Background to the task force

The International Society for Pharmacoeconomics and Outcomes Research (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-direct-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 articles, “Interpreting Indirect Treatment Comparisons & 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-Comparison and Network-Meta-Analysis Studies: Report of the ISPOR Task Force on Indirect Treatment Comparisons Good Research Practices: Part 2.” Summaries were presented for comment at the 15th Annual International Meeting in Atlanta, GA, USA, in May 2010. Drafts 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 considered the comments from the Task Force Review Group, and the revised drafts were sent for comment to the ISPOR membership (5550) in September 2010. Altogether, Part 1 received 23 comments, and Part 2 received 13 comments. All written comments are published at the ISPOR Web site. The authors considered all comments (many of which were substantive and constructive), made revisions, and submitted them to Value in Health.
Introduction

The ISPOR Task Force on Indirect Treatment Comparisons Good Research Practices is publishing its report as two articles. This article relies on Part 1 of the report [1] for motivation, concepts, and a variety of definitions (e.g., indirect treatment comparison [ITC], mixed treatment comparison, network meta-analysis, heterogeneity, similarity, and consistency).

Terminology for indirect treatment comparisons, mixed treatment comparisons, and network meta-analysis varies in the literature. In practice, all of these methods involve meta-analysis in the context of a network of evidence. Thus it seems beneficial to use a single term for all except the simplest analyses: “network meta-analysis” applies when the evidence network involves more than two randomized controlled trials (RCTs) and more than two interventions.

Part 1 of the report emphasizes aspects of network meta-analysis that are of most importance to decision makers and others who must appreciate and apply the results of such studies. We encourage readers to study it before proceeding with this article, which focuses on more-technical aspects of conducting a network meta-analysis. The sections that follow discuss strategies for developing the network of evidence, assumptions, and statistical methods (objectives, models, frequentist vs. Bayesian approaches, and model validation). We then present a checklist for good research practices, discuss illustrative examples, and conclude by mentioning eight areas of current and needed research.

Identifying the evidence network

A network meta-analysis starts with a network of evidence: the relevant treatments and the clinical trials that have compared those treatments directly. Its structure is often readily apparent from a diagram in which each node represents a treatment (or perhaps a class of treatments), and each link or edge connects treatments that have been directly compared in one or more RCTs. The structure of the network may have implications for the interpretation of the evidence [2]. Part 1 of the report contains diagrams for several types of evidence networks.

The literature search for a network meta-analysis builds the network, applying the same basic standards as for a meta-analysis leading to a direct comparison [3–6]. If the focus is a comparison of two treatments, say B and C, the search aims to locate all studies that have included B and another comparator, all studies that have included C and another comparator, and any studies that have compared B and C directly.

If no studies have compared B and C, but each has been compared with a common comparator, say A, then it may be appropriate to base an indirect comparison of B and C on the direct comparison of B and A and the direct comparison of C and A. Beyond this straightforward situation, B and C may have other common comparators—or none at all.

In the absence of a common comparator for B and C, the treatments with which they have been compared may have a common comparator, or the connection with a common comparator may require additional links. If more than a few links separate B and C, an indirect comparison may be unreliable, because each additional link tends to increase the standard error of the indirect comparison [7]. The contribution of a link depends on a number of factors, including the number of studies underlying that direct comparison, the sample sizes in those studies, and the homogeneity of the study-specific estimates. A decision on proceeding with an indirect comparison should, therefore, consider more information than the number of links.

It may be difficult to identify all relevant comparators for the treatments of interest, and any search involves costs and trade-offs. It may be efficient to proceed in stages, using one of the strategies developed by Hawkins et al. [7]. They refer to the treatments that one wishes to compare as primary comparators. Treatments that have been directly compared with a primary comparator are secondary comparators, and treatments that have been directly compared with a secondary comparator are tertiary comparators, and so on. The order of a comparison is determined by the number of intermediate comparators. Thus, first-order comparisons are direct comparisons, second-order indirect comparisons are based on trials that share a single intermediate comparator, third-order indirect comparisons involve two intermediate comparators, and so on.

Table 1 lists the first five searches for their breadth-first strategy. In the searches that exclude a comparator, one can minimize the burden of searching by excluding the comparator that is likely to produce the largest number of hits (e.g., placebo). Each search in the sequence will locate, in the chosen database(s) (e.g., MEDLINE, EMBASE), all clinical trials that contribute to an indirect comparison of the corresponding order: Search 1, all first-order (direct) comparisons; Search 2, all second-order indirect comparisons; and so on. Through these searches, additional treatments that may not be viewed as comparators or within the scope of the appraisal (e.g., unlicensed treatments, in some instances) may contribute to the evidence network.

The depth-first strategy begins with the same search as the breadth-first strategy, but it does not include the omitted comparator in any subsequent search. Search 2 targets all secondary comparators identified in Search 1, Search 3 targets all tertiary comparators identified in Search 2, and so on. Hawkins et al. [7,8] give a detailed discussion of the searches and resulting comparisons for two second-line treatments of advanced/metastatic non-small-cell lung cancer.

The literature review should also search for any meta-analyses that have already produced direct (or even indirect) comparisons of potentially relevant treatments, to provide empirical validation for the analysis.

Table 1 – The first five searches in the breadth-first strategy. Each search uses the Boolean odds ratio to combine comparators.

<table>
<thead>
<tr>
<th>Search</th>
<th>Comparators</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>All primary comparators except one</td>
</tr>
<tr>
<td>2</td>
<td>All primary comparators</td>
</tr>
<tr>
<td>3</td>
<td>All secondary comparators except one</td>
</tr>
<tr>
<td>4</td>
<td>All secondary comparators</td>
</tr>
<tr>
<td>5</td>
<td>All tertiary comparators except one</td>
</tr>
</tbody>
</table>

Adapted from Table 1 of Hawkins et al. [7].

Assumptions

As discussed in Part 1 [1], network meta-analysis relies on the randomization in the RCTs that compared the treatments directly. It also involves a similarity assumption: “Combining studies should only be considered if they are clinically and methodologically similar” [9]. Nevertheless, “no commonly accepted standard [defines] which studies are ‘similar enough’” [9]. For network meta-analysis, covariates that act as relative treatment effect modifiers must be similar across trials (or adjusted for using meta-regression). And, when it combines indirect evidence with direct evidence, network meta-analysis adds the assumption of consistency: The indirect evidence must be consistent with the direct evidence.

A meta-analysis for a direct comparison usually requires that the individual studies estimate either a common treatment effect or study-specific treatment effects distributed around a typical value
The choice between a common effect and a distribution of effects gives rise to fixed-effects and random-effects approaches, respectively. Heterogeneity among studies within a direct comparison is acceptable, as long as their treatment effects share a common typical value, and it may even increase generalizability. On the other hand, heterogeneity between the sets of studies that contribute direct comparisons to an indirect comparison or a network meta-analysis would indicate a lack of similarity.

The discussion of models that follows uses the assumptions of homogeneity, similarity, and consistency as needed. In practice, one must check these assumptions, to the extent possible. "Investigators should base decisions about combining studies on thorough investigations of clinical and methodological diversity as well as variation in effect size" [9]. Agreement is seldom perfect, and both statistical and clinical judgment may be required (e.g., re-examining information in the reports on some trials, calculating direct and indirect estimates separately before proceeding to a network meta-analysis). It may be possible to adjust for differences on study-level characteristics (via meta-regression), but such adjustments are unlikely to overcome substantial disparities among the studies. Interpretations of results should acknowledge this limitation.

**Statistical methods**

**Objectives**

Objectives of network meta-analysis may include considering all relevant evidence, answering research questions in the absence of direct evidence, improving the precision of estimates by combining direct and indirect evidence, ranking treatments, and assessing the impact of certain components of the evidence network.

The choice of effect measure should be determined by the clinical question and the nature of the data. Common measures of relative effect include odds ratio, risk ratio (or relative risk), mean difference, and hazard ratio. The models described below apply to any measure of relative effect, as long as the necessary quantities (e.g., likelihood and link function) are properly defined.

**Models**

We present a sequence of models, starting with ordinary meta-analysis by fixed effects and random effects, to show the natural progression to fixed- and random-effects models for networks and then meta-regression models that allow treatment-by-covariate interactions [11,12]. Some models for networks can accommodate data from multiarm trials, but the presence of such trials adds complexity to the analysis.

The models below relate the underlying outcome to the effects of treatments and covariates on a scale that is appropriate for the particular analysis (e.g., log of odds, change from baseline, or log hazard rate). The probability distribution for the observed outcome is the same for each study. The random-effects model, equation 2, replaces $d$ with $\delta$, the trial-specific effect of treatment $B$ relative to treatment $A$, and $d$ becomes the mean of the distribution (usually normal) of random effects, which has variance $\sigma^2$.

$$\eta_k = \begin{cases} \mu_j & k = A \\ \mu_j + \delta_j & k = B \end{cases}$$

$$\delta_j \sim N(d, \sigma^2)$$

**Fixed-effects network meta-analysis.** When the evidence network consists of multiple pairwise comparisons (i.e., AB trials, AC trials, BC trials, and so on), the set of comparators usually varies among studies, complicating the notation. One approach labels the treatments $A$, $B$, $C$, and so on, and uses $A$ for the primary reference treatment in the analysis. In each study it designates one treatment, $b$, as the base treatment. The labels can be assigned to treatments in the network in such a way that the base treatments follow $A$ (i.e., $A$, $B$, $C$, and so on) and the non-base treatments follow all the base treatments in the alphabet. In the various models, “after” refers to this alphabetical ordering. The general fixed-effects model for network meta-analysis can be specified as follows:

$$\eta_k = \begin{cases} \mu_{jb} & b = A, B, C, \text{ if } k = b \\ \mu_{jb} + d_{jk} - d_{jk} & k = B, C, D, \text{ if } k \text{ is 'after' } b \\ 0 & \text{ if } k = A \end{cases}$$

$\mu_{jb}$ is the outcome for treatment $b$ in study $j$, and $d_{jk}$ is the fixed effect of treatment $k$ relative to treatment $b$. The $d_{jk}$ are identified by expressing them in terms of effects relative to treatment $A$: $d_{jk} = d_{jk} - d_{ja}$ with $d_{ja} = 0$ (the order of the substrings on $d_{jk}$ is conventional, but counterintuitive). For the underlying effects, this relation is a statement of consistency: the “direct” effect $d_{jk}$ and the “indirect” effect $d_{jk} - d_{ja}$ are equal.

In the simplest indirect treatment comparison, $A$ is the primary reference treatment and also the base treatment for the AB trials and the AC trials. The term “adjusted indirect comparison” has been applied to such analyses, but it is an infelicitous choice, because the comparison involves no adjustment in any of the usual senses. We prefer “anchored indirect comparison”; the indirect comparison of $B$ and $C$ is anchored on $A$.

**Random-effects network meta-analysis.** As above, the random-effects model replaces $d_{jk}$ with $\delta_{jk}$, the trial-specific effect of treatment $k$ relative to treatment $b$. These trial-specific effects are drawn from a random-effects distribution: $\delta_{jk} \sim N(d_{jk}, \sigma^2)$. Again, the $d_{jk}$ are identified by expressing them in terms of the primary reference treatment, $A$. This model assumes the same random-effect variance $\sigma^2$ for all treatment comparisons, but the constraint can be relaxed. (A fixed-effects model results if $\sigma^2 = 0$.)

$$\eta_k = \begin{cases} \mu_{jb} & b = A, B, C, \text{ if } k = b \\ \mu_{jb} + \delta_{jk} - \delta_{jk} & k = B, C, D, \text{ if } k \text{ is 'after' } b \\ 0 & \text{ if } k = A \end{cases}$$

$$\delta_{jk} \sim N(d_{jk}, \sigma^2) = N(d_{jk} - d_{ja}, \sigma^2)$$

$\delta_{jk} = 0$

**Meta-regression models with treatment-by-covariate interactions.** Meta-regression models include study-level covariates and provide a way to evaluate the extent to which covariates account for heterogeneity of treatment effects. They can also reduce bias and inconsistency between treatment comparisons when covariates are distributed unevenly [13,14]. The covariates enter the model via the mean of the distribution of random effects. The model below uses study-level values of a single covariate, $X$, and allows its coefficient, $b$, to vary among the comparisons [12]. It is a good...
idea to center a covariate (e.g., at its overall mean), to make the results more interpretable.

\[ n_k = \begin{cases} \mu_k & \text{if } k = b \\ \mu_k + \delta_{ik} & \text{if } k \text{ is ‘after’ } b \end{cases} \]

\[ \delta_{ik} \sim N(d_{ik} + \beta_0 X_i, \sigma^2) = N(d_{ik} - d_{ak} + (\beta_{ak} - \beta_{ak}) X_i, \sigma^2) \]

(5)

\[ d_{ak} = 0, \beta_{ak} = 0 \]

Other models can use values of the covariate for the combination of study and treatment. A simplification of the model above uses the same coefficient for all comparisons:

\[ n_k = \begin{cases} \mu_k & \text{if } k = b \\ \mu_k + \delta_{ik} & \text{if } k \text{ is ‘after’ } b \end{cases} \]

\[ \delta_{ik} \sim N(d_{ik} + \beta_0 X_i, \sigma^2) = N(d_{ik} - d_{ak} + (\beta_{ak} - \beta_{ak}) X_i, \sigma^2) \]

(6)

\[ d_{ak} = 0 \]

If a model with a constant \( \beta \) is satisfactory, comparisons among treatments (adjusted for the contribution of the covariate) are straightforward, because \( d_{ak} - d_{ak} \) applies at any value of the covariate (study-level or study-by-treatment-level). In the model with \( \beta_0 \), the adjustment for the covariate can improve internal validity, but the analyst must choose a value of \( \chi \) at which to make comparisons among treatments. Meta-regression also has the drawback that the relation between the outcome and the covariate in patient-level data can differ, and even be opposite in direction, from the corresponding relation in study-level summaries.

If the network appropriately includes a multiarm trial, omitting it from the analysis may introduce bias. The analysis, then, must take into account the correlation among the estimates for the pairs of arms; some methods can do this more easily than others. It is incorrect to analyze the pairwise effects in a multiarm trial as if they came from separate studies. This error is not rare; an example in a Canadian Agency for Drugs and Technologies in Health report shows the input data for the three direct comparisons in a three-arm trial as three separate studies [15]. Salanti et al. [16] illustrate analyses that incorporate correlation among arms.

### Analysis framework: Frequentist versus Bayesian approaches

**Frequentist approach.** The label “frequentist” applies to most of the traditional statistical methods applied in making comparisons, including the weighted means with confidence intervals (based on an assumed normal distribution) in fixed-effects and random-effects meta-analysis and the Mantel-Haenszel estimate (e.g., for an odds ratio) (a fixed-effects procedure). In models such as network meta-analysis and meta-regression, estimation and inference are based on some form of maximum likelihood.

**Bayesian approach.** Bayesian methods combine the likelihood (roughly, the probability of the data as a function of the parameters) with a prior probability distribution (which reflects prior belief about possible values of those parameters) to obtain a posterior probability distribution of the parameters [17]. The posterior probabilities provide a straightforward way to make predictions, and the prior distribution can incorporate various sources of uncertainty. For parameters such as treatment effects, the customary prior distributions are noninformative. The assumption that, before seeing the data, all values of the parameter are equally likely minimizes the influence of the prior distribution on the posterior results. When information on the parameter is available (e.g., from observational studies or from a previous analysis), however, the prior distribution provides a natural way to incorporate it.

For one random-effects network model in which the outcome measure is log odds, the Bayesian analysis has the following components:

\[ r_{jk} \sim \text{binomial}(p_{jk}, n_{jk}) \]

model:

\[ \logit(p_{jk}) = \begin{cases} \mu_k & \text{if } k = b \\ \mu_k + \delta_{ik} & \text{if } k \text{ is ‘after’ } b \end{cases} \]

\[ \delta_{ik} \sim N(d_{ik}, \sigma^2) = N(d_{ik} - d_{ak}, \sigma^2) \]

(7)

\[ d_{ak} = 0 \]

priors:

\[ d_{ak} \sim \text{normal}(0, 10^6) \]

\[ \sigma \sim \text{uniform}(0, 2) \]

In the likelihood the observed number of events for treatment \( k \) in study \( j \), \( r_{jk} \), has a binomial distribution (independent of other treatments in study \( j \) and of other studies) whose event probability is \( p_{jk} \), described by the model. The parameters in the distributions of random effects have vague prior distributions: \( N(0, 10^6) \) for the \( d_{ak} \) (independently) and \( \text{Uniform}(0, 2) \) for \( \sigma \). These priors are common choices in such models.

**Differences.** For most of the models an analysis could follow either a frequentist approach or a Bayesian approach. In current practice most meta-analyses for direct comparisons use frequentist methods. For the more-complicated models, particularly networks involving mixed treatment comparisons, Bayesian methods have undergone substantially greater development, facilitated by advances in computing that support their computational intensity and the need to monitor convergence to the posterior distribution. Lumley [18] described a maximum-likelihood approach using linear mixed models; his method has been applied in relatively few articles [19–21].

In brief, a frequentist analysis yields point estimates and confidence intervals. The typical Bayesian analysis produces an empirical version of the joint posterior distribution of the parameters, from which one can derive summary measures for individual parameters, such as the posterior mean and a “credible interval” (CI) (e.g., the endpoints of the 95% CI are the 2.5 and 97.5 percentiles of the posterior distribution), as well as posterior distributions for functions of parameters (e.g., estimates of the probability that each treatment is best). The Bayesian CIs reflect the uncertainty in estimating heterogeneity, whereas frequentist random-effects models do not propagate that uncertainty.

Choices of prior distributions are, to some extent, arbitrary, so they are often subjected to sensitivity analysis, which may be especially important for priors on heterogeneity in random-effects models. Lambert et al. [22] discuss sensitivity analysis for exploring the effect of the use of vague priors. On the other hand, some frequentist methods involve approximations and assumptions that are not stated explicitly or verified when the methods are applied. Therefore both insight into the sensitivity of results from a Bayesian analysis to assumptions on priors and transparent reporting of assumptions underlying a frequentist analysis are highly important.

**Model validation**

**Evaluation of assumptions.** As mentioned above, assumptions require checking. In assessing similarity, researchers may be able to use statistical information, but they must rely primarily on clinical judgment of whether differences among studies may affect the comparisons of treatments or make some comparisons inappropriate.

Evaluation of homogeneity and consistency (if the network supports both direct and indirect comparisons) should be specified as components of the analysis and should reflect the risks and benefits of combining data for the particular research question.
The discussion that follows sketches the current state of some approaches for examining homogeneity and consistency.

Ordinary meta-analyses (especially by frequentist methods) customarily evaluate heterogeneity of effects, as a basis for choosing between a fixed-effect or a random-effects procedure. Borenstein et al. [23] advise against considering a fixed-effect model as a presumptive choice; they explain that fixed-effect models and random-effects models reflect fundamentally different assumptions about the distribution of the data. Thus, the choice should reflect an understanding of whether the studies share a common effect and the goals of the analysis. These issues are at least as important in a network meta-analysis. When information on heterogeneity within the direct comparisons is available, consideration of it can form a preliminary step in a network meta-analysis, but one should first examine potential effect modifiers, because disparities among studies may preclude analysis of the network. The common measures of heterogeneity in a direct comparison use the studies’ estimates, \( y_i \) (\( i = 1, \ldots, k \)), of effect and the estimated variances of those effect estimates, \( s_i^2 \), to form the weighted mean, \( \bar{y}_w = \sum w_i y_i \) (\( \sum w_i \)), where \( w_i = 1/s_i^2 \) (\( \bar{y}_w \) is the fixed-effect estimate of the common effect). They then calculate the statistic \( Q = \sum w_i (y_i - \bar{y}_w)^2 \) and either refer it to the chi-squared distribution on \( k - 1 \) degrees of freedom or quantify the heterogeneity in terms of the measure \( I^2 = \frac{Q - (k - 1)}{Q} \). \( I^2 \) (usually expressed as a percentage) estimates the proportion of total variation among the effect estimates that is due to heterogeneity; \( I^2 < 30\% \) is considered mild heterogeneity, and \( I^2 > 50\% \), notable heterogeneity. Deeks et al. [25] discuss strategies for addressing heterogeneity. It is important to examine the effects graphically for the presence of groups and outliers; these numerical measures or tests should have only a secondary role. If a model fits poorly, graphical techniques can aid more-detailed examination.

A departure from consistency arises when the direct and indirect estimates of an effect differ (e.g., the direct estimate of \( d_{AC} \) does not equal the indirect estimate obtained from \( d_{AC} - d_{AB} \)). The treatments involved (and their comparisons) must belong to a loop in the network of evidence. Thus, consistency is a property of loops, rather than of individual comparisons. Researchers must evaluate departures from consistency and determine how to interpret them. Salanti et al. [16] provide much valuable guidance. Most agencies to which the results of a network meta-analysis could be submitted currently require that direct estimates and indirect estimates be calculated separately and shown to be consistent before direct evidence and indirect evidence are combined.

Methods for evaluating consistency have been an active area of research. In a network containing a single loop (and no multiarm trials) Bucher et al. [28] compared the indirect estimate of a treatment effect with the corresponding direct estimate (the resulting test of consistency, however, has shortcomings that may make it unreliable). For networks of two-arm trials that contain loops, Lumley [18] introduced a frequentist model that uses one variance parameter to summarize inconsistency in the network as a whole. Lu and Ader [13] focused on the structure of networks and expanded a hierarchical Bayesian model by adding one parameter for each independent consistency relation. By comparing the models with and without those parameters, one can assess overall inconsistency, and the posterior distributions of the added parameters show the extent of inconsistency in the various loops. A graphical assessment can use the forest plot of Lam and Owen [29] to examine consistency between direct and indirect estimates. In a hierarchical Bayesian setting, Dias et al. [30] extended the approach of Bucher et al. [28] to general networks (but not using indirect evidence from multiarm trials), by deriving a weighted difference between the estimate from the network and the direct estimate. By plotting the posterior densities of the direct, indirect, and network estimates, they show how the direct evidence and the indirect evidence are combined in the network estimate. For each effect that has direct evidence Dias et al. [30] also split the information into direct and indirect information and examine the posterior distribution of the difference between the resulting direct and indirect estimates. They discuss how to handle multiarm trials in this analysis.

Assessment of model fit. In frequentist analyses, measures of model fit are similar to those for direct evidence and depend on the particular outcome measure. Bayesian analyses customarily use deviance (a likelihood-based measure)—the lower the residual deviance, the better the fit. For comparing models, the deviance information criterion (DIC) adds a penalty term, equal to the effective number of parameters in the model. If a model fits poorly, graphical techniques can aid more-detailed examination.

Sensitivity analysis. Sensitivity analyses should focus on the areas of greatest uncertainty. Potential effect modifiers can be explored by stratifying on variations in study design or population. Comparisons between random-effects and fixed-effects analyses may be appropriate. Bayesian analyses should also explore the influence of choosing different prior distributions.

### Checklist for good research practices

Practices for conducting and reporting systematic reviews and meta-analyses have received extensive discussion since the mid-1980s and have been the subject of guidelines and recommendations, most recently the PRISMA statement [6]. Because network-meta-analysis studies have many features in common with systematic reviews and meta-analyses, we recommend that they follow all applicable parts of the PRISMA checklist. In addition, Table 2 supplements the checklist in Part 1 [1] by highlighting key components of network-meta-analysis studies, areas in which they have distinct (often additional) requirements, and recent developments. We intend this checklist for use in light of ongoing, dynamic research on network meta-analysis. Improved methods and their application will lead to changes in the checklist.

### Illustrative examples

Separate analyses of a portion of the data from two extensive meta-analyses provide illustrations of ITC (both frequentist and Bayesian approaches) and network meta-analysis. This section also discusses available software.

Stettler et al. [32] reported on a collaborative network meta-analysis of outcomes associated with two drug-eluting stents (the paclitaxel-eluting stent [PES] and the sirolimus-eluting stent [SES]) and bare-metal stents (BMS). The literature search yielded 7 RCTs comparing PES with BMS, 15 RCTs comparing SES with BMS, 14 RCTs comparing SES with PES, and 1 RCT comparing all three. Their analysis of overall mortality and other primary safety endpoints is an instructive application of Bayesian random-effects models for network meta-analysis involving multiple follow-up times [33]. This example focuses on the rate, at 1 year, of target lesion revascularization (TLR), a secondary effectiveness endpoint involving subsequent percutaneous intervention.

The Appendix for this article (found at doi:10.1016/j.ijv.2011.01.011) gives the details of the Bayesian analysis of target lesion revascularization on the log-odds scale: direct meta-analyses of PES versus BMS, SES versus BMS, and PES versus SES; the corresponding separate indirect comparisons; the comparisons from the network meta-analysis; and WinBUGS code and the data.

The forest plot in Figure 1 shows the data from the studies that compared PES and SES, the study-specific odds ratios for PES ver-
sus SES (from the Bayesian analysis), and five estimates of the overall odds ratio: traditional fixed-effects pairwise (using the Mantel-Haenszel method), traditional random-effects pairwise (using the DerSimonian-Laird method), Bayesian direct, Bayesian indirect, and Bayesian network (Fig. 3 of the Appendix, found at doi:10.1016/j.jval.2011.01.011, shows odds ratios for SES versus PES). To facilitate numerical comparisons, Table 3 presents the estimates. The traditional fixed-effects and random-effects estimates and their 95% confidence intervals differ very little: 1.39 (1.17 to 1.66) and 1.38 (1.16 to 1.65). The random-effects interval is no wider, because the estimated heterogeneity is small ($I^2 = 0.74\%$).

To make an anchored indirect estimate of the odds ratio for PES versus SES, using the direct estimates for PES versus BMS and SES versus BMS, it is necessary to omit the three-arm trial (or do a more-complicated calculation that recognizes the contributions of that trial to both direct estimates). The right-hand column of Table 3 gives the estimates without the three-arm trial. The fixed-effects and random-effects indirect estimates are similar, 1.74 (1.28 to 2.36) and 1.70 (1.09 to 2.65) and are considerably higher than their direct counterparts. The confidence interval for the random-effects indirect estimate is considerably wider than that for the fixed-effects estimate, and both are substantially wider than those for the direct estimates. All the confidence intervals are wide enough that further analysis would be required to assess consistency between the indirect estimates and direct estimates.

The three Bayesian estimates, 1.55 (1.17 to 2.05), 1.60 (0.98 to 2.47), and 1.55 (1.22 to 1.96), are similar, and somewhat higher than the traditional fixed-effects and random-effects estimates. The indirect estimate has the widest credible interval. More importantly, the estimate from the Bayesian network meta-analysis has a somewhat narrower credible interval than the Bayesian direct estimate, as a consequence of taking into account all the evidence in the network. We would report that estimate, 1.55 (1.22 to 1.96), as the odds ratio for PES versus SES.

To provide a richer illustration of some aspects of network meta-analysis (e.g., correlations between estimates and probability of a treatment’s being best), the Appendix (found at doi:10.1016/j.jval.2011.01.011) also analyzes data from a network of 112 RCTs that compared (in the aggregate) the acceptability of 12 new-generation antidepressants in the acute-phase treatment of unipolar major depression [34].

### Table 2 – Checklist of good research practices for conducting and reporting network-meta-analysis studies.

<table>
<thead>
<tr>
<th>Checklist item</th>
<th>Recommendation(s)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Search strategies</td>
<td>● Follow conventional guidelines for systematic literature searches; be explicit about search terms, literature, and time frames, and avoid use of ad hoc data</td>
</tr>
<tr>
<td></td>
<td>● Consider iterative search methods to identify higher-order indirect comparisons that do not come up in the initial search focusing on lower-order indirect comparisons</td>
</tr>
<tr>
<td>Data collection</td>
<td>● Set forth evidence network demonstrating direct and indirect linkages between treatments, based on identified study reports</td>
</tr>
<tr>
<td></td>
<td>● Follow conventional guidelines for data collection; use a pre-specified protocol and data extraction form</td>
</tr>
<tr>
<td></td>
<td>● Include sufficient study detail in data extraction to permit assessment of comparability and homogeneity (e.g., patient and study characteristics, comparators, and outcome measures)</td>
</tr>
<tr>
<td>Statistical analysis plan</td>
<td>● Prepare statistical analysis plan prior to data analysis, but permit modifications during data analysis, if necessary</td>
</tr>
<tr>
<td></td>
<td>● Provide step-by-step descriptions of all analyses, including explicit statements of all assumptions and procedures for checking them</td>
</tr>
<tr>
<td></td>
<td>● Describe analytic features specific to network meta-analysis, including comparability and homogeneity, synthesis, sensitivity analysis, subgroup analysis and meta-regression, and special types of outcomes</td>
</tr>
<tr>
<td>Data analysis</td>
<td>● Follow conventional guidelines for statistical model diagnostics</td>
</tr>
<tr>
<td></td>
<td>● Evaluate violations of similarity or consistency assumption in evidence network</td>
</tr>
<tr>
<td></td>
<td>● If similarity or consistency is a problem, consider use of meta-regression models with treatment × covariate interactions to reduce bias</td>
</tr>
<tr>
<td>Reporting</td>
<td>● Follow PRISMA statement for reporting of meta-analysis</td>
</tr>
<tr>
<td></td>
<td>● Explicitly state the study research questions (e.g., in Introduction or Objectives section of report)</td>
</tr>
<tr>
<td></td>
<td>● Provide graphical depiction of evidence network</td>
</tr>
<tr>
<td></td>
<td>● Indicate software package used in the analysis and provide code (at least in an online appendix)</td>
</tr>
</tbody>
</table>

### Software

We used WinBUGS for the Bayesian analyses in this example [35]. Other software for Bayesian analysis includes OpenBUGS [36]. Version 9.2 of SAS [37] includes a procedure for Markov chain Monte Carlo, but the examples in the documentation appear not to include applications in meta-analysis. Frequentist approaches have been developed by Lumley [18] in R using linear mixed models and are also feasible in SAS using PROC NLMIXED [38].

### Areas for future research

Research has produced a substantial body of methods for network meta-analysis, but a number of topics would benefit from further investigation. This section discusses eight of them, roughly in order of priority.

### Adjustments to improve the similarity of trials and empirical assessment of their benefit

Patients studied in the various trials may not be comparable. If these differences among trials are treatment-effect modifiers, an ITC is biased [39]. In addition, in a network meta-analysis different types of comparisons may be inconsistent [12]. Because the number of studies is often limited, adjustment by incorporating covariates with “conventional” meta-regression approaches may be unsatisfactory [38–40]. In addition, limited or inconsistent reporting of the study-level covariates (or effects for levels of a possible effect modifier) complicates analysis. Future studies should examine the value of existing methods to enhance comparability of trials, as well as the level of consistency in reporting of study-level covariates.

### Structure and properties of networks

To explore the geometry of treatment networks for four or more interventions, Salanti et al. [2] used two measures from the ecological literature, diversity and co-occurrence. Diversity characterizes the number of interventions and, for a set of networks with the same number of interventions, the frequency distribution of interventions. Co-occurrence measures the appearance of specific pairwise combinations of interventions. These measures allow 1) comparing networks of different sizes within a particular network shape and 2) signaling the presence of potential biases in the assessment or reporting of specific comparisons. It
is not known, however, whether these are the best measures to characterize evidence networks, or how they correlate with effect estimates, heterogeneity, similarity, or consistency.

**Methods for verifying the assumption of consistency**

As discussed above, a number of approaches, including graphical representations, have been developed to assess consistency in network meta-analysis. Salanti et al. [16] and Dias et al. [30] have discussed the strengths and weaknesses of these approaches. Further research is needed, to improve understanding of these methods and encourage their use. As Salanti et al. [16] point out, a measure for inconsistency analogous to $I^2$ would be a welcome addition.

**Multiple outcomes**

Meta-analyses that consider multiple outcomes usually do so in separate analyses. Intuitively, however, because “multiple endpoints are usually correlated, a simultaneous analysis that takes their correlation into account should add efficiency and accuracy” [41]. Though this approach has received considerable attention in recent years [42–44], it has not yet been extended to network meta-analysis for studies that have analyzed multiple endpoints [45].

Lu et al. [33] illustrated how to combine data recorded at multiple time points; these methods can be extended to estimate treatment effects at multiple time points using the network for each time point and the correlation between multiple observations. This would be beneficial to cost-effectiveness models where transition probabilities vary over time.

Historically meta-analysts have combined studies using a common summary measure (e.g., mean change from baseline) and have excluded studies that presented the same measure in a slightly different way (e.g., mean difference). Methods for combining alternative summary measures have recently been proposed in order to minimize potential selection or reporting bias: Welton et al. [46] presented methods for combining mean change from baseline and mean difference data from different studies, and Woods et al. [47] presented...
between fixed-effects and random-effects models. These methods need estimations for the heterogeneity parameter would offer a compromise between fixed-effects and random-effects models incorporating covariate adjustment that ensures the same summary measures of the same endpoint.

When heterogeneity may be present. Cooper et al. [12] proposed alternative random-effects models for meta-regression [12]. The resulting overspecification (i.e., the number of parameters to be estimated exceeds the number of studies) can compel the meta-analyst to use a fixed-effects model (i.e., the number of parameters to be estimated exceeds the number of studies) can compel the meta-analyst to use a fixed-effects model or meta-regression [12].

The practitioner may ask, “Is there a direct way of deciding on the size of a network?” Is there an optimum? What size is sufficient?” Size involves the estimated effects. The fixed-effect and random-effects methods are based on IPD can be considered superior. When IPD are available, meta-regression models usually have sufficient power to evaluate heterogeneity and improve consistency. Apart from potentially questionable fit, adjustments based on aggregate-level covariates can produce ecological bias [52]. This problem can be avoided by using IPD, or minimized by using IPD for a subset of studies [53]. The added value of using IPD in network meta-analysis should be further evaluated to develop recommendations on its relevance and practicability.

### Conclusion

This article, the second part of the report from the Task Force, sets out general best-practice guidelines for practitioners who are interested in performing network meta-analyses. The underpinning statistical methods are presented and supported by examples that suggest ways of presenting results to nonspecialized audiences. The techniques are not always applicable and may be subject to the biases and limitations discussed in this report. Failure of the assumptions of similarity and consistency may render results questionable. Where the techniques are applicable, care and transparency should be employed, and we encourage adherence to our checklist. Standardization of methods would enhance the overall credibility and applicability of indirect treatment comparisons and network meta-analysis. Finally, ongoing research in many areas should lead to periodic revisions of the recommendations.

### Supplemental Material

Supplemental material accompanying this article can be found in the online version as a hyperlink at [doi:10.1016/j.jval.2011.01.011](http://www.valueinhealthjournal.com/issues) at [www.valueinhealthjournal.com/issues](http://www.valueinhealthjournal.com/issues) (select volume, issue, and article).