Available online at www.sciencedirect.com

ScienceDirect

journal homepage: www.elsevier.com/locate/jval

Indirect Treatment Comparison/Network Meta-Analysis Study Questionnaire to Assess Relevance and Credibility to Inform Health Care Decision Making: An ISPOR-AMCP-NPC Good Practice Task Force Report

Jeroen P. Jansen, PhD[1,2,*], *Thomas Trikalinos, MD, PhD*[3], *Joseph C. Cappelleri, PhD, MS, MPH*[4],
Jessica Daw, PharmD, MBA[5], *Sherry Andes, BSPharm, PharmD, BCPS, BCPP, PAHM*[6],
Randa Eldessouki, MBBCH, MSc, MD[7], *Georgia Salanti, PhD*[8]

[1]*Redwood Outcomes, Boston, MA, USA;* [2]*Tufts University School of Medicine, Boston, MA, USA;* [3]*Program in Public Health, Center for Evidence-based Medicine, Brown University, Providence, RI, USA;* [4]*Pfizer, Inc., New London, CT, USA;* [5]*Clinical Pharmacy, UPMC Health Plan, Pittsburgh, PA, USA;* [6]*Catamaran, Louisville, KY, USA;* [7]*Scientific & Health Policy Initiatives, ISPOR, Lawrenceville, NJ, USA;* [8]*Department of Hygiene and Epidemiology, School of Medicine University Campus, University of Ioannina, Ioannina, Greece*

ABSTRACT

Despite the great realized or potential value of network meta-analysis of randomized controlled trial evidence to inform health care decision making, many decision makers might not be familiar with these techniques. The Task Force developed a consensus-based 26-item questionnaire to help decision makers assess the relevance and credibility of indirect treatment comparisons and network meta-analysis to help inform health care decision making. The relevance domain of the questionnaire (4 questions) calls for assessments about the applicability of network meta-analysis results to the setting of interest to the decision maker. The remaining 22 questions belong to an overall credibility domain and pertain to assessments about whether the network meta-analysis results provide a valid answer to the question they are designed to answer by examining 1) the used evidence base, 2) analysis methods, 3) reporting quality and transparency, 4) interpretation of findings, and 5) conflicts of interest. The questionnaire aims to help readers of network meta-analysis opine about their confidence in the credibility and applicability of the results of a network meta-analysis, and help make decision makers aware of the subtleties involved in the analysis of networks of randomized trial evidence. It is anticipated that user feedback will permit periodic evaluation and modification of the questionnaire.

Keywords: bias, checklist, credibility, decision making, indirect comparisons, mixed treatment comparisons, multiple treatment comparison, network meta-analysis, questionnaire, relevance, validity.

Introduction

Four Good Practices task forces developed a consensus-based set of questionnaires to help decision makers evaluate 1) prospective and 2) retrospective observational studies, 3) network meta-analysis (indirect treatment comparison), and 4) decision analytic modeling studies with greater uniformity and transparency [1,2]. The primary audiences of these questionnaires are assessors and reviewers of health care research studies for health technology assessment, drug formulary, and health care services decisions requiring varying levels of knowledge and expertise. This report focuses on the questionnaire to assess the relevance and credibility of network meta-analysis (including indirect treatment comparison).

Systematic reviews of randomized controlled trials (RCTs) are considered a key summary of evidence for informing clinical practice guidelines, formulary management, and reimbursement policies. Many systematic reviews use meta-analysis to synthesize evidence from several RCTs addressing the same question [3]. Sound comprehensive decision making requires comparisons of all relevant competing interventions. Ideally, well-designed and conducted RCTs would simultaneously compare all interventions of interest. Such studies are almost never available, thereby complicating decision making [4–7]. New drugs are often compared with placebo or standard care, but not against each other, in trials aiming to contribute toward obtaining approval for drug licensing; there may be no commercial incentive to compare the new treatment with an active control treatment. Even if there was an incentive to incorporate competing interventions in an RCT, the interventions of interest may vary by country or have changed over time because of new evidence and treatment insights. Finally, for some indications, the relatively large number of competing interventions makes a trial incorporating all of them impractical.

In the absence of trials involving a direct comparison of treatments of interest, an indirect comparison can provide useful

* *Address correspondence to:* Jeroen Jansen, Redwood Outcomes, Boston, MA 02108.
E-mail: jjansen@redwoodoutcomes.com; jeroen.jansen@tufts.edu.

1098-3015/$36.00 – see front matter Copyright © 2014, International Society for Pharmacoeconomics and Outcomes Research (ISPOR).
Published by Elsevier Inc.
http://dx.doi.org/10.1016/j.jval.2014.01.004

Background to the Task Force

On May 21, 2011, the Board of Directors approved, in principle, ISPOR's participation with the Academy of Managed Care Pharmacy (AMCP) and the National Pharmaceutical Council (NPC) in the Comparative Effectiveness Research Collaborative Initiative (CER-CI) for advancing appropriate use of outcomes research evidence to improve patient health outcomes. ISPOR's contribution to the CER-CI was to develop articles on how to assess prospective and retrospective observational studies, indirect treatment comparison (network meta-analysis), and decision analytic modeling studies to inform health care decision making. Good Practice task forces were created to develop these articles. Task Force Chairs were identified from leaders of ISPOR Good Research Practices task forces. Each Task Force consisted of two members from the AMCP, the NPC, and the ISPOR.

Each Task Force met independently via teleconference. In addition, the Task Force Chairs met via teleconferences and face-to-face meetings held on April 20, 2012 (San Francisco, CA, USA), June 3, 2012 (Washington, DC, USA), June 28-29, 2012 (Boston, MA, USA), November 4, 2012 (Berlin, Germany), and May 21, 2013 (New Orleans, LA, USA), to coordinate a common outline and format for these articles. A focus group representing the US formulary decision-making community (22 participants) was convened April 19, 2012, at the AMCP Meeting, San Francisco, CA, USA, for feedback on the draft outline, format, and content of the assessment articles. The content of these reports was presented for comment at the ISPOR Annual International Meetings held June 10, 2012, and May 22, 2013, and during the European Congress held November 5 and 6, 2012.

A draft indirect treatment/network meta-analysis Task Force report was sent for comment to the Task Force review group on August 5, 2013. Written comments were considered, and a final draft report was sent for comment to the ISPOR membership on September 24, 2013. A total of 54 written comments were received. All written comments are published on the ISPOR Web site, which can be accessed via the Research menu on ISPOR's home page: http://www.ispor.org. The final report was submitted to *Value in Health*.

evidence for the difference in treatment effects between competing interventions (which otherwise would be lacking) and for judiciously selecting the best choice(s) of treatment. For example, we can indirectly compare two treatments, which have never been compared against each other in an RCT, if each has been compared against the same comparator [7–14].

Although it is often argued that indirect comparisons are needed only when direct comparisons are not available, it is important to realize that to use all available evidence, one should combine information from both direct and indirect comparisons. A collection of RCTs informing on several treatments constitutes a network of evidence, in which each RCT directly compares a subset, but not necessarily all, of treatments. Such a network involving treatments compared directly, indirectly, or both can be synthesized by means of network meta-analysis [10–14]. In traditional meta-analysis, all included studies compare the same intervention with the same comparator. Network meta-analysis extends this concept by including multiple pairwise comparisons across a range of interventions and provides estimates of relative treatment effects on multiple treatment comparisons for comparative effectiveness purposes based on direct and/or indirect evidence. Even when results of the direct evidence are conclusive, combining them with results of similar indirect estimates in a mixed treatment comparison may yield a more precise estimate for the interventions directly compared [4,5,12]. When we use the term "network meta-analysis" in this report, we include indirect treatment comparisons as well.

Despite the great realized and potential value of network meta-analysis to inform health care decision making and its increasing acceptance (e.g., Canadian Agency for Drugs and Technologies in Health, National Institute for Health and Care Excellence in the United Kingdom, Institute for Quality and Efficiency in Health Care in Germany, and Haute Autorité de Santé in France], it is not commonly used to inform health care decisions, and many decision makers are not familiar with it. There is a critical need for education on network meta-analysis, as well as transparent and uniform ways to assess the quality of reported network meta-analyses to help inform decision making.

In creating a questionnaire for health care decision makers to assess network meta-analyses, the Task Force was asked to work toward two goals. The first was to provide a guide for gauging one's confidence in the findings of a network meta-analysis. The aim was to create a questionnaire for use by individuals with understanding of principles of clinical research, but without in-depth knowledge of design and statistics. The second goal was for the questionnaire to have educational and instructional value to prospective users of network meta-analyses. We anticipate modifications to the structure, content, or wording of the questionnaire based on the feedback, after it has been put to use.

Questionnaire Development

One issue in creating questionnaires for decision makers is whether they should be linked to scorecards, annotated scorecards, or checklists. Concerns were raised by the Task Force that a scorecard with an accompanying scoring system may be misleading; it may not have adequate validity and measurement properties. Scoring systems may also provide users with a false sense of precision and have been shown to be problematic in the interpretation of randomized trials [15].

An alternative to a scorecard is a checklist. However, it was deemed by the Task Force that checklists might mislead users in their assessments because a network meta-analysis may satisfy all the elements of a checklist and still harbor "fatal flaws" in the methods applied in the publication. Moreover, users might have the tendency to count up the number of elements present, convert it into an implicit score, and then apply that implicit scoring to their overall assessment of the evidence. In addition, the applicability of a network meta-analysis may depend on evidence external to the meta-analysis and specific to the setting of interest. In general, a decision maker should be aware of the strengths and weaknesses of each piece of information available (one of which is a network meta-analysis) and apply his or her own reasoning. Furthermore, a checklist format may undermine the educational potential of the questionnaire.

The Task Force decided to develop a questionnaire characterized by two principal concepts: relevance and credibility. *Relevance* is the extent to which results of a network meta-analysis, if trustworthy, apply to the setting of interest. The relevance domain includes questions related to the population, comparators, end points, time frame, and other policy-relevant differences. *Credibility* is the extent to which the network meta-analysis provides valid answers to the question it is designed to answer. To generate questions for the credibility domain, the Task Force relied on the expertise of its members and the scientific literature including the reports of the ISPOR Task Force and other pertinent publications on indirect treatment comparisons and network meta-analysis [9–14]. Items and suggested wording were also informed by earlier or recent efforts that provided guidance to evidence reviewers [16,17]. Questions guiding the assessment of credibility were grouped into five subdomains: evidence base used for the indirect comparison or network meta-analysis, analysis, reporting quality and transparency, interpretation, and conflict of interest.

The developed questionnaire has 26 questions guiding the reader in assessing the relevance (4 questions) and credibility (22 questions) of a network meta-analysis. Each question can be answered with "Yes," "No," and "Can't Answer." "Can't Answer" can be used if the item is not reported in sufficient detail or at all, or if the assessor does not have sufficient training to answer the question. For one question ("Were statistical methods used that preserve within-study randomization? (No naive comparisons)"), a "No" will imply a fatal flaw. This fatal flaw suggests that findings can be misleading and that the decision maker should use caution in applying the findings to inform decisions. However, the occurrence of the fatal flaw neither prevents a user from completing the questionnaire nor does it mandate that such results are deemed inappropriate for decision making.

On the basis of how questions are answered, the user would make an overall judgment about each credibility subdomain. A designation of "Strength" implies confidence that the network meta-analysis was conducted well and without influential shortcomings. A designation of "Neutral" suggests that potentially important concerns are raised, but the user deemed them unlikely to affect their credibility, as applicable. A designation of "Weakness" suggests that the findings are likely to be biased and misleading because of numerous or important shortcomings in the design or conduct of the meta-analysis. Finally, a designation of "Fatal flaw" implies that the user believes that findings of the network meta-analysis are likely to be biased and conclusions misleading.

Following on, the user will consider the domain judgments to opine about the overall relevance and credibility of the network meta-analysis for decision making, as either "sufficient" or "insufficient." We would like to remind that the questionnaire is not a score or a checklist, in which one would count the frequency of each designation to reach an overall judgment. We consider such mechanistic approaches simplistic to say the least. The user will judge whether and how to incorporate results from a network meta-analysis in the decision-making process, after considering all inputs.

Following internal testing by the Task Force members during September and October of 2012 and subsequent modification of the questionnaire, the revised questionnaire was further tested by volunteers not involved in its development. Each volunteer was asked to evaluate three published network meta-analysis with the questionnaires being developed during April and May of 2013. Based on the feedback received, the current version of the questionnaire was found helpful in assisting users systematically opine about the relevance and credibility of network meta-analysis studies to their setting of interest. The questionnaire is provided in Appendix 1 in Supplemental Materials found at http://dx.doi.org/10.1016/j.jval.2014.01.004.

Questionnaire Items

This section provides detailed information to facilitate answering the 26 questions. A glossary is provided in Appendix 2 in Supplemental Materials found at http://dx.doi.org/10.1016/j.jval.2014.01.004.

Relevance

Relevance addresses the extent to which the results of the network meta-analysis apply to the setting of interest to the decision maker. A relative term, relevance has to be determined by each decision maker and the respective judgment determined by one decision maker will not necessarily apply to other decision makers.

1. Is the population relevant?

This question addresses whether the populations of the RCTs that form the basis for the network meta-analysis sufficiently match the population of interest to the decision maker (i.e., there is no clinical reason to assume that the relative treatment effects obtained with the network meta-analysis would not translate to the population of interest to the decision maker). Relevant population characteristics are not only limited to the specific disease of interest but also pertinent to disease stage, severity, comorbidities, treatment history, race, age, sex, and possibly other demographic characteristics. Typically, RCTs included in the network analysis are identified by means of a systematic literature search with the relevant studies in terms of population predefined by study selection criteria. If these criteria are reported, this is a good starting point to judge relevance in terms of population. Evidence tables with inclusion criteria and baseline patient characteristics for each study provide the most relevant information to judge population relevance; exclusion criteria are also noteworthy. For example, if a decision involves covering a Medicare Part D population (e.g., those aged 65 years or older), studies with few patients at or above 65 years of age may be less relevant.

2. Are any relevant interventions missing?

This question gets at whether the intervention(s) included in the network meta-analysis matches the one(s) of interest to the decision maker and whether all relevant comparators have been considered. Depending on the included RCTs, the network meta-analysis may include additional interventions not necessarily of interest for the decision maker. This does not, however, compromise relevance. Aspects to consider when judging the relevance of included biopharmaceuticals are dose and schedule of a drug, mode of administration, and background treatment. A question whether the drug is used as induction or maintenance treatment can be of relevance as well. For other medical technologies, one can consider whether the procedure or technique in the trials is the same as the procedure or technique of interest to the decision maker.

3. Are any relevant outcomes missing?

This question asks what outcomes are assessed in the network meta-analysis and whether the outcomes are meaningful to the decision maker. There has been increasing emphasis on outcomes that are directly meaningful to the patient or the health system such as cardiovascular events (e.g., myocardial infarction and stroke) or patient functioning or health-related quality of life (e.g., short-form 36 health survey and EuroQol five-dimensional questionnaire) and decreasing emphasis on surrogate outcomes (e.g., cholesterol levels) unless validated. Other considerations include the feasibility of measuring relevant (i.e., final) outcomes, the predictive relationship between surrogate outcomes and final outcomes, and what kind of evidence will be considered "good-enough," given the patient population, the burden of the condition, and the availability of alternative treatments (along with the evidence supporting those treatments). Not only are the outcomes themselves of interest, the timing of their assessment is of interest as well. For example, a network meta-analysis that included RCTs with a longer follow-up may be more relevant to help inform treatment decisions for a chronic disease than a network meta-analysis limited to studies with a short follow-up (if follow-up is related to treatment effect).

4. Is the context (settings and circumstances) applicable?
This question addresses whether there are any differences between the RCTs included in the network meta-analysis versus the setting and circumstances the decision maker is interested in. These situations may not necessarily concern the population, interventions, and outcomes but may still render the findings not applicable. For example, the year when the studies included in the network meta-analysis were performed can be of interest when background medical care of a certain disease has dramatically changed over time as when the standard treatment changes over time. Many of the RCTs underlying network meta-analyses comparing biopharmaceuticals are often designed for efficacy (can it work?) purposes and therefore the setting or circumstances may be different from the real-world intent (does it work?). A relevant question to ask, therefore, is whether the relative treatment effects and their rank ordering of interventions as obtained with the network meta-analysis would still hold if real-world compliance or adherence would have been taken into consideration. The answer might be "no" if some of the interventions are associated with a much lower compliance in the real-world setting than other interventions.

Credibility

Once the network meta-analysis is considered sufficiently relevant, its credibility will be assessed. *Credibility* is defined as the extent to which the network meta-analysis or indirect comparison accurately or validly answers the question it is designed to answer. For the assessment questionnaire, we take the position that credibility is not limited to internal validity (i.e., the observed treatment effects resulting from the analysis reflect the true treatment effects) but also concerns reporting quality and transparency, interpretation, and conflict of interest (see Fig. 1). The internal validity of the network meta-analysis can be compromised as a result of the presence of bias in the identification and selection of studies, bias in the individual studies included, and bias introduced by the use of inappropriate statistical methods.

Evidence Base Used for the Indirect Comparison or Network Meta-Analysis

The first six questions of the credibility domain pertain to the validity of the evidence base feeding information into the network meta-analysis.

1. Did the researchers attempt to identify and include all relevant RCTs?
To have a network meta-analysis that reflects the available evidence base, a systematic literature search needs to be performed. Although there is no guarantee that all relevant studies can be identified with the search, it is important that the researchers at least attempt to achieve this goal. Important things to consider include the following:

- *Did the* ***search strategy*** *target RCTs between all interventions of interest?*
- *Were* ***multiple databases*** *searched (e.g., MEDLINE, EMBASE, and Cochrane Central Registry of Trials)?*
- *Would review* ***selection criteria*** *admit all RCTs of interest (if identified by the literature search)?*

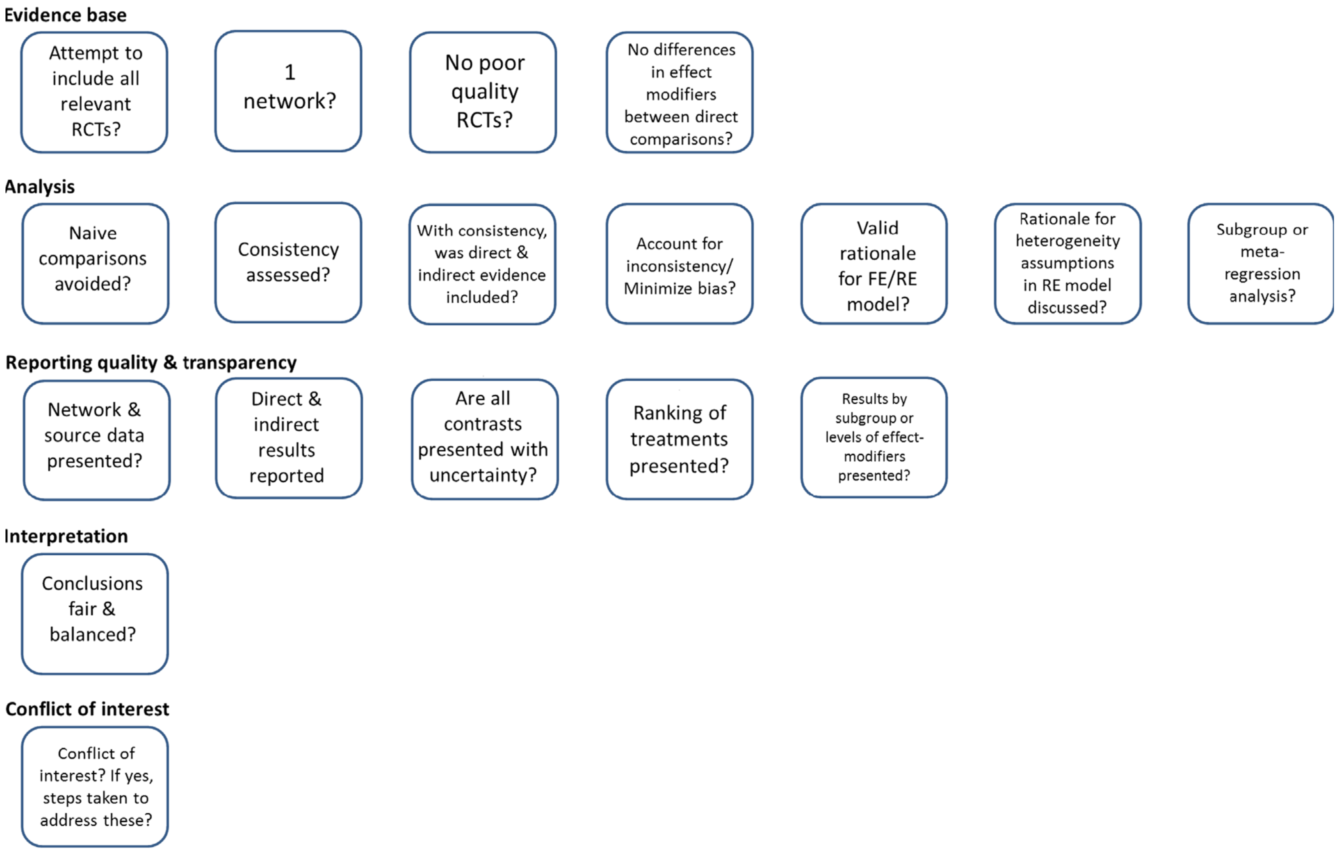
Evidence base
Attempt to include all relevant RCTs?
1 network?
No poor quality RCTs?
No differences in effect modifiers between direct comparisons?
Analysis
Naive comparisons avoided?
Consistency assessed?
With consistency, was direct & indirect evidence included?
Account for inconsistency/ Minimize bias?
Valid rationale for FE/RE model?
Rationale for heterogeneity assumptions in RE model discussed?
Subgroup or meta-regression analysis?
Reporting quality & transparency
Network & source data presented?
Direct & indirect results reported
Are all contrasts presented with uncertainty?
Ranking of treatments presented?
Results by subgroup or levels of effect-modifiers presented?
Interpretation
Conclusions fair & balanced?
Conflict of interest
Conflict of interest? If yes, steps taken to address these?

Fig. 1 – Overview of domains related to assessment of the credibility of a network meta-analysis. FE, fixed effects; RCTs, randomized controlled trials; RE, random effects.

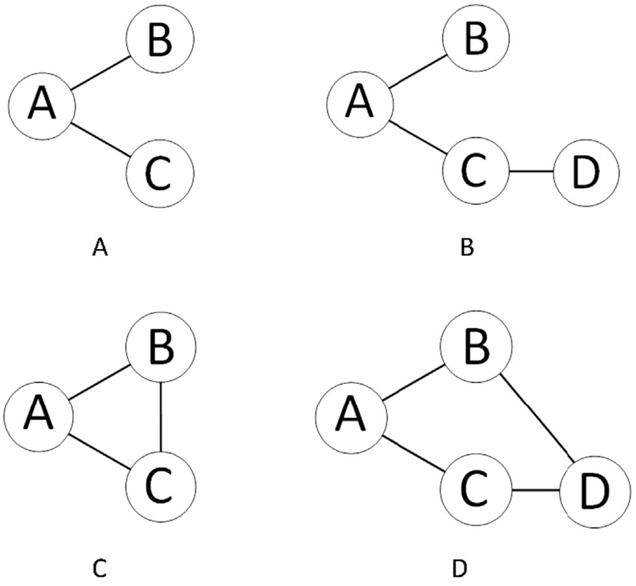

Fig. 2 – Connected networks of randomized controlled trials to allow for network meta-analysis and indirect comparisons. All interventions that can be indirectly compared are part of one network.

A "yes" to the above inquiries probably implies a good effort to include all available relevant published RCTs (of course one would have to review the syntax of the actual search strategies to authoritatively discuss their adequacy). An additional step to identify missing or unpublished key studies would be to search clinical trial databases, such as http://dx.doi.org/clinicaltrials.gov. Systematic reviews that follow the Preferred Reporting Items for Systematic Reviews and Meta-Analyses reporting guidelines are easier to assess [18].

2. Do the trials for the interventions of interest form one connected network of RCTs?

To allow comparisons of treatment effects across all interventions in the network meta-analysis, the evidence base used for the meta-analysis should correspond to a connected network. In simple terms, this means that any two treatments can be compared either directly (head to head) or indirectly, through (one or more) intermediate common referents. Figure 2 depicts four connected evidence networks. The nodes represent interventions, and the edges (i.e., connections) imply that one or more RCTs have compared the respective treatments directly. Any two treatments that have not been compared head to head are amenable to an indirect comparison. For example, the network in Figure 2A includes AB studies comparing intervention B with A and AC studies comparing intervention C with A. The relative treatment effect of C versus B can be obtained with the indirect comparison of the AB and AC studies. The network in Figure 2B also includes CD studies, and can inform on any pairwise comparison (i.e., contrasts) among A, B, C, and D. The networks in Figure 2C,D have a closed loop, which implies that for some of the treatment comparisons there is both direct and indirect evidence. In Figure 2C, there is both direct and indirect evidence for the AB, AC, and BC comparisons. In Figure 2D, there is direct and indirect evidence for all comparisons with the exception of the AD and BC contrast, for which there is only indirect evidence.

In RCTs, the observed outcome with an intervention is the result of study characteristics, patient characteristics, and the treatment itself. In a placebo-controlled trial, the result of the placebo arm reflects the effect of study and patient characteristics on the outcome of interest, say, outcome y (see Fig. 3). In other words, the placebo response is the result of all known and unknown prognostic factors other than active treatment. We can call this the study effect. In the active intervention arm of the trial, the observed outcome y is a consequence of the study effect and a treatment effect. By randomly allocating patients to the intervention and placebo groups, both known and unknown prognostic factors (as well as both measured and unmeasured prognostic factors) between the different groups within a trial are on average balanced. Hence, the study effect as observed in the placebo intervention arm is expected to be the same in the active intervention arm and, therefore, the difference between the active intervention arm and the placebo intervention arm (say delta y) is attributable to the active intervention itself, resulting in a treatment effect.

With an indirect comparison, interest centers on the comparison of the treatment effects of interventions that are not studied in a head-to-head fashion. To ensure that the indirect comparisons of interventions are not affected by differences in study effects between studies, we want to only consider the treatment effects of each trial. This consideration implies that all interventions indirectly compared have to be part of one network of trials in which each trial has at least one intervention (such as placebo) in common with another trial, as illustrated in Figure 4. If some interventions of interest are not part of the same network, then it

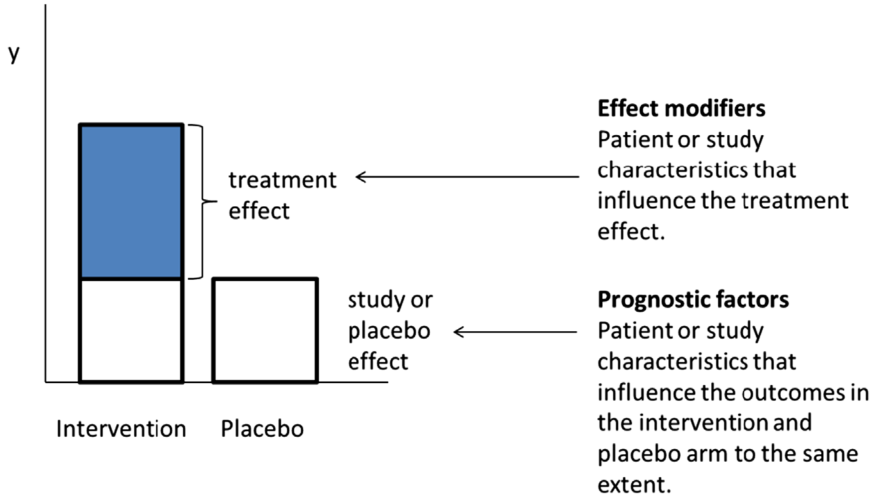

Fig. 3 – Treatment effects, study effects, effect modifiers, and prognostic factors in a randomized placebo-controlled trial.

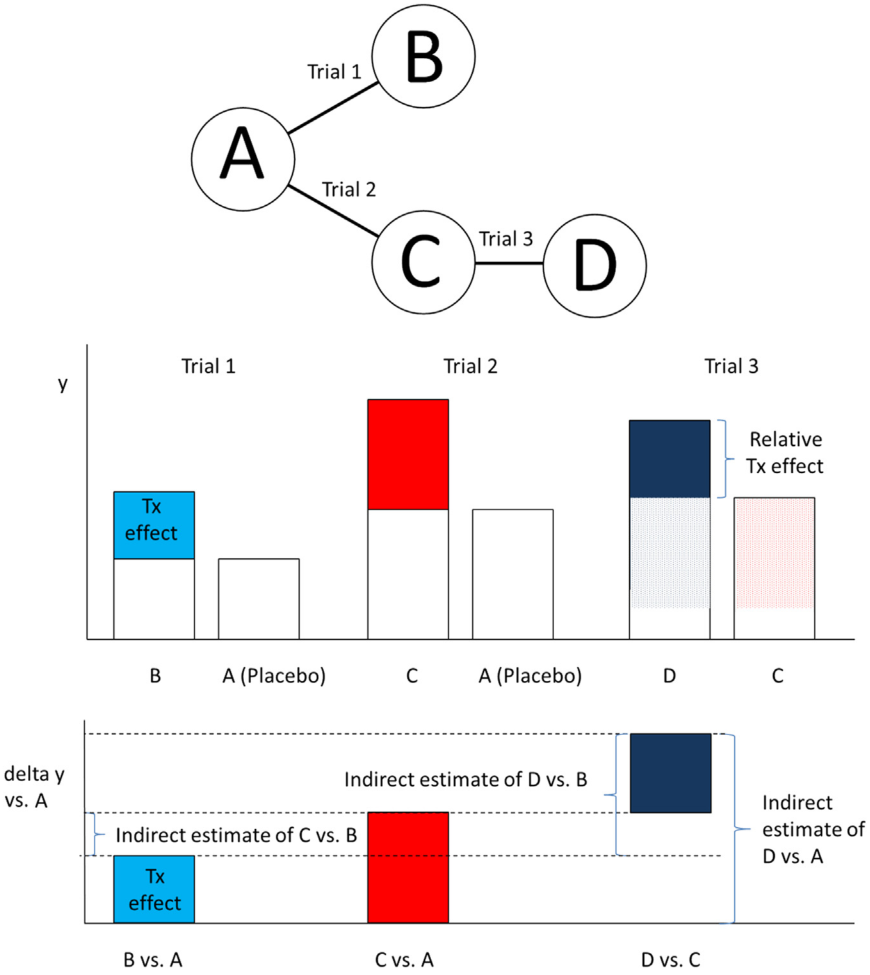

Fig. 4 – Indirect comparison of AB, AC, and CD studies in which each trial is part of one network. y refers to the outcome of interest on a continuous scale, for example, change from baseline in pain, or log odds of a response; delta y vs. A reflects the difference in the outcome of interest with treatment B, C, and D relative to treatment A. *Note.* In terms of treatment effects, B is more efficacious than A, C is more efficacious than B, and D is more efficacious than C. Tx, treatment.

is not possible to perform an indirect comparison of treatment effects of these interventions without a substantial risk of bias, as illustrated in Figure 5.

3. Is it apparent that poor quality studies were included, thereby leading to bias?

The validity of the network meta-analysis is at risk not only when certain studies are not identified but also when the internal validity of individual RCTs is compromised. To answer this credibility question, the network meta-analysis report should have provided summary information on key study characteristics of each RCT, such as method of randomization, treatment allocation concealment, blinding of the outcome assessor, and dropout. Frequently, a network meta-analysis report provides an overview of bias in individual studies as assessed with a specific checklist for individual study validity, such as the Cochrane Collaboration's tool for assessing risk of bias in randomized trials [19].

4. Is it likely that bias was induced by selective reporting of outcomes in the studies?

Outcome reporting bias occurs when outcomes of the individual trials are selected for publication on the basis of their findings [20]. This can result in biased treatment effect estimates obtained with the network meta-analysis. As an assessment of the likelihood of bias due to selective reporting of outcomes, a determination can be made whether there is consistency in the studies used for the network meta-analysis with respect to the different outcomes. In other words, a check should be placed whether any of the selected studies do not report some of the outcomes of interest and were therefore not included in some of the network meta-analyses of the different end points. It can be informative to look at the original publication of the suspect study whether there is any additional information. Furthermore, if reported in the network meta-analysis report, a check on the reasons studies were excluded from the systematic review would be beneficial to ensure no eligible studies were excluded *only* because the outcome of interest was not reported [20]. If this were the case, a related risk of bias—publication bias—in the overall findings of the network meta-analysis would manifest itself.

5. Are there systematic differences in treatment effect modifiers (i.e., baseline patient or study characteristics that have an impact on the treatment effects) across the different treatment comparisons in the network?

Study and patient characteristics can have an effect on the observed outcomes in the intervention and control arms of an

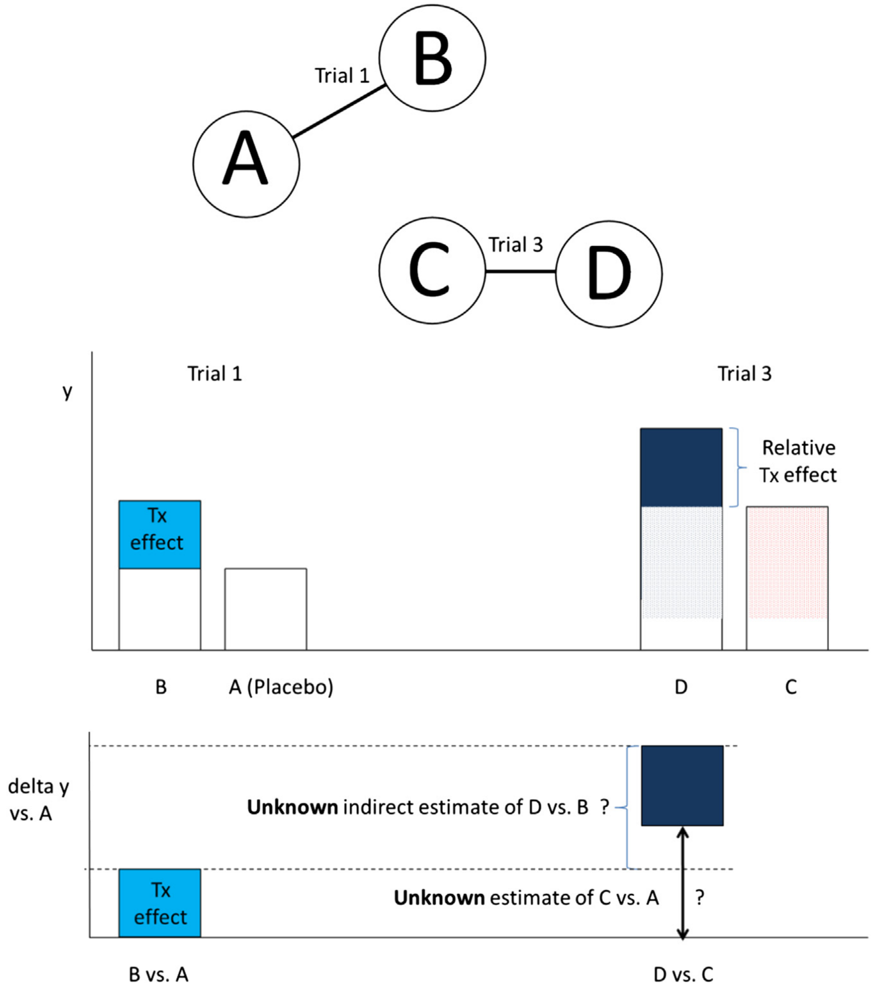

Fig. 5 – AB and CD studies that do not have an intervention in common, making an indirect comparison without a substantial risk of bias not feasible. Tx, treatment.

RCT. As mentioned previously, study and patient characteristics that affect outcome to the same extent in the active intervention and placebo intervention arms are called *prognostic factors*. More specifically, the placebo response, which serves as a barometer of study effect, captures the effect of all study and patient characteristics that are prognostic factors on the outcome of interest.

Study and patient characteristics that affect the difference between the active intervention and the placebo intervention regarding the outcome of interest are treatment effect modifiers (Fig. 3). For example, if a medical intervention works only for men and not for women, a trial among men will demonstrate a positive treatment effect relative to placebo, whereas a trial only among women would not. Sex is a treatment effect modifier for that intervention. As another example, if the outcome of interest (e.g., improvement in pain) is greater in a 24-week trial than in a 12-week trial, and there is no difference in the treatment effect of the intervention relative to placebo between the 24-week trial and the 12-week trial, then trial duration is only a prognostic factor and not an effect modifier. If a variable is both a prognostic factor of the outcome of interest and a treatment effect modifier, then the placebo response (or baseline risk) of a placebo-controlled trial is associated with the treatment effect.

Although the indirect comparison or network meta-analysis is based on RCTs, randomization does not hold across the set of trials used for the analysis because patients are not randomized to different trials. As a result, systematic differences in the distribution of patient characteristics across trials can ensue. In general, if there is an imbalance in study and patient characteristic–related effect modifiers across the different types of direct comparisons in a network meta-analysis, the corresponding indirect comparisons are biased [21,22]. In Figures 6 to 8, examples of valid and biased indirect comparisons from a network meta-analysis are provided. It is important to acknowledge that there is always some risk of imbalances in unknown or unmeasured effect modifiers between studies evaluating different interventions. Accordingly, there is always a small risk of residual confounding bias, even if all observed effect modifiers are balanced across the direct comparisons.

The answer to question 5 is a “yes” if there are substantial (or systematic) differences in effect modifiers, which can be judged by comparing study-specific inclusion and exclusion criteria, baseline patient characteristics, and study characteristics that are expected to be effect modifiers.

6. If yes (i.e., there are such systematic differences in treatment effect modifiers), were these imbalances in effect modifiers across the different treatment comparisons identified before comparing individual study results?

Frequently, there are several trial and patient characteristics that are different across the different direct comparisons. Deciding

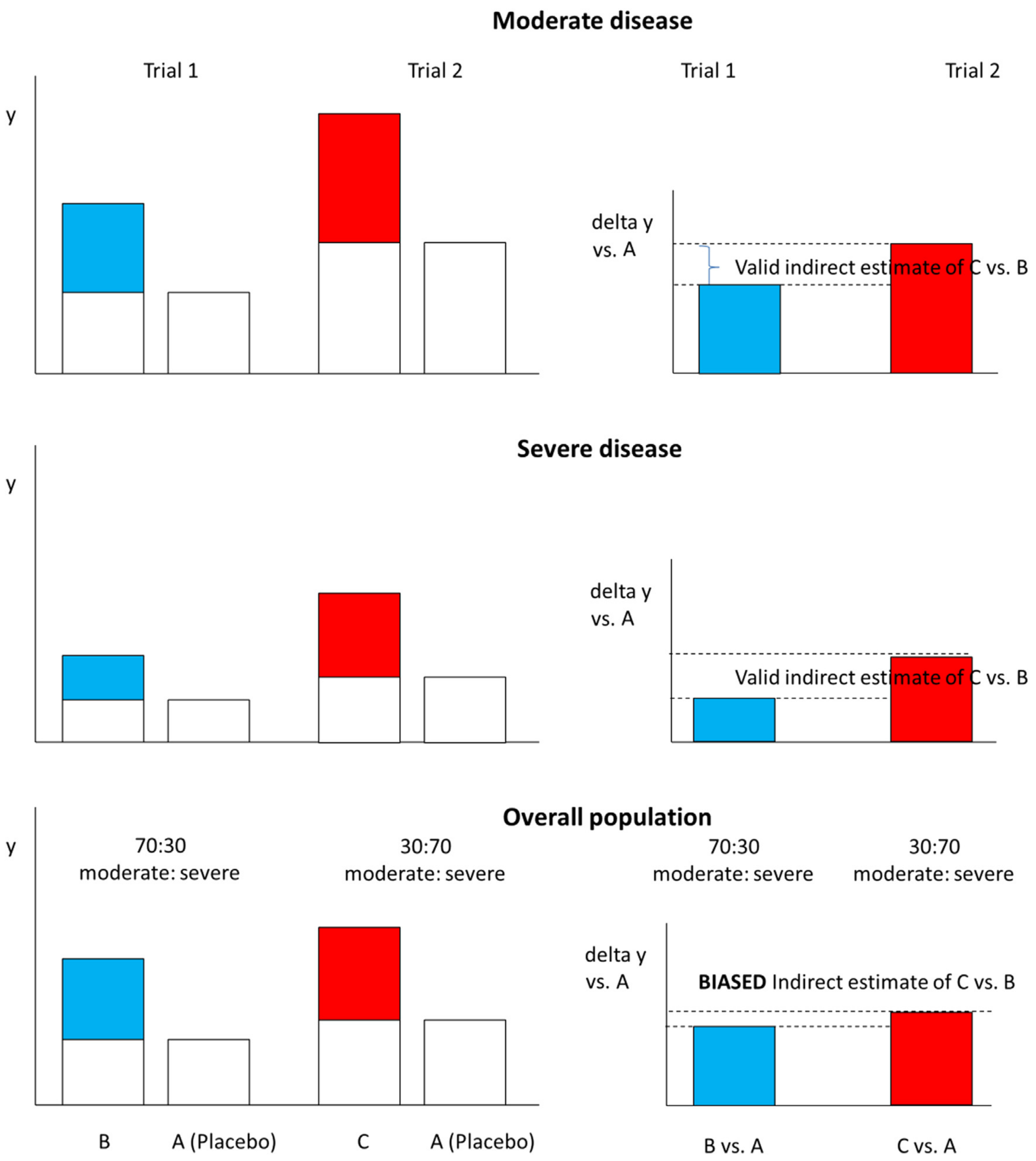

Fig. 6 – Indirect comparison of an AB and AC study with different proportions of patients with moderate and severe disease. *Note.* Disease severity is an effect modifier. The indirect comparison for the moderate disease subgroup is valid, as is the indirect comparison for the severe subgroup. The indirect comparison of the results for the overall population of both studies is biased because the distribution of the effect modifier severity is different for the AB and AC studies.

which covariates are effect modifiers based on observed patterns in the results across trials can lead to false conclusions regarding the sources of inconsistency and biased indirect comparisons [22,23]. It is recommended that researchers undertaking the network meta-analysis first generate a list of potential treatment effect modifiers for the interventions of interest on the basis of previous knowledge or reported subgroup results *within* individual studies before comparing results *between* studies. Next, the study and patient characteristics that are determined to be likely effect modifiers should be compared across studies to identify any imbalances between the different types of direct comparisons in the network.

Analysis

The next seven questions pertain to the statistical methods used for the network meta-analysis.

7. Were statistical methods used that preserve within-study randomization? (No naive comparisons)

To acknowledge the randomization of treatment allocation within RCTs and thereby minimize the risk of bias as much as possible, a network meta-analysis of RCTs should be based on their relative treatment effects [4–14]. In other words, statistical methods need to be used that preserve within-study randomization. An invalid or

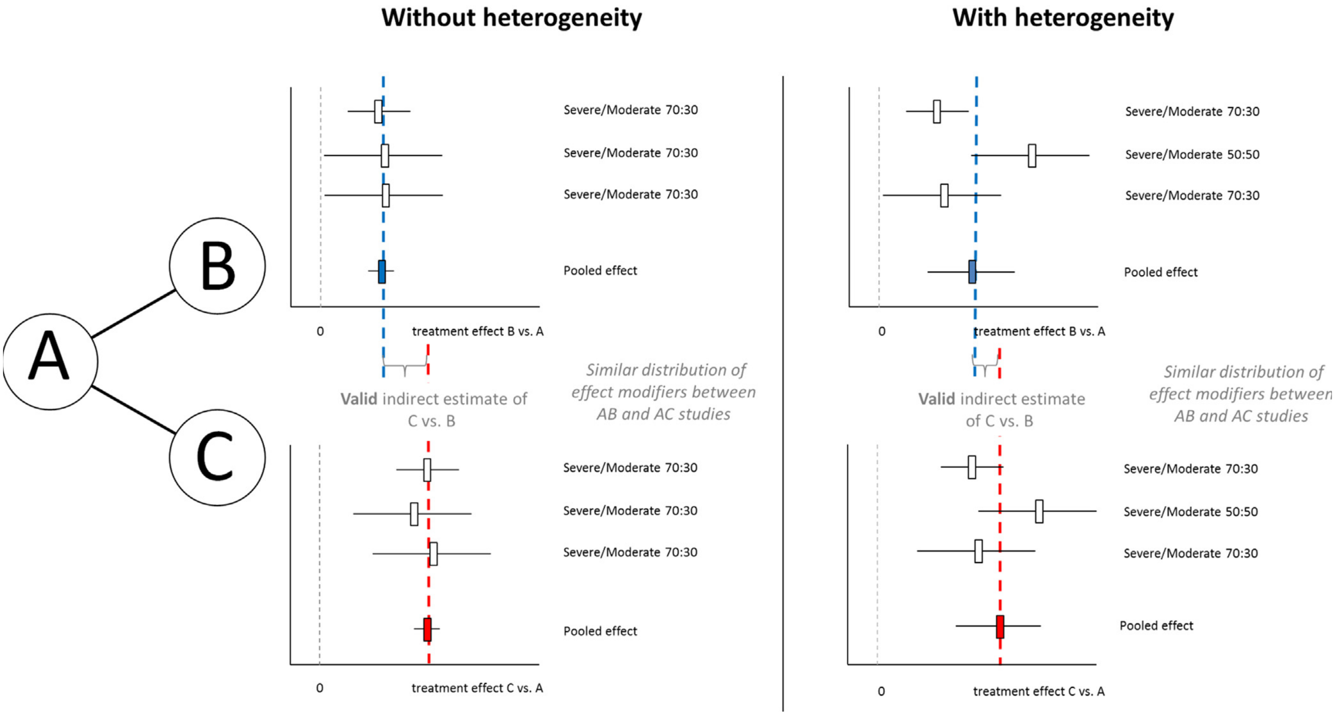

Fig. 7 – Valid network meta-analysis of AB and AC studies with and without heterogeneity: No imbalance in the distribution of the effect modifier disease severity between AB and AC comparisons.

naive indirect comparison of RCTs that does not preserve randomization is presented in Figure 9. The naive indirect comparison does not take any differences in study effects (as represented with the white boxes representing placebo) across trials into account. With RCTs available that are part of one evidence network, the naive indirect comparison can be considered a fatal flaw.

8. If both direct and indirect comparisons are available for pairwise contrasts (i.e., closed loops), was agreement in treatment effects (i.e., consistency) evaluated or discussed?
If a network has a closed loop, there is both direct evidence and indirect evidence for some treatment contrasts (Fig. 10). For

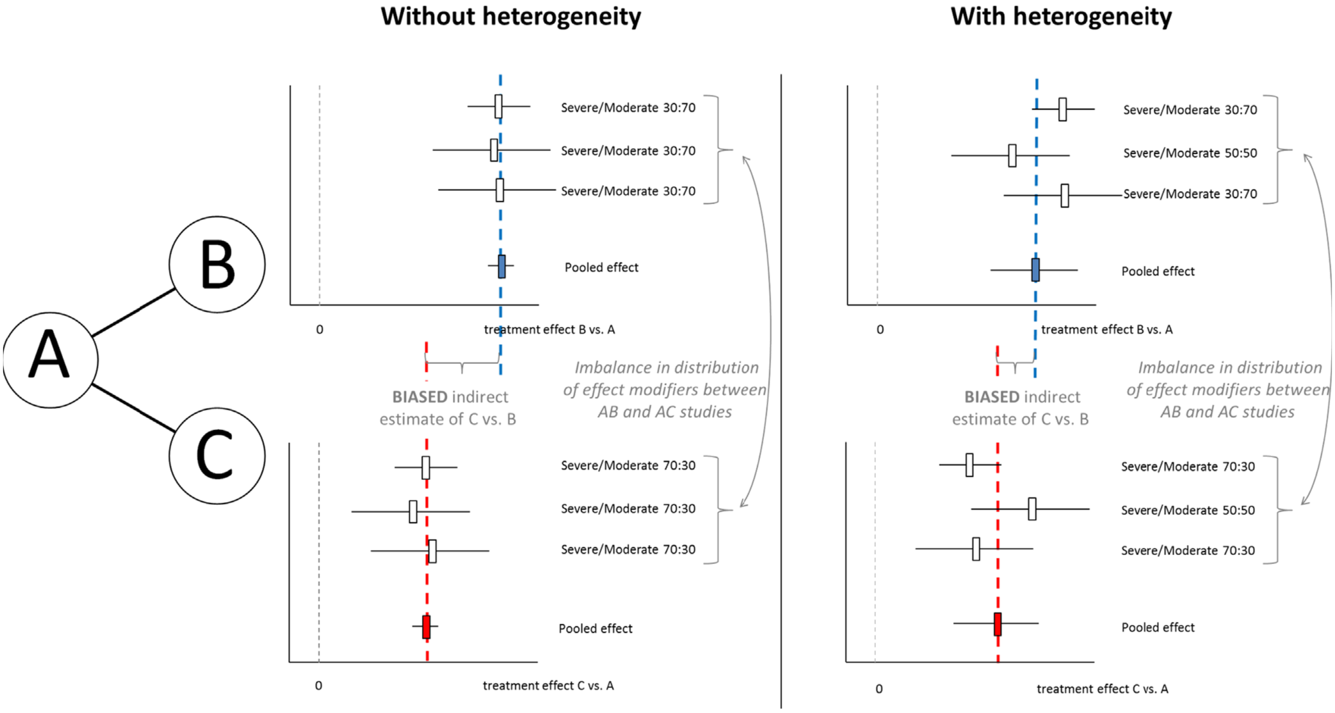

Fig. 8 – Biased network meta-analysis of AB and AC studies with and without heterogeneity: Imbalance in the distribution of the effect modifier disease severity between AB and AC comparisons.

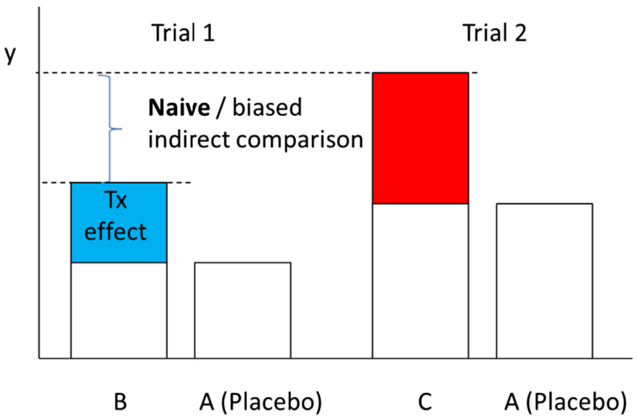

Fig. 9 – Naive indirect comparison that is invalid because differences in study effects are not acknowledged. *Note.* In this example, the difference in the treatment effect of C versus B is overestimated. Tx, treatment.

example, in an ABC network that consists of AB trials, AC trials, and BC trials, direct evidence for the BC contrast is provided by the BC trials and indirect evidence for the BC contrast is provided by the indirect comparison of AC and AB trials. If there are no systematic differences in treatment effect modifiers across the different direct comparisons that form the loop, then there will be no systematic differences in the direct and indirect estimate for each of the contrasts that are part of the loop [4,10,22]. Combining direct estimates with indirect estimates is valid, and the pooled (i.e., mixed) result will reflect a greater evidence base and one with increased precision regarding relative treatment effects. However, if there are systematic differences in effect modifiers across the different direct comparisons of the network loop, the direct estimates will be different from the corresponding indirect estimates and combining these may be inappropriate (Fig. 11). Hence, it is important that in the presence of a closed loop any direct comparisons are compared with the corresponding indirect comparisons regarding effects size or distribution of treatment effect modifiers. However, statistical tests for inconsistency should not be overinterpreted and should include knowledge of the clinical area.

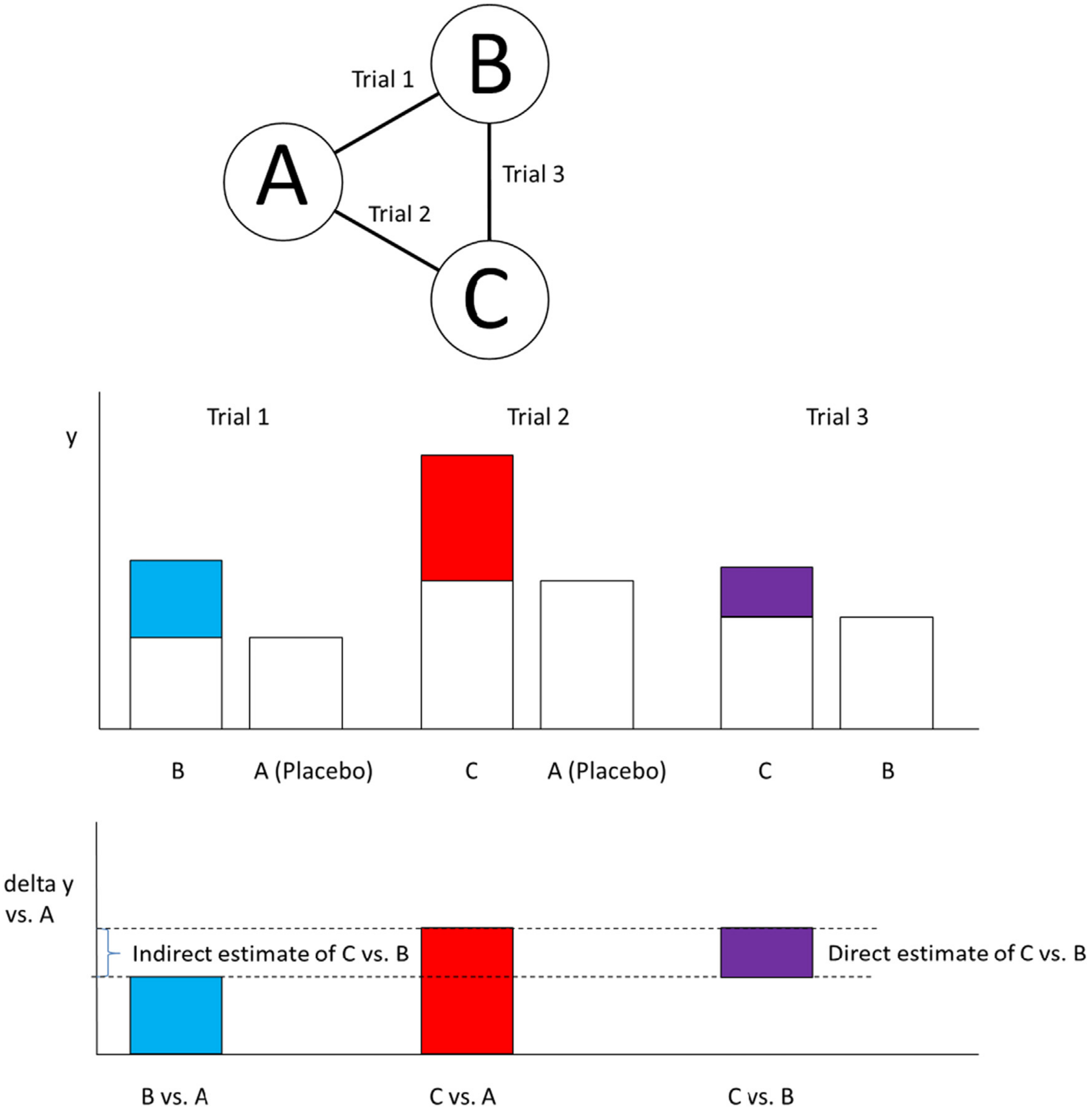

Fig. 10 – Example of a closed loop network that exhibits consistency between direct and indirect comparisons.

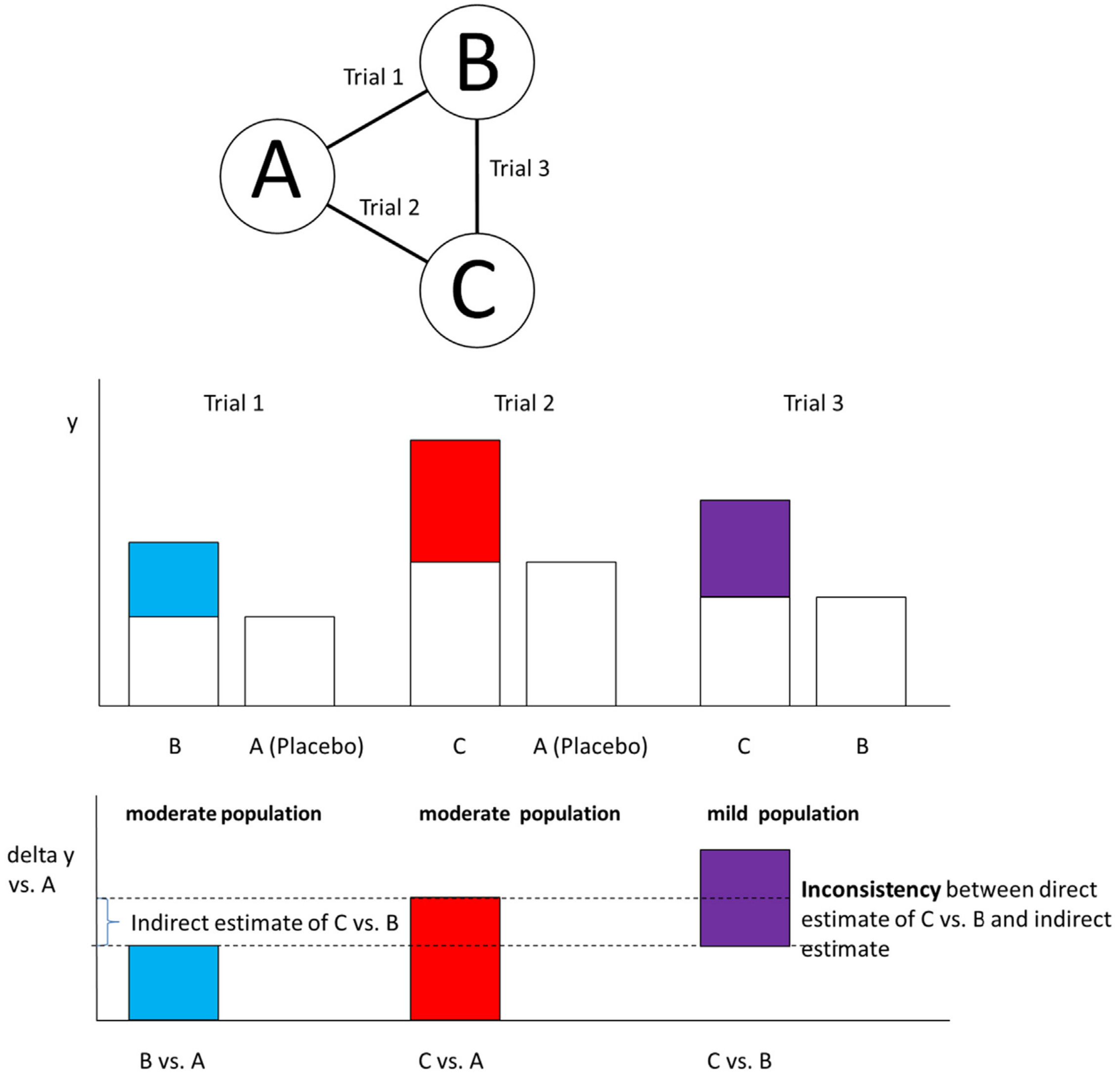

Fig. 11 – Example of a closed loop network that exhibits inconsistency between direct and indirect comparisons.

9. In the presence of consistency between direct and indirect comparisons, were both direct and indirect evidence included in the network meta-analysis?

If there is a closed loop in an evidence network, the relative treatment effect estimates obtained with direct comparisons are comparable to those obtained with the corresponding indirect comparisons, and there is no (substantial) imbalance in the distribution of the effect modifiers, then it is of interest to combine results of direct and indirect comparisons of the same treatment contrast in a network meta-analysis. As a result, the pooled result will be based on a greater evidence base with increased precision for relative treatment effects than when only direct evidence for the comparison of interest would be considered [4,5,10,12].

10. With inconsistency or an imbalance in the distribution of treatment effect modifiers across the different types of comparisons in the network of trials, did the researchers attempt to minimize this bias with the analysis?

In general, if there is an imbalance in the distribution of effect modifiers across the different types of direct comparisons, transitivity (e.g., if C is more efficacious than B, and B more efficacious than A, then C is more efficacious than A) does not hold and the corresponding indirect comparison is biased and/or there is inconsistency between direct and indirect evidence (Figs 8 and 11).

If there are a sufficient number of studies included in the network meta-analysis, it may be possible to perform a meta-regression analysis in which the relative treatment effect of each study is a function of not only a treatment comparison of that study but also an effect modifier [23–27]. In other words, with a meta-regression model we estimate the pooled relative treatment effect for a certain comparison on the basis of available studies, adjusted for differences in the level of the effect modifier between studies. This will allow indirect comparisons of different treatments for the same level of the covariate, assuming that the estimated relationship between effect modifier and treatment effect is not greatly affected by ecological bias (see Fig. 12) with a relationship at the aggregated study level not necessarily equating to the true relationship found in the individual patient data. For example, if there are differences in the proportion of subjects with severe disease among trials and disease severity affects the efficacy of at least one of the compared interventions, a meta-regression analysis can be used to adjust the indirect comparison

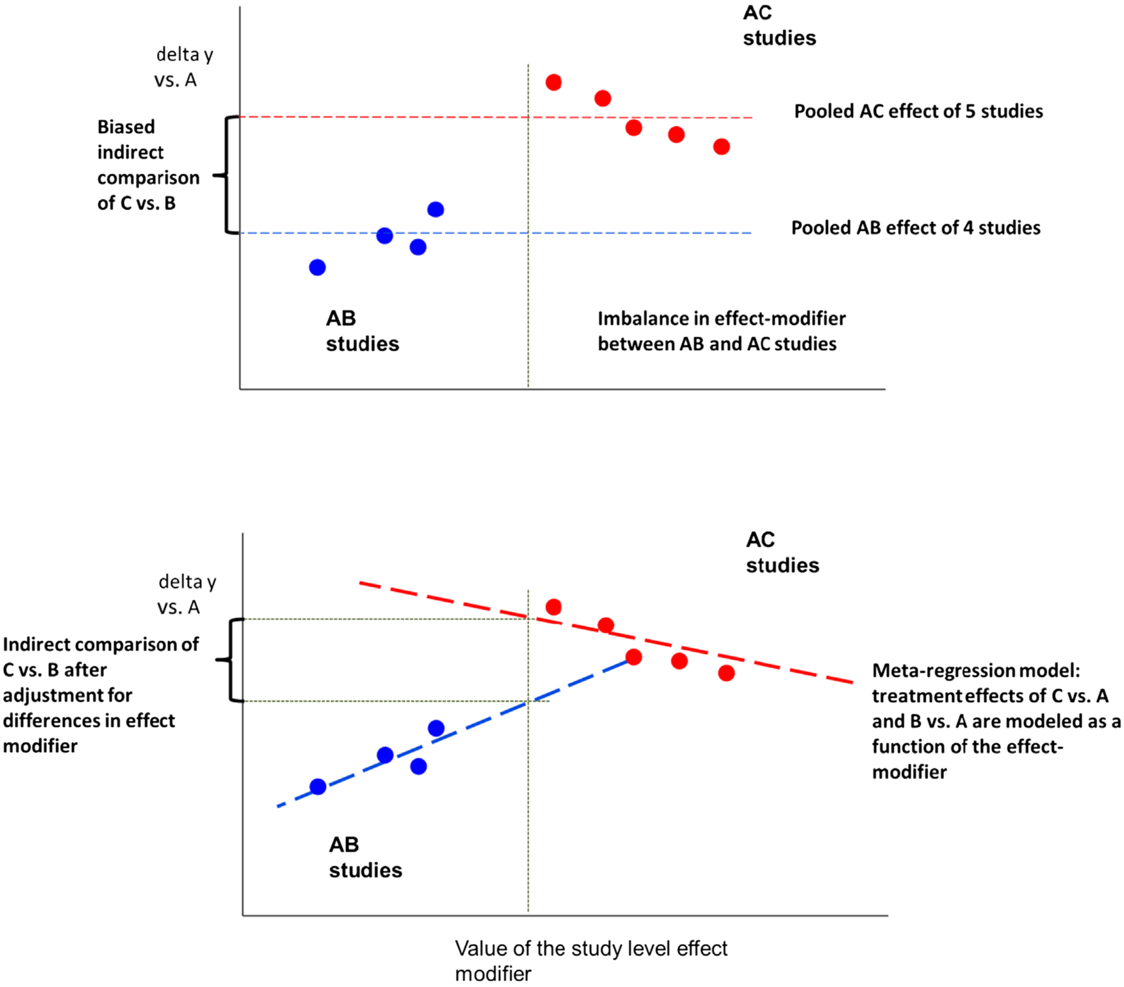

Fig. 12 – Network meta-analysis without and with adjustment for imbalance in effect modifier using a meta-regression model. *Note*. With this approach, the treatment effects for AC and AB depend on the value of the effect modifier.

estimates for the differences in the proportion of severe disease. In addition, the model can predict indirect treatment comparison estimates for different proportions of patients with severe disease in a population.

A challenge with meta-regression analysis is the low power that depends on the number of studies. Having access to patient-level data for a subset of the evidence base improves parameter estimation of meta-regression network meta-analysis models [28–30]. With these models, it is important to understand that we aim to explain the effect of modifiers of the relative treatment effects; given the network, there is no need to adjust for differences in the study effects (or prognostic factors in general) between RCTs. That being said, one can use the placebo response of a study as the covariate to adjust for any inconsistency, but this relies on the assumption that study and patient characteristics (such as severity of illness) that are effect modifiers of the relative treatment effect are also prognostic factors of the outcome with placebo [22]. However, this assumption is not trivial to examine and requires appropriate model specification.

As an alternative to a meta-regression analysis, researchers can also attempt to use models with so-called inconsistency factors [31,32]. These network meta-analysis models explicitly acknowledge any difference between direct and indirect relative treatment effect estimates of two interventions and thereby are less prone to bias estimates when direct and indirect estimates are pooled. However, the interpretation of the treatment effects obtained with these models is not useful for decision making.

In the absence of inconsistency and absence of differences in effect modifiers between different types of direct comparisons, this question should be answered with a “yes”. If there are inconsistencies or systematic differences in effect modifiers across comparisons, this question will also be answered with “yes” if meta-regression models that are expected to explain or adjust for the consistency or bias were used. (Yes, the researchers attempted to minimize the bias with this analysis). Of course, one needs to be aware of the risk of model misspecification, instable estimates, and ecological bias when using only study-level data. In the presence of inconsistency, but the researchers did not attempt to adjust for this, the question should be answered with “no.”

11. Was a valid rationale provided for the use of random-effects or fixed-effect models?

Most, if not all, meta-analyses include studies that are clinically and methodologically diverse, and thus one expects that between-study heterogeneity in treatment effects will be present. Thus, we generally advocate using a random-effects model to combine data [33]. A random-effects model assumes that each study has its own true treatment effect, because study characteristics and the distribution of patient-related effect modifiers differ across studies. The study-specific true effects are then assumed to follow a distribution around an overall mean (the meta-analysis mean), and with a variance (between-study heterogeneity) that reflects how different the true treatment effects are between them. Especially for the network meta-analysis case, several random-effect model variants have been proposed. In contrast, a fixed-effect (equal-effect) model assumes that the true treatment effect is common in all studies comparing the same treatments. This implies that there are no effect modifiers, or that they have the same distribution across all studies in the meta-analysis. The plausibility of model assumptions should guide model choice, and, in general, we deem that the assumptions of random-effects models are much more plausible than of fixed-effect models.

Often, model fit criteria are invoked for choosing between models, where the model with the better trade-off between fit and parsimony (the fixed-effect model being the most parsimonious) is preferred. However, model fit is not a proxy for the plausibility of model assumptions. Model fit criteria have a role in choosing between variants of random-effects models, but even in these cases they represent an operationalization. Solely relying on a statistical test for homogeneity to argue the use of the fixed-effect model instead of the random-effects model cannot be considered sufficient, either.

An argument for the use of a fixed-effect model instead of a random-effects model should include a judgment about the similarity of studies according to important effect modifiers and the prior belief, based on experience with the relevant clinical field, that the intervention is likely to have a fixed relative effect irrespective of the populations studied.

If it is technically not feasible to estimate the heterogeneity parameter of a random-effects model, which is, for example, the case in a star-shaped evidence network with only one study for each direct comparison, one may have used a fixed-effect model. However, it is important that the effect of ignoring heterogeneity on the findings is acknowledged. In such a situation, it is arguably still preferable to use a random-effects model and make assumptions about the extent of heterogeneity (i.e., assuming a value for the heterogeneity parameter).

12. *If a random-effects model was used, were assumptions about heterogeneity explored or discussed?*

With random-effects models, the between-study variation in treatment effects for the direct comparisons is explicitly taken into consideration. In a network meta-analysis, variants of the random-effects model exist. Two common variants differ in their assumptions about between-study heterogeneity for each comparison among treatments. One assumes that between-study heterogeneity is the same for all comparisons, and another allows between-study heterogeneity to differ by comparison. Exploration or, at least, a discussion of the choice between random-effects variants is desirable.

This question is not applicable if the network meta-analysis used a fixed-effect model.

13. *If there are indications of heterogeneity, were subgroup analyses or meta-regression analysis with prespecified covariates performed?*

Heterogeneity in relative treatment effects (i.e., true variation in relative treatment effects across studies comparing the same interventions) can be captured with random-effects models, but the analysis will provide the average relative treatment effect across the different levels of the responsible effect modifier(s). This finding may not be very informative for decision making, especially if there are great differences in relative treatment effects for the different levels of the effect modifiers [33]. It is more informative to estimate relative treatment effects for the different levels of the effect modifier, either with subgroup analysis or with a meta-regression analyses in which treatment effects are modeled as a function of the covariate, as illustrated in Figure 12.

Often there are a limited number of trials in a network meta-analysis, but many trial and patient characteristics may be different across studies. Deciding which covariates to include in the meta-regression models used for the network meta-analysis based on observed patterns in the data of the trials can lead to false conclusions regarding the sources of heterogeneity [23,24]. To avoid data dredging, it is strongly recommended to prespecify the potential treatment effect modifiers that will be investigated.

This question is not applicable if the network meta-analysis used a fixed-effect model or if there was no indication of between-study heterogeneity.

Reporting Quality and Transparency

The next six questions pertain to the transparency in the presentation of the evidence base and the results of the network meta-analysis. With a sufficiently transparent presentation, the credibility of the findings given the available studies can be accurately assessed and, if desired, replicated.

14. *Is a graphical or tabular representation of the evidence network provided with information on the number of RCTs per direct comparison*

To help understand the findings of a network meta-analysis, an overview of the included RCTs is required. The evidence base can be summarized with an evidence network in which the available direct comparisons are reflected with edges (i.e., connections) between the different interventions along with the number of RCTs per direct comparison. It is recommended that any trial that compares more than two interventions (i.e., more than two arms) is highlighted. With such a network, it is immediately clear for which treatment contrasts there is direct evidence, indirect evidence, or both [34]. A table in which studies are presented in the rows, the interventions in the columns, and observed results with each intervention of each study in the cells can prove informative as well.

15. *Are the individual study results reported?*

To assess the (face) validity of the results of the network meta-analysis, the individual study results need to be provided, either in the publication or in an online supplement. More specifically, presentation of the individual study results allows reviewers to compare these with the results of the network meta-analysis. It will also facilitate replication of the analysis, if desired.

16. *Are results of direct comparisons reported separately from results of the indirect comparisons or network meta-analysis?*

To judge whether the assumption of consistency between direct and indirect evidence holds, estimates of (pooled) direct comparisons can be compared with estimates obtained from the corresponding indirect comparisons [32]. However, this is not a trivial task. A more pragmatic approach is to present (pooled) direct evidence separately from results of the network meta-analysis in which direct and indirect evidence for some comparisons (i.e., presence of closed loops) are combined. Although the absence of a difference between these two sets of results does not guarantee there is no inconsistency, the opposite does hold: If the results based on direct evidence are systematically different from results based on the combination of direct and indirect evidence, then the indirect evidence has to be inconsistent with the direct evidence.

17. Are all pairwise contrasts between interventions as obtained with the network meta-analysis reported along with measures of uncertainty?

With a network meta-analysis, relative treatment effect estimates between all the interventions included in the analysis can be obtained. For decision making it is very informative when all these possible contrasts are presented. Equally important, for every relative treatment effect that is estimated, measures of uncertainty need to be presented (i.e., 95% confidence intervals (CI) or 95% credible intervals (CrI), which will be defined in the next section).

18. Is a ranking of interventions provided given the reported treatment effects and its uncertainty by outcome?

A network meta-analysis can be performed in a frequentist or a Bayesian framework. The result of a frequentist network meta-analysis comparing treatments is an estimate of the relative treatment effect along with a *P* value and 95% CI. The *P* value indicates whether the results are statistically "significant" or "nonsignificant." The *P* value reflects the probability of having a test statistic at least as extreme as the one that was actually observed assuming that the null hypotheses (e.g., no difference between treatments) are true [35]. The usefulness of the *P* value, though, is limited for decision making because it does not provide information on the probability that a hypothesis (e.g., one treatment is better than the other) is true or false. Moreover, when the decision maker is faced with a choice between more than two treatments, the interpretation of a *P* value associated with each pairwise comparison in a network meta-analysis becomes even more difficult because it does not provide information about the ranking of treatments. The 95% CIs corresponding to the effect estimate as obtained within a frequentist framework cannot be interpreted in terms of probabilities either; the 95% CI does not mean that there is a 95% probability that the true value is between the boundaries of the interval.

Within the Bayesian framework, the belief regarding the treatment effect size before looking at data can be conveyed with a probability distribution: the prior distribution. This probability distribution will be updated after having observed the data, resulting in the posterior distribution summarizing the updated belief regarding the likely values for this effect size [36]. The output of the Bayesian network meta-analysis is a joint posterior distribution of all relative treatment effects between interventions included in the network. The posterior distribution for each relative treatment effect can be summarized with a mean or median to reflect the most likely value for the effect size, as well as the 2.5th and 97.5th percentile: the 95% CrI. In Figure 13, the output obtained with a frequentist and Bayesian network meta-analysis is summarized. In the frequentist framework, we obtain relative treatment effects of each intervention relative to a control, along with a 95% CI (and *P* value). In the Bayesian framework, however, we obtain posterior probability distributions summarizing the likely values for the treatment effect of each intervention relative to a control, which are typically reported as a "point estimate" and a 95% CrI.

Because the posterior distribution is a probability distribution, it allows for probability statements. Unlike the 95% CI obtained with the frequentist framework, the 95% CrI can be interpreted in terms of probabilities: there is a 95% chance that the true effect size falls between the boundaries of the CrI. Consequently, in a network meta-analysis fitted within a Bayesian framework, the multiple inferences based on CIs or *P* values can be replaced with probability statements (see Fig. 14). For example, " there is *x*% probability that treatment C is better than B," or "there is a *y*% probability that treatment D is the most efficacious out of treatment A, B, C, D, and E regarding this outcome," or "there is *z*% probability that intervention E is the least efficacious" [37].

For each outcome of interest, the probability that each treatment ranks first, second, third, and so on out of all interventions compared can be called rank probabilities and are based on the location, spread, and overlap of the posterior distributions of the relative treatment effects. Rank probabilities can be summarized with a graph in which on the horizontal axis the rank from 1 to the number of treatments in the analysis is provided, and the vertical axis reflects a probability. Now, for each treatment the probability against the rank is plotted and these dots connected by treatment: a rankogram as illustrated in Figure 15 [37]. Alternatively, the ranking probabilities can be presented with bar charts. Note that solely presenting the probability of being the best can result in erroneous conclusions regarding the relative ranking of treatments because interventions for which there is a lot of uncertainty (i.e., wide CrI) are more likely to be ranked best. The benefit of having rank probabilities is that these "summarize" the distribution of effects, thereby acknowledging both location and uncertainty. Alternative summary measures of rank probabilities, such as the surface under the cumulative ranking curve, have also been proposed [37].

Technically, one can approximate the results of a Bayesian analysis that has uninformative priors using the numerical results from a model fit with maximum likelihood in a frequentist setting. This slight of hand corresponds to interpreting the likelihood function as a probability distribution and implies a Bayesian interpretation of a frequentist analysis. Thus, one could "obtain" approximations of rank probabilities even from a network meta-analysis performed in a frequentist framework, assuming noninformative priors.

19. Is the effect of important patient characteristics on treatment effects reported?

If it has been determined that certain patient characteristics are effect modifiers and differ across studies, then it is of interest to report relative treatment effects for different levels of the effect modifier as obtained with meta-regression analysis or subgroup analyses. Factors of interest can include sex, severity of disease, distribution of biomarkers, and treatment history, for example.

Interpretation

20. Are the conclusions fair and balanced?

If the conclusions are in line with reported results of the network meta-analysis, the available evidence base, credibility of the analysis methods, and any concerns of bias, then the conclusions can be considered to be fair and balanced.

Conflict of Interest

21. Were there any potential conflicts of interest?

Conflicts of interest may exist when an author (or author's institution or employer) has financial or personal relationships or affiliations that could affect (or bias) the author's decisions, work, or manuscript.

22. If yes, were steps taken to address these?

To address potential conflicts of interest, all aspects of conflicts of interest should be noted (including the specific type and relationship of the conflict of interest), and the publication should be peer reviewed. The contribution of each author should be clearly noted to document full disclosure of activities. Also, a fair and

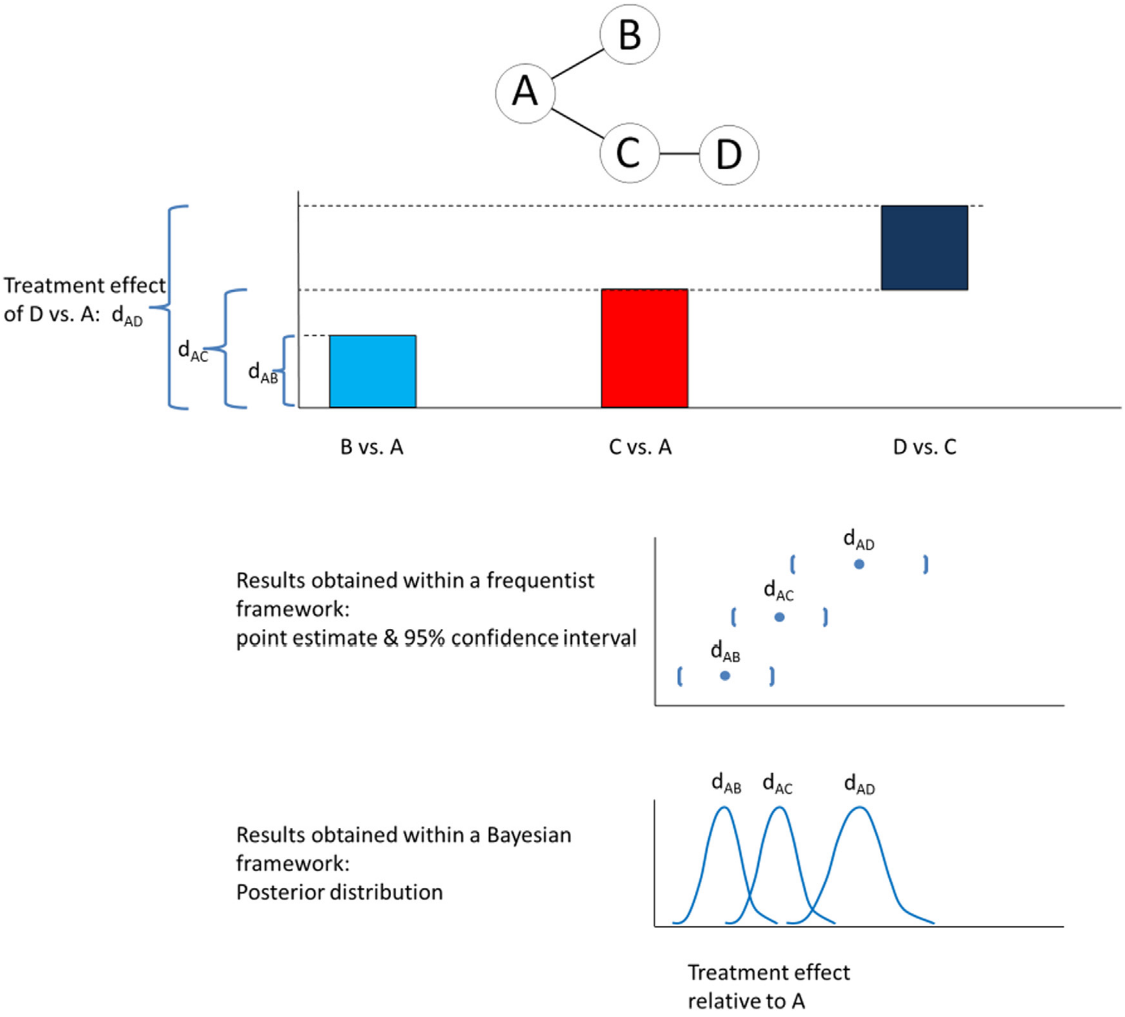
B
A
C
D
Treatment effect of D vs. A: d_{AD}
d_{AC}
d_{AB}
B vs. A
C vs. A
D vs. C
Results obtained within a frequentist framework: point estimate & 95% confidence interval
d_{AD}
d_{AC}
d_{AB}
Results obtained within a Bayesian framework: Posterior distribution
d_{AB}
d_{AC}
d_{AD}
Treatment effect relative to A

Fig. 13 – Frequentist versus Bayesian output of a network meta-analysis.

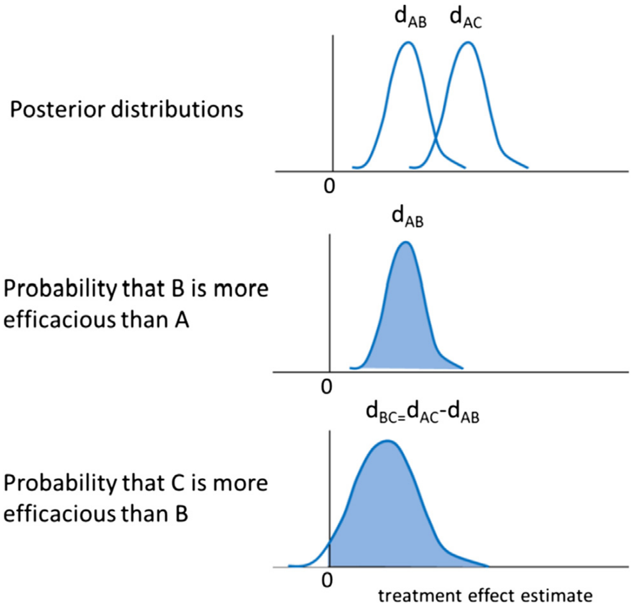
d_{AB}
d_{AC}
Posterior distributions
0
d_{AB}
Probability that B is more efficacious than A
0
$d_{BC=}d_{AC}$-d_{AB}
Probability that C is more efficacious than B
0
treatment effect estimate

Fig. 14 – Probabilistic interpretation of posterior distribution with Bayesian framework.

balanced exposition, including the breadth and depth of the study's limitations, should be accurately discussed.

Discussion

The objective of this study was to develop a questionnaire to help evidence evaluators form their opinions on the relevance and credibility of a network meta-analysis to help inform health care decision making. Relevance has to do with the extent to which the network meta-analysis is applicable to the problem faced by the decision maker. Credibility has to do with the extent to which the findings of the network meta-analysis are valid and trustworthy. The questionnaire also has an educational purpose: to raise awareness that evidence evaluation can be challenging and important elements may not be obvious to all potential users. Furthermore, the questionnaire may provide guidance to researchers when performing a network meta-analysis.

The developed questionnaire, building upon earlier work [13,16,17], assists evidence evaluators in applying a structured and consistent approach. The questionnaire does not determine an overall impression or summary score. Although some may be interested in such scores to facilitate overall evidence synthesis, the use of such scores can be misleading. In addition, the applicability of a study may depend on whether there is any other evidence that addresses the specific issue or the decision

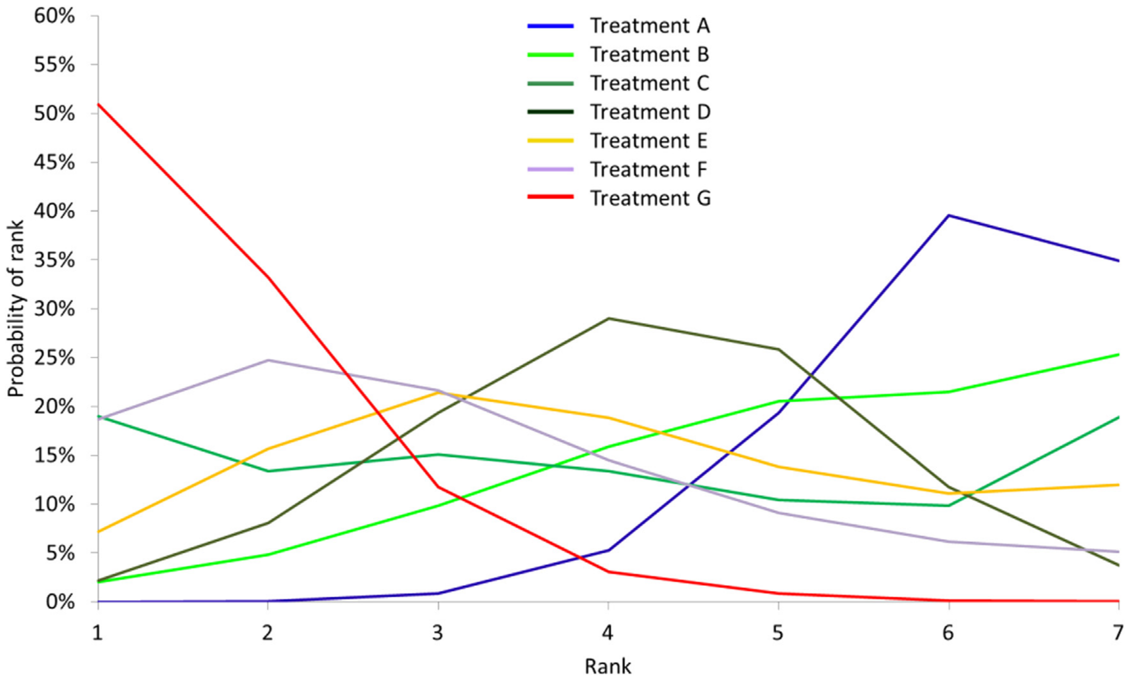

Fig. 15 – "Rankograms" showing the probability for each treatment to be at a specific rank in the treatment hierarchy.

being made. In general, an evidence evaluator needs to be aware of the strengths and weaknesses of each piece of evidence and apply his or her own reasoning.

Our questionnaire is reminiscent of the Grading of Recommendations, Assessment, Development and Evaluation (GRADE) Working Group framework for systematic reviews or for clinical practice guidelines but has important differences [38]. The GRADE system aims to characterize the likelihood that a clinically important causal effect exists primarily when two treatments are compared. GRADE evaluates five dimensions that it calls "directness," "consistency," "precision," "risk of bias," and "miscellaneous issues." "Directness" in GRADE is similar to our relevance domain, and "risk of bias" is subsumed by our credibility domain. In GRADE, "precision" is related to the decision maker's confidence that a clinically important true effect exists and "consistency" refers to the congruence of the various sources of evidence that are being considered by the decision maker. However, a formal decision-making process (e.g., a decision analysis) subsumes the assessments of "precision" and "consistency" by calculating expectations for the decision-relevant quantities and by propagating uncertainty. Thus, we deemed that the appropriate scope for the questionnaire is to include only the relevance and credibility domains.

User Testing

The Task Force wanted to strike a balance between simplicity and comprehensiveness to ensure a questionnaire useful for the end user who is not necessarily a methodologist. Approximately 22 persons, with different levels of epidemiological, statistical, and clinical expertise, were solicited to participate in user testing. Each volunteer was asked to evaluate three published network meta-analysis, rated by the Task Force as "good quality," "medium quality," and "poor quality," respectively. The response was 82%. There were not enough users to perform a formal psychometric evaluation of the questionnaire. However, some insightful and interesting descriptive observations could be made.

We calculated multirater agreement (calculated as the percentage agreement over the response categories "yes," "no," "not applicable," "not reported," and "insufficient information" for each credibility question) as an indirect indication of the clarity of the phrasing of the credibility domain (people can legitimately disagree in their assessments/interpretations). Agreement exceeded 80% for 8 of the 22 questions. The average agreement score was 72%, with a range of 42% to 91%. As expected, the lowest agreement scores were observed for the credibility questions 5, 6, 10, 13, and 19 relating to the key concepts "effect modifiers," "inconsistency," and "heterogeneity." Agreement in responses was greater for the study the Task Force selected as a good example. Agreement about the overall credibility was 83%. The good-quality study was generally rated as sufficient with respect to relevance and credibility, while the poor-quality study was generally rated not sufficient.

These descriptive results are congruent with the Task Force's a priori expectations, and thus the current version of the questionnaire was deemed acceptable for wider release. Over time, we expect to change the content or phrasing of the questionnaire, based on feedback from users. In practice, the questionnaire does not have to be completed by a single individual. Completion by a cross-disciplinary team consisting of a clinician and a statistician can be very useful. The clinician can answer questions related to interventions, outcomes, and effect modifiers, whereas the statistician can answer questions related to analytical issues.

Educational Needs

Across many international jurisdictions, the resources and expertise available to inform health care decision makers vary widely. Although there is broad experience in evaluating evidence from RCTs, there is less experience with network meta-analysis among decision makers. ISPOR has provided Good Research Practice recommendations on network meta-analysis [14,15]. This questionnaire is an extension of those recommendations and serves as a platform to assist the decision maker in understanding what a systematic evaluation of this research requires. By understanding what a systematic structured approach to the appraisal of network meta-analysis entails, it is hoped that this will lead to a general increase in sophistication by decision makers in the use of this evidence. To that end, we anticipate additional educational efforts and promotion of this questionnaire and that it will be made available to an increasing number of health care decision makers. In addition, an interactive (i.e., web-based) questionnaire has been developed at https://www.healthstudyassessment.org/ that will facilitate uptake and support the educational goal of the questionnaire.

Conclusions

The Task Force developed a consensus-based questionnaire to help decision makers assess the relevance and credibility of meta-analysis to help inform health care decision making. The questionnaire aims to provide a guide for assessing the degree of confidence that should be placed in a network meta-analysis, and enables decision makers to gain awareness of the subtleties involved in evaluating these kinds of studies. It is anticipated that user feedback will permit periodic evaluation and modification to the questionnaires, with the ensuing improvement to it. The goal is to make these questionnaires as useful as possible to the health care decision-making community.

Source of financial support: The authors have no other financial relationships to disclose.

Supplemental Materials

Supplementary Materials accompanying this article can be found in the online version as a hyperlink at http://dx.doi.org/10.1016/j.jval.2014.01.004 or, if a hard copy of article, at www.valueinhealthjournal.com/issues (select volume, issue, and article).

REFERENCES

[1] Berger M, Martin B, Husereau D, et al. A questionnaire to assess the relevance and credibility of observational studies to inform health care decision making: an ISPOR-AMCP-NPC Good Practice Task Force Report. Value Health 2014;17: 143–156.

[2] Caro JJ, Eddy DM, Kan H, et al. A modeling study questionnaire to assess study relevance and credibility to inform health care decision making: an ISPOR-AMCP-NPC Good Practice Task Force report. Value Health 2014;17: 174–182.

[3] Higgins JPT, Green S, eds. Cochrane Handbook for Systematic Reviews of Interventions Version 5.1.0 [updated March 2011]. Baltimore, MD: The Cochrane Collaboration, 2011. Available from: http://www.cochrane-handbook.org. [Accessed February 11, 2013].

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

[5] Ioannidis JPA. Indirect comparisons: the mesh and mess of clinical trials. Lancet 2006;368:1470–2.

[6] Sutton A, Ades AE, Cooper N, Abrams K. Use of indirect and mixed treatment comparisons for technology assessment. Pharmacoeconomics 2008;26:753–67.

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

[8] Bucher HC, Guyatt GH, Griffith LE, Walter SD. The results of direct and indirect treatment comparisons in meta-analysis of randomized controlled trials. J Clin Epidemiol 1997;50:683–91.

[9] Song F, Altman DG, Glenny A, Deeks JJ. Validity of indirect comparison for estimating efficacy of competing interventions: empirical evidence from published meta-analyses. BMJ 2003;326:472.

[10] Lu G, Ades AE. Combination of direct and indirect evidence in mixed treatment comparisons. Stat Med 2004;23:3105–24.

[11] Mills E, Thorlund K, Ioannidis J. Demystifying trial networks and network meta-analysis. BMJ 2013;346:f2914.

[12] Dias S, Sutton AJ, Ades AE, Welton NJ. Evidence synthesis for decision making, 2: a generalized linear modeling framework for pairwise and network meta-analysis of randomized controlled trials. Med Decis Making 2013;33:607–17.

[13] Jansen JP, Fleurence R, Devine B, et al. Interpreting indirect treatment comparisons & network meta-analysis for health care decision making: report of the ISPOR Task Force on Good Research Practices—part 1. Value Health 2011;14:417–28.

[14] Hoaglin DC, Hawkins N, Jansen JP, et al. Conducting indirect treatment comparison and network meta-analysis studies: report of the ISPOR Task Force on Indirect Treatment Comparisons Good Research Practices—part 2. Value Health 2011;14:429–37.

[15] Jüni P, Witschi A, Bloch R, Egger M. The hazards of scoring the quality of clinical trials for meta-analysis. JAMA 1995;282:1054–60.

[16] Ades AE, Caldwell DM, Reken S, et al. Evidence synthesis for decision making, 7: a reviewer's checklist. Med Decis Making 2013;33:679–91.

[17] Donegan S, Williamson P, Gamble C, Tudur-Smith C. Indirect comparisons: a review of reporting and methodological quality. PLoS One 2010;5:e11054.

[18] Moher D, Liberati A, Tetzlaff J, Altman DG, the PRISMA Group. Preferred reporting items for systematic reviews and meta-analyses: the PRISMA statement. Ann Intern Med 2009;151:264–9: W64.

[19] Higgins JP, Altman DG, Gøtzsche, et al; Cochrane Bias Methods Group; Cochrane Statistical Methods Group. The Cochrane Collaboration's tool for assessing risk of bias in randomised trials. BMJ 2011;343:d5928.

[20] Dwan K, Gamble C, Kolamunnage-Dona R, et al. Assessing the potential for outcome reporting bias in a review: a tutorial. Trials 2010;11:52.

[21] Jansen JP, Crawford B, Bergman G, Stam W. Bayesian meta-analysis of multiple treatment comparisons: an introduction to mixed treatment comparisons. Value Health 2008;11:956–64.

[22] Jansen JP, Schmid CH, Salanti G. Direct acyclic graphs can help understand bias in indirect and mixed treatment comparisons. J Clin Epidemiol 2012;65:798–807.

[23] Thompson SG, Higgins JP. How should meta-regression analyses be undertaken and interpreted? Stat Med 2002;21:1559–73.

[24] Schmid CH, Stark PC, Berlin JA, et al. Meta-regression detected associations between heterogeneous treatment effects and study-level, but not patient-level, factors. J Clin Epidemiol 2004;57:683–97.

[25] 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;28:1861–81.

[26] Salanti G, Marinho V, Higgins JP. A case study of multiple-treatments meta-analysis demonstrates that covariates should be considered. J Clin Epidemiol 2009;62:857–64.

[27] Dias S, Welton N, Marinho VCC, et al. Estimation and adjustment of bias in randomised evidence using mixed treatment comparison meta-analysis. J Royal Stat Soc A 2010;173:613–29.

[28] Jansen JP. Network meta-analysis of individual and aggregate level data. Res Synth Meth 2012;3:177–90.

[29] Saramago P, Sutton AJ, Cooper NJ, Manca A. Mixed treatment comparisons using aggregate and individual participant level data. Stat Med 2012;31:3516–36.

[30] Donegan S, Williamson P, D'Alessandro U, Tudur Smith C. Assessing the consistency assumption by exploring treatment by covariate interactions in mixed treatment comparison meta-analysis: individual patient-level covariates versus aggregate trial-level covariates. Stat Med 2012;31:3840–57.

[31] Lu G, Ades AE. Assessing evidence inconsistency in mixed treatment comparisons. J Am Stat Assoc 2006;101:447–59.

[32] Dias S, Welton NJ, Caldwell DM, Ades AE. Checking consistency in mixed treatment comparison meta-analysis. Stat Med 2010;29:932–44.

[33] Ades AE, Lu G, Higgins JP. The interpretation of random-effects meta-analysis in decision models. Med Dec Making 2005;25:646–54.

[34] Salanti G, Kavvoura FK, Ioannidis JPA. Exploring the geometry of treatment networks. Ann Intern Med 2008;148:544–53.

[35] Goodman SN. Towards evidence based medical statistics, 1: the P value fallacy. Ann Intern Med 1999;120:995–1004.

[36] Gelman AB, Carlin JS, Stern HS, Rubin DB. Bayesian Data Analysis (2nd ed.). Boca Raton: Chapman and Hall–CRC, 2003.

[37] Salanti G, Ades AE, Ioannidis JP. Graphical methods and numerical summaries for presenting results from multiple-treatment meta-analysis: an overview and tutorial. J Clin Epidemiol 2011;64:163–71.

[38] Guyatt GH, Oxman AD, Vist, et al; for the GRADE Working Group. GRADE: an emerging consensus on rating quality of evidence and strength of recommendations. BMJ 2008;336:924–6.