VALUE IN HEALTH 14 (2011) 417–428

available at www.sciencedirect.com

SCIENCE DIRECT®

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

SCIENTIFIC REPORT

Interpreting Indirect Treatment Comparisons and Network Meta-Analysis for Health-Care Decision Making: Report of the ISPOR Task Force on Indirect Treatment Comparisons Good Research Practices: Part 1

Jeroen P. Jansen, PhD[1,], Rachael Fleurence, PhD[2], Beth Devine, PharmD, MBA, PhD[3], Robbin Itzler, PhD[4], Annabel Barrett, BSc[5], Neil Hawkins, PhD[6], Karen Lee, MA[7], Cornelis Boersma, PhD, MSc[8], Lieven Annemans, PhD[9], Joseph C. Cappelleri, PhD, MPH[10]*

[1]*Mapi Values, Boston, MA, USA;* [2]*Oxford Outcomes, Bethesda, MD, USA;* [3]*Pharmaceutical Outcomes Research and Policy Program, School of Pharmacy, School of Medicine, University of Washington, Seattle, WA, USA;* [4]*Merck Research Laboratories, North Wales, PA, USA;* [5]*Eli Lilly and Company Ltd., Windlesham, Surrey, UK;* [6]*Oxford Outcomes Ltd., Oxford, UK;* [7]*Canadian Agency for Drugs and Technologies in Health (CADTH), Ottawa, ON, Canada;* [8]*University of Groningen / HECTA, Groningen, The Netherlands;* [9]*University of Ghent, Ghent, Belgium;* [10]*Pfizer Inc., New London, CT, USA*

A B S T R A C T

Evidence-based health-care decision making requires comparisons of all relevant competing interventions. In the absence of randomized, controlled trials involving a direct comparison of all treatments of interest, indirect treatment comparisons and network meta-analysis provide useful evidence for judiciously selecting the best choice(s) of treatment. Mixed treatment comparisons, a special case of network meta-analysis, combine direct and indirect evidence for particular pairwise comparisons, thereby synthesizing a greater share of the available evidence than a traditional meta-analysis. This report from the ISPOR Indirect Treatment Comparisons Good Research Practices Task Force provides guidance on the interpretation of indirect treatment comparisons and network meta-analysis to assist policymakers and health-care professionals in using its findings for decision making. We start with an overview of how networks of randomized, controlled trials allow multiple treatment comparisons of competing interventions. Next, an introduction to the synthesis of the available evidence with a focus on terminology, assumptions, validity, and statistical methods is provided, followed by advice on critically reviewing and interpreting an indirect treatment comparison or network meta-analysis to inform decision making. We finish with a discussion of what to do if there are no direct or indirect treatment comparisons of randomized, controlled trials possible and a health-care decision still needs to be made.

Keywords: Bayesian, decision making, comparative effectiveness, indirect treatment comparison, mixed treatment comparison, network meta-analysis.

Background to the task force

The ISPOR Board of Directors approved the formation of an Indirect Treatment Comparisons Good Research Practices Task Force to develop good research practices document(s) for indirect treatment comparisons in January 2009. Researchers, experienced in systematic reviews, network meta-analysis, synthesis of evidence, and related statistical methods working in academia, research organizations, the pharmaceutical industry, or government from the United States, Canada, and Europe, were invited to join the Task Force Leadership Group. Several health-care decision makers who use indirect/mixed treatment comparison evidence in health-care decisions were also invited. The Task Force met, primarily by teleconference with an ongoing exchange of email, and face to face in April 2010 to develop the topics to be addressed, agree on the outline, and draft the report. The Leadership Group determined that to adequately address good research practices for indirect treatment comparisons and the use of these comparisons in health-care decisions, the Task Force Report would comprise two papers: "Interpreting Indirect Treatment Comparisons and Network Meta-Analysis for Health-Care Decision Making: Report of the ISPOR Task Force on Indirect Treatment Comparisons Good Research Practices: Part 1" and "Conducting Indirect Treatment Comparisons and Network Meta-Analysis Studies: Report of the ISPOR Task Force on Indirect Treatment Comparisons Good Research Practices: Part 2." Summaries of the papers were presented for comment at the 15th Annual International Meeting in Atlanta, GA, USA, in May 2010. Drafts of the two papers were sent for comment to the Task Force Review Group (103 invited and self-selected individuals interested in this topic) in July 2010. The authors of the papers considered the comments from the Task Force Review Group, and the revised drafts of the two papers were sent for comment to the ISPOR membership (5550 members) in September 2010. Altogether, the Part 1 paper received 23 comments, and the Part 2 paper received 13 comments. All written comments are published on the ISPOR Web site. The authors of each paper considered all comments (many of which were substantive and constructive), revised the papers further, and submitted them to *Value in Health*.

* *Address correspondence to:* Jeroen P. Jansen, Mapi Values, 133 Portland Street, Boston, MA 02114 USA.
E-mail: jeroen.jansen@mapivalues.com.

1098-3015/$36.00 – see front matter

doi:10.1016/j.jval.2011.04.002

Introduction

The ISPOR Task Force on Indirect Comparisons Good Research Practices is publishing its report as two papers. This paper relies on Part 2 of the report [1] for the good research practices for conducting indirect treatments comparisons and network meta-analysis studies.

Systematic reviews of randomized, controlled trials (RCTs) are considered the standard basis for evidence-based health-care decision making for clinical treatment guidelines and reimbursement policies. Many systematic reviews use meta-analysis to combine quantitative results of several similar and comparable studies and summarize the available evidence [2]. Sound decision making requires comparisons of all relevant competing interventions. Ideally, robustly designed RCTs would simultaneously compare all interventions of interest. Unfortunately, such studies are almost never available, thereby complicating decision making [3–6]. New drugs are often compared to placebo or standard care, but not against each other, in trials aimed to contribute (as expeditiously as possible) to obtaining approval for drug licensing; there may be no commercial incentive to compare the new treatment to an active control treatment [5,6]. Even if there were an incentive to incorporate competing interventions in an RCT, the interventions of interest may vary by country or have changed over time due to new evidence and treatment insights. Therefore, for some indications, the 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 evidence of the difference in treatment effects among competing interventions (which otherwise would be lacking) and for judiciously selecting the best choice(s) of treatment. For example, if two particular treatments have never been compared against each other head to head, but these two treatments have been compared to a common comparator, then an indirect treatment comparison (ITC) can use the relative effects of the two treatments versus the common comparator [7–10].

Although it is often argued that indirect comparisons are needed when direct comparisons are not available, it is important to realize that both direct and indirect evidence contributes to the total body of evidence. The results from indirect evidence combined with the direct evidence may strengthen the assessment between treatments directly evaluated [3]. Even when the results of the direct evidence are conclusive, combining them with the results of indirect estimates in a mixed treatment comparison (MTC) may yield a more refined and precise estimate of the interventions directly compared and broaden inference to the population sampled because it links and maximizes existing information within the network of treatment comparisons [9].

If the available evidence consists of a network of multiple RCTs involving treatments compared directly or indirectly or both, it can be synthesized by means of so-called network meta-analysis [11]. In a 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 effect on multiple treatment comparisons for comparative effectiveness purposes. (In this report, the term comparative effectiveness is used to refer to any comparison of outcomes between interventions called relative effectiveness in European jargon) [12]. We have used comparative effectiveness and relative treatment effect without making a distinction whether the evidence base consists of RCTs designed for drug licensing (efficacy) or real-world pragmatic randomized studies (effectiveness). Network meta-analysis is about estimating relative treatment effects between competing interventions).

Given the great value of ITC and network meta-analysis for health-care decision making and its increasing acceptance (e.g., Pharmaceutical Benefits Advisory Committee in Australia, Canadian Agency for Drugs and Technologies in Health, National Institute for Health and Clinical Excellence [NICE] in the United Kingdom), this report provides practical guidance for policymakers and other health-care practitioners to enrich their understanding of these evidence synthesis methods [6,13]. We start with an overview of how RCTs of competing interventions form networks of evidence that allow multiple treatment comparisons. We then discuss the synthesis of the available evidence with a focus on terminology, assumptions, validity, and statistical methods, followed by some advice on critically reviewing and interpreting an ITC or network meta-analysis. The last section discusses what to do if there are no direct or indirect treatment comparisons of RCTs possible and a health-care decision still needs to be made.

Multiple treatment comparisons and evidence networks

Figure 1 shows networks of increasing complexity in which multiple treatments have been compared. Each node reflects an intervention, and a line connecting two nodes reflects one or more RCTs. For every intervention in a connected network, a relative treatment effect can be estimated versus another intervention. Suppose that the main comparison of interest is between intervention C and intervention B, but no direct assessment has compared them. In the first network on the left in Figure 1, intervention B has been compared to intervention A in an AB trial, and C has been compared to A in an AC trial, so an indirect comparison can estimate the relative treatment effect of C versus B. The ITC of C versus B is "anchored" on A (we favor this more descriptive term rather than "adjusted," which appears in the literature as well). A could represent an active treatment comparator or placebo. Of key importance in an ITC is not to "break randomization" [5,10,14]. For example, if A, B, and C are interventions for rheumatoid arthritis patients, it is incorrect to simply compare the observed fraction of responders on drug B in the AB trials to the observed fraction of responders on drug C in the AC trials. Using the data in this way fails to separate the efficacy of the drugs from possible placebo effects (RCTs are designed to separate drug effects from other effects). Another reason to avoid breaking randomization is that differences in response may reflect different baseline risks, even where the relative risk is consistent between trials [5,15]. Using data only from the treatment arms of interest to draw comparisons, omitting the data from the control or placebo arms, is called a "naïve indirect comparison," results in bias, and should be avoided [8]. To preserve the randomization within each trial, one must compare the relative treatment effects (e.g., compare the odds ratio for B versus A from the AB trials to the odds ratio for C versus A from the AC trials).

The second network in Figure 1 would permit an ITC of interventions B, C, D, and E, anchored on the common comparator A. Because these interventions are all connected in the network (i.e., each pair has a path from one to the other), indirect comparisons can be performed for C versus B, D versus B, E versus B, D versus C, E versus C, and E versus. D. An example of such a "star-shaped" network is a recent comparison of bisphosphonate therapies for osteoporosis in which four competing interventions were all studied in placebo-controlled trials [16]. For some of the interventions, multiple placebo-controlled trials were available, and the analysis can be labeled a network meta-analysis. Another example is the ITC of intracoronary drug-eluting stents by Biondi-Zoccai et al. [17].

In the third network, not all trials have a common comparator, but all interventions are still connected. The additional interventions F and G are connected with A, B, C, D, and E by the EF trials and the FG trials, and an indirect comparison of each intervention with any other is possible (although comparisons with longer paths will have

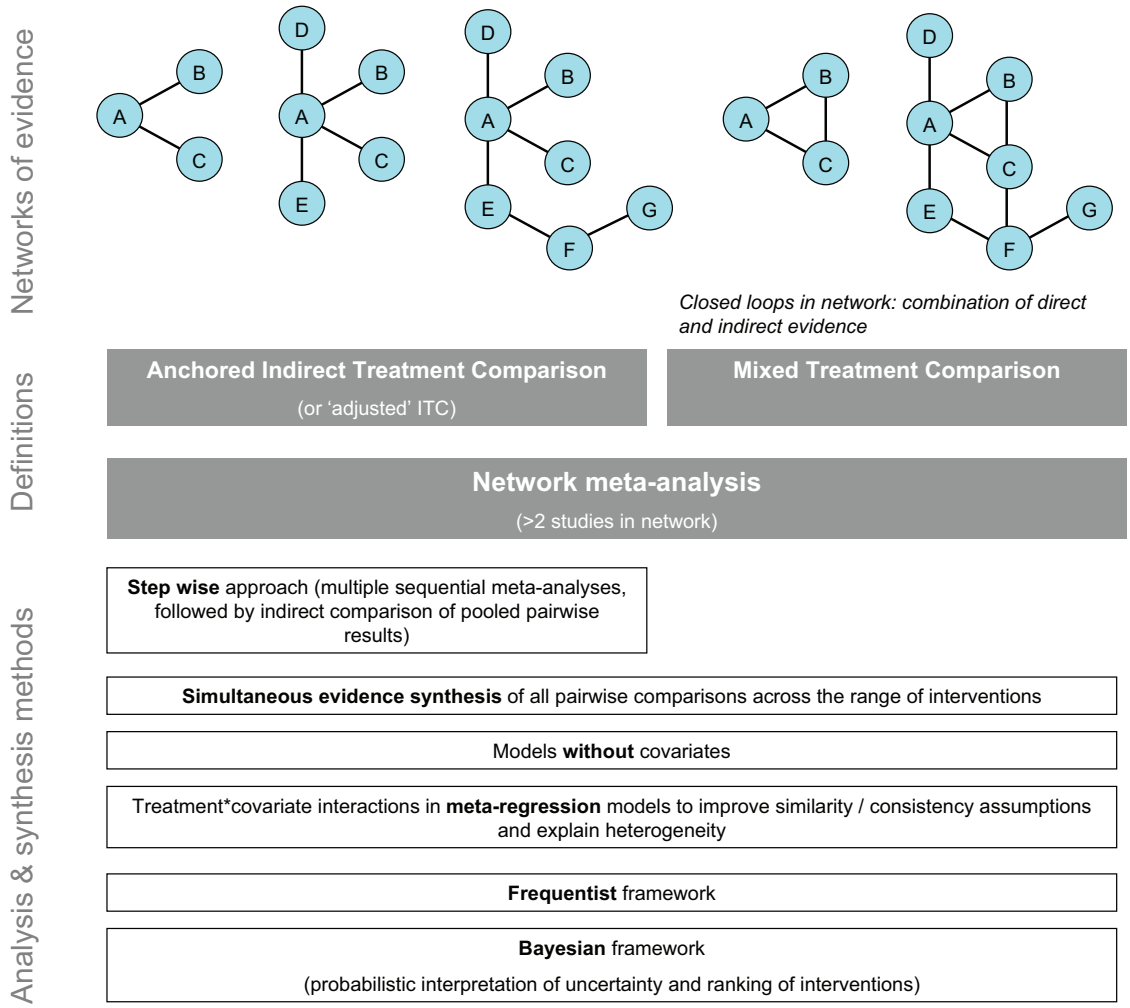

Fig. 1 – Network meta-analysis of RCTs to perform multiple treatment comparisons.

less precision) [9]. An example is the network meta-analysis of antifungal treatment for confirmed systemic fungal infections [18].

The fourth network structure consists of interventions A, B, and C (as in the first network), but now head-to-head RCT data are available for every comparison; the network of evidence consists of AB trials, AC trials, and BC trials. An important characteristic of this network is the "closed loop": each comparison has both direct and indirect evidence. For example, the BC comparison has direct evidence from the BC trials and indirect evidence from the AB and AC trials (and similarly for the AB and AC comparisons). A network in which some of the pairwise comparisons have both direct and indirect evidence is called an MTC [3,9]. A recent example of an MTC comparing three interventions is the study by Stettler et al. [19] regarding drug-eluting and bare-metal stents.

The fifth network also involves an MTC for interventions A, B, and C, but interventions A, C, E, and F form another, longer loop. For networks that contain loops, it is important that the indirect comparisons be consistent with the direct comparisons, as discussed in the next section [9,20,21]. Recent examples of network meta-analysis with loops include the network meta-analysis of first line antihypertensive therapies by Psaty et al. [22], the study of stroke prevention among patients with atrial fibrillation by Cooper et al. [23], a network meta-analysis of opioids for breakthrough cancer pain by Vissers et al. [24], and the network meta-analysis of new-antidepressants for major depression by Cipriani et al. [25]. Salanti et al. [26] provide an overview of different network structures of some of the recently published studies.

Whatever the structure of the network, pairwise comparisons, either direct or indirect or both, can be made between interventions that are connected. The terms ITC, MTC, and network meta-analysis are sometimes used interchangeably. We propose using network meta-analysis when the evidence base consists of more than two RCTs connecting more than two interventions. If the network consists of at least one closed loop, labeling the analysis an MTC is appropriate. Any analysis of an open-loop network can be called an ITC. In the remainder of this paper, we use the term network meta-analysis to refer to synthesis of a network of trials and only explicitly use ITC or MTC when it facilitates the explanation and discussion of concepts and assumptions.

Synthesis of the evidence

Assumptions

Given a network of interventions and RCTs comparing them, the objective of the analysis is to synthesize the results from the individual RCTs, thereby obtaining (pooled) estimates of relative treatment effects for pairwise comparisons. Although the comparators shared by the RCTs form the basis of the network, the key question

is whether the trials in the network are sufficiently similar to yield meaningful results for the ITC and MTC.

A traditional meta-analysis combines the results of several RCTs that compared the same interventions, say A and B, to get an overall estimate of relative effect (e.g., odds ratio, relative risk, or difference in change from baseline) and a corresponding estimate of uncertainty. It is important to realize that randomization holds within each RCT of A and B, but not across the RCTs. Thus, the trials may differ on study and patient characteristics. If these characteristics are modifiers of the relative treatment effect of B versus A, then the studies are said to be heterogeneous.

Similarly, in a network meta-analysis of RCTs involving multiple treatment comparisons, the randomization holds only within the individual trials. Relative treatment effects for a particular pairwise comparison may exhibit heterogeneity. Also, if the trials differ among the direct comparisons (e.g., AB trials differ from AC trials) and these differences are modifiers of the relative treatment effects, then the estimate of the indirect comparison is biased [8,15,21,27]. Examples of effect modifiers are patient characteristics, the way in which the outcomes are defined and/or measured, protocol requirements such as allowed cotreatment, and the length of follow up. In other words, if the distribution of interactions between relative treatment effects and covariates is not balanced across trials that are comparing different sets of interventions, the *similarity assumption* of an ITC is violated, and confounding biases the analysis [15,21]. Figure 2 depicts the comparisons involved in the similarity assumption of an ITC. If the AB trials and the AC trials are comparable in effect modifiers, then an indirect estimate for the relative effect of C versus B (d_{BC}, which can be a difference in normally distributed data, or a log odds ratio [OR], or log hazards ratio, etc.) can be obtained from the estimates of the effect of B versus A (d_{AB}) and the effect of C versus A (d_{AC}): $d_{BC} = d_{AC} - d_{AB}$. In essence, this implies that the same true d_{BC} is obtained as would have been estimated in a three-arm ABC trial [9].

When direct evidence and indirect evidence are combined for a particular pairwise comparison, it is important that the indirect estimate is not biased and there is no discrepancy between the direct and indirect comparisons [21,26,28,29]. Therefore, consistency between these direct and indirect comparisons should be accounted for. Figure 3 depicts the components involved in the *consistency assumption*. The network has both direct and indirect evidence of every pairwise comparison of interventions A, B, and C. (For example, d_{BC} can be obtained from the BC trials, but also indirectly from the AC trials and the AB trials.) For consistency, the following equation needs to be satisfied: $d_{BC} = d_{AC} - d_{AB}$ [21,28]. If there is an imbalance in modifiers of the relative treatment effects across studies for one or more of the comparisons, the consistency assumption may not be justifiable. Consistency only applies to the loops of evidence. It is not meaningful to say, for example, that the AB comparison is consistent with the AC comparison. We can only say that the AB, AC, and BC comparisons are consistent. As a simple example of inconsistency in an ABC network with an AB trial, an AC trial, and a BC trial, let us assume that the population OR (ignoring sampling error) of response with treatment B relative to A is 0.4 ($OR_{AB} = 0.4$) and the OR with C versus A is 0.5 ($OR_{AC} = 0.5$), then we would expect the OR of C versus B to be $OR_{BC} = OR_{AC}/OR_{AB} = 0.5/0.4 = 1.25$. There is inconsistency if the BC trial shows $OR_{BC} \neq 1.25$ (of course, in actual MTC analysis, there is always sampling error, and this kind of strict evaluation of consistency based on the point estimates is not appropriate. Here it aims to illustrate the inconsistency concept). In summary, heterogeneity pertains to variation in the same treatment effect among studies, whereas evidence inconsistency is the discrepancy between direct and indirect comparisons.

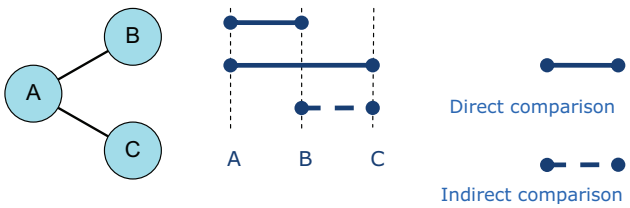

Fig. 2 – Similarity assumption in an indirect treatment comparison. AB trials and the AC trials are comparable on effect modifiers, and an unbiased indirect estimate for the relative effect of C versus B can be obtained from the estimates of the effect of B versus A and the effect of C versus A.

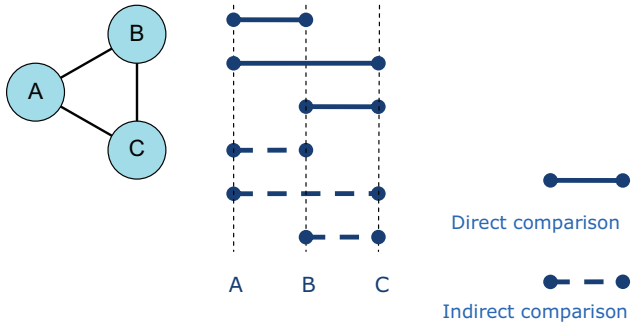

Fig. 3 – Consistency assumption in a mixed treatment comparison. AB trials, AC trials, and BC trials are comparable in effect modifiers, and for each pairwise comparison, the direct and indirect estimates are consistent.

Analysis

To synthesize the results of the different RCTs in the network and obtain relative effect estimates for all possible pairwise comparisons, an analysis method needs to be used that preserves randomization within trials and minimizes bias due to lack of randomization across trials.

In Figure 1, an overview of the analysis methods is presented and discussed in more detail here below. Whatever the method of analysis, the pooling of individual study results and indirect comparisons should be based on relative effect measures (e.g., OR, difference in change from baseline, hazards ratio) to preserve randomization. If the network does not consist of loops, the results of the RCTs available for each of the direct comparisons can be combined by means of multiple traditional meta-analyses (e.g., a meta-analysis of the AB trials and a meta-analysis of the AC trials) followed by an indirect comparison of the pooled results of each of these meta-analyses [7,11].

If the network of interventions consists of a loop, then the analysis method needs to combine estimates of the direct comparisons with estimates of the indirect comparisons. In the ABC network in which for each of the pairwise comparisons we have RCTs (network 4 in Fig. 1), the pooled relative treatment effect of the BC comparison from the BC trials needs to be combined with the indirect estimate based on the AB trials and the AC trials. The same applies to the AB and AC comparisons. It is clear that the more complex the network is, the more burdensome and potentially confusing such a stepwise approach is.

As an alternative to multiple sequential meta-analyses and indirect comparisons, a statistical model can be defined that reflects the mathematical relationships between the relative effect estimates of the direct and indirect comparisons in the complete network [9]. Given a network of A, B, and C comparisons, the relative effect estimates can be expressed as follows: $d_{BC} = d_{AC} - d_{AB}$ (assuming similarity/consistency assumptions hold). When this expression is generalized to any network with multiple different in-

terventions, the following is obtained: $d_{bk} = d_{Ak} - d_{Ab}$, with k the intervention and b the comparator. Depending on the network, k can be intervention B, C, D, E, etc. Comparator b can be A, B, C, D, etc, as long as k is alphabetically after b. This expression implies that any estimate for a particular pairwise comparison can be expressed by the relative effect estimates of the intervention and comparator relative to an overall reference treatment A as long as all interventions are connected in one network. $d_{AB}, d_{AC}, d_{AD}, \ldots, d_{Ak}$, are called *basic parameters* of the model that are estimated based on the available studies. d_{BC}, d_{BD}, d_{CD}, and so on are *functional parameters* and can be calculated based on the (pooled) estimates for the basic parameters [28]. For a network involving K treatments and T types of comparisons, there are K-1 basic parameters and $T - K + 1$ functional parameters. To summarize, a network meta-analysis model is an extension of a traditional meta-analysis model consisting of not one, but $K - 1$ parameters that need to be estimated to allow for multiple pairwise comparisons across a range of K interventions. Such a network meta-analysis model applies to networks with and without loops, i.e., both ITCs and MTCs.

Network meta-analysis can be performed with fixed- or random-effects models. With a fixed-effects model, it is assumed that there is no variation in relative treatment effects across studies for a particular pairwise comparison [15,30]. Observed differences for a particular comparison among study results are solely due to chance. For any given treatment comparison in a fixed-effects model, the following question arises: "What is the true treatment effect?" [2]. If there is heterogeneity, however—variation in true (or underlying) relative treatment effects for a particular pairwise comparison—random-effects models must be used. A random-effects approach typically assumes that true relative effects across studies are considered *exchangeable* (i.e., the prior position of expecting underlying effects to be similar but not identical) and can be described as a sample from a normal distribution whose mean is the pooled relative effect and whose SD reflects the heterogeneity [2,30–33]. One could argue that with a random-effects model, the question asked is "What is the average of the true treatment effects, and how much do these effects vary across trials?" [2]. With a random-effects model for a network meta-analysis, the variance reflecting heterogeneity is often assumed to be constant for all pairwise comparisons [9].

Although a random-effects model explicitly models heterogeneity, it does not explain heterogeneity. Extending network meta-analysis models with treatment-by-covariate interactions attempts to explain heterogeneity in relative treatment effects and estimates relative treatment effects for different levels of the covariate. As outlined previously, network meta-analysis will be biased if there are differences in covariates across those studies that are indirectly compared and act as modifiers of the relative treatment effect [15,21,27]. This implies that by taking into account these covariates with treatment-by-covariate interactions in a meta-regression model (i.e., a model that includes study-level covariates), the impact of bias due to similarity and/or consistency violations can be reduced [21]. (Covariates that vary across studies but are not effect modifiers do not need to be taken into consideration in a meta-regression model.)

Unfortunately, the number of studies in a network is often limited, and in such cases, adjustment by incorporating study-level covariates with meta-regression models may sometimes be questionable [15,34]. In addition, aggregate-level covariate adjustment might produce ecological bias, limiting the interpretation of estimated results for subgroups [34–36]. In contrast, patient-level network meta-analyses usually have sufficient power to estimate meta-regression models, thereby reducing inconsistency and providing the opportunity to explore differences in effect among subgroups. However, obtaining patient-level data for all RCTs in the network may be considered unrealistic. As an alternative, one could use patient-level data when available and aggregate-level data for studies in the network for which such data are not available, thereby improving parameter estimation over aggregate data–only models.

Because with a random-effects model the study-specific treatment effects are explicitly modeled, a random-effects model "fits" the data better than a fixed-effects model. Similarly, extending a fixed- or random-effects model by incorporating treatment-by-covariate interaction terms can also improve model fit. For any given data set, however, the more parameters that need to be estimated, the more uncertain the estimates of these parameters will be. Hence, the objective is to use a model that sufficiently fits the data (and reduces confounding bias) but that provides stable parameter estimates. The choice of a fixed- or random-effects meta-analysis model, with or without covariate interactions, can be made by comparing different competing models regarding their goodness-of-fit to the data. The goodness-of-fit can be estimated by calculating the difference between the deviance for the fitted model and the deviance for the saturated model (which fits the data perfectly). For example, the Akaike information criterion, which uses the likelihood function, the Bayesian information criterion, or deviance information criterion can all be used for model selection [37–39].

Network meta-analysis can be performed within a frequentist or Bayesian framework. With a frequentist approach, the result of the analysis is a point estimate with a 95% confidence interval (CI). The 95% CI, under repeated sampling, would contain the true population parameter 95% of the time. It must be noted that CIs obtained with a frequentist approach cannot be interpreted in terms of probabilities; the 95% CI does not mean that there is 95% probability that "true" or population value is between the boundaries of the interval [40].

Bayesian methods involve a formal combination of a prior probability distribution, which reflects a prior belief of the possible values of the model parameter of interest, with a likelihood distribution of these parameters based on the observed data, to obtain a corresponding posterior probability distribution [41]. The likelihood informs us about the extent to which different values for the parameter of interest are supported by the data [42]. The posterior distribution, as obtained with the Bayesian approach, can be interpreted in terms of probabilities (e.g., "There is an x% probability that treatment A results in a greater response than treatment B"). This is different from the interpretation of the findings within a conventional frequentist approach. To not influence the observed results by the prior distribution, an often-heard critique of the Bayesian approach, a noninformative prior distribution can be used for the treatment effect parameter(s). With such a "flat" prior distribution, it is assumed that before seeing the data, any parameter value is "equally" likely. As a consequence, posterior results are not influenced by the prior distribution but are driven by the data as in a conventional frequentist meta-analysis.

A major advantage of the Bayesian approach is that the method naturally leads to a decision framework that supports decision making [41–43]. For a network meta-analysis, a specific advantage is that the posterior probability distribution allows calculating the probability of which of the competing interventions is best and other probability statements [40]. This aspect of a Bayesian analysis is providing information that is directly relevant to health-care decision makers (e.g., policymakers and health-care professionals/clinicians). As discussed later, however, there is a risk of overinterpreting this probability. Other advantages of a Bayesian meta-analysis include the straightforward way to make predictions and the possibility to incorporate different sources of uncertainty [41,42].

Critically reviewing and interpreting a network meta-analysis

To assist health-care decision makers in using the findings of network meta-analyses, we describe in this section how to critically review and interpret such studies. The importance of correctly

assessing results of network meta-analyses cannot be overstated because these are intended to inform comparative effectiveness choices and are likely to have coverage implications. Understanding the validity of these studies is therefore critical. In the following section, we briefly review issues related to internal and external validity of network meta-analyses. We provide a list of items that we recommend be reported for a network meta-analysis to allow proper evaluation and interpretation of findings to inform decision making.

Internal and external validity

Decision makers making use of results of network meta-analyses will need to assess whether the differences between treatments are most likely true or whether they can be explained by bias in the analysis. The internal validity of the analyses is contingent on three factors: 1) the appropriate identification of the studies that make up the evidence network, 2) the quality of the individual RCTs, and 3) the extent of confounding bias due to similarity and consistency violations.

Appropriate search and selection methods of all relevant RCTs must be conducted, although the delimitation of what constitutes the evidence network is a matter of current research [44,45]. Nevertheless, even with rigorous and extensive literature search methods, the extent of publication bias must be assessed. It is well-known that negative or small trials are less likely to be published, so the evidence network may be limited accordingly [46]. Furthermore, in a network of RCTs, specific comparisons can heavily outweigh less compared interventions, resulting in asymmetrical networks [26]. The validity of a network meta-analysis will also be contingent on the internal validity of the single RCTs included in the evidence network. The inclusion of poor-quality trials may be an issue. Randomization does not guarantee that an RCT is unbiased [8,47,48]. There may be a lack of adequate allocation concealment; patients may be excluded after randomization, which may result in an imbalance between groups; or a lack of blinding of the outcome may overestimate the treatment effect [49]. Thus, each RCT included in a network meta-analysis should be critically evaluated for bias.

After addressing the threats to internal validity associated with the development of the evidence network, the similarity between the trials included in the network will also be a determinant of the internal validity of the analyses. Studies may differ with respect to the characteristics of the patients, the way in which the outcomes were measured or defined, the protocol requirements including the concomitant interventions allowed, the length of follow-up as well as differential loss to follow-up, and the time frame during which the studies were conducted [14].

As outlined earlier, a network meta-analysis is affected by confounding bias if there are differences across trials that are indirectly compared regarding relative treatment effect modifiers. This bias may be reduced by adjusting for these differences by incorporating treatment-by-covariate interactions in the statistical models used. One can only judge, however, the similarity of trials and potentially adjust for bias regarding study level covariates that are measured. Hence, differences in baseline risks and placebo responses across trials should be assessed because these can reflect additional important differences in study or patient characteristics across studies.

The external validity of the network meta-analysis will naturally be limited by the external validity of the RCTs included in the evidence network, and health-care decision makers will need to review whether results can be extrapolated to the population of interest. It is important to remember that registration trials for regulatory purposes are more likely to include selective homogeneous populations, which compromises external validity [50,51]. From a decision-making perspective, a certain degree of variation in the patient populations may be welcome for comparative and cost-effectiveness evaluations if it reflects real-world practice. Hence, some heterogeneity across trials in the evidence network may arguably increase external validity as long as the heterogeneity within direct comparisons are greater than the variation of effect modifiers across studies that are indirectly compared to avoid similarity violations as much as possible. Although we are not aware of any network meta-analysis that evaluated this explicitly, a possible approach is by means of an analysis of variance of the within versus between direct comparisons relative treatment effects.

Reporting

In Table 1, we present a simplified checklist of items that should be included in a report of a network meta-analysis to enable health-care decision makers to interpret the findings on comparative health outcomes. The checklist is not exhaustive but is intended as a general guide. Some caution should be exercised when using this list to judge the quality of published network meta-analyses because this list focuses on reporting quality and does not capture explicit items to judge or score the internal and external validity of a network meta-analysis.

In the introductory section, a clear statement of the objectives should clarify what the decision problem is, with a specific focus on the patient population and the competing interventions of interest. The treatments that will be compared in the network meta-analysis may be limited to all drugs in a class, but can also include competing drugs of different classes and, in some cases, other medical interventions. Whatever the scope of interventions, a clear rationale for the choice should be described.

In the methods section, the development of the evidence network should be described and should follow systematic review procedures that include an explicit search strategy in a variety of databases and prespecified inclusion and exclusion criteria for the study selection process. A protocol is recommended to describe these elements as well as prespecify the outcomes to be analyzed to avoid outcome selection bias [53]. Rigorous data extraction methods should be used, and authors should indicate whether double data extraction was performed, how disagreements were resolved, and how missing data were handled. These methods have been described in detail elsewhere (Centre for Reviews and Dissemination Handbook) and should be reported following the PRISMA statement [54,55].

The data analysis section should provide a comprehensive overview of the statistical methods used, including the justification of the choice of outcomes and endpoints, relative effect estimates, the choice of fixed- or random-effects models, Authors should also specify whether the models were extended with study-level covariates to improve similarity and reduce inconsistency. If the analyses were performed within a Bayesian framework, the choice of prior distributions for the model parameters should be defined. A description of different sensitivity analyses pertaining to studies included in the networks and prior distributions (if applicable) should be reported.

It is not the mandate of the Task Force to be prescriptive in recommending elements to be reported in the results section. Nevertheless, we recommend that, at a minimum, the elements in the following section be reported for users of a network meta-analysis to be able to judge the internal validity of the analyses.

A list of the studies identified by the systematic review and those included in the network meta-analysis should be provided. In some instances, these will differ if there were insufficient data reported in particular studies to include in the actual analysis. A flow diagram that illustrates the way in which trials were selected can be helpful. The reader is referred to the PRISMA statement for specific recommendations on how to report the results of a systematic review [54]. A list of key patient and study characteristics of each study should be provided in table format. This is essential to judge whether there are differences across trials that might act

Table 1 – Simplified checklist to assist decision makers in evaluating a reported network meta-analysis.

Report section	Checklist Item	What to look for in the paper
Introduction	Are the rationale for the study and the study objectives stated clearly?	A clear rationale for the review A clear objective or research question that pertains to the network meta-analysis
Methods	Does the methods section include the following? Description of eligibility criteria Information sources Search strategy Study selection process Data extraction (validity/quality assessment of individual studies)	A systematic review of the literature in accordance with Centre for Reviews and Dissemination (CRD) guidelines and PRISMA [52,54,55]
	Are the outcome measures described?	Justification of outcome measures selected for analysis
	Is there a description of methods for analysis/synthesis of evidence? Do the methods described include the following? Description of analyses methods/models Handling of potential bias/inconsistency Analysis framework	Description and justification of statistical model(s) used: multiple meta-analysis of pairwise comparisons vs. network meta-analysis models; fixed- vs. random-effects models; models without or with covariate (interactions) Description of whether analyses were performed with a frequentist or Bayesian approach Description of how possible bias/inconsistency was evaluated (either qualitative or quantitative, e.g., comparison of direct evidence with the indirect evidence). If meta-regression models are used, rationale for selection of covariates in models Description of relative-effect estimates used for presentation of findings (e.g., odds ratio, relative risk, hazard ratio, difference in change from baseline) Description of whether relative-effect measures were transformed into expected (absolute) outcomes (e.g., proportion of responders)
	Are sensitivity analyses presented?	Rationale for and description of sensitivity analyses Studies included Prior distributions for model parameters in Bayesian framework
Results	Do the results include a summary of the studies included in the network of evidence? Individual study data? Network of studies?	Description of results of study identification and selection process Table/list of studies with information regarding study design and patient characteristics (that might act as effect modifiers); these are important to judge potential similarity/consistency issues Figure of network of studies Table with raw data by study and treatment as used for the analysis/model. (Optionally present relative effects of available direct comparisons of each study)
	Does the study describe an assessment of model fit? Are competing models being compared?	Justification of model results
	Are the results of the evidence synthesis (ITC/MTC) presented clearly?	Table/ figure with results for the pairwise comparisons as obtained with analyses; Point estimates and measure of uncertainty (95% CIs) In Bayesian framework, probability to reflect decision uncertainty (i.e., probability of which treatment is best if multiple treatments are being compared and probability that one treatment is better than the comparator)
	Sensitivity/scenario analyses	Description of (different) findings with sensitivity/scenario analysis
Discussion	Does the discussion include the following? Description/summary of main findings Internal validity of analysis External validity Implications of results for target audience	Summary of findings Internal validity (individual trials, publication bias, differences across trials that might violate similarity and consistency assumptions) Discussion regarding generalizability of findings (given patient population within and across trials in network.) Interpretation of results from a biological and clinical perspective

as effect modifiers, thereby causing bias in the analysis. For example, differences in patient age, length of time with a disease, or history of treatment may constitute effect modifiers. Also, the geographic regions where the studies were performed may reflect additional differences between patient populations not reflected in reported patient characteristics. A graphic representation of the

evidence network with labels of the different RCTs can be helpful and will improve transparency of the analyses.

Point estimates and the corresponding measures of uncertainty should be reported for each (treatment arm) of the individual trials. Although the network meta-analysis uses the relative-effect measures of the different trials, outcomes by treatment arm for the individual studies provide important information. As illustrated in the example in Table 2 (based on Cipriani et al. [25]), this facilitates the understanding of the network and provides a comparison of the common reference treatment (or placebo) outcomes across trials, which may help assess key differences among the trials. Presenting relative treatment effects for each of the RCTs in a table or a figure such as a forest plot is also helpful and allows comparisons between the pooled results of the network meta-analysis and the individual study results.

In the section of the report where the results of the network meta-analysis are presented, competing models should be compared in terms of their goodness-of-fit to the data, and residual deviance calculations may be provided to justify the study's choice of the base case model. As a minimum, the estimates of relative treatment effects (e.g., ORs, hazard ratios, differences in means) along with 95% CI or credible intervals (depending on the framework of analysis) compared to a common reference treatment or anchor should be reported (Table 3). In order to appreciate the value of a network meta-analysis, it is recommended that results of all (relevant) pairwise comparisons (as a reflection of the functional parameters) are presented as well (Table 4). Forest plots can be very informative in presenting pairwise comparisons, as illustrated by Vissers et al. [24]. (Comment: Although the data of Cipriani et al. [25] were used to illustrate how source data and results of a network meta-analysis can be presented, we like to point out that the use of these data does not imply endorsement of the findings by the ISPOR Task Force.)

It may sometimes be useful to decision makers to report estimates of relative treatment effect on a different scale than that used for the model analysis. For example, it may be useful to the report results from a network meta-analysis conducted on an OR as relative risks, absolute risk differences, and numbers needed to treat. These estimates will depend on the estimated probability of response for the reference treatment. Analyses using the (Bayesian) statistical software package WinBUGS facilitate the derivation of estimates of relative treatment effects on different scales [56].

If the analyses are performed within a Bayesian framework, the uncertainty in the relative-effect estimates can be translated into probabilities of decision uncertainty. For example, the OR along with the 95% credible intervals for each of the interventions relative to a common anchor enables the calculation of the probability that each treatment is the most efficacious out of all treatments compared. For example, in Table 3, there is a 39.2% probability that escitalopram shows the greatest acceptance (i.e., lowest dropout rate) out of 12 antidepressants compared (before considering the available evidence, each treatment would have an a priori change of 100%/12 = 8.3%). Although this illustrates an important advantage of the use of the Bayesian framework, caution should be applied when only the probability of a treatment being best or ranked first is provided. This is because information of the "spread" of rankings for a treatment is also important. For example, a treatment for which there are few trial data and consequently a wide CI may have a probability approaching 50% of being the best treatment, but may nevertheless have a probability of 50% of being the worst treatment. It is therefore also useful to calculate the expected ranking of efficacy for all treatments based on the probabilities of all treatment rankings (i.e., probability of being the best, probability of second best, and so on), as illustrated in Table 3 [25].

In addition to estimates of relative treatment effects, it may be useful to report estimates of the absolute probability of outcome for binary outcomes. This will require an estimate of the baseline probability for the anchor treatment. This may be derived from the trial data or other sources and may be subject to sensitivity analyses. The method used to estimate the baseline probability should be clearly stated. In Table 3, we report the expected dropout rate based on the results of the network meta-analysis by Cipriani et al. in combination with a fixed-effects estimate for the dropout rate with fluoxetine as a reference.

The discussion section of a report should present a critical assessment of the results with respect to internal and external validity. Authors should provide a thoughtful discussion of the assumptions of similarity and consistency and whether these can be assumed to hold for the analysis at hand. The discussion should also address whether the network meta-analysis results are in line with expectations based on previous meta-analyses and other (observational) evidence available [57]. Furthermore, an explanation for observed differences between compared interventions from both biological and clinical perspectives is recommended. Apart from the appropriateness of the results, the relevance of the findings for real-world clinical and reimbursement decision making should be addressed.

Interpretation of findings

After assessing the validity and results of a network meta-analysis, decision makers will want to carefully consider whether the findings can be applied to their decision problems. Is one treatment better than another and do these results apply to the population of interest to the decision maker?

Frequently, a small number of studies in a network meta-analysis limits the possibility to adjust for possible bias due to similarity issues by means of statistical techniques. Rather than immediately ignoring the results of the analysis by claiming that trials are not comparable, the decision maker should make an attempt to hypothesize the possible direction of bias in the indirect estimates. An important question to ask is how different a nonbiased indirect comparison would be and whether this would lead to a different conclusion and decision.

An issue to consider is whether a treatment can be considered more effective than another when only a limited number of outcomes have been analyzed. Selection of outcomes analyzed must be clearly justified at the outset of the analysis (for example, do these reflect primary outcomes used in clinical trials?). Synthesizing the results of multiple network meta-analyses must be considered. How do we interpret situations in which drug A is better on a number of clinical outcomes but not all of these outcomes? How is the decision made in these cases? A possible approach is to weigh the different end points based on the relative importance according to the decision maker and calculate the probabilities of which treatment is best, taking into account these weights [58]. These are not issues specific to a network meta-analysis. Indeed, the development of measures such as the quality-adjusted life-year has been fueled by the need to compare disparate health outcomes using a common metric by means of decision models. Nevertheless, it is an issue to consider when interpreting a network meta-analysis as well.

Furthermore identification of the "best" or most appropriate treatment cannot be made on the basis of efficacy end points alone. To inform health-care decision making for clinical treatment guidelines and reimbursement policies, the efficacy findings of a network meta-analysis must be interpreted in light of other available (observational) evidence and other characteristics of the competing interventions, such as safety and convenience.

The general development of evidence synthesis methods to provide systematic exhaustive evidence on which to base decisions has a natural place in comprehensive decision models that include both costs and effects and are used to determine the cost-effectiveness of interventions by such bodies as the NICE [5,59–

Table 2 – Example table how source data as used in a network meta-analysis can be presented - New-generation antidepressants for major depression, dropouts.

Study	Fluoxetine		Bupropion		Citalopram		Duloxetine		Escitalopram		Fluvoxamine		Milnacipran		Mirtazapine		Paroxetine		Reboxetine		Sertraline		Venlafaxine	
	r	n	r	n	r	n	r	n	r	n	r	n	r	n	r	n	r	n	r	n	r	n	r	n
2906/421	45	119															50	123						
29060/365	27	70															21	68						
29060/785					43	207											41	199						
Aberg-Wisted, 2000																	26	177			33	176		
Agren, 1999					8	133									18	137								
Aguglia, 1993	31	56																			17	52		
AK130939			45	204																			46	198
Akkaya, 2003																			7	57			7	50
Alves, 1999	9	47																					10	40
Amini, 2005	3	18													2	18								
Annseaau, 1993											23	64					16	56						
Annseau, 1994	18	93											23	97										
Baldwin, 2005									15	166							14	159						
Benkert, 1999															30	139	33	136						
Bennie, 1995	23	144																			24	142		
.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
Detke, 2004							21	188									10	86						
.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
Zanardi, 1996																	9	22			0	24		

For complete dataset see Cipriani et al., 2009.

r, dropouts; n, number of patients. Reprinted from The Lancet, 373 (9665), Cipriani A, Furukawa TA, Salanti G, Geddes JR, Higgins JP, Churchill R, Watanabe N, Nakagawa A, Omori IM, McGuire H, Tansella M, Barbui C, Comparative efficacy and acceptability of 12 new-generation antidepressants: a multiple-treatments meta-analysis, 746-58, 2009, with permission from Elsevier.

Table 3 – Sample table showing how results of a network meta-analysis can be presented.

	Odds ratio for acceptability*	95% Credible interval	Estimated dropout, %†	95% Credible interval	Probability of being among the four best treatments, %‡	Probability of being the best out of all compared, %	Rank
Fluoxetine	1.0	Reference	27.0		3.4	0.0	5
Bupropion	1.12	(0.92–1.36)	24.8	(21.4%–28.6%)	19.3	16.9	3
Citalopram	1.11	(0.91–1.37)	25.0	(21.4%–29.0%)	18.7	15.3	4
Duloxetine	0.84	(0.64–1.10)	30.6	(25.2%–36.5%)	0.7	0.3	10
Escitalopram	1.19	(0.99–1.44)	23.7	(20.5%–27.2%)	27.6	39.2	1
Fluvoxamine	0.82	(0.62–1.07)	31.1	(25.6%–37.1%)	0.4	0.2	11
Milnacipran	0.97	(0.69–1.32)	27.6	(21.7%–34.5%)	7.1	6.4	6
Mirtazapine	0.97	(0.77–1.21)	27.6	(23.4%–32.3%)	4.4	1.9	7
Paroxetine	0.91	(0.79–1.05)	28.9	(26.1%–31.9%)	0.2	0.0	9
Reboxetine	0.70	(0.53–0.92)	34.6	(28.7%–41.0%)	0.1	0.0	12
Sertraline	1.14	(0.96–1.36)	24.5	(21.4%–27.8%)	21.3	19.8	2
Venlafaxine	0.94	(0.81–1.09)	28.2	(25.4%–31.3%)	0.9	0.1	8

Relative treatment effect of each treatment relative to reference comparator expressed as odds ratios (with 95% credible intervals), expected outcome (with 95% credible intervals), and probability best as a measure of decision uncertainty; new-generation antidepressants for major depression, acceptability (dropouts).

Based on Cipriani et al. [25].

Reprinted from The Lancet, 373 (9665), Cipriani A, Furukawa TA, Salanti G, Geddes JR, Higgins JP, Churchill R, Watanabe N, Nakagawa A, Omori IM, McGuire H, Tansella M, Barbui C, Comparative efficacy and acceptability of 12 new-generation antidepressants: a multiple-treatments meta-analysis, 746-58, 2009, with permission from Elsevier.

* Odds ratio <1 favors fluoxetine; odds ratio reflects odds of acceptability (i.e., not dropping out).

† Reference dropout was based on fixed-effects meta-analysis of fluoxetine arms of all trials.

‡ Reported by Cipriani et al. [25].

61]. Network meta-analysis represents a valuable set of analytical tools to inform clinical evidence in cost-effectiveness analysis.

Decision making in the absence of direct and indirect treatment comparisons of RCTs

Pragmatic, randomized, naturalistic, head-to-head trials are arguably the gold standard to obtain comparative effectiveness estimates given their high internal and external validity [62]. These trials, however, take a long time to complete and can never provide relative effectiveness information for all competing interventions, especially when new treatments are developed continuously. Hence, an ITC or network meta-analysis can be considered a useful and realistic alternative. To minimize bias, a network meta-analysis requires RCT evidence. Evidence from RCTs, however, may not be available in a significant proportion of situations that decision makers face. For example, in oncology, Phase II trials often have a single arm. A review of practice guidelines found that few recommendations were based on high-quality evidence, and many were based on expert opinion, individual case studies, and standards of care [63]. There are often good reasons for the absence of RCTs. Time may be too short to conduct RCTs of rapidly emerging technologies [64]. RCTs may be unethical if clinicians believe that there is a causal association between the intervention and the outcome, for example, between sleeping position and sudden infant death. Limited resources can also be a factor for the lack of RCT evidence.

Whatever the reasons for the absence of RCTs and therefore the absence of indirect treatment comparisons based on randomized evidence, health-care payers, health-care professionals, and patients may need to make decisions. It is wrong to assume that these stakeholders can postpone the decision and wait for the "appropriate" evidence. In particular, decisions have to be made based on the available set of possible choices. In reality, not covering or prescribing an intervention is a tacit decision to stay with the status quo. This decision has societal implications. It may or may not maximize health benefits for the population (if made by the decision maker) or health benefits for the patient (if made by the health-care provider and patient) [65].

A critical question for decision makers, then, is whether to use observational comparative studies if RCTs or indirect comparisons of RCTs are not available. To answer this question, it is important to remember that in a network meta-analysis of RCTs, the value of randomization does not hold across trials. If study or patient characteristics differ among trials for interventions indirectly compared and are modifiers of relative treatment effects, the analysis will be biased. Hence, an ITC or network meta-analysis of RCTs is a form of observational evidence, but arguably less prone to confounding bias than is a cohort study (or any other observational design). A cohort study is biased if differences in unmeasured covariates affect both the intervention and the outcome, whereas an ITC or network meta-analysis of RCTs is biased only if differences in unmeasured covariates among trials are modifiers of relative treatment effects, which is arguably much more unlikely. Thus, asking whether comparative observational studies should be used in the absence of an ITC or network meta-analysis of RCTs is synonymous with asking what level of observational evidence can be considered to have sufficient internal validity to inform decision making; or, more specifically, with what level of observational evidence are decision makers comfortable? Is the minimum acceptable level of observational evidence an ITC or network meta-analysis of RCTs or is a cohort study sufficient?

To answer such questions, decision makers must recognize that the lower the internal validity, the greater the risk of biased results and therefore the greater the risk of making an inferior decision. If the new treatment is chosen over the standard treatment because of biased estimates of comparative effectiveness, and the true outcomes favor the standard treatment, then health benefits are foregone.

The debate over the proper use of RCT evidence, indirect comparison of RCTs, and "traditional" observational studies is likely to continue. Observational studies can be considered complementary evidence to RCTs [66]. Fortunately, the needs for comparative effectiveness research are driving developments in evidence syn-

Table 4 – Sample table showing how results of a network meta-analysis can be presented.

Intervention	Comparator											
	Fluoxetine	Bupropion	Citalopram	Duloxetine	Escitalopram	Fluvoxamine	Milnacipran	Mirtazapine	Paroxetine	Reboxetine	Sertraline
Fluoxetine	1										
Bupropion	1.12 (0.92–1.36)	1									
Citalopram	1.11 (0.91–1.37)	1.00 (0.78–1.28)	1								
Duloxetine	0.84 (0.64–1.10)	0.75 (0.55–1.01)	0.75 (0.55–1.02)	1							
Escitalopram	1.19 (0.99–1.44)	1.06 (0.86–1.32)	1.07 (0.86–1.31)	1.43 (1.09–1.85)	1						
Fluvoxamine	0.82 (0.62–1.07)	0.73 (0.53–1.00)	0.73 (0.54–0.99)	0.98 (0.67–1.41)	0.69 (0.50–0.94)	1					
Milnacipran	0.97 (0.69–1.32)	0.87 (0.58–1.24)	0.87 (0.60–1.24)	1.16 (0.77–1.73)	0.81 (0.55–1.15)	1.18 (0.76–1.75)	1				
Mirtazapine	0.97 (0.77–1.21)	0.87 (0.66–1.14)	0.87 (0.66–1.15)	1.16 (0.83–1.61)	0.81 (0.62–1.07)	1.18 (0.87–1.61)	0.99 (0.69–1.53)	1			
Paroxetine	0.91 (0.79–1.05)	0.81 (0.65–1.00)	0.81 (0.65–1.01)	1.08 (0.84–1.40)	0.76 (0.62–0.93)	1.1 (0.84–1.47)	0.94 (0.68–1.31)	0.93 (0.75–1.17)	1		
Reboxetine	0.70 (0.53–0.92)	0.62 (0.45–0.86)	0.62 (0.45–0.84)	0.83 (0.57–1.22)	0.58 (0.43–0.81)	0.85 (0.57–1.26)	0.72 (0.48–1.10)	0.72 (0.51–1.03)	0.77 (0.56–1.05)	1	
Sertraline	1.14 (0.96–1.36)	1.01 (0.82–1.27)	1.02 (0.81–1.28)	1.36 (1.01–1.83)	0.95 (0.77–1.19)	1.38 (1.03–1.89)	1.17 (0.84–1.72)	1.17 (0.91–1.51)	1.25 (1.04–1.52)	1.63 (1.19–2.24)	1
Venlafaxine	0.94 (0.81–1.09)	0.84 (0.68–1.02)	0.84 (0.67–1.06)	1.12 (0.84–1.50)	0.78 (0.64–0.97)	1.14 (0.86–1.54)	0.97 (0.69–1.40)	0.97 (0.76–1.23)	1.03 (0.86–1.24)	1.34 (0.99–1.83)	0.82 (0.67–1.00)

Relative treatment effect of pairwise comparisons expressed as odds ratios (with 95% credible intervals); new-generation antidepressants for major depression, acceptability (dropouts).
Based on Cipriani et al. [25]. Reprinted from The Lancet, 373 (9665), Cipriani A, Furukawa TA, Salanti G, Geddes JR, Higgins JP, Churchill R, Watanabe N, Nakagawa A, Omori IM, McGuire H, Tansella M, Barbui C, Comparative efficacy and acceptability of 12 new-generation antidepressants: a multiple-treatments meta-analysis, 746-58, 2009, with permission from Elsevier.

thesis and its understanding. At this stage, we conclude that, in the absence of (head-to-head) RCTs, decision makers can use observational evidence as long as they are aware of the potential risks in using evidence of lower quality and are comfortable taking these risks. If decision makers prefer to wait for head-to-head RCTs or, the next best thing, an ITC of RCTs, they must realize that they are choosing the "old" treatment over the "new" treatment, with potential societal implications.

In essence, for every appraisal of a new intervention, a decision maker must make the trade-off between, on the one hand, the risk of making the wrong decision and therefore losing health benefits by relying on lower quality evidence and, on the other hand, postponing the decision and therefore possibly also forgoing health benefits. This trade-off is influenced by considerations including the burden of disease and the number of currently available treatments. Whatever the outcome of the debate, the quality of decision making will be increased by being transparent and explicit about which type of evidence is being used and evaluating its limitations and consequences.

Conclusion

This report, the first part of the report from the Task Force, outlines the key concepts of ITC and MTC and provides guidance for reviewing and interpreting these studies to inform decision making. Network meta-analysis can be considered an extension of traditional meta-analysis by including multiple different pairwise comparisons across a range of different interventions to allow multiple treatment comparisons in the absence of head-to-head evidence. Furthermore, the methodology can combine direct and indirect treatment comparisons, thereby synthesizing a greater share of the available evidence than a traditional meta-analysis. Although the evidence networks underlying network meta-analysis typically include RCTs, randomization does not hold across trials and there is a risk of confounding bias, compromising internal validity. Accordingly, a network meta-analysis must be considered observational evidence, but is arguably less prone to confounding bias than an observational comparative (prospective) cohort study. Although the methodological issues regarding indirect comparisons and network meta-analysis are recognized, application of this method is expected to continue and grow, simply because failing to view accumulation of information as an evolving process would undermine the role played by scientific evidence in shaping health-care decision making. For that reason, the goal of the Task Force is to help educate policymakers and health-care professionals about these studies and identify areas for future research.

REFERENCES

[1] Hoaglin DC, Hawkins N, Jansen JP, et al. Conducting indirect treatment comparisons and network meta-analysis studies: report of the ISPOR task force on indirect treatment comparisons good research practices—Part 2. Value Health 2011;14:xx-xx.

[2] Higgins JPT, Green S, eds. Cochrane Handbook for Systematic Reviews of Interventions Version 5.0.2 [updated September 2009]. The Cochrane Collaboration, 2009. Available from: http://www.cochrane-handbook.org. [Accessed February 11, 2010].

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

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

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

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

[7] 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.
[8] 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.
[9] Lu G, Ades AE. Combination of direct and indirect evidence in mixed treatment comparisons. Stat Med 2004;23:3105–24.
[10] Song F, Loke YK, Walsh T, et al. Methodological problems in the use of indirect comparisons for evaluating healthcare interventions: survey of published systematic reviews. BMJ 2009;338:b1147.
[11] Lumley T. Network meta-analysis for indirect treatment comparisons. Stat Med 2002;21:2313–24.
[12] Working Group on Relative Effectiveness The Pharmaceutical Forum.[online]. Available from: http://ec.europa.eu/pharmaforum/effectiveness_en.htm. [Accessed December 6, 2010].
[13] Pharmaceutical Benefits Advisory Committee. Guidelines for preparing submissions to the Pharmaceutical Benefits Advisory Committee. (Version 4.3). Australian Government, Department of Health and Ageing, December 2008.
[14] Glenny AM. Statistical methods for indirect treatment comparisons. Health Technol Assess 2005;9:17–49.
[15] 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.
[16] Jansen JP, Bergman GJ, Huels J, Olson M. Prevention of vertebral fractures in osteoporosis: mixed treatment comparison of bisphosphonate therapies. Curr Med Res Opin 2009;25:1861–8.
[17] Biondi-Zoccai GG, Agostoni P, Abbate A, et al. Adjusted indirect comparison of intracoronary drug-eluting stents: evidence from a metaanalysis of randomized bare-metal-stent-controlled trials. Int J Cardiol 2005;100:119–23.
[18] Mills EJ, Perri D, Cooper C, et al. Antifungal treatment for invasive Candida infections: a mixed treatment comparison meta-analysis. Ann Clin Microbiol Antimicrob 2009;8:23.
[19] Stettler C, Wandel S, Alleman S, et al. Outcomes associated with drug-eluting and bare-metal stents: a collaborative network meta-analysis. Lancet 2007;370:937–48.
[20] Ades AE. A chain of evidence with mixed comparisons: models for multi-parameter synthesis and consistency of evidence. Stat Med 2003;22:2995–3016.
[21] 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.
[22] Psaty BM, Lumley T, Furberg CD, et al. Outcomes associated with various antihypertensive therapies used as first-line agents: a network meta-analysis. JAMA;289:2534–44.
[23] Cooper NJ, Sutton AJ, Lu G. Mixed comparison of stroke prevention treatments in individuals with nonrheumatic atrial fibrillation. Arch Intern Med 2006;166:1269–75.
[24] Vissers D, Stam W, Nolte T, et al. Efficacy of intranasal fentanyl spray versus other opioids for breakthrough pain in cancer. Curr Med Res Opin 2010;26:1037–45.
[25] Cipriani A, Furukawa TA, Salanti G, et al. Comparative efficacy and acceptability of 12 new-generation antidepressants: a multiple-treatments meta-analysis. Lancet 2009;373:746–58.
[26] Salanti G, Kavvoura FK, Ioannidis JPA. Exploring the geometry of treatment networks. Ann Intern Med 2008;148:544–53.
[27] Coory M, Jordan S. Frequency of treatment-effect modification affecting indirect comparisons: a systematic review. Pharmacoeconomics 2010;28:723–32.
[28] Lu G, Ades AE. Assessing evidence inconsistency in mixed treatment comparisons. J Am Stat Assoc 2006;101:447–59.
[29] Dias S, Welton NJ, Caldwell DM, Ades AE. Checking consistency in mixed treatment comparison meta-analysis. Stat Med 2010;29:932–44.
[30] Borenstein M, Hedges LV, Higgins JPT, Rothstein H. Introduction to Meta-Analysis. Chichester, England: John Wiley & Sons, Ltd., 2009.
[31] Skene AM, Wakefield JC. Hierarchical models for multicentre binary response studies. Stat Med 1990;9:919–29.
[32] Gelman AB, Carlin JS, Stern HS, Rubin DB. Bayesian Data Analysis. Boca Raton, FL: Chapman and Hall–CRC; 1995.
[33] Cappelleri JC, Ioannidis JPA, Lau, J. Meta-analysis of therapeutic trials. In: Chow S-C, ed., Encyclopedia of Biopharmaceutical Statistics (3rd ed.), Revised and Expanded. New York, NY: Informa Healthcare, 2010.
[34] Berlin JA, Santanna J. Schmid CH, et al. Individual patient-versus group-level data meta-regression for the investigation of treatment effect modifiers: ecological bias rears its ugly head. Stat Med 2002;21:371–87.
[35] Greenland S, Morgenstern H. Ecological bias, confounding, and effect modification. Int J Epidemiol 1989;18:269–74.
[36] Lambert PC, Sutton AJ, Jones DR. A comparison of summary patient-level covariates in metaregression with individual patient data meta-analysis. J Clin Epidemiol 2002;55:86–94.
[37] McCullagh P, Nelder J. Generalized Linear Models, Second Edition. Chapman & Hall/CRC, 1989.
[38] Dempster AP. The direct use of likelihood for significance testing. Stat Comput 1997;7:247–52.
[39] Spiegelhalter DJ, Best NG, Carlin BP, van der Linde A. Bayesian measures of model complexity and fit. J R Stat Soc (Series B) 2002;64:583–639.
[40] Goodman, SN. Towards evidence based medical statistics: 1. The P value fallacy. Ann Intern Med 1999;120:995–1004.
[41] Sutton AJ, Abrams KR, Jones DR, et al. Methods for Meta-Analysis in Medical Research. London, UK: Wiley, 2000.
[42] Spiegelhalter D, Abrams K, Myles J. Bayesian Approaches to Clinical Trials and Health Care Evaluation. Chichester, UK: John Wiley & Sons, 2004.
[43] Luce BR, Claxton K. Redefining the analytical approach to pharmacoeconomics. Health Econ 1999;8:187–9.
[44] Hawkins N, Scott DA, Woods BS, Thather N. No study left behind: a network meta-analysis in non-small-cell lung cancer demonstrating the importance of considering all relevant data. Value Health 2009;12:996–1003.
[45] Hawkins N, Scott DA, Wood BS. How far do you go? Efficient searching for indirect evidence. Med Decis Making 2009;29:273–81.
[46] Salanti G, Higgins JPT, Ades AE, Ioannidis JPA. Evaluation of networks of randomized trials. Stat Methods Med Res 2008;17:279–301.
[47] Cappelleri JC, Ioannidis JP, Schmid CH, et al. Large trials vs meta-analysis of smaller trials: how do their results compare? JAMA 1996;276:1332–8.
[48] Abel U, Koch A. The role of randomization in clinical studies: myths and beliefs. J Clin Epidemiol 1999;52:487–97.
[49] Schulz KF, Chalmers I, Hayes RJ, Altman DG. Empirical evidence of bias. Dimensions of methodological quality associated with estimates of treatment effects in controlled trials. JAMA 1995;273:408–12.
[50] Schwartz D, Lellouch J. Explanatory and pragmatic attitudes in therapeutical trials. J Chronic Dis 1967;20:637–48.
[51] Rothwell PM. Factors that can affect the external validity of randomized controlled trials. PLOS Clin Trials 2006;1:e9.
[52] Systematic Reviews: CRD's guidance for undertaking reviews in health care Centre for Reviews and Dissemination. Centre for Reviews and Dissemination, University of York; January 2009.
[53] Williamson PR, Outcome selection bias in meta-analysis Stat Methods Med Res 2005;14:515–24.
[54] Liberati A, Altman DG. Tetzlaff J, et al. The PRISMA Statement for reporting systematic reviews and meta-analyses of studies that evaluate health care interventions: explanation and elaboration. Ann Intern Med 2009;151:W65–94.
[55] Moher D, Liberati A, Tetzlaff J, Altman DG, and the PRISMA Group. Preferred reporting items for systematic reviews and meta-analyses: the PRISMA Statement. Ann Intern Med 2009;151:264–9.
[56] Spiegelhalter D, Thomas A, Best N, Lunn D. WinBUGS User Manual: Version 1.4. MRC Biostatistics Unit: Cambridge, 2003.
[57] Atkins D, Best D, Briss PA, et al. Grading quality of evidence and strength of recommendations. BMJ 2004;328:1490.
[58] Jansen JP. Bergman GJD, Huels J, Olson M. The efficacy of bisphosphonates in the prevention of vertebral, hip, and non-vertebral-non-hip fractures in osteoporosis: a network meta-analysis. Semin Arthritis Rheum 2011;40:275–84.
[59] Petitti DB. Meta-analysis, Decision Analysis, and Cost-effectiveness Analysis. Methods for Quantitative Synthesis in Medicine (2nd ed.). New York, NY: Oxford University Press, 2000.
[60] Cooper NJ, Sutton AJ, Abrams KR. Comprehensive decision analytic modelling in economic evaluation: a bayesian approach. Health Econ 2004;13;203–26.
[61] Ades AE, Sculpher M, Sutton A, et al. Bayesian methods for evidence synthesis in cost-effectiveness analysis. Pharmacoeconomics 2006;24:1–19.
[62] Tunis SR, Stryer DB, Clancy CM. Practical clinical trials: increasing the value of clinical research for decision making in clinical and health policy. JAMA 2003;290:1624–32.
[63] Tricoci P, Allen JM, Kramer JM, et al. Scientific evidence underlying the ACC/AHA clinical practice guidelines. JAMA 2009;301:831–41.
[64] Chambers D, Rodgers M, Woolacott N. Not only randomized controlled trials, but also case series should be considered in systematic reviews of rapidly developing technologies. J Clin Epidemiol 2009;62:1253–60.
[65] Claxton K. The irrelevance of inference: a decision-making approach to the stochastic evaluation of health care technologies. J Health Econ 1999;18:341–64.
[66] Fleurence RL, Naci H, Jansen JP. The critical role of observational evidence in comparative effectiveness research. Health Aff (Millwood) 2010;29:1826–33.