC	Matching-adjusted indirect comparison
MD	Mean difference
ML-NMR	Multilevel network meta-regression
MS	Member State
NMA	Network meta-analysis
OR	Odds ratio
PH	Proportional hazards
PICO	Population, intervention, comparator, outcome
RCT	Randomised controlled trial
RMST	Restricted mean survival time
RR	Risk ratio
RoB	Risk of bias
ROBINS-I	Risk of bias in non-randomised studies - of interventions
SAP	Statistical analysis plan
STC	Simulated treatment comparison
SUCRA	Surface under the cumulative ranking curve

62
63

1 INTRODUCTION

The European Network for Health Technology Assessment (EUnetHTA) 21 Methodological Guideline D4.3.2 *Direct and Indirect Comparisons* is an update of a previous EUnetHTA guideline (2013, updated in 2015); it describes the currently available method for direct and indirect comparisons regarding their underlying assumptions, strengths, and weaknesses, and specifies the appropriateness of methods to the data situation. This Practical Guideline is intended for assessor/co-assessors and gives additional, more-detailed advice for use in practice. As indicated in the criteria for selection of assessors and co-assessors, it is expected that statistical expertise will be available in the assessment team (see the EUnetHTA 21 Guideline D5.3.1 *Procedural Guideline for Appointing Assessors and Co-Assessors for JCA/CA)*.

1.1 Definitions

The terms used in this document might be used with a slightly different meaning in other contexts. Below, we define the terms as they are used in this guideline.

Direct comparison: comparison of treatments either by means of a single comparative study or a pairwise meta-analysis or other method for synthesis of comparative studies without indirect comparisons;

Effectiveness: describes how well a treatment works in practice; includes efficacy and safety;

Exchangeability: if patients from one treatment group are substituted to another, the same treatment effect is expected; contains the components similarity, homogeneity, and, in the case of indirect comparisons, consistency;

Health technology: treatment, intervention, and health technology are terms for any health technology that can be assessed;

Indirect comparison: evidence synthesis in which inference about the relative effectiveness of two treatments is made without the use of trials comparing both treatments head-to-head; indirect comparisons are also made when more general methods of network meta-analysis are applied, even when direct evidence for the comparison of interest is available;

Network meta-analysis (NMA): generalisation of meta-analysis to include more-complex evidence networks, which can include both direct evidence and indirect evidence; NMA incorporates other terms used in the literature to describe the synthesis of both direct and indirect evidence, such as mixed treatment comparisons and indirect treatment comparisons;

Population-adjusted method for indirect comparisons: method for indirect comparisons with adjustment for imbalances in effect modifiers between studies and additionally, in the case of disconnected networks, for imbalances in prognostic variables between studies.

1.2 Relevant articles in Regulation (EU) 2021/2282

Articles from Regulation (EU) 2021/2282 directly relevant to the content of this practical guideline are:

- Article 9: Joint Clinical Assessment (JCA) reports and the dossier of the Health Technology Developer (HTD);
- Article 18: Preparation of the joint scientific consultations outcome document.

2 SCOPE AND OBJECTIVE OF THE GUIDELINE

This Practical Guideline describes how to deal in practice with evidence syntheses in JCA reports and provides guidance for assessors and co-assessors dealing with submitted results of direct and indirect treatment comparisons from HTDs. Each Section of this Guideline contains a list of requirements that should be reported in the JCA reports in cases in which a direct or indirect treatment comparison was submitted. It is not the objective of this Guideline to make explicit recommendations about whether a submitted direct and indirect treatment comparison should be accepted by the Member States (MSs). Each MS should be enabled to decide on the validity of direct or indirect treatment comparisons itself based on the JCA report, which should include all methodological details needed to do so.

In the EUnetHTA 21 Methodological Guideline *Direct and Indirect Comparisons*, the methods for evidence syntheses are summarised, and general guidance is provided on which method(s) are appropriate in a particular situation. This Practical Guideline gives more practical advice for assessors/co-assessors within the framework described in the Methodological Guideline. Often, when using evidence synthesis methodology, some assumptions will be made, which might affect the certainty of results. The aim of this Guideline is to enable HTA assessors and developers to identify potential issues and reduce bias and uncertainty as much as possible. However, we recognise that there is an element of subjectivity in the assessment of many assumptions and that decisions might vary between MSs. There might be exceptional circumstances in which methods of evidence synthesis will need to be applied despite uncertainty or doubt as to their validity. We believe that these should be kept to a minimum and only used in circumstances in which there is a lack of other options to produce an estimate of relative treatment effect. In these scenarios, the HTD has to provide convincing evidence to support the claim that the corresponding results still produce meaningful estimates of relative treatment effectiveness, and acknowledge any additional sources of bias and uncertainty. The argument that the required data to apply preferable methods are not available, is insufficient to demonstrate the validity of the results coming from less appropriate methods with high uncertainty.

3 GENERAL CONSIDERATIONS

JCAs can use results from multiple trials, which are combined through evidence synthesis. A rigorous systematic review of the relevant literature with explicit inclusion and exclusion criteria is a prerequisite before conducting any evidence synthesis. Evidence synthesis can allow researchers to obtain a more-robust estimate of the treatment effect and, in the case of indirect treatment comparisons (ITCs), provide relative treatment effects for interventions that have not been studied in the same trial. However, it is important that the selection of trials and modelling choices is made with caution and is rigorously examined by assessors in collaboration with healthcare professionals and statisticians. Importantly, according to the European Union (EU) regulation, assessors must ensure that estimates are obtained by pooling relative treatment effects from each trial (i.e., compared with an appropriate comparator) and no inference is based on pooling the absolute effect of a particular treatment in a trial (i.e., regarding the mean outcome in one group only). The rest of this Section details how to assess whether trials are sufficiently similar to be combined, the main modelling choices to consider and scrutinise, and the inferences that can or cannot be made based on the methods and data used.

3.1 Initial feasibility questions

For direct and indirect comparisons by means of evidence syntheses, the aspects of the population, intervention, control, outcome (PICO) framework and the study design of the included studies have to be examined. Depending on the research goal, the patient population of interest, the intervention, and the control are prespecified and studies have accordingly been searched and selected. Only studies relevant for the given research question and fitting to the PICO scheme should be included in the evidence synthesis. Here, we assume that all studies included in a considered evidence synthesis are relevant for the research question and the corresponding PICO.

However, patient characteristics, such as distributions of age, sex, disease duration, measurement, and operationalisation of the outcome of interest, and features of the experimental design still need to be assessed in detail. Additional aspects, such as year and region of study conduct or forms of treatment application, also have to be assessed if they potentially represent possible effect modifiers (see following Sections).

Requirements for reporting:

- A determination that the studies included in the evidence synthesis match the established PICO based on all information described above.

3.2 Assessment of exchangeability

The fundamental assumption of exchangeability, which is required for (network) meta-analysis, is operationalised by assessing the properties of similarity, homogeneity, and, in the case of indirect comparisons, consistency. We emphasise here that these three properties are not, strictly speaking, distinct assumptions, because a failure of homogeneity or consistency is often the result of an imbalance in effect modifiers between studies (i.e., a violation of similarity). However, in many cases, not all effect modifiers will be known or reported across all studies and, therefore, assessment of homogeneity and consistency (if relevant) could detect an imbalance in unknown effect modifiers that would not be identified through assessment of similarity alone. In situations in which few studies are available for one or more pairwise comparisons, statistical tests might be underpowered to detect violations of homogeneity or consistency and, therefore, the assessment of exchangeability will depend entirely on the similarity of the included studies in terms of observed characteristics. Thus, assessors should be aware that such assessments cannot explore the potential impact of unknown effect modifiers.

3.2.1 Assessment of similarity

The similarity assumption states that all studies considered are comparable with respect to possible effect modifiers across all interventions. This is tested by means of the PICO scheme (see above). The PICO scheme chosen and the resulting inclusion and exclusion criteria apply to all studies included in

the evidence synthesis. For similarity, the following aspects should always be evaluated to identify possible effect modifiers [7].

1. **Study and patient characteristics** (including duration of follow-up): a list of potential effect modifiers should be drawn up *a priori*. The basis for this can be not only clinical considerations, but also findings from other studies on the therapeutic indication. The following characteristics are generally relevant: age, sex, disease severity, region, and study duration. Only those factors that are identified as potential effect modifiers should be included in this list. Effect modifiers can be identified through a literature search, input from healthcare professionals and, other methods;
2. **Characteristics of the intervention** (e.g., dosage or application, concomitant treatments): an intervention should be considered as a comparator even if it has not yet been granted European Medicines Authority (EMA) marketing authorisation.

The evaluation of similarity should also consider methodological factors that should not differ substantially between studies. Consideration of the observed values of relevant outcomes has also been shown to be helpful in evaluating similarity.

3. **Characteristics of outcomes** (e.g., definitions of outcomes): an *a priori* definition of what is considered sufficiently similar for each characteristic will usually be difficult. It will often also depend on what is present in the studies included;
4. **Observed values of relevant outcomes at baseline:** an examination of the observed values of relevant outcomes at baseline can provide information on the similarity of the individual studies, especially the study arms in which the comparator is used. However, to determine similarity, it is not a standard prerequisite that the observed values have to be identical, because the distribution of prognostic variables might well differ between studies. Nevertheless, extreme differences that even lead to floor or ceiling effects regarding the range of possible outcome values should not exist. If the corresponding information is not available at baseline, the values recorded during the course of the study or at the time of analysis can be used instead.

It is important to note the following issues regarding effect modification:

- Not all prognostic variables are effect modifiers;
- Effect modification is a property of the relative effect between a pair of treatments. As such, it is possible for a variable to modify the relative treatment effect of A versus B, but not the effect of treatment A versus treatment C. This could occur, for example, when A is placebo, B is a therapy targeting a particular genetic mutation [e.g., an epidermal growth factor receptor (EGFR) tyrosine kinase inhibitor (TKI)], and C is another active treatment that does not specifically target this mutation (e.g., chemotherapy): in this case, the presence of this genetic mutation in an individual could be an effect modifier for A versus B but not A versus C. More generally, a variable could be an effect modifier for both A versus B and A versus C, but the magnitude and even the direction of this effect could differ between comparisons (e.g., if patients responded less well to chemotherapy in the presence of this genetic mutation);
- The status of a variable as an effect modifier, and the magnitude and direction of this effect, is specific to the scale on which the treatment effect is measured. For example, in a hypothetical placebo-controlled study of an influenza vaccine, female participants experience a reduction in risk from 10% to 5% and male participants from 6% to 3%, with vaccination compared with placebo. On the relative risk scale, sex is not an effect modifier [relative risk (RR) = 0.5 in both groups], but it is on the risk-difference scale (–5% for females versus –3% for males).

It is essential that the process used to identify relevant effect modifiers is comprehensive and transparently reported. This process should include a comprehensive review of the literature and consultation of healthcare professionals with knowledge of the disease area. The set of all potentially relevant effect modifiers should be reported in the submission.

The assessment of similarity should include a quantitative analysis of the impact on all observed patient covariates. However, statistical tests for effect modification using subgroup data from clinical trials (e.g., testing for the significance of interaction terms) will often be underpowered and suffer from issues with multiplicity. Given that the risk of both type 1 and type 2 errors is typically high, statistical tests for effect modification should not be used in isolation to justify the selection of covariates as potential effect modifiers [19,21]. The assessor should also obtain opinions from healthcare professionals to assess whether there are missing effect modifiers.

After assessment of all these aspects, a decision has to be made about whether all studies considered in the evidence synthesis are comparable with respect to possible effect modifiers across all interventions (sufficient similarity) or not (insufficient similarity).

Requirements for reporting:

- Description of methodology used to identify potential effect modifiers and whether it sufficiently captures all possible effect modifiers;
- Assessment of the list of all potential effect modifiers identified and whether this list is likely to be complete; where possible, estimates of the magnitude and direction of the interaction effects;
- Description of any likely missing effect modifiers and the magnitude of their effect;
- The final conclusion about whether the assumption of sufficient similarity is expected to hold or not, with reasoning.

3.2.2 Assessment of homogeneity

The homogeneity assumption states that there is no meaningful heterogeneity between the effect estimates of the individual studies of each possible direct comparison. Even if studies are sufficiently similar, it is still possible that the data show meaningful heterogeneity. Heterogeneity can be caused by unknown effect modifiers and also by factors initially judged to be sufficiently similar or not judged to be potential effect modifiers. To test the homogeneity assumption for a pairwise comparison, at least two direct studies must be available for this comparison in principle, although typically at least five studies are required for a reliable assessment [9]. If only one study is available for each pairwise comparison, the homogeneity assumption cannot be tested. However, this does not prevent the performance of an indirect comparison. The heterogeneity between the studies has to be assessed to determine whether a pooling of the results is meaningful at all and to choose between the fixed-effect and random-effects approach for the evidence synthesis. It is important to use statistical methods as well as design features of the included studies to assess heterogeneity.

Two widely used statistical approaches to assess heterogeneity are given by the statistical test based on the Q statistic (Q-test) [8,53] and the heterogeneity measure I^2 [8,25]. As a rough guide for the interpretation of I^2, the following overlapping categories were proposed [9]:

- 0–40%: might not be important;
- 30–60%: might represent moderate heterogeneity;
- 50–90%: might represent substantial heterogeneity;
- 75–100%: Considerable heterogeneity.

However, the importance of observed I^2 values depends on the magnitude and direction of treatment effects and the strength of evidence for heterogeneity (p-value from the Q-test, uncertainty of the I^2, or number of studies) [9].

One easy and objective criterion to decide whether the studies should not be pooled is given by the statistical significance of the Q-test (p <0.05). However, the current data situation should always be considered when interpreting the results of the Q-test. On the one hand, the Q-test suffers from low power, especially in the situation of few studies [8], which means that a non-significant Q-test does not necessarily indicate that there is no relevant heterogeneity. On the other hand, in the case of a large number of studies, the Q-test might be statistically significant although only low heterogeneity is shown

in the forest plot. In such instances, the I^2 measure can help to describe the amount of heterogeneity. For example, if $I^2 < 50\%$, it might be useful to decide that there is no substantial heterogeneity even if the Q-test is statistically significant. In any case, a graphical inspection of the forest plot is advisable in addition to the use of the Q-test and the heterogeneity measure I^2 for the assessment of heterogeneity [8].

After assessing heterogeneity, it must be determined whether there is meaningful heterogeneity between the effect estimates of the individual studies of each possible direct comparison (insufficient homogeneity) or not (sufficient homogeneity). If it can be decided that there is sufficient homogeneity and it is meaningful to pool the included studies, it has to be determined whether a fixed-effect or a random-effects model should be used for the evidence synthesis. In the case of indirect comparisons, the assessment results of the consistency assumption also have to be considered (see below). A fixed-effect model assumes a common treatment effect in all studies, which is implausible in many situations and requires rigorous justification from the HTD. Therefore, the standard approach is to use the random-effects approach. However, if there is a marked consistency of the PICO and design properties of all studies, for instance in the case of evidence syntheses when only few studies are available, a fixed-effect model might be appropriate. An example of where the fixed-effect model can regularly be assumed to be valid is the situation of two studies with identical design, which can be found after drug approval. In general, however, the random-effects model is the appropriate choice, although its use might not be feasible in practice (see Section 4.1.3).

Requirements for reporting:

- The complete evaluation of whether the analyses provided to support the homogeneity assumption (including the forest plots, the *p*-values for the heterogeneity test, and the I^2 values) for all pairwise comparisons are sufficient to demonstrate that it is likely to hold;
- The final conclusion of whether the assumption of sufficient homogeneity holds or not with reasoning (including sensitivity analyses);
- The final conclusion of whether it is meaningful to pool the included studies, with reasoning;
- The decision of whether a fixed-effect or random-effects approach is adequate, with reasoning.

3.2.3 Assessment of consistency

General remarks

Under the consistency assumption, the same treatment effect is estimated through both the direct and indirect pathways for a particular contrast in the network. Inconsistency is a form of heterogeneity that is linked to the structure of the network but concerns the contrasts between treatments. Thus, inconsistency is between-trial variation comparing different treatment contrasts, and heterogeneity is between-trial variation within treatment contrasts.

The choice of methods to assess consistency depends on the network structure. Not all methods are suitable for networks of any complexity, but simpler methods are preferred where they are suitable. The methods used to test for inconsistency should be clearly identified and justification provided for this choice, with reference to the network structure.

Although any indirect comparison relies on the consistency assumption, it cannot be tested in networks without a loop structure. Therefore, the first step is to examine the network diagram for loops. It is also important to identify multi-arm trials, because these represent a loop that is consistent by definition.

In practice, NMAs often contain too few studies and sparse data to assess inconsistency adequately. Failure to detect inconsistency does not imply that the evidence is consistent. Statistical detection of inconsistency requires more data than are required to establish a treatment effect. Inconsistency can be caused by imbalance in the distributions of effect modifiers in the direct and indirect evidence, commonly factors such as age, severity, and line of treatment, which might be confounded with each other. Therefore, the check for inconsistency should be done alongside the assessment of similarity and heterogeneity in the NMA.

To minimise the risk of drawing incorrect conclusions, more empirical indicators are also suggested. Empirical assessment of heterogeneity and the between-trials variation in trial baseline can be used to assess the risk of inconsistency. Comparison of events and responses in the placebo arms might be useful in this context, although we emphasise that, although differences between placebo arms might indicate an imbalance in prognostic variables across studies, this need not result in inconsistency unless these variables are also effect modifiers.

Bucher method for single loops

The Bucher method is a two-stage method for testing consistency, in which the first step is to synthesise each pair-wise contrast and the second is to test whether the direct and indirect evidence are in conflict. The estimate of inconsistency comes from subtracting the direct and indirect estimates and referring the null hypothesis of no inconsistency to the normal distribution. This test can be applied on three independent sources of data but not on multi-arm trials, because the effect estimates in multi-arm trials are correlated.

The Bucher method can also be extended to networks with multiple loops calculating the statistic referring to a chi-square distribution. However, repeated use of the Bucher test in large complex networks with multiple loops can be problematic and, instead, an inconsistency model could be applied for assessing consistency in complex networks. Furthermore, the use of the Bucher method to test for consistency is not advisable when random-effects models are used to synthesise one or more of the pairwise comparisons [12].

Inconsistency models: Bayesian NMA

The principle of the inconsistency model is to assume no consistency, that all contrasts in the network are unrelated, and that the relative treatment effects are estimated directly from all contrasts (unrelated mean effect). In a consistency model, effects of all included treatments are estimated relative to the reference treatment. To test consistency, the deviance and deviance information criterion (DIC) statistics of the consistency and inconsistency models are compared. Plots of the posterior mean deviance of the individual data points in the inconsistency model against the corresponding posterior mean deviance in the consistency model can help identify loops in which inconsistency is present [16].

Further assessment of inconsistency will be a comparison of the posterior estimates of the treatment effect between the consistency and inconsistency models and assessing whether credible intervals overlap.

Node-splitting methods: Bayesian

The node-splitting method [15] can be applied to any contrast in any network of different complexities in which there is both direct and indirect evidence. In this method, the information contributing to the estimates of a parameter (a so-called ‘node’) is split into evidence that is direct only and indirect, which is based on the remaining evidence in the network meta-analysis. The indirect estimates in the node-splitting method use not only the indirect evidence of a specific loop, but also the whole evidence base in the network. Comparison of the residual deviance and DIC and the estimate of the heterogeneity parameter for random-effect models, of the full NMA, and of the model with a split node can then be used to assess potential inconsistency between the evidence for a particular node. Reduction of these parameters in the node-split model can be an indicator of inconsistency. The assumption that a split results in equal treatment effects for direct and indirect evidence can be tested in the same way as an inferential hypothesis. Therefore, *p*-values can be calculated to indicate that the hypothesis of equal treatment effects for direct and indirect evidence can be rejected. Although a smaller *p*-value would indicate inconsistency, interpretation of these *p*-values is context dependent and no formal framework for the required significance level exists.

Requirements for reporting:

- Methods used to test for inconsistency and justification for this choice with reference to the network structure; the report should highlight whether the methods used are likely to be appropriate;
- Criteria used to determine whether a meaningful violation of consistency has been detected;
- Summary of the results of statistical tests and/or models used to investigate consistency, stating whether these indicate the presence of inconsistency, and describing the extent of the inconsistency and resulting uncertainty in these results;
- In cases in which inconsistency is detected, description of the possible sources of inconsistency in terms of effect modifiers and, if possible, estimates of the magnitude of effect modification;
- If methods have been used to explore the qualitative aspects of node splits, resulting *p*-values should be reported as well as an explanation of the assumptions underlying the analysis;
- The final conclusion of whether the assumption of sufficient consistency holds, with reasoning.

3.3 Possible approaches when the assumptions are violated

If at least one of the components of the exchangeability assumption is not valid for a considered data situation, the following consequences are possible.

1. **Splitting into subgroups:** if dissimilarity is shown for a potential effect modifier or heterogeneity is shown that can be explained by the effect modifier, it might be useful to divide the entire study pool into several subpools and draw separate conclusions (e.g., for men and women). The limitations of subgroup analyses based upon aggregated data should be taken into account [14];
2. **Use of (network) meta-regression:** potential effect modifiers can be included as covariates in a (network) meta-regression model. This requires a sufficient number of data points (= number of studies) so that all parameters can be estimated in the model. The limitations and assumptions of meta-regression based upon aggregated data should be taken into account [4,6,24];
3. **Exclusion of studies:** in the case that only very few studies are responsible for dissimilarity or heterogeneity, whether it is useful to exclude these studies from the analysis should be discussed. However, in this case, sensitivity analyses should also be performed in which at least some of these studies are included (see below);
4. **Sensitivity analyses:** if a clearly useful procedure is not possible, sensitivity analyses at least should be performed that allow assessment of the impact of the violated assumptions (e.g., consideration of study results with unexplained heterogeneity in two separate study pools with homogeneous study results). Given that there are many areas of uncertainty regarding the 'right' methods for meta-analysis, sensitivity analyses are also an important aid in all further decisions in the process to estimate their impact on the results;
5. **Population-adjusted indirect comparisons:** when there is a suspected violation of the similarity assumption via one of more observed (patient-level) effect modifiers, it might be possible to apply population-adjusted methods, such as matching-adjusted indirect comparison (MAIC), simulated treatment comparison (STC), or multilevel network meta-regression (ML-NMR), to obtain indirect estimates of treatment effects. However, these methods have numerous limitations and might not generate results that are applicable to the research question (see Section 5).

The options described above could lead to the formation of new networks and study pools (e.g., for two different subgroups) and, thus, to a separate performance of a direct or indirect comparison. In this case, subsequent testing of the assumptions in the respective new networks and study pools is necessary.

Requirements for reporting:

- The complete evaluation results regarding potential effect modification;
- Approach to, and reasoning for, handling dissimilarity and heterogeneity;
- The complete results of all sensitivity analyses;
- If the entire study pool was split into several subpools:
  - A complete description of the subpools;
  - The complete evaluation results of the similarity and homogeneity assumptions of all pairwise comparisons within all subpools (see Sections 3.2.1 and 3.2.2).

4 METHODS APPLICABLE TO DIRECT OR INDIRECT COMPARISONS

4.1 Methods for direct comparisons

4.1.1 Standard approaches

Standard approaches for meta-analyses according to the fixed-effect model (with the assumption of a common effect in all included studies) are given by the inverse variance method for continuous data and the Mantel–Haenszel method for binary data [3]. Other useful methods are available in special situations, such as rare events (see below).

Given that the assumption of a common effect in all included studies can be implausible, the standard model for meta-analyses is usually given by the random-effects model. In accordance to the recommendations of the Cochrane Collaboration, the Knapp–Hartung (KH) method should be used with the Paule–Mandel estimator for the heterogeneity parameter for frequentist meta-analyses with five or more studies [55].

With sufficient justification, other methods for meta-analysis can be used in special situations. In the situation of binary data with rare events, the Peto method [9] can be applied. However, this method should not be used when treatment effects are large and the trial arm sizes are unbalanced [5,54]. In the situation of many double-zero studies (i.e., no observed events in both treatment arms), the beta-binomial model can be applied [31,33]. This model allows the inclusion of double-zero studies and contains a random effect for the baseline risk. Nevertheless, the treatment–study interaction is included as a fixed effect, which means that the standard beta-binomial model is a fixed-effect meta-analytic model. As a general alternative to frequentist methods, a Bayesian approach can be used for meta-analysis provided that the required prior distributions are available and can be justified [52].

In general, the choice of methods for direct comparisons must be justified. This includes, but is not limited to, justification for the use of a fixed-effect model over a random-effects model, the choice of informative, non-informative, or vague priors (Bayesian), and baseline risk adjustment models. Additionally, any subgroup or meta-regression analysis for different levels of identified effect modifiers must be described and justified. Further considerations must be given to the number and heterogeneity of the included studies, number of events (rare versus common events), scale [odds ratio (OR), RR, hazard ratio (HR), or mean difference (MD)], quality of evidence etc. when assessing the appropriateness of the method and model choices.

4.1.2 Application of the Knapp-Hartung method

For direct comparisons based on the random-effects model, the general standard approach is given by the KH method. In general, this method holds the type 1 error even in the case of few studies [55]. However, in homogeneous data situations, the standard error of the estimated treatment effect according to the KH method might be arbitrarily small. In this case, the calculated confidence interval is misleadingly narrow [3]. To avoid such misleading results, a simple *ad hoc* variance correction was proposed [30]. In practice, a check is required to decide whether the use of the *ad hoc* variance correction is required. A comparison with the confidence interval calculated by means of the DerSimonian-Laird (DSL) method [10] should be used for this purpose. If the confidence interval of the KH method is narrower than that of the DSL method, the use of the *ad hoc* variance correction is required [29,58]. However, the application of the KH method with *ad hoc* variance correction can reduce the power of the KH method. Thus, in the case of very few studies, the KH method can lead to non-informative results (see Section 4.1.3).

4.1.3 Direct comparisons with very few studies

Meta-analyses with fewer than five studies are problematic in most cases. First, a reliable assessment of heterogeneity is frequently not possible and, therefore, the choice between the fixed-effect and the random-effects model is difficult. Second, the standard random-effects KH approach has frequently very low power. The power might be so low that the KH confidence interval is wider than the union of all confidence intervals of the included studies [48]. In such cases, the KH method is not useful because the results are non-informative and, thus, alternative approaches are required.

In general, a random-effects model should be applied even for meta-analyses with very few studies. However, the chance that the assumption of the fixed-effect approach is valid is greater in the case of very few studies compared with situations with a large number of studies. Especially in the situation with only two studies, it might be justified to apply the fixed-effect model by default. This means that the fixed-effect model should always be applied when there are only two studies, unless there are clear reasons against its use.

In the situation in which a random-effects model is indicated (i.e., 2 studies and clear reasons against the fixed-effect model and in the case of three or four studies without clear reasons in favour of the fixed-effect model), the first approach should be to use the random-effects model by means of the KH method (with or without *ad hoc* variance correction). However, because of the low power, a comparison with the DSL method is helpful to find a valid conclusion. If both methods (DSL and KH) yield the same result regarding statistical significance, the corresponding conclusion can be drawn for the treatment effect. If the estimated treatment effect resulting from the DSL method is statistically significant but that of the KH method is not, the situation is less clear. In this case, a qualitative summary of the study results can be performed. If there are at least two statistically significant studies in the same direction and most of the available evidence supports this direction, the conclusion of a significant effect can be drawn, although this effect cannot be quantified. To explore whether most of the evidence supports this direction, the study weights of the performed random-effects model according to the KH method can be used. As an example, clear thresholds for the required study weights are proposed in the methods paper of the Institute for Quality and Efficiency in Health Care (Institut für Qualität und Wirtschaftlichkeit im Gesundheitswesen, IQWiG) [27]. More details about the model choice for direct comparisons with very few studies are given elsewhere [48].

Alternatively, a random-effects Bayesian meta-analysis with weakly informative prior distribution for the heterogeneity parameter might be useful in the case of very few studies, because external heterogeneity information decreases the problem of estimating heterogeneity with insufficient data [44]. A clear rationale is required for the choice of the prior distributions, because this choice can have substantial effects on the final results in Bayesian meta-analyses with very few studies [50]. Moreover, the impact of the chosen prior distribution should be explored in sensitivity analyses.

Requirements for reporting:

- Assessment of whether the assumptions of the chosen method for meta-analysis are justified; in the case of deviations from standard meta-analytic approaches, a thorough justification for the chosen approach should be given and assessed in the JCA;
- Assessment of the forest plot with point estimates and confidence intervals of all included studies, the *p*-value of the heterogeneity test, the I^2 value, and the pooled effect estimate with confidence interval; in the case of a random-effects model, an assessment of the prediction intervals;
- Determination of whether a fixed- or random-effects model is appropriate;
- In the case of a Bayesian meta-analysis, an assessment of the chosen prior distributions with justification and sensitivity analyses (see also Section 4.2 for reporting requirements for Bayesian approaches); a clear indication of the extent to which the estimated treatment effects are sensitive to the choice of prior distribution; where informative prior distributions are used, the justification for this should be assessed;
- In the case of a qualitative summary of the study results, a description of the chosen approach with criteria used for the decision whether there is an overall effect (e.g., thresholds for study weights).

4.2 Indirect comparisons

In this section, important domains in assessment of the credibility of indirect comparison methods are described.

The first step in the assessment of the statistical analysis is to consider whether the method used is correct for the network of evidence. The Bucher indirect treatment comparison is appropriate for a network comprising two treatments indirectly compared through a common comparator. The Bucher method can also be applied in star-shaped and ladder networks, but then as multiple pairwise comparisons. Multi-arm trials can only be included as pairwise comparisons, but the generated effect estimates are correlated and the corresponding standard errors are inappropriate. This correlation will be problematic if the aim is to use the estimates in a decision model because the method assumes independence between pairwise comparisons. In cases with several different pairwise comparisons, a network meta-analysis encompassing all this evidence should be considered. Frequentist and Bayesian methods are equally applicable. Naive comparisons should not be used because they do not preserve randomisation. Unanchored and disconnected evidence networks cannot be analysed with these methods (see Section 6).

If the method for analysis is deemed appropriate (assumptions met), the appropriateness of the model used must be validated. This includes, but is not limited to, justification for the use of a fixed-effect model over a random-effects model, the choice of informed, uninformed, or vague priors (Bayesian), and baseline risk adjustment models. Additionally, any subgroup or meta-regression analysis for different levels of identified effect modifiers must be described and justified. Further considerations must be given to the number and heterogeneity of studies informing each contrast, number of events (rare versus common events), scale (OR, RR, HR, or MD), quality of evidence, and so on, when assessing the appropriateness of the method and model choices. In networks with large discrepancies in the number of studies informing each contrast, the estimated treatment effect might change from being significant in a clinical trial study to non-significant in the evidence synthesis.

In addition to the methods of NMA described here and in the EUnetHTA 21 Methodological Guideline D4.3.2 *Direct and Indirect Comparisons*, many other approaches have been proposed in the literature, such as the original method of Lumley [32] and the 'arm-based' NMA introduced by Hong *et al.* [26]. However, many of these methods make different fundamental assumptions to those described in this document and are, in general, unlikely to be suitable for use in JCAs [13,46,57]. If such an analysis is presented in a JCA, it is essential that assessors carefully examine the underlying assumptions and assess their plausibility, as well as the relevance of the results obtained.

Requirements for reporting:

- Determination of whether pooling of the studies is meaningful, and justification for this determination (will be informed by the assessment of exchangeability);
- Assessment of whether the chosen method for the network meta-analysis is appropriate given the evidence base (including assumptions regarding the variances of the effects);
- Assessment of the graphical and tabular presentations of the evidence network, including the information on the number of randomised controlled trials (RCTs) per contrast;
- Assessment of the separate results from direct and indirect comparisons, including measures of uncertainty; where both direct and indirect estimates for a particular comparison are available, any discrepancies between them should be highlighted;
- If possible, assessment of rankograms [surface under the cumulative ranking curve (SUCRA), cumulative probability curves, and probability of being the best treatment];
- In the case of a Bayesian NMA, an assessment of the following issues:
  - The chosen prior distributions with justification and sensitivity analyses;
  - Plots of the posterior mean deviance of individual data points for the original model versus the inconsistency model;
  - Convergence of the Markov chains

4.3 Evidence synthesis of time-to-event data

4.3.1 Assessment of the proportional hazards assumption

A (network) meta-analysis of HRs requires that the proportional hazards (PH) assumption holds for all pairwise comparisons in the network. This means that the assumption must be tested for all included studies, which requires either access to individual patient-level data (IPD) for all studies or the construction of pseudo-IPD from digitised Kaplan–Meier curves (e.g., by using the algorithm proposed by Guyot [22]). The PH assumption is also required for comparisons for which there is no direct evidence available; this cannot be assessed directly.

Failure of the PH assumption occurs when the HR between treatment arms is non-constant, which can be interpreted as a time-varying treatment effect. An example of this is the delayed effect on survival observed in studies of immunotherapies in the treatment of advanced cancers [34]. When the PH assumption fails, the average HR will vary according to the length of follow-up, which can differ across studies in the network. Furthermore, the HR obtained from the Cox model might be biased as an estimate of this average because of censoring (unless a suitable adjustment is performed to account for this) [47]. Therefore, if the PH assumption is deemed to be implausible for one or more comparisons in the network, then (network) meta-analysis of HRs should not be carried out. In this scenario, there are two alternative approaches that may be undertaken:

- (Network) meta-analysis of restricted mean survival times (see Section 4.3.2);
- (Network) meta-analysis of flexible survival models [fractional polynomials (FPs) or piecewise exponential models] (see Section 4.3.3).

Requirements for reporting:

- Assessment of whether the PH assumption has been thoroughly evaluated in the submission, with particular reference to the following criteria:
  - Log-cumulative hazard plots for all studies (the lines representing the intervention and comparator should be parallel if PH holds);
  - Plots of Schoenfeld residuals (these should show no trend over time if PH holds);
  - Results of any statistical tests used to assess the PH assumption;
  - Any opinions from healthcare professionals received on the plausibility of the PH assumption; for example, if a delayed treatment effect is expected, then PH might not hold.

4.3.2 (Network) meta-analysis of restricted mean survival time

When the PH assumption does not hold, it is possible to carry out a (network) meta-analysis of restricted mean survival time (RMST) [45]. This involves the selection of a relevant time-point for follow-up and then calculation of the area under the Kaplan–Meier curve between randomisation and this time. Relative treatment effects are then computed as either the difference or ratio of RMSTs between treatment arms. These effects can then be synthesised in a fixed or random-effects meta-analysis using methods previously described.

When RMST is used, a key consideration is the choice of follow-up time, because different choices can produce different results. Possible values are limited by the available data, and some higher values might be more uncertain because of the limited numbers at risk. Therefore, it might be necessary to consider the duration of follow-up of the included studies to select an appropriate time-point. It is important that prespecified criteria for selecting the base case follow-up time are clearly reported, and that a range of follow-up times be presented in sensitivity analysis.

4.3.3 (Network) meta-analysis with flexible survival time models

The use of flexible models for the hazard function allows (network) meta-analysis to be carried out without the assumption of PHs. These methods require the use of IPD or, more commonly, pseudo-IPD, whereby published survival curves for the endpoints of interest are scanned and digitalised (e.g., by using the algorithm proposed by Guyot [22]).

FP (network) meta-analysis involves modelling time-dependent hazard rates for each intervention separately (as linear combinations of positive and negative powers of time), allowing for a wide range of different-shaped hazards. Treatment effects comprise multiple correlated parameters, and can be synthesised using fixed or random-effects models. A similar approach is possible using restricted cubic spline models [20].

Evidence synthesis using FP requires selection of the most appropriate model for the hazard rates; that is, the most appropriate combination of powers of the time variable. This can be assessed using measures of statistical fit [e.g., Akaike information criterion (AIC), Bayesian information criterion (BIC), or, in a Bayesian framework, DIC], visual fit to the observed hazards and survival functions, and/or clinical plausibility. Assessors should be aware that the use of different FP models can lead to different conclusions regarding relative treatment effects; therefore, sensitivity analysis is important.

Piecewise exponential models also allow for a relaxation of the proportional hazards assumption. With this approach, the follow-up period for all treatments is split into a fixed number of pieces, and the hazard rate for each intervention in the network is assumed to be constant within each piece. Treatment effects estimated using this method comprise piecewise HRs; thus, the PH assumption is required within each piece, but not over the entire follow-up period. Such an assumption might be plausible in situations in which a delayed treatment effect is expected (e.g., immunotherapies in oncology), but where PH is expected to hold thereafter. These piecewise hazard ratios can be incorporated into an (network) meta-analysis in the usual way, using fixed- or random-effects models.

To carry out piecewise exponential (network) meta-analysis, it is necessary to choose the number and location of the cut-points of the pieces. This can be done using visual and statistical fit to the observed data, and external opinion might also be helpful. Furthermore, the plausibility of the PH assumption within each piece should be assessed using the methods described previously. Assessors should again be aware that choosing different numbers and locations of cut-points can alter the estimated treatment effect; thus, sensitivity analysis is important.

A limitation that applies to both FPs and piecewise exponential models is that the estimated treatment effects they produce are multidimensional and not easily interpretable. There is no obvious way to perform statistical inference in this setting (i.e., testing for statistically significant treatment effects). The usual method for addressing this is to compare either restricted mean survival time or extrapolated mean survival time, obtained from the chosen models. This is usually carried out in a Bayesian framework, in which posterior distributions of the model parameters are used to obtain posterior estimates of (restricted/extrapolated) mean survival times. When extrapolation is used, the plausibility of long-term extrapolations should also be considered as part of the model selection process. When restricted means are used, consideration must be given to the chosen time-point.

Requirements for reporting:

- For flexible parametric models: assessment of model choice with reference to measures of statistical fit and any other information used to inform this choice (e.g., clinical opinion);
- For RMST: assessment of the rationale for the choice of follow-up time; sensitivity of the results to this choice should be assessed;
- Comparison of observed and modelled HRs over time (e.g., table of the HRs at different time points and/or the plot of HR over time to indicate whether the chosen method is appropriate);
- Comparison of the survival time distribution implied by the chosen (best-fitting or most plausible) model along with the alternative models and the study KM data; evaluation of visual fit to the observed data;
- Where multiple model choices are comparable in terms of fit and/or plausibility, the results obtained from these alternative models should be compared and assessed.

5 Assessment of population-adjusted methods

5.1 General considerations: is population adjustment for indirect comparisons appropriate?

Population-adjusted methods are used in the context of an ITC or more general NMA, in which there is concern that the similarity assumption might not hold. These methods aim to adjust for this imbalance to obtain an unbiased estimate of the relative treatment effect in the scenario in which IPD is available for one or more trials in the network, and only aggregate data (AgD) for others. MAIC and STC should not be used when full IPD is available for all studies; IPD network meta-regression is generally the appropriate method to adjust for covariate imbalances in this case.

The most common examples of population-adjusted methods are MAIC, STC, ML-NMR, and other mixed IPD and aggregate data regression methods. The MAIC method reweights patients in the IPD study to match the characteristics of the AgD study, whereas STC and ML-NMR fit outcome regression models to the IPD studies, which can be extrapolated to other populations A consideration when selecting among MAIC, STC, and ML-NMR is that MAIC adjustments are only applicable to the population of the AgD study, whereas STC and ML-NMR can be used to extrapolate treatment effects to any population with known covariate values for the effect modifiers. Both MAIC and STC are limited to simple networks with two studies, whereas ML-NMR can be applied to any connected network.

When assessing a population-adjusted indirect comparison, the problem of multiplicity arising from 'researcher degrees of freedom' must be considered. Indeed, the number of methods and potential covariate combinations available to the modeller raises the possibility of selecting the method that produces the most favourable results for the intervention under assessment. For this reason, these methods are often more suitable as an exploratory analysis rather than as the primary analysis. In the case of anchored comparisons, it should be demonstrated that bias will be reduced by the use of a population-adjusted methods.

MAIC and STC methods are also sometimes used in the case of disconnected networks; in this context, absolute outcomes (rather than relative effects) are adjusted and, therefore, adjustment must account for all potential confounders in addition to effect modifiers. The EUnetHTA 21 Methodological Guideline *Direct and Indirect Comparisons* details the many issues regarding population adjustment methodologies for unanchored ITCs. By describing these methods here, we are not endorsing them, and once again reiterate that estimates arising from unanchored ITCs are unreliable.

Requirements for reporting:

- Assessment of the justification for population adjustment as a means of estimating treatment effectiveness;
- A complete description of the method and/or model(s) used for population adjustment and estimation of the treatment effects, and an assessment of the appropriateness of this choice.

5.2 Assessing covariate selection (all population-adjusted methods)

The validity of all population-adjusted methods depends on the inclusion of all effect modifiers as covariates in the relevant model. These should be identified using the methods described in Section 3.2.1, ideally before conducting the analysis. In the case of unanchored comparisons, all prognostic variables must also be included (see Sections 5.5 and 6).

In the case of both MAIC and STC, only effect modifiers of the relative effect being estimated in the IPD trial are needed to carry out the adjustment. However, interpretation of the results also requires knowledge of effect modifiers for the AgD trial. In the case of more-complex networks of evidence (e.g., using ML-NMR), knowledge of effect modifiers for all pairwise comparisons is typically needed.

Covariates that are initially balanced (or approximately balanced) between study populations at baseline should not be omitted because the adjustment procedure could create an imbalance where none existed before. Methods of covariate selection based upon statistical significance or model fit are of limited use for STC and MAIC, given that, when limited IPD is available, these methods will typically be underpowered to detect relevant effects.

When effect modifiers are omitted for any reason, population-adjusted treatment effects obtained using the methods described here will necessarily be biased. The magnitude of this bias depends on both the magnitude of effect modification associated with the missing covariate(s) and the extent of the imbalance between treatment groups in terms of this characteristic (after adjustment). It is often unknown whether covariates are missing or which covariates these might be. When multiple relevant-effect modifiers are missing, the combined impact becomes difficult to predict. Assessors should highlight the potential for residual bias in the resulting estimate and give an indication of the size and direction of that bias where possible.

Assessors should be aware that, when population-adjusted indirect comparisons are carried out despite relevant covariates being unavailable, bias in the estimated treatment effects could be increased as a result of adjustment compared with the results of a standard NMA. Consider an example in which the relevant effect modifiers are background statin use and history of cardiovascular disease and where, in one study, there is a strong positive association between these two variables (e.g., because of statin therapy being initiated following a cardiovascular event), but there is no such association in the other study (e.g., because of statin being used for primary prevention of cardiovascular disease among these patients). In such a scenario, adjustment for background statin use but not cardiovascular disease in a MAIC could increase the imbalance in the proportion of patients with a history of cardiovascular disease across studies.

To account for the risk of bias (RoB) because of missing or unknown effect modifiers, it is possible to perform statistical inference on the estimated treatment effect by testing against a shifted null hypothesis; that is, a null hypothesis of some non-zero relative treatment effect of a magnitude large enough to account for any plausible bias arising from missing covariates. If this is done, the shifted hypothesis to be tested should be prespecified and its magnitude clearly justified.

Requirements for reporting:

- Assessment of the methodology used to identify relevant-effect modifiers;
- Assessment of the adequacy of the set of included effect modifiers to generate an unbiased estimate of the treatment effect;
- When relevant-effect modifiers have not been included in the assessment model, a quantification of the potential magnitude and likely direction of the resulting bias;
- If shifted hypothesis testing has been used, an assessment whether this is sufficient to account for the likely magnitude of residual bias arising from missing covariates.

5.3 Additional considerations for outcome regression approaches

The STC and ML-NMR methods involve fitting an outcome regression model (e.g., a generalised linear model or Cox PH model) to the available IPD to obtain an estimate of the outcome at each level of the included covariates. The chosen model must estimate treatment effects on the same scale as the indirect treatment comparison. For example, if the indirect comparison is to be carried out on the log OR scale, an appropriate choice for the outcome regression model would be a generalised linear model with a binomial likelihood and logit link. It is not appropriate to use logistic regression to adjust absolute risks in each arm and then carry out the indirect treatment comparison on the risk-difference or log-risk ratio scales.

A fundamental assumption of STC and ML-NMR is that the effect of the covariates is additive on the outcome measure scale (i.e., that the functional form of the outcome regression model is appropriate).

For example, if a Cox PH model is used for the treatment effect (log hazard ratio), then the effect of the covariates is assumed to be linear on the log hazard ratio scale (i.e., PHs). In the case of the IPD study, this should be assessed and reported using standard model diagnostics (e.g., analysis of residuals). External data could also be helpful here; for example, the effect of LDL cholesterol levels on cardiovascular event rates has been characterised as approximately linear on a log-rate scale [18].

In the case of anchored STCs, the inclusion of additional prognostic variables (that are not also effect modifiers) in the outcome model will not reduce bias, but could improve precision of the estimated treatment effect and, therefore, can be considered. In this case, standard measures of model fit, such as AIC/BIC, residual deviance, and so on, can be used to select these additional covariates.

The STC and ML-NMR methods can generate estimates of the treatment effect in any target population by substituting the relevant mean covariate values into the outcome regression model. This can be useful if the population of interest differs from the trial populations. However, the validity of these estimates is unknown outside the range of covariate values included in the IPD study; extrapolation beyond this region might not generate meaningful estimates of the treatment effect. For example, if the age range of the IPD study population is 40–55 years, it would not be appropriate to use STC/ML-NMR to extrapolate treatment effects to a population with a mean age of 60 years, because the relationship between age and treatment effect cannot be assessed outside the range of the IPD study. More generally, treatment effects for covariate combinations that are not well represented in the IPD study will be uncertain. Therefore, the degree of overlap in baseline covariates should be reported and assessed by, for example, plotting the distributions of baseline characteristics in the IPD trial(s) together with the mean and confidence intervals from the AgD trials.

The usual approach to STC involves substituting mean covariate values from the AgD population into the outcome regression model, which estimates the conditional treatment effect at this level of the covariates (i.e., the predicted individual-level response) [39,42]. However, the summary effect estimate from the AgD study is typically a marginal treatment effect (i.e., population average) or, in some cases, a conditional effect but typically adjusted for a different set of covariates. As a result, STC, conducted using the substitution of mean covariate values, combines incompatible effect estimates, potentially leading to bias in the estimation for both estimands (conditional and marginal) when the outcome regression model is nonlinear and there are invalid standard errors in all cases [39,42]. To overcome this, approaches to STC targeting marginal treatment effects have been proposed [28,43]. These approaches generally require additional assumptions to estimate the joint covariate distribution from the AgD study; therefore, the plausibility of these assumptions should be assessed. The JCA report should clarify the STC approach used and the target estimand.

Additional assumptions and/or data requirements for ML-NMR depend on the targeted treatment effect. In a simple network of one IPD study and one AgD study, the (population average) marginal treatment effect in the AgD population can be estimated with no further assumptions beyond those required for STC [35]. However, estimation of conditional treatment effects, marginal effects in any other population, or any application of ML-NMR in a network with two or more AgD studies requires the estimation of additional treatment–covariate interactions. To achieve this, the available data must include, for each treatment in the network, either full IPD from at least one study investigating that treatment or enough AgD studies investigating that treatment to estimate the relevant interactions. If such data are not available, then the ‘shared effect modifier’ assumption is required (see Section 5.6) for certain treatment classes within the network and, therefore, the plausibility of this assumption must be assessed [35,38]. Furthermore, specification of the joint covariate distributions for the AgD studies is required, which typically necessitates additional assumptions.

Requirements for reporting:

- An assessment of the model fit and appropriateness of the outcome regression model to capture the effect of covariates (including treatments) on outcomes;
- An assessment of the covariate overlap between the IPD study (or studies) and the populations to which relative treatment effects are adjusted (e.g., the AgD study or studies);
- For STC, a description of the method used to estimate outcomes (e.g., substitution of mean covariate values, simulation, or numerical integration) and the treatment effect that is targeted by the chosen approach; assessment of whether the estimands that have been combined are compatible, highlighting any potential for bias;
- For ML-NMR, clear statement as to whether the available data are sufficient to estimate treatment–covariate interactions; statement of any additional assumptions (e.g., shared effect modifier) that have been made to estimate the model;
- An assessment of the method used to estimate the joint covariate distributions in the AgD studies, if required (applies to ML-NMR and certain approaches to STC).

5.4 Additional considerations for matching-adjusted indirect comparisons

When MAIC is used to carry out population adjustment, the principal concern is whether the weighted pseudo-population has the same distribution of effect modifiers (anchored and unanchored comparisons) and prognostic variables (unanchored only) as the target population. These distributions should be reported and their similarity assessed; if nontrivial differences exist for one or more variables after matching, then the results of the MAIC will likely be biased. The use of hypothesis tests for the equality of means after matching is not recommended as a method to decide whether sufficient balance has been achieved, because multiple tests are typically required (increasing the risk of type 1 error) and statistical power might be low (type 2 error).

The distribution of weights should be examined to assess the extent of overlap between the two populations. The approximate effective sample size should also be reported. If this is considerably smaller than the original sample size, then statistical power will be reduced accordingly. The presence of extreme weights and/or low ESS also indicates that the target population of the MAIC is considerably different from the source population. This could be problematic in the context of a JCA because it is likely that the source population is of greater interest to the assessment than the target population (see also Section 5.6).

The assessor should ensure that an appropriate method, such as robust standard errors or bootstrapping, has been used to estimate the confidence interval associated with the treatment effect. Failure to do so will result in confidence intervals that are artificially narrow and do not capture the full extent of (statistical) uncertainty in the estimated treatment effect.

Requirements for reporting:

- Assessment of covariate balance achieved after matching, and of potential impact of any residual imbalance on the results (if this can be estimated);
- Assessment of the distribution of weights and effective sample size after matching to assess the extent of overlap between the two populations;
- Statement as to whether the reported confidence interval for the treatment effect appropriately captures the additional uncertainty arising from reweighting (e.g., whether the confidence interval has been estimated using an appropriate method, such as robust standard errors or bootstrapping).

5.5 Dealing with unanchored MAICs and STCs: additional challenges

Population-adjusted methods for indirect comparisons are also used when considering disconnected networks. Comparing treatments in an unanchored network is essentially a comparison of absolute effects rather than of relative effects, which is not the goal of the JCA. The validity of the results depends on all relevant prognostic variables (as well as effect modifiers) being included as covariates in the relevant model, which is unlikely to be satisfied in practice. In general, this will substantially increase the amount of adjustment required. The process used to identify prognostic variables is analogous to that described previously for effect modifiers and should be reported transparently in the submission.

Differences in patient characteristics are typically more likely to affect absolute outcomes than they are relative outcomes, which means that more covariates must be included in the adjustment model to obtain an unbiased estimate the treatment effect. For example, if two hypothetical treatments, A and B, aimed at lowering blood pressure were to be compared in an unanchored comparison, then adjustment would need to be carried out for all covariates potentially affecting blood pressure, such as age, sex, smoking status, race, geographical location, body mass index, diabetes status, and many others that might not have been recorded. By contrast, an anchored comparison of an A and B via a common (e.g., placebo) comparator would only require adjustment for covariates affecting response to treatment.

There will inevitably be differences in the trials other than patient characteristics. Interventions will be administered under different conditions and endpoints might be recorded in different ways (e.g., investigator versus independent assessment of tumour progression). Again, these differences typically have a greater impact on unanchored comparisons compared with anchored comparisons, because absolute effects are being compared. An assessor should assess these carefully, using opinions from healthcare professionals again if required, to decide whether it is appropriate to undertake an unanchored MAIC or STC.

In summary, although these methods are often presented as the only way of quantifying a relative treatment effect, this does not mean that the method will be of sufficient standard to confidently estimate a relative treatment effect. A better, although still problematic, option is the use of methods for the analysis of non-randomised data, which require access to the full IPD information (see Section 6).

Requirements for reporting:

- An assessment of the methodology used to identify all relevant prognostic variables;
- An assessment of the appropriateness of carrying out an unanchored indirect comparison, with reference to data availability, definitions of outcomes, comparability of study characteristics, and other considerations; if full IPD were available for all studies, then this should be clearly highlighted because, under this scenario, other IPD-based methods (e.g., propensity score matching) would likely be more appropriate;
- An assessment of whether the set of included covariates is likely to be sufficient to generate an unbiased comparison of outcomes; quantification of the magnitude and direction of potential bias arising from missing prognostic variables in the analysis.

5.6 Interpretation and use of population-adjusted results

Population adjustment estimates treatment effects in the population of the AgD study, which might not be generalisable outside of that population. However, in the context of JCAs, it is likely that estimation of the treatment effect in the population of the 'source' (IPD) study is of interest. Phillippo *et al.* [37] highlight this issue in relation to two MAIC analyses of the same two trials comparing secukinumab and adalimumab to placebo as treatments for ankylosing spondylitis. The relative treatment effect differed depending on which trial the IPD was taken from, which is explained by patient differences in the target studies. Although STC and ML-NMR have the advantage that treatment effects from the IPD study can be extrapolated to any population, none of the available population adjustment methods can estimate the relevant (indirect) treatment effect outside of the population of the target AgD study unless additional assumptions are made [37].

In some situations, it might be reasonable to 'transpose' effect estimates obtained from anchored population-adjusted indirect comparisons to other populations, such as that of the source trial. Doing so requires the additional 'shared effect modifier' assumption proposed by Phillippo *et al.* [36]. This assumption applies to a set of active treatments and states that, relative to a common comparator: (i) the covariates that are effect modifiers and (ii) the change in treatment effect for each effect modifier (i.e., the magnitude and direction of the interaction terms), are the same for all active treatments in this set. When this holds, the relative effect between any pair of treatments in this set will be the same in any population, which means, in particular, that treatment effects obtained from population-adjusted indirect comparisons can be transposed to the population of the source (IPD) trial or indeed any other relevant population. The shared-effect modifier assumption is more likely to hold for treatments with a similar mode of action (e.g., an ITC of two angiotensin-converting enzyme inhibitors) than for those in different classes (e.g., in an ITC of and angiotensin-converting enzyme inhibitor versus an angiotensin receptor blocker). Strong biological and/or clinical justification must be provided to justify its use in a JCA.

Different population adjustment methods target different estimands. The MAIC method targets the marginal treatment effect (population average effect over the AgD population), whereas STC, performed using substitution of mean covariate values, targets the conditional treatment effect at the specified level of the covariates (individual-level treatment effect for the 'average' patient in the AgD population). In its most general form, ML-NMR can target either estimand in any target population [40-42].

To incorporate results of an MAIC or STC into a wider NMA, it is necessary to assume similarity of effect modifiers across the network after adjustment; in other words, the distribution of effect modifiers across all studies in the network is similar to that of the target study rather than of the source study.

Population adjustment aims to reduce bias arising from an imbalance of effect modifiers (or prognostic variables for unanchored comparisons) but does so at the cost of increased variance. The result is a loss of precision when estimating treatment effects (i.e., wider confidence intervals) or, equivalently, a loss of statistical power. When inference is made on the basis of population-adjusted comparisons, assessors should take into account that these comparisons are typically underpowered.

Requirements for reporting:

- A clear description of the population in which the treatment effect has been estimated, and its relevance to the assessment question; any limitations should be clearly outlined, and potential biases arising from population differences should be reported (including an assessment of the likely magnitude and direction of any bias, if possible);
- Clear statement as to whether the 'shared-effect modifier' assumption is required to estimate the treatment effect in the target population; if this assumption is invoked, the biological and/or clinical basis for this assumption should be scrutinised and the strengths and limitations clearly described;
- A comparison between the population-adjusted estimates of treatment effects with those obtained from standard methods of (network) meta-analysis; if the magnitude, direction, and/or precision of these effects differ considerably, then assessors should discuss likely explanations for this (e.g., covariate adjustment, loss of effective sample size, or underlying assumptions);
- The target estimand of the chosen population-adjustment method, that is, marginal (population-average relative treatment effect) or conditional (individual level treatment effect for the 'average' patient) relative effects, and its relevance to the assessment question.

6 ASSESSMENT OF COMPARISONS BASED UPON NON-RANDOMISED EVIDENCE

6.1 General considerations

All commonly encountered sources of evidence outside of RCTs are non-randomised (i.e., single-arm trials, cohort studies, case-control studies, other observational studies, and the use of historical controls). Any such study has much greater potential to include material bias in the estimate of treatment effect compared with an appropriate RCT, and this is likely to carry through when combining evidence from these sources. A key concern is that the underlying assumption of exchangeability is unlikely to hold because there is a very high risk of confounding bias, meaning that the association between intervention and outcome differs from its causal effect.

Therefore, treatment comparisons based upon non-randomised evidence require careful consideration of its validity. The inclusion and exclusion criteria for each study should be carefully examined, because these criteria are typically more restrictive for clinical trials than for observational studies, leading to potential violations of the positivity assumption (e.g., individuals with very poor prognosis are often excluded from clinical trials but not from cohort studies). The potential for unmeasured confounding arising from 'volunteer bias' should also be considered when interpreting the results: willingness to participate in a clinical trial might be associated with several prognostic variables that might be unmeasured, such as access to medicine, socioeconomic status, location, educational attainment, and overall health status. The RoB in the results should be assessed by using an appropriate tool, such as Risk of Bias in Non-Randomised Studies – of Interventions (ROBINS-I) [49] (see also the EUnetHTA 21 Practical Guideline *Validity of Clinical Studies* for practical guidance on the assessment of the validity of individual studies).

In some cases, it might be that the lack of randomisation can be compensated for by rigorous adjustment for confounding. However, this requires that all confounders and effect modifiers relevant for adjustment are measured and that the model and covariate selection strategies for adjustment are prespecified and based upon transparent criteria [23]. The requirement of all confounders and effect modifiers being measured is unlikely to be met, given that unknown modifiers and confounders are assumed to be always present. These adjustment methods require access to the full IPD information. Aggregated data alone are not sufficient to reliably estimate treatment effects. A statistical analysis plan (SAP) is required to describe the methods planned to adjust for confounding.

Requirements for reporting:

- Assessment of the inclusion and exclusion criteria for the relevant non-randomised data;
- Assessment of the RoB and the validity of the results of all included trials;
- Comparison of baseline characteristics of all included trials;
- An assessment of the methodology used to identify all relevant prognostic variables and effect modifiers;
- An assessment of the SAP with the methods used to adjust for confounding;
- An assessment of whether the set of included covariates is likely sufficient to generate an unbiased comparison of outcomes; quantification of the magnitude and direction of potential bias arising from missing prognostic variables and effect modifiers in the analysis.

6.2 Propensity scores

6.2.1 Checking that the assumptions of propensity score matching and/or weighting are valid

An important method to adjust for confounding in non-randomised studies is by using propensity scores. This method requires careful planning of all possible modelling options in the form of an SAP [56]. As mentioned in EUnetHTA 21 Methodological Guideline on *Direct and Indirect Comparisons*, three

assumptions must be met when using non-randomised data and propensity scores or another method to adjust for confounding: positivity, overlap, and balance. In the JCA context, the assessor/co-assessor must check and report the validity of these assumptions.

Checking the positivity assumption

The positivity assumption means that patients in both groups must be theoretically eligible for both treatments of interest. In randomised evidence, positivity is guaranteed by randomisation. In non-randomised evidence, the positivity assumption concerns the probability of receiving treatment, but this probability needs to be modelled (e.g., by propensity score) because of the absence of randomisation. Suspicion of violation for positivity assumption (e.g., inclusion of patients in one treatment group, with contraindication to the other treatment group, or patients with a propensity score equal or close to zero or one) should be systematically reported.

Checking the overlap assumption

Sufficient overlap means that the distribution of patients among the different propensity scores must be similar. To allow this assumption to be checked, propensity score distribution (using histograms or density plot), among samples if applicable (i.e., whole population, matched population, and/or population created by weighting), should be reported in the JCA and discussed. The overlap depends on the matching performed and the techniques used [17]. In the case of trimming, if a large proportion of the sample is lost after trimming regions of non-overlap, then it could indicate insufficient overlap [11]. When trimming is performed, the selected population should be described in detail to investigate whether it sufficiently represents the original research population.

Checking the balance assumption

The populations in the compared groups must be sufficiently balanced after adjustment for confounding. The achieved balance must be assessed before and after matching, weighting, or stratification. Absolute standardised differences between the treatment groups should be used to compare the balance for each covariate [2]. Cut-offs for acceptable absolute standardised difference vary (0.1–0.25) [51]. Therefore, the final conclusion regarding the balance assumption would be left to MS for absolute standardised differences <0.25; if any absolute standardised difference is ≥0.25, violation of the balance assumption should be stated. Doubly robust methods combining propensity scores and outcome regression can be used to reduce bias arising from residual covariate imbalance after matching or weighting.

Checking the inferential goal

The inferential goal (i.e., target of inference) determines, in part, the choice of a specific propensity score method. The most common estimands are the average treatment effect (ATE) and the average treatment effect among treated (ATT). Adequation between inferential goal and chosen estimand (ATT or ATE) should be evaluated by the assessor/co-assessor. Adequation between propensity score method (matching, stratification, adjustment using the propensity score, or weighting) and the chosen estimand should also be assessed (e.g., matching primarily estimates ATT) [1].

Requirements for reporting:

- An assessment of the SAP with the propensity score methods used to adjust for confounding;
- An assessment of the required assumptions of sufficient positivity, overlap, and balance;
- The final decision whether the assumptions of positivity, overlap, and balance hold, with reasoning based on the analyses submitted by the HTD.

6.2.2 Interpreting results of propensity score

In the JCA report, when propensity score methods are used, qualitative analysis must be first performed by the assessor/co-assessor, assessing the validity of assumptions (see Section 6.2.1). If underlying

assumptions are considered to be violated, this must be explicitly reported before quantitative results are interpreted, because they would be biased.

Quantitative results (effect estimates with confidence intervals) should be presented for both crude and propensity score analyses. Results of sensitivity analyses should always be presented to evaluate the robustness of results. Quantitative results assess the degree of statistical association, but a statistically significant association does not necessarily imply a causal relationship. The JCA report should be factual and the assessor/co-assessor is not supposed to conclude on causality.

Requirements for reporting:

- An assessment of the models used for confounder adjustment and estimation of the treatment effects and whether any limitations exist with regard to model choice;
- A clear description of the population in which the treatment effect has been estimated, and its relevance to the assessment question; any limitations should be clearly outlined, and potential biases arising from population differences should be reported (including an assessment of the likely magnitude and direction of any bias if possible);
- An assessment of the effect estimates with confidence intervals for the crude data and after adjustment;
- An assessment of the results of all sensitivity analyses;
- If shifted hypothesis testing has been used, an assessment whether this is sufficient to account for the likely magnitude of residual bias arising from missing covariates.

7 FURTHER RELEVANT DOCUMENTS (UNDER DEVELOPMENT)

- **EUnetHTA 21 Methodological Guideline D.4.3.2: *Direct and Indirect Comparisons***

  A methodological guideline for assessors and co-assessors that describes the currently available usual methods for direct and indirect comparisons regarding their underlying assumptions, strengths, and weaknesses.

- **EUnetHTA 21 Practical Guideline D.4.2.1: *Scoping Process***

  A practical guideline for assessors and co-assessors that describes the methods and principal steps of the scoping process.

- **EUnetHTA 21 Practical Guideline D4.5.1: Applicability of Evidence: *Practical Guideline on Multiplicity, Subgroup, Sensitivity and* Post-hoc *Analyses***

  A practical guideline for assessors and co-assessors that describes how to consider complementary analyses and how to handle multiplicity issues.

- **EUnetHTA 21 Practical Guideline D4.6.1: *Validity of Clinical Studies***

  A practical guideline for assessors and co-assessors that describes possible approaches and specific instructions for action when assessing the certainty of results coming from individual studies, regardless of whether they are RCTs or other types of study.

- **EUnetHTA 21 Guideline D5.3.1: *Procedural Guideline for Appointing Assessors and Co-Assessors for JCA/CA***

  A guidance document that describes the minimum selection criteria for the appointment of assessors and co-assessors of JCAs/collaborative assessments (CAs) for pharmaceuticals and MDs.

8 FUTURE RECOMMENDATION

This document was written in combination with the EUnetHTA 21 Methodological Guideline *Direct and Indirect Comparisons*. The Hands-on Group recommends that consideration be given to merge these documents at a later date under the direction of the future Methodological Subgroup of the HTA Regulation Coordination Group.

Given that the methods used for direct and indirect comparisons might be updated more frequently compared with other methodological areas, it is advisable that these guidelines be reviewed under the direction of the future Methodological Subgroup every 3 years.

9 References

1 Ali MS, Prieto-Alhambra D, Lopes LC *et al.* Propensity score methods in health technology assessment: principles, extended applications, and recent advances. *Front Pharmacol.* 2019;10:973.

2 Austin PC. Balance diagnostics for comparing the distribution of baseline covariates between treatment groups in propensity-score matched samples. *Stat Med.* 2009;28(25):3083-107.

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):382-92.

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):371-87.

5 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):1428-44.

6 Cooper NJ, Sutton AJ, Morris D *et al.* Addressing between-study heterogeneity and inconsistency in mixed treatment comparisons: application to stroke prevention treatments in individuals with non-rheumatic atrial fibrillation. *Stat Med.* 2009;28(14):1861-81.

7 Cope S, Zhang J, Saletan S *et al.* A process for assessing the feasibility of a network meta-analysis: a case study of everolimus in combination with hormonal therapy versus chemotherapy for advanced breast cancer. *BMC Med.* 2014;12:93.

8 Cordero CP, Dans AL. Key concepts in clinical epidemiology: detecting and dealing with heterogeneity in meta-analyses. *J Clin Epidemiol.* 2021;130:149-51.

9 Deeks JJ, Higgins JPT, Altman DG *et al.* Analysing data and undertaking meta-analyses. In: Higgins JPT, Thomas JJC, Cumpston M *et al.* (eds). *Cochrane Handbook for Systematic Reviews of Interventions*, 2nd edition. Hoboken, NJ: Wiley; 2019, pp. 241-84.

10 DerSimonian R, Laird N. Meta-analysis in clinical trials. *Control Clin Trials.* 1986;7(3):177-88.

11 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.

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

13 Dias S, Ades AE. Absolute or relative effects? Arm-based synthesis of trial data. *Res Synth Methods.* 2016;7(1):23-8.

14 Dias S, Sutton AJ, Welton NJ *et al.* Evidence synthesis for decision making 3: heterogeneity – subgroups, meta-regression, bias, and bias-adjustment. *Med Decis Making.* 2013;33(5):618-40.

15 Dias S, Welton NJ, Caldwell DM *et al.* Checking consistency in mixed treatment comparison meta-analysis. *Stat Med.* 2010;29(7-8):932-44.

16 Dias S, Welton NJ, Sutton AJ *et al. NICE DSU Technical Support Document 4: Inconsistency in Networks of Evidence Based Upon Randomised Controlled Trials*. London, UK: National Institute for Health and Care Excellence; 2011.

17 Faria R, Hernandez Alava M, Manca A et al. NICE DSU Technical Support Document 17: The Use of Observational Data to Inform Estimates of Treatment Effectiveness for Technology Appraisal: Methods for Comparative Individual Patient Data. London, UK: National Institute for Health and Care Excellence; 2015.

18 Ference BA, Ginsberg HN, Graham I *et al.* Low-density lipoproteins cause atherosclerotic cardiovascular disease. 1. Evidence from genetic, epidemiologic, and clinical studies. A consensus statement from the European Atherosclerosis Society Consensus Panel. *Eur Heart J.* 2017;38(32):2459-72.

19 Fisher DJ, Carpenter JR, Morris TP *et al.* Meta-analytical methods to identify who benefits most from treatments: daft, deluded, or deft approach? *BMJ* 2017;356:j573.

20 Freeman SC, Carpenter JR. Bayesian one-step IPD network meta-analysis of time-to-event data using Royston-Parmar models. *Res Synth Methods.* 2017;8(4):451-64.

21 Greenland S. Tests for interaction in epidemiologic studies: a review and a study of power. *Stat Med.* 1983;2(2):243-51.

22 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.

23 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):758-64.
24 Higgins JPT, Thompson SG. Controlling the risk of spurious findings from meta-regression. *Stat Med.* 2004;23(11):1663-82.
25 Higgins JPT, Thompson SG. Quantifying heterogeneity in a meta-analysis. *Stat Med.* 2002;21(11):1539-58.
26 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.
27 IQWiG. *General Methods, Version 6.1*. Cologne, Germany: Institute for Quality and Efficiency in Health Care; 2022.
28 Ishak KJ, Proskorovsky I, Benedict A. Simulation and matching-based approaches for indirect comparison of treatments. *Pharmacoeconomics* 2015;33(6):537-49.
29 Jackson D, Law M, Rücker G *et al.* The Hartung-Knapp modification for random-effects meta-analysis: a useful refinement but are there any residual concerns? *Stat Med.* 2017;36(25):3923-34.
30 Knapp G, Hartung J. Improved tests for a random effects meta-regression with a single covariate. *Stat Med.* 2003;22(17):2693-710.
31 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):1097-116.
32 Lumley T. Network meta-analysis for indirect treatment comparisons. *Stat Med.* 2002;21(16):2313-24.
33 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):366-81.
34 Mick R, Chen TT. Statistical challenges in the design of late-stage cancer immunotherapy studies. *Cancer Immunol Res.* 2015;3(12):1292-8.
35 Phillippo DM. *Calibration of Treatment Effects in Network Meta-Analysis using Individual Patient Data*, *PhD Thesis*. Bristol, UK: Bristol Medical School; 2019.
36 Phillippo DM, Ades AE, Dias S *et al.* Methods for population-adjusted indirect comparisons in health technology appraisal. *Med Decis Making.* 2018;38(2):200-11.
37 Phillippo DM, Ades AE, Dias S *et al. NICE DSU Technical Support Document 18: Methods for Population-Adjusted Indirect Comparisons in Submission to NICE*. London, UK: National Institute for Health and Care Excellence; 2016.
38 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):1189-210.
39 Phillippo DM, Dias S, Ades AE *et al.* Assessing the performance of population adjustment methods for anchored indirect comparisons: a simulation study. *Stat Med.* 2020;39(30):4885-911.
40 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):2759-63.
41 Remiro-Azócar A, Heath A, Baio G. Conflating marginal and conditional treatment effects: comments on "Assessing the performance of population adjustment methods for anchored indirect comparisons: a simulation study". *Stat Med* 2021;40(11):2753-8.
42 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):750-75.
43 Remiro-Azócar A, Heath A, Baio G. Parametric G-computation for compatible indirect treatment comparisons with limited individual patient data. *Res Synth Methods.* Published online April 29, 2022. https://dx.doi.org/10.1002/jrsm.1565.
44 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):448-74.
45 Royston P, Parmar MK. Restricted mean survival time: an alternative to the hazard ratio for the design and analysis of randomized trials with a time-to-event outcome. *BMC Med Res Methodol.* 2013;13:152.
46 Salanti G, Higgins JP, Ades AE *et al.* Evaluation of networks of randomized trials. *Stat Methods Med Res.* 2008;17(3):279-301.
47 Schemper M. Cox analysis of survival data with non-proportional hazard functions. *Statistician.* 1992;41:455-65.

48 Schulz A, Schürmann C, Skipka G *et al.* Performing meta-analyses with very few studies. In: Evangelou E, Veroniki AA (eds). *Meta-Research: Methods and Protocols*. New York: Humana; 2022, pp. 91-102.
49 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.
50 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):3046-67.
51 Stuart EA, Lee BK, Leacy FP. Prognostic score-based balance measures can be a useful diagnostic for propensity score methods in comparative effectiveness research. *J Clin Epidemiol.* 2013;66(8 Suppl):S84-S90.
52 Sutton AJ, Abrams KR. Bayesian methods in meta-analysis and evidence synthesis. *Stat Methods Med Res.* 2001;10(4):277-303.
53 Sutton AJ, Abrams KR, Jones DR *et al. Methods for Meta-Analysis in Medical Research*. Chichester, UK: Wiley; 2000.
54 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):1351-75.
55 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.
56 Wang SV, Pinheiro S, Hua W *et al.* STaRT-RWE: structured template for planning and reporting on the implementation of real world evidence studies. *BMJ* 2021;372:m4856.
57 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):5197-213.
58 Wiksten A, Rücker G, Schwarzer G. Hartung-Knapp method is not always conservative compared with fixed-effect meta-analysis. *Stat Med* 2016;35(15):2503-15.