Special Issue Paper

Research Synthesis Methods

Received 10 August 2011, Revised 12 March 2012, Accepted 14 March 2012 Published online 11 June 2012 in Wiley Online Library

(wileyonlinelibrary.com) DOI: 10.1002/jrsm.1037

Indirect and mixed-treatment comparison, network, or multiple-treatments meta-analysis: many names, many benefits, many concerns for the next generation evidence synthesis tool‡

Georgia Salanti[*†]

The ever increasing number of alternative treatment options and the plethora of clinical trials have put systematic reviews and meta-analysis under a new perspective by emphasizing the need to make inferences about competing treatments for the same condition. The statistical component in reviews that compare multiple interventions, network meta-analysis, is the next generation evidence synthesis toolkit which, when properly applied, can serve decision-making better than the established pairwise meta-analysis. The criticism and enthusiasm for network meta-analysis echo those that greeted the advent of simple meta-analysis. The main criticism is associated with the difficulty in evaluating the assumption underlying the statistical synthesis of direct and indirect evidence. In the present article, the assumption of the network meta-analysis are presented using various formulations, the statistical and nonstatistical methodological considerations are elucidated, and the progress achieved in this field is summarized. Throughout, focus is put on highlighting the analogy between the concerns and difficulties that the scientific community had some time ago when advancing from individual trials to their quantitative synthesis via meta-analysis and those currently expressed about the transition from head-to-head meta-analyses to network meta-analysis. Copyright © 2012 John Wiley & Sons, Ltd.

Keywords: assumptions; transitivity; consistency; systematic reviews; evidence-based practice

1. Introduction

When meta-analysis was introduced over 30 years ago, it generated a considerable level of debate. Clinicians and epidemiologists criticized the approach for mixing incomparable interventions, often resulting in confusing results (Hunt, 1997). It took its time before the scientific community was accustomed to the idea of heterogeneity and realized that its exploration can be informative. Methods to measure heterogeneity (such as the I^2 (Higgins *et al.*, 2003)) and approaches to reflect it in the meta-analysis result (using the random effects models and predictive intervals (Higgins *et al.*, 2009)) improved understanding and boosted confidence in the meta-analysis methods. Approaches to deal with publication bias (Rothstein *et al.*, 2005) and to account for effect modifiers or to evaluate the risk of bias in the body of evidence were developed. Eventually, meta-analysis became widely recognized as a useful tool by national and international policy-making bodies such as the World Health Organization, the National Institute of Clinical Excellence (NICE) in the UK, and the Agency for Healthcare Research and Quality (AHRQ) Evidence-Based Practice program in the USA. Individual researchers and organizations, such as the Cochrane Collaboration, publish hundreds of systematic reviews per year, and meta-analysis is now an integral part of evidence-based medicine (Sackett *et al.*, 2000) and evidence based practice (Gibbs, 2003). From a researcher's perspective, it has been shown that meta-analyses are the most highly cited and therefore influential publications in health research (Patsopoulos *et al.*, 2005).

Department of Hygiene and Epidemiology, University of Ioannina School of Medicine, Ioannina, Greece

**Correspondence to: Georgia Salanti, Department of Hygiene and Epidemiology, University of Ioannina School of Medicine, University Campus, Ioannina 45110, Greece.*

[†]E-mail: gsalanti@cc.uoi.gr

*[‡]This article is published in **Research Synthesis Methods** as a special issue on **Network Meta-analysis**, edited by **Georgia Salanti**, **University of Ioannina**, **Greece.***

Perhaps one of the most debated issues in meta-analysis is the broadness of the research question in the systematic review regarding patient and trial characteristics. Reviews that take a broader approach ('lumping') have wider inclusion criteria regarding the various trial settings, treatment and control definitions, and patient groups. More narrow research questions are favored by other researchers ('splitting') as they produce a more homogeneous evidence base. A broader approach is typically associated with increased power and greater potential for generalizability, whereas extreme splitting suffers from multiple testing problems and may overemphasize minor differences between interventions (Gotzsche, 2000); (Grimshaw *et al*., 2003).

The ever increasing number of alternative treatment options and the plethora of trials has put the dilemma of 'lumping' versus 'splitting' under a new light by emphasizing the need to make inferences about all indicated treatments for the same condition and patient characteristics. It is increasingly recognized that an important drawback of the current state of the art in systematic reviews and meta-analysis is that they are focused on comparing only two interventions at the time. When many competing treatments exist, meta-analyses produce a plethora of information and, they do not directly help making decisions of relevance. Moreover, although the clinical and policy-making interest lies in comparing active agents, new drugs are often compared with placebo or no-treatment in order to obtain approval for drug licensing (Sutton and Higgins, 2008). Given that clinical practice changes over time and that licensed and reference treatments differ across countries, it is unrealistic to expect that evidence for all interventions of interest will be provided from individual trials and meta-analyses of direct comparisons.

The need for a broader, objective, and inclusive view of the available evidence increased the interest of researchers and funding bodies for 'comparative effectiveness reviews' (Longworth *et al*., 2011; Mitka, 2010) or 'comparing multiple interventions review'. Such reviews provide an evidence base that reflects the network of comparisons that arises when collating studies involving different subsets of competing treatments (Caldwell *et al*., 2005; Jansen *et al*., 2011; Lu and Ades, 2004; Salanti *et al*., 2008a; Welton *et al*., 2008). The need for a new robust framework that answers critical decision-making questions directly has been identified by core national and international agencies. The updated NICE report is very receptive to evidence from indirect sources and network meta-analysis, compared with the previous report (NICE, 2008). The Cochrane Collaboration introduced a new form of reviews, the Overviews of Reviews, which summarizes evidence for the relative effectiveness of several interventions for the same condition (Higgins and Green, 2008) and has established a new methods group to support and enhance the publication of reviews with multiple interventions (cmimg.cochrane.org).

Comparing multiple intervention reviews have an optional quantitative part, which is variably known as network meta-analysis, multiple-treatments meta-analysis, or mixed-treatment comparison. The terms are often used interchangeably, and they refer to the same framework that combines direct and indirect information across a network of randomized trials to infer about the relative effectiveness of multiple interventions. The idea of indirect comparison, which underlines the methods, is a simple one: we can compare treatment A to treatment B via a common comparator C, by statistically combining the information from A versus C (AC) and B versus C (BC) studies.

Although network meta-analysis is not necessarily a 'lumping' approach, with respect to the study populations and settings, it aspires to answer questions broader than the head-to-head meta-analysis by including many competing treatments and making use of a wider evidence space by including direct and indirect comparisons. The research question can be focused on three interventions only or can be as wide as to address all treatments for the same condition as long as these are 'jointly randomizable'; a patient could, in principle, be randomized in any of the alternative treatment options. Several application and method papers have outlined the benefits of a joint analysis (Caldwell *et al*., 2005; Cooper *et al*., 2011; Hoaglin *et al*., 2011; Mills *et al*., 2011). The advantages include improvement in precision for the estimated effect sizes and the ability to compare treatments that have not been directly compared in any trial. Figure 1 shows the number of published reviews that have used statistical methods (ranging from simple

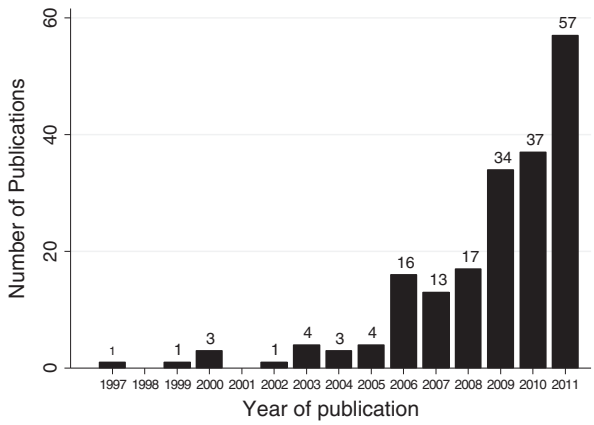

Figure 1. Number of systematic reviews with at least three interventions that derived at least one indirect or mixed estimate (search PubMed until January 2011)

calculation of an indirect effect size to full network meta-analysis) to synthesize evidence about the relative effectiveness of more than two interventions.

Despite the advantages of network meta-analysis, it is far from being an established practice in the research literature. Applications of network meta-analysis appear to be very promising but often controversial (Piccini and Kong, 2011; Thijs *et al*., 2008). The recent evaluation of the relative effectiveness of 12 new-generation antidepressants attracted supporters as well as skeptics (Barbui *et al*., 2009; Cipriani *et al*., 2009). Although researchers recognize the benefits of network meta-analysis, they often use the indirect evidence as a 'second choice', giving priority to direct evidence (Edwards *et al*., 2009). Overall, the majority of health care practitioners are skeptical towards this emerging technique (Ioannidis, 2006). The assumptions underlying the models, the statistical expertise required to fit them, the observational nature of the indirect comparisons, the issue of inconsistency (disagreement between various sources of evidence), and the lack of an interpretable and simple measure to summarize the results and to evaluate the risk of bias contribute to the skepticism.

2. Aims and a roadmap to this article

The criticism and enthusiasm for network meta-analysis echo those that greeted the advent of simple meta-analysis. From my perspective, network meta-analysis is the next-generation evidence synthesis tool, which when properly applied, can serve medical decision-making better than the established pairwise meta-analysis. The aim of this paper is to elucidate the methodological considerations associated with comparing multiple intervention reviews and network meta-analysis, to summarize the progress achieved in the field, and to highlight areas that need further research. Throughout, focus is put on the terminology, and I emphasize the analogy between the concerns and difficulties that the scientific community had some time ago when advancing from individual trials to their quantitative synthesis via basic head-to-head meta-analysis and those currently expressed about the transition from head-to-head meta-analysis to network meta-analysis.

The article is structured so that different sections are relevant to readers with different interests, backgrounds, and degrees of previous experience with network meta-analysis.

- Section 3 starts with an introduction to the concept of indirect and mixed comparisons between two treatments, and it is suitable for *readers without previous knowledge* of the field. The section continuous with an extensive discussion of the assumption underlying indirect and mixed comparisons and can be relevant to *researchers of all backgrounds* (statistical, clinical, or epidemiological) and all levels of experience.
- Section 4 describes the extension of the idea of indirect and mixed comparisons into a full network of trials. Sections 4.1 and 4.2 introduce network meta-analysis without going into technicalities. Discussion of the statistical models in Sections 4.3 to 4.6 will interest more *methodologists familiar with network meta-analysis* and *statisticians*. However, the potential of the methodology to account for covariates and address bias might be relevant to all researchers; interested readers can skim through Sections 4.6.1 and 4.6.2. Finally, paragraph 4.7 discusses ways to present results in a clinically useful manner.
- Section 5 discusses the implications of the methods for the early steps of a systematic review and, particularly, in association with framing the research question and defining the inclusion criteria. *Researchers who plan a network meta-analysis* might find this section useful.
- The paper concludes with Section 6 by highlighting practices that might increase the proper application of network meta-analysis in the published literature.

3. Indirect and mixed treatment comparison of treatments: concepts and assumptions

Two treatments, A and B, can be compared either directly (in studies that synthesized in meta-analysis produce the estimate μ^D_{AB} where the superscript denotes the 'direct' evidence) or indirectly in trials via a common comparator treatment C. The indirect estimate μ^I_{AB} can be derived by combining the meta-analyses of all A versus C studies and B versus C studies as $\mu^I_{AB} = \mu^D_{AC} - \mu^D_{BC}$. This process is often referred to as the 'adjusted indirect comparison' because within-trial randomization is preserved by using the meta-analysis estimates μ^D_{AC} and μ^D_{BC} (Bucher *et al*., 1997). If both direct and indirect estimates are available for the same comparison, one can calculate a 'mixed' effect size μ^M_{AB} by taking the weighted average of μ^D_{AB} and μ^I_{AB}. The mixed effect size μ^M_{AB} is estimated with more precision, compared with μ^D_{AB} (Bucher *et al*., 1997). In summary, in a simple evidence network formed by AB, AC, and BC studies (a triangular network, Figure 2a), for every comparison between two treatments, one can estimate direct and indirect estimates and combine them into a mixed estimate.

In more complex evidence networks, indirect estimates can be derived via many possible routes and not necessarily using a single common comparator C. If AD and DB trials also exist, we can obtain two indirect estimates for μ^I_{AB}'anchored' to C or D. These can also be synthesized together and with the direct estimate μ^D_{AB} if available.

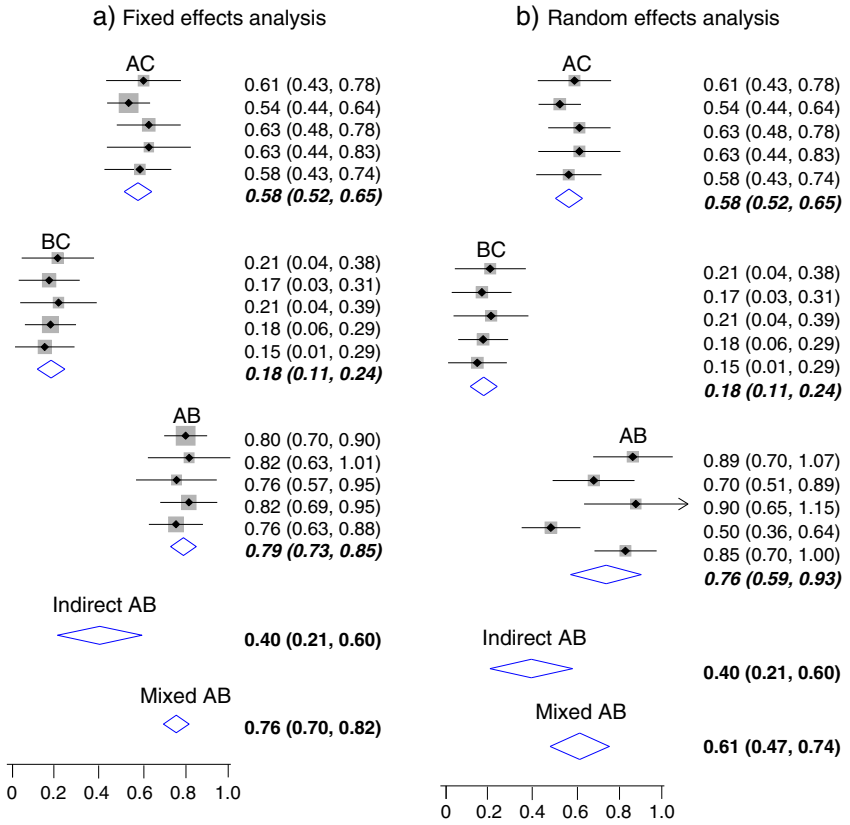

Figure 2. Three fictional sets of trials that compare pairs of interventions A, B, and C. (a) There is no heterogeneity in the relative effects of the interventions within each comparison, but there is statistically significant inconsistency in the ABC loop. (b) There is no statistically significant inconsistency because of the heterogeneity in the AB comparison which produces wider confidence intervals

Although many terms have been used in the literature, we favor the distinction between *indirect* estimates, for its clarity, when one or more μ^I_{AB} is available (but no direct evidence is available for A versus B), and *mixed* estimates when one or more μ^I_{AB} are synthesized with μ^D_{AB} (Jansen *et al.*, 2011).

The direct summary effects can be estimated by using either fixed or random effects meta-analysis. Under a fixed effect model, it is assumed that each study estimates the same comparison-specific mean effect μ^D_{AC}, μ^D_{BC}, and μ^D_{AB} (Figure 2a). Under the random effects model (Figure 2b), it is assumed that each study estimates different, yet related, study-specific effects which come from the same comparison-specific distribution with means μ^D_{AC}, μ^D_{BC}, and μ^D_{AB} and with a variance reflecting the heterogeneity (Figure 3).

Let us assume the approach as a two-step process: first, the indirect estimate μ^I_{AB} is derived (via one or many intermediate treatments), and then, it is synthesized with the direct estimate μ^D_{AB} to provide the mixed estimate. The first step is associated with the assumption of *transitivity*, and the second with the extension of transitivity over a loop of evidence, called consistency. The assumption of *transitivity* – that indirect comparison validly estimates the unobserved head-to-head comparison – cannot be tested statistically, but its plausibility can be evaluated conceptually and epidemiologically. The assumption of *consistency* – that the direct and indirect estimates are in agreement – is a prerequisite to calculate a valid mixed estimate. Consistency is the extension of transitivity across a closed 'loop of evidence', where both direct and indirect evidence are available, and statistical methods can be used to evaluate it. Transitivity and consistency are the same, single assumption. However, separate examination will facilitate the identification of as many potential scenarios of violation as possible. In the following sections, some examples are presented, where one might hold but not the other.

3.1. Transitivity

An underlying assumption when μ^I_{AB} is calculated is that one can learn about A versus B via C. We call this 'the transitivity assumption', and the common comparator C is 'transitive' when it allows valid comparison of the treatments to which it is linked.

Indirect comparison can be viewed as a special case of subgroup meta-analysis. The studies form two groups according to the comparison (e.g. AC and BC), and the difference between the two subgroup summary estimates is the indirect estimate of A versus B (see Figures 2 and 3). This highlights the observational nature of the approach. Although studies are randomized, the choice of the comparisons being made in each study has not been randomized.

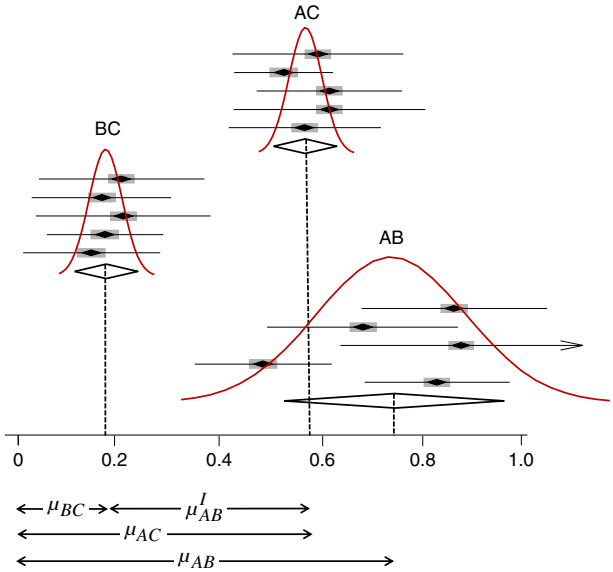

Figure 3. Three fictional sets of trials that compare pairs of interventions A, B, and C. The study-specific blocks represent the study-specific observed relative effects y_{iAC}, y_{iBC} and y_{iAB}. The random effects distributions of the underlying true effects δ_{iAC}, δ_{iBC} and $\delta_{iAB}$are shown for each comparison. For the AB meta-analysis, the heterogeneity variance is $\tau^2 = 0.03$ (estimated using the method of moments), whereas for AC and BC it is close to zero. The arrows at the bottom of the figure show the direct and indirect comparisons for A versus B

So, the assumption of transitivity relates to the assumptions being made while undertaking a subgroup meta-analysis or meta-regression.

In the literature, this assumption has been referred to as the similarity assumption (Donegan *et al*., 2010). The more descriptive term 'transitivity' first used by Baker and Kramer (2002) seems preferable for three main reasons. First, 'transitivity' describes better the aim of the assumption (to compare two treatments via a third one). Second, the term 'similarity' reduces to homogeneity when we refer to a single head-to-head comparison, whereas transitivity clearly refers to more than two comparisons. Finally, 'similarity' may wrongly suggest that similarity is required for all characteristics of trials and patients across the evidence base, when in reality, valid indirect comparison can be obtained even when studies are dissimilar in characteristics which are not effect modifiers (see *Transitivity interpretation 4*). Next, I describe the notion of transitivity by using five equivalent expressions in order to assist clinicians and methodologists in predicting scenarios where the assumption can be violated.

Transitivity interpretation 1: The treatment C is similar when it appears in AC and BC trials. When comparing different fluoride treatments, comparison between fluoride toothpaste and fluoride rinse can be made via placebo. However, placebo toothpaste and placebo rinse might not be comparable as the mechanical function of brushing might have a different effect on the prevention of caries. If this is the case, the transitivity assumption is doubtful (Salanti *et al*., 2009). More generally, the definition of the nodes in the treatment network is a challenging issue with important implications for the joint analysis as, very often, treatments are given at various doses, administrations routes, frequencies, etc. Note that transitivity is violated when the treatment in question differs systematically between trials, not randomly. For example, consider that the common comparator C is a treatment given at different doses, but there is no systematic difference on the average dose of C between AC and BC trials. In this case, the transitivity assumption can hold, although there might be heterogeneity within the AC and BC comparisons. Consequently, the 'anchor' treatment C can be represented by a single node (see Figure 4a), allowing an indirect comparison of A and B. If, however, C is given via a different administration route in all AC (e.g. rinsing) and BC trials (e.g. brushing), then it is questionable whether the two types of Cs can form a common node; if not, then indirect comparison of A versus B via C is impossible (Figure 4b).

Transitivity Interpretation 2: The 'missing' treatment in each trial is missing at random. AC trials do not have a B arm, and BC trials do not have an A arm. Another way to define the transitivity assumption is to consider these 'missing' arms as missing at random (Lu and Ades, 2006). Figure 4 shows how the different scenarios in the context of missing data apply to the transitivity assumption by using the standard graphical representation of the 'missing at random' (Figure 4c) and 'missing not at random' (Figure 4d) assumptions. Consider, for example, the scenario where baseline risk alters the relative effectiveness of A versus B but also suggests which arms shall be included in a study if trialists prefer the more effective intervention, A, in high-risk populations (Figure 4d). Transitivity is also violated when the arm missingness is directly associated with the true relative effectiveness of the interventions (Figure 4d). Evidence from some research areas showed that the choice of comparator is not always random. Very often, trialists prefer to compare their experimental treatment to placebo or a suboptimal intervention rather than a realistic alternative such as an established effective treatment (Heres *et al*., 2006; Rizos *et al*., 2011; Salanti *et al*.,

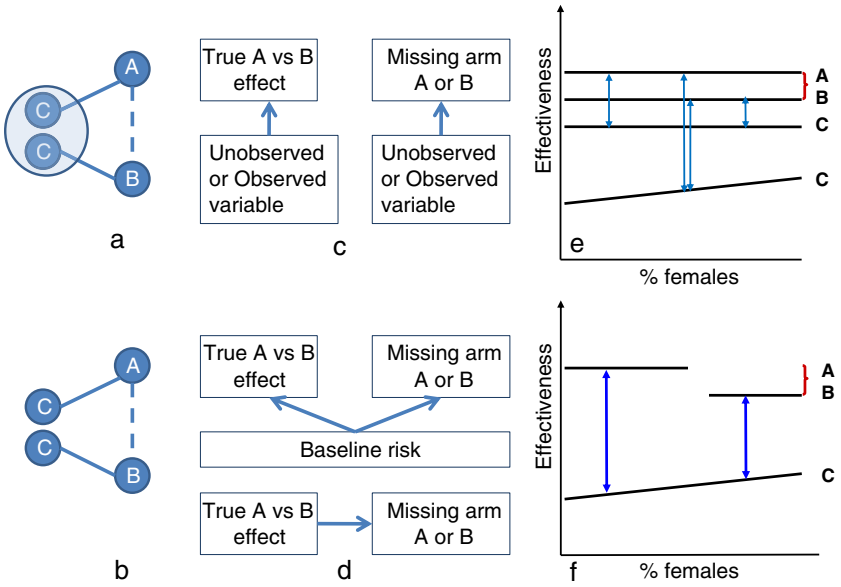

Figure 4. (a and b) Transitivity interpretation 1, (c and d) transitivity Interpretation 2, (e and f) transitivity interpretation 4. The top panels show cases that the transitivity holds for comparison A versus B (existence of a common comparator, arms missing at random, same distribution of effect modifiers). The bottom panels show scenarios of violations of the transitivity assumption.

2008b). If the choice of the comparison is associated, directly or indirectly, with the relative effectiveness of the interventions, then the assumption of transitivity is violated.

Transitivity interpretation 3: There are no differences between observed and unobserved relative effects of AC and BC beyond what can be explained by heterogeneity. A study *i*, which compares treatments B and C, provides an estimate of their relative effectiveness denoted by y_{iBC}. This can be, for example, the mean difference of the outcome between arms B and C. Under the random effects assumption, the *observed* relative treatment effects y_{iBC} relate to underlying *true* effects δ_{iBC} (Table 1). For the same comparison (here, BC), the true relative treatment effects are said to be *exchangeable* across studies: this mean that all δ_{iBC} (directly estimable from the observed y_{iBC}) belong to the same distribution as presented in Figure 3. The width of the distribution reflects the magnitude of the heterogeneity and accounts for the differences in the true relative effects across studies. To see the relation between heterogeneity and transitivity, consider that a latent δ^*_{iBC} underlies an AC study; this is the true relative treatment effect that one could estimate if this study had a B arm. Transitivity claims that the unobserved δ^*_{iBC} is exchangeable with the relative treatment effects δ_{iBC} directly estimable in BC studies. This means that all δ^*_{iBC} and δ_{iBC} come from the same distribution, and consequently, they cannot differ beyond what is expected by the heterogeneity variance.

Transitivity interpretation 4: The two sets of trials, AC and BC, do not differ with respect to the distribution of effect modifiers. This formulation facilitates the conceptual evaluation of the transitivity assumption. Distribution of effect modifiers of the relative treatment effects should be similar in AC and BC trials in order to obtain a valid AB indirect comparison. Clinicians and methodologists who aim to synthesize evidence from many comparisons should identify a priori possible effect modifiers and compare their distributions across comparisons.It is important to note that the transitivity assumption is assumed for the mean effect sizes (that is, μ^D_{AC}, μ^D_{BC}) and not for the individual study results. Consequently, an effect modifier, which differs across studies that belong to the same comparison but has a similar distribution across comparisons, will not violate the transitivity assumption. For example, if sex is an effect modifier and AC trials differ in terms of representation of women in the sample (heterogeneity in AC studies) but the same variability is observed in the set of BC trials, then transitivity may hold even if sex is an effect modifier. In Figure 4e, the effectiveness of A and B relative to C change according to the proportion of women in the trial. Still, the combination of AC and BC studies will provide a valid indirect estimate

Table 1. Relative effects: Observed, underlying true (directly estimable) and unobserved true (unobserved) relative effects in the studies.

Comparison in study *i*	Observed relative effects	True underlying (directly estimable) relative effects	'Missing' arm	Unobserved true (latent) relative effects
AC	y_{iAC}	δ_{iAC}	B	$\delta^*_{iAB}, \delta^*_{iBC}$
BC	y_{iBC}	δ_{iBC}	A	$\delta^*_{iAB}, \delta^*_{iAC}$
AB	y_{iAB}	δ_{iAB}	C	$\delta^*_{iAC}, \delta^*_{iBC}$

of A versus B because we have AC and BC studies that cover the entire spectrum of the effect modifier, sex. In contrast, if the distribution of the variable differs across comparisons (as it is the case in Figure 4f), so that AC studies include, on average, less women, then transitivity does not hold. Baker and Kramer (2002) give a collection of examples where the transitivity assumption is violated because of effect modifiers that differ across comparisons. Adjustment can be used to account for small differences in the distribution of effect modifiers and improve transitivity (see Section 4.6.1). Only effect modifiers, and not 'colliders' (variables that are influenced by the choice of comparison and the relative effectiveness), should be considered for adjustment. Adjustment for colliders, as in classical epidemiology, will introduce bias rather than improve the plausibility of the transitivity assumption (Greenland and Rothman, 2008).

Transitivity Interpretation 5: Participants included in the network could in principle be randomized to any of the three treatments A, B, C. The assumption of transitivity could be violated if interventions have different indications. For example, if treatment A is a chemotherapy regimen typically administered as a second line treatment, whereas treatments B and C can be used either as first or second line treatments, we cannot assume that participants in a BC trial could have been randomized in an AC trial and would have given the same results. Although this consideration is fundamental and should be addressed when building the evidence network (see Sections 1 and 5), it might be the case that treatments are comparable in theory but not in practice. For example, interferon, glatiramer acetate, and natalizumab are commonly used in clinical practice for patients with relapsing–remitting multiple scleroses, whereas mitoxantrone, methotrexate, cyclophosphamide, or azathioprine are more frequently given to patients with a progressive disease. However, evidence to support this clinical 'tradition' is not solid, and it would be appealing to compare all these treatments. In practice however, transitivity will be violated, as comparisons might differ with respect to disease severity (Casetta *et al*., 2007; Sudlow and Counsell, 2003).

3.2. Consistency

Consistency refers to the agreement between direct and one or more indirect sources of evidence. It is a property of a 'closed loop' (a path that starts and ends at the same node) which in the case of three interventions, is simply a triangle. The assumption of consistency is linked to the assumption of transitivity, as the former is the extension of the latter across all indirect comparisons in the loop. In a simple triangular loop, consistency holds when transitivity can be assumed for at least two out of the three nodes because if A and B are transitive, then C is transitive as well.

The assumption of consistency can be evaluated statistically by comparing μ^D_{AB} and μ^I_{AB} in a simple *z*-test (often called the Bucher method) (Bucher *et al*.,1997). Alternatively, one could estimate the inconsistency in the ABC evidence triangle as $I_{ABC} = \left|\mu^D_{AB} - \mu^I_{AB}\right|$ (often called 'inconsistency factors') and its 95% confidence interval as $I_{ABC} \pm 1.96\sqrt{\left(SE\left(\mu^D_{AB}\right)\right)^2 + \left(SE\left(\mu^I_{AB}\right)\right)^2}$ (Bucher *et al*., 1997; Caldwell *et al*., 2010; Salanti *et al*., 2009). If consistency holds, it seems reasonable to pool μ^D_{AB} and μ^I_{AB} to obtain μ^M_{AB}. Clinical interest may lie more in comparing μ^D_{AB} and μ^M_{AB}, but these two quantities are correlated, and some sort of predefined criteria should be employed to define disagreement such as those suggested in O'Regan *et al*. (2009).

Considering that the consistency assumption can take the form of the 'extension of transitivity' to hold across all treatments in a loop, interpretations of consistency can be formulated as

Consistency interpretation 1: Each treatment in the loop pertains to a similar definition independent of its comparator in studies. Treatment A is the same when compared to B or C; treatment B is the same when compared to A or C, and so on.

Consistency interpretation 2: The 'missing' treatments in each trial in the loop are missing at random. Table 1 shows the three sets of studies in the ABC loop and the 'missing' arms in each set. Under consistency, these missing arms are missing at random.

Consistency interpretation 3: There are no differences between observed and unobserved effects for every comparison in the loop beyond those attributed to heterogeneity. Table 1 shows the directly estimable and the unobserved relative treatment effects for the three comparisons in the ABC loop. Consistency claims that all true underlying effects belong to the same comparison-specific distribution (as shown in Figure 4) irrespective of whether they are directly observed or not.

Consistency interpretation 4: All sets of trials grouped by comparison are similar with respect to the distribution of effect modifiers. This means that the distributions of an effect modifier in AC, BC, and AB studies are similar.

The assumption of consistency can be reflected mathematically in the consistency equation $\mu_{AB} = \mu_{AC} - \mu_{BC}$. If there is no direct evidence for the relative effectiveness of A versus B, then the consistency assumption reduces to transitivity, and the equation suggests that we can derive μ_{AB} indirectly via C (right part of the equation). If there is direct evidence to estimate μ_{AB} (left part of the equation), consistency claims that the two pieces of evidence give

the same result. Re-arranging the parts of the equation shows that one consistency equation is enough to reflect consistency for all three comparisons.

Note that the absence of consistency does not necessarily mean that all indirect comparisons in the loop are invalid. Consider, for example, that treatment C is transitive, and AC and BC trials are similar regarding the distribution of effect modifiers (e.g. all studies are carried out in adults with a similar distribution in age), so that μ^I_{AB} is a valid estimate of the relative effectiveness of A versus B for the given setting and population. If now, the AB studies have all being carried out in younger populations (e.g in adolescents), then the consistency assumption may not hold; both μ^I_{AB} and μ^D_{AB} are valid, but μ^M_{AB} is meaningless.

3.3. Conceptual and statistical connections between transitivity, consistency, and heterogeneity

Heterogeneity can be perceived as the property of a pairwise meta-analysis and is defined as the 'disagreement' between the study-specific relative treatment effects beyond what can be explained by chance. This 'disagreement' is measured by the heterogeneity variance in the distribution of the true relative effects (Figure 3). Inconsistency can be defined in similar terms; it is the disagreement between the directly estimable $\delta_{iAC}, \delta_{iBC}, \delta_{iAB}$ relative effects in AC, BC, and AB studies and the unobserved $\delta^*_{iAC}, \delta^*_{iBC}, \delta^*_{iAB}$ true relative effects beyond what can be explained by heterogeneity (see Consistency interpretation 3). Alternatively, inconsistency can be seen as disagreement between direct and indirect mean relative effects beyond what chance can explain (heterogeneity is accounted for in the estimation of the 'mean' indirect and direct relative effects).

Extension of the notion of homogeneity across many different comparisons is a *sufficient*, but *not necessary*, condition of consistency. This means that homogeneity within each set of trials AC, BC, and AB does not guard against inconsistency in an ABC loop. Figure 2a shows three (fictional) homogeneous sets of studies that create an inconsistent ABC loop, as shown by the 95% confidence intervals of the inconsistency factor I_{ABC}.

The assumption of consistency is often equated with its statistical manifestation (e.g. an inconsistency factor compatible with zero or a statistically nonsignificant *z*-test). There are two reasons why results from the statistical evaluation of the consistency assumption using the Bucher method or more advanced methods should be interpreted with great caution. First, indirect comparison is often used when direct evidence is limited and only a few studies contribute to the estimation of μ^D_{AB}; this makes statistical testing underpowered. Second, there is an inverse association between the amount of heterogeneity in the pairwise comparisons and the power of testing for inconsistency. The mean effects μ^D_{AC} and μ^D_{BC} are often estimated using the random effects models in order to encompass a small to moderate amount of heterogeneity. The forest plot in Figure 2b makes clear the trade-off between statistical detection of inconsistency and heterogeneity when the random effects model is used: large heterogeneity will increase the imprecision in the pairwise meta-analysis estimates, resulting in larger confidence intervals for μ^I_{AB}, which will be compatible with μ^D_{AB}, thus, concealing a potentially important inconsistency (Figure 2b). This is useful to bear in mind when the heterogeneities in the pairwise meta-analyses are large. As a statistically nonsignificant *Q* test should not be interpreted as evidence of homogeneity in a traditional meta-analysis, insignificant *z*-test results or I_{ABC} compatible with zero should not be taken as proof for the absence of inconsistency. Finally, the lack of direct evidence of any comparison makes the testing of the for consistency impossible, but the transitivity assumption is still needed to derive indirect estimate.

The crucial issue of how to address inconsistency, if present, is addressed in Section 4.5 after extending the statistical methods and the assumption for more general evidence structures.

4. Extending indirect and mixed treatment comparison in complex evidence structures network meta-analysis

4.1. Network meta-analysis

In complex evidence networks, indirect estimates can be derived via many possible routes and not necessarily by using a single common comparator. Similarly, mixed estimates can be obtained by synthesizing several indirect estimates with a direct estimate. By focusing, for example, on a particular comparison A versus B, one can derive indirect estimates via all 'anchor' treatments, that is, treatments that link to A and B (for example, in Figure 5 μ^I_{ABviaC}, μ^I_{ABviaD}) and synthesize them with μ^D_{AB} by using the inverse variance method. The transitivity assumption is underlying the process for all 'anchor' treatments (here, C and D), and consistency between direct and indirect estimates can be evaluated. However, with an increased number of treatments and comparisons, this multistep process becomes a very tedious task. Moreover, the focus of the analysis might not be on a particular comparison but in comparing all treatments. Statistical methods that synthesize all pieces of evidence for all comparisons in the network in one step exist. In agreement with the previous use of the term, network meta-analysis refers to the joint analysis of a trial network, whereas the term 'mixed' characterizes the evidence (and estimates) for a specific comparison that is located in a closed loop, and 'indirect' the evidence about a comparison with which trials do not exist to compare directly (Jansen *et al.*, 2011). Network meta-analysis, by synthesizing simultaneously all evidence for all comparisons in the network, offers several advantages. The issues of increased power and the ability to

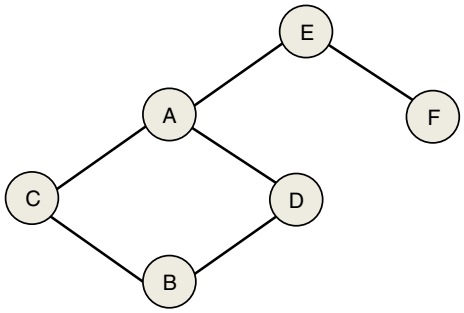

Figure 5. Network of interventions

compare all treatments even in the absence of direct evidence have been extensively discussed in the literature (Caldwell *et al*., 2005; Cooper *et al*., 2011; Jansen *et al*., 2008; Salanti *et al*., 2008a; Sutton *et al*., 2008). The advantage of the nonselective use of the evidence should also be highlighted in contrast to multistep approach (Ades, 2004; Hawkins *et al*., 2009b; Jansen *et al*., 2008; Salanti *et al*., 2008a; Sutton *et al*., 2008) and is discussed in the next section.

4.2. Why considering network meta-analysis is the best practice in comparing multiple interventions reviews

It is important to note that many reviews do compare several treatments, but they do not explicitly use a network meta-analysis. Some reviews perform separate comparison-specific meta-analyses and synthesize pieces of evidence in a multistep process, as described previously. Many reviews present the head-to-head meta-analyses results and conduct informal indirect comparisons, such as by using a statistical test to compare μ^D_{AC} and μ^D_{BC} in order to infer about the AB comparison, or compare the magnitude and significance of μ^D_{AC} and μ^D_{BC} in a narrative way. This is as undesirable and misleading as inferring about subgroup meta-analysis on the basis of the statistical significance of effects within subgroups. In a recent evaluation of 62 systematic reviews that used some form of indirect inference, 35 were found to have used methods that do not estimate an indirect relative effect but used informal methods instead (Edwards *et al*., 2009).

Valid criticisms of network meta-analysis are equally applicable to informal, nonquantitative, narrative, indirect comparison of several seemingly independent meta-analyses results. In fact, informal approaches can be dangerous as they are not associated with any of the 'safeguarding' methods put in place to express formally the disagreement between the different sources of evidence, such as the evaluation of inconsistency. The narrative presentation of several meta-analyses and the selective use of indirect comparison can be as confusing and misleading as a qualitative description of individual study results, which does not explicate contradicting findings. Considerations for the statistical synthesis in a comparing multiple interventions review should follow the same principles with traditional systematic reviews. The selective presentation of individual study results without their synthesis or without addressing the observed disagreements (e.g. heterogeneity) that potentially prohibit meta-analysis is widely accepted as a poor practice for systematic reviews. It is equally unwise to ignore (altogether or selectively) indirect evidence without justification. In systematic reviews, exclusion of a study from the meta-analysis is justified only when this study is believed to have a high risk of bias in the results. Similarly, the choice of the pieces of evidence to synthesize in a comparing multiple interventions review should be done after considering the potential of bias and the risk of violating the transitivity assumption. Considering a network meta-analysis will naturally lead reviewers in evaluating the assumption, will encourage exploration of potential disagreements between pieces of evidence giving better insight to the research question, and will add transparency to the choices being made regarding data synthesis. If, finally, a network meta-analysis is deemed appropriate, it will provide the most precise evidence about the relative effects of the interventions.

4.3. Core statistical network meta-analysis models

Network meta-analysis can be seen as a special case of meta-regression on dummy variables that reflect the comparisons. The idea is that, when consistency is assumed, all present comparisons can be written as functions of a minimum set of comparisons called 'basic parameters' (Lu & Ades, 2004). With T treatments in the network, only $T-1$ comparisons are needed to represent all possible comparisons, once the consistency equations are employed. For example, in the simple case of three treatments A, B, and C, the effect sizes will be regressed on two dummy variables $V1$ and $V2$ that represent, say, comparisons AC and BC. An AC study will have $V1=1$ and $V2=0$, whereas a BC study will be modeled with covariate values $V1=1$ and $V2=-1$ (see Salanti *et al*., 2008a). Note that the choice of the $T-1$ comparisons to enter the model does not affect the results; a meta-regression using AC and BC as covariates will give results equivalent to the meta-regression using AC and AB covariates. Standard meta-regression routines (such as metareg in STATA (StataCorp LP, 2011)) can be used to fit the model.

In the presence of multi-arm trials, more than one effect size per study is extracted. The observed effects within a study and their underlying random effects are correlated. Multivariate meta-regression, which is often used to model jointly multiple outcomes, can be used for this purpose, treating the different comparisons as different outcomes

(Jackson *et al*., 2011; Mavridis and Salanti, 2012). The freely available R packages mvmeta and metasem can be used for this purpose (Cheung, 2011; Gasparrini, 2012), and a new version of mvmeta in STATA is available (White, 2011). However, the use of the standard multiple meta-regression routines to fit network meta-analysis should be carried out with caution. The model involves the between-studies variance–covariance matrix, a matrix with the heterogeneity parameter in the diagonal, and the between-studies covariance c in the off-diagonal. When the observations do not refer to different outcomes but to different comparisons, the between-studies covariance c is a function of the heterogeneity, and this relation needs to be embedded in the model. Standard software routines that use a general structure for the covariance will not model the data properly, although in the presence of only few multiple arm trials, ignoring the relation between heterogeneity and between-trials covariance should not have much impact on the results. The current special issue of *Research Synthesis Methods* includes an article which shows how to fit network meta-analysis using the new mvmeta STATA command, accounting properly for correlations induced by multi-arm trials (White *et al*., 2012).

Despite the fact that network meta-analysis can be fit as a frequentist meta-regression model, the vast majority of applications have been carried out within a Bayesian context (Lu and Ades, 2004). Extension of the models into complex types of data, incorporation of prior beliefs, easier estimation of predictions, and expressing results using probabilistic statements make the Bayesian modeling more attractive. For an overview of the models, see Salanti *et al*. (2008a) All Bayesian applications have been carried out in WinBUGS (Lunn *et al*., 2000), and codes for various types of data and models have been made available online (see bris.ac.uk/cobm/research/mpes/mtc.html, www.mtm.uoi.gr, nicedsu.org.uk/index.htm).

A common feature in most applications of network meta-analysis is that they use a common heterogeneity parameter for all treatment comparisons. This offers an advantage as estimation of the heterogeneity for comparisons, which are informed directly by only a few studies, borrows strength from comparisons with many studies leading to a better estimation of the heterogeneity. On the other hand, the assumption of common heterogeneity can lead to paradoxes, such as network estimates being less precise than pairwise estimates for some comparisons. This can happen for comparisons that are informed directly by trials with no or little heterogeneity. In a network meta-analysis, the 'homogeneous' comparisons share the same heterogeneity parameter with comparisons exhibiting larger heterogeneity, and this will lead into more imprecise network estimates for the former. An alternative model, which allows for different heterogeneity parameters across comparisons and various structures for the between-studies variance–covariance matrix, has been discussed in Lu and Ades (2009). The authors derived the exact structure of the matrix as imposed by the consistency equations and suggested an approach based on the Cholesky decomposition to ensure positive definitiveness. However, the added benefit of this complex approach and its feasibility in the presence of multi-arm trials remains to be shown in applications.

4.4. The assumptions of transitivity and consistency for a network meta-analysis

Distinguishing between direct and different sources of indirect evidence in a network of interventions is often impossible but also unnecessary. Indirect information 'flows' in the network via several routes using different, often multiple, 'anchor' treatments. It is obvious that, in a network meta-analysis, transitivity is assumed for all 'anchor' treatments (that is, for all treatments that link two others), in the sense that they link comparable sets of studies and provide valid indirect estimates. Therefore, investigators should carefully consider the various scenarios in Section 3.1 and evaluate potential sources of violation. This implies looking at the distribution of potential confounders across all comparisons, considering whether all missing arms can be assumed to be missing at random, and so forth. A network for which transitivity can be assumed in all 'anchor' treatments is, by definition, consistent. Inconsistency in a network can be manifested as a disagreement between different sources of evidence for the same comparison. For example, inconsistency can be created by the disagreement between two indirect estimates μ^{I}_{ABviaC} and μ^{I}_{ABviaD} for the same comparison AB 'anchored' at different treatments C and D or between indirect and direct estimates μ^{I}_{ABviaC} and μ^{D}_{ABviaC}. Both conceptual and statistical evaluation of the consistency assumption is required to obtain valid network meta-analysis estimates. For instance, in a star network (a network where all treatments have been compared with a common treatment but not between themselves), there are no multiple sources of evidence for the same comparisons, so no inconsistency can occur; still, the assumption of transitivity should be evaluated conceptually for all indirect comparisons to derive valid network meta-analysis estimates. In the network of Figure 5, testing for consistency in the closed ABCD loop will give some indication about the transitivity in nodes A, B, C, and D, but the plausibility of the transitivity assumption in node E needs to be evaluated as well.

Simple methods to detect inconsistency are based on extending the idea of statistical comparison between direct and indirect estimates. It has been suggested that inconsistency factors I_{ABC} and their 95% CIs are calculated for all loops formed in a network and presented as forest plots to identify loops for which $I_{ABC} \neq 0$ (Salanti *et al*., 2009). This approach suffers from problems related to multiple testing of correlated quantities; if transitivity does not hold for a given 'anchor' treatment, this will be shown as inconsistency in several loops involving this treatment. Caldwell *et al*. (2010) presented a test for simultaneously testing for inconsistency in several loops that share only one comparisons, so is only applicable to a specific type of networks. There is currently limited empirical evidence about the occurrence of statistical inconsistency. A study evaluated 44 triangular networks,

and only three were found inconsistent (Song *et al*., 2003). An update of this study reported a higher proportion of inconstant loops (16 out of 112 triangles) (Song *et al*., 2011). O'Regan *et al*. (2009) empirically evaluated the agreement between indirect and mixed estimates that appear in networks of at least four treatments. Using fixed effects approaches, they concluded that indirect and mixed estimates do not differ, although the 51 comparisons they examined came from seven reviews only.

Approaches in evaluating inconsistency in a network as a whole, rather than splitting it into loops, have gained popularity but are more cumbersome to apply, and they also have their own drawbacks. For network models fitted within a Bayesian framework, the consistency assumption can be evaluated by comparing a model assuming consistency with one that does not, using the Deviance Information Criterion, a measure similar to Akaike's criterion (Spiegelhalter *et al*., 2002). The model without consistency does not use the consistency equations to derive indirect and mixed estimates and is similar to a series of independent and pairwise meta-analysis (which share the same heterogeneity parameter). The assumption of consistency is challenged when the inconsistency model presents, for the same data, a better trade-off between model fit and complexity. The most important drawback with this method is that results may depend on the parameterization of the multi-arm trials. Approaches that test as well as account for inconsistency are discussed in the following paragraph.

4.5. How to deal with inconsistent networks

Despite the best efforts of investigators to construct a consistent network, significant inconsistency may arise. In this case, investigators should address it by considering a strategy similar to the one followed once heterogeneity is detected in a simple meta-analysis. Meta-analysis aims to address some of the challenges created by often contradicting individual experiments. Exploration of heterogeneity has proven to be a key feature of systematic reviews, often providing insight into the clinical question under investigation. It has become an established practice for review protocols to list potential sources of heterogeneity and use them to form more homogeneous subgroups and generate hypotheses for effect modifiers. Similar practices should be followed in network meta-analysis with a clear strategy to deal with the inconsistency developed a priori and described in the

Table 2. Possible strategies that respond to heterogeneity and inconsistency.

Action	Heterogeneity	Inconsistency
Check the data	Studies that 'stand out' in the forest plot are checked for data extraction errors (Gotzsche *et al*., 2007; Horton *et al*., 2010).	Simple loop inconsistency graphs can indicate studies with data extraction errors. Inconsistency in loops where a comparison is informed by a single study is particularly suspicious for data errors.
Try to bypass	There is empirical evidence that some measures are associated with larger heterogeneity than others (Deeks, 2002; Friedrich *et al*., 2011).	Empirical evidence is needed to evaluate the impact of the effect measure on statistical inconsistency, particularly given the trade-off between heterogeneity and statistical inconsistency when random effects models are used.
Resign to it	Investigators may decide not to undertake meta-analysis in the presence of excessive heterogeneity.	Investigators may decide not to synthesize the network data in the presence of excessive inconsistency.
Encompass it	Apply random effects meta-analysis (DerSimonian and Laird, 1986; Higgins *et al*., 2009).	Apply models that relax the consistency assumption by adding an 'extra' inconsistency-specific random effect (Lu and Ades, 2006; White *et al*., 2012). However, as random effects are not a remedy for excessive heterogeneity and should be applied only for unexplained heterogeneity, inconsistency models should be employed to reflect inconsistency in the results, not to adjust for it.
Explore it	Use prespecified variables in a subgroup analysis or meta-regression (Thompson and Higgins, 2002).	Split the network into subgroups or use network meta-regression to account for differences across studies and comparisons. Specify the variables in the protocol, including bias-related characteristics. See for examples Cooper *et al*. (2009); Dias *et al*. (2010c) Salanti *et al*. (2009).

protocol. Table 2 summarizes the strategies currently followed in meta-analysis to address heterogeneity and their adaptation for the problem of inconsistency.

The issue of choosing between direct and indirect evidence may arise if important inconsistency is identified. There are arguments for giving priority to direct evidence; the strongest one is that it does not rely on the transitivity assumption. However, there can be cases where indirect evidence may be preferable if the direct evidence is subject to bias. Empirical data have shown that indirect estimates may be free of sponsorship bias that can exaggerate the result of direct comparisons (Song *et al*., 2008). Rather than arbitrarily excluding studies or entire comparisons that do not seem to add up, investigators should try to explain the observed inconsistency and explore causes of disagreement, which can be enlightening. Dias *et al*. (2010b) introduced the 'node-splitting' approach which separates the evidence conveyed by the network for any given comparison into direct and indirect. The process is computationally intensive and becomes cumbersome for networks with many treatments and comparisons but offers the advantage of identifying potential 'anchor' treatments for which transitivity may not hold.

When inconsistency is evident, researchers may decide to synthesize the data in a way that reflects the extra uncertainty because of inconsistency. Lu and Ades (2006) introduced a model that accounts for inconsistency by adding an extra random effect to each loop in which inconsistency can occur. In the presence of inconsistency, the source-specific mean relative treatment effects (i.e. the direct and indirect) are allowed to differ by a small random quantity instead of being claimed identical. This model has an elegant analogy to random effects model which postulates that the true study-specific relative treatment effects are not identical but are exchangeable. The variance of the 'loop' inconsistencies is called inconsistency variance, in analogy to the heterogeneity variance. One can then compare inconsistency variance with heterogeneity variance to evaluate the assumption that the heterogeneity may be sufficient to explain differences between sources of evidence (see interpretation of Transitivity 3 in Section 3.1). The main disadvantages of the model are that results may depend on the parameterization of the multi-arm trials and that it might be sensitive to the prior distributions for the variance parameters (Lambert *et al*., 2005).

4.6. Extending the network meta-analysis

4.6.1. Network meta-regression. Adjusting for factors that can vary across comparisons may improve the plausibility of the transitivity assumption while reducing heterogeneity. Characteristics such as differences in baseline risk (if there is a common comparator) and sample size (as a single proxy for study quality) can be considered. Several studies have extended the network meta-analysis model by including covariates to account for variation in dose, adjuvant treatments, or sponsorship (Cipriani *et al*., 2009; Salanti *et al*., 2009; Nixon *et al*., 2006; Cipriani *et al*., 2011; Cooper *et al*., 2009). An important application of network meta-regression is to account for differences in trial characteristics associated with the risk of bias, such as allocation concealment (Dias *et al*., 2010a). The issue of bias is particularly challenging, but also promising, and is discussed in Section 4.6.2.

Network meta-regression is expected to be subject to the same problems as simple meta-regression; ecological bias when aggregated data are used as covariates, low power with few studies and high false positive rates if fixed effects are used in the presence of heterogeneity (Higgins and Thompson, 2004). The suitability of the adjustment can be judged by monitoring changes in the heterogeneity parameter, by contrasting adjusted mixed estimates to the direct estimates and by comparing the goodness of fit of the models.

4.6.2. The issue of bias revisited when comparing multiple interventions. The credibility of the summary estimate from a conventional meta-analysis is challenged when studies with a high risk of bias are included. This concern carries over when comparing multiple interventions. Meta-regression can be used to adjust for bias in a way similar to the standard pairwise meta-analysis. For example, an indicator variable can be created to define the 'appropriate', 'unclear', and 'inappropriate' method of allocation concealment and can be included in the network meta-regression. This approach (with a modification for probabilistic modeling of the 'unclear' risk of bias) has been used by Dias *et al*. (2010c). However, in a network meta-analysis, it is necessary to make assumptions about the direction of the bias. If all trials are placebo-controlled trials (so that the network is star-shaped), it is expected that the bias will favor the active treatment. When two active treatments are compared, stronger assumptions are required for the direction of bias, such that the newer or the sponsored treatment is favored. Directionality assumptions can be embedded in the model in the form of an index variable.

Adjusting for bias in a network of interventions offers the advantage of increased power compared with the traditional meta-analysis. Consider, for example, that comparison AB is informed by very few studies or studies that all fall in the same quality category, that is, they all have poor allocation concealment. Then, conducting sensitivity analysis or adjusting via meta-regression is suboptimal or impossible. However, if these studies are part of a network meta-regression model, the bias coefficient for allocation concealment estimated in the network from all involved comparisons is imposed to AB studies as well, and the summary estimate for AB can be adjusted. This rests on the assumption that the magnitude of bias is similar across comparisons, which can be defendable in many clinical settings.

A similar approach has been taken to address small study effects and publication bias in a network of antidepressants (Moreno *et al*., 2011). The association between sample size, heterogeneity, and the probability of publication is often manifested as a funnel plot asymmetry and has long been a very challenging issue in

meta-analysis. The same problem applies to networks of interventions and can take a more extended form: comparisons that do not give significant results may be underrepresented or completely missing from the network, and their relative effects will primarily be informed by indirect evidence. As publication bias and selective reporting might affect interventions and comparisons in different ways depending on the clinical context, the problem of publication bias in the network should be considered carefully. Attempts have been made to associate the possibility of selection bias with asymmetry measures of the network, borrowing the methodology used in ecology (Salanti *et al*., 2008b).

Network meta-analysis also offers the opportunity to evaluate biases that can affect an entire research field. Characteristics associated with the design of the study (e.g. blinding) can be studied jointly in a *network meta-epidemiology*. Regression bias coefficients estimated within a network can be assumed exchangeable across a collection of networks extending the idea of meta-epidemiology (Sterne *et al*., 2002) and providing large-scale evidence for potential sources of bias (Salanti *et al*., 2010b).

Because network meta-analysis combines studies that compare a treatment against a variety of comparators, it enables researchers to explore biases that are not identifiable in a head-to-head meta-analysis. Optimism bias, associated with the use of novel interventions, has been a concern difficult to address, so far (Djulbegovic *et al*., 2011; Heres *et al*., 2006; Soares *et al*., 2005). However, in a network of interventions, the same treatment C can be the newer and, hence, the 'favored' in a comparison AC, but the older in another comparison BC. This allows us to study apparent changes in the effectiveness of C because of optimism. In a network meta-epidemiology model, three networks on different cancer treatments were linked to estimate the novelty bias effect (Salanti *et al*., 2010b).

The presence of bias can manifest itself as an inconsistency. Going back to one of our starting arguments, exploration of possible sources of inconsistency may enhance understanding of the way that interventions appear to work under different conditions but also give better insight into how the present evidence is formed and whether certain sources are more reliable than others.

4.6.3. Multivariate versions of the models and applications to specific types of data. Treatments are typically compared with respect to many outcomes, which are often correlated. Extensions of multivariate meta-analysis will be desirable to account simultaneously for multiple treatments and outcomes. The expected benefits would be increased precision and lower risk of introducing bias because of selective reporting, particularly for safety outcomes (Jackson *et al*., 2011; Riley *et al*., 2004). Although a general framework has not been developed yet, several models for specific types of correlated outcomes are available. A flexible Bayesian model for competing outcomes (such as effectiveness and safety) is described in Ades *et al*. (2010). Network meta-analysis models for an outcome measured at multiple follow-ups or for outcomes in the form of transition probability between various health states are also available in the literature (Lu *et al*., 2007; Price *et al*., 2010). Note that the majority of these models are tailored to the data structure at hand, and consequently, investigators who want to use them need to adjust them to the particular problem. Applications of the network meta-analysis to other types of outcome data, except from binary and continuous, include so far Poisson (Dias *et al*., 2010c), time to event data (Jansen, 2011; Kyrgiou *et al*., 2006; Welton *et al*., 2008), and mixing count with survival data (Woods *et al*., 2010).

4.7. Presentation of results and reporting of network meta-analysis

When presenting results from a network meta-analysis, it is useful to show both the direct and the mixed estimates, along with their 95% confidence intervals and comment on possible disagreements. In a consistency model, all pairwise comparisons can be obtained, and the effect sizes are often presented in the form of a 'league table' (see, for example, Padwal *et al*. (2011) and Cipriani *et al*. (2011)) or in a forest plot against a common comparator (see for example Sciarretta *et al*. (2011)). A review of several of these methods can be found in Salanti *et al*. (2010a).

Presenting the results of a network meta-analysis using absolute measures is valid as long as there is a common comparator of interest. Examples can be found in Mills *et al*. (2009), where efficacy of each treatment was presented using the expected absolute mortality given a baseline death probability, and in Manzoli *et al*. (2009), where hemagglutination inhibition with various influenza vaccines was presented as the absolute change from a baseline vaccine for which a prior distribution for the outcome was assumed. Presentation of the results using predictive intervals (Higgins *et al*., 2009), although infrequent, best conveys the uncertainty due to heterogeneity (Salanti *et al*., 2010a).

Ranking measures and probabilities have become popular as they provide an understandable gateway to the results, particularly in the presence of many competing treatments. The probability for each treatment of being the best is often calculated when the network model is fitted within the Bayesian framework, but methods are also available for frequentist approaches. The probability of being the best has the disadvantage that it does not reflect the spread of rankings for the treatments and may be misleading (Jansen *et al*., 2011; Salanti *et al*., 2010a). An obvious solution to this is to calculate the probabilities for all ranks. The probability of each treatment to achieve each possible rank can be plotted to yield 'rankograms' (Cipriani *et al*., 2011; Cipriani *et al*., 2009) or may be presented in a table (Hawkins *et al*., 2009b). Presentation of the cumulative ranking curves in a single plot and a numerical summary of

the area below the cumulative raking curve for each treatment is useful as it gives a clear ordering of all treatments on the basis of a summary of the rank probabilities, see Salanti *et al*. (2010a). Probabilistic statements can also be used to facilitate clinical interpretation of the results; for example, the probability that the relative effectiveness of each treatment reaches a particular threshold of clinical importance has been used to present the results from a network meta-analysis of nonsteroidal anti-inflamatory drugs (Trelle *et al*., 2011). Presentation of results on the basis of the statistical significance of pairwise comparisons, as suggested by Fadda *et al*. (2011), may be misleading as it overemphasizes the importance of *p*-values.

Reporting of the assumptions underlying the indirect and mixed comparison of multiple interventions varies considerably across reviews. In a recent summary, only 12 out of 30 reviews that compared or combined indirect and direct evidence commented on the consistency assumption (Song *et al*., 2009). The authors identified several problems in both conducting and reporting reviews that use indirect evidence, such as inappropriate statistical methods, nonsystematic search of the literature, and lack of transparency as to whether presented results pertained to direct, indirect, or mixed evidence. A more recent study evaluated 41 reviews that include indirect and mixed comparisons against a list of quality criteria for conduct and reporting (Donegan *et al*., 2010). They concluded that the assumptions are not routinely assessed, and reporting of all necessary information to judge the appropriateness of the mixed or indirect estimate is suboptimal. It is interesting to note, however, that both empirical studies by Song *et al*. (2009) and Donegan *et al*. (2010) implicitly assume that direct evidence is preferable to indirect. For example, Donegan and colleagues included a quality item 'Does the review state that more trials providing direct evidence are needed?'. It is not clear, however, whether it is indeed a good practice to call *by default* for more randomized trials if mixed evidence is conclusive. This assumption of superiority of the direct evidence has impacted on the conclusions of both empirical studies. Two useful lists of recommendations for undertaking, reporting, and evaluating a network meta-analysis can be found in Jansen *et al*. (2011) and Ades *et al*. (2012). These lists could be used as a starting point to extend the PRISMA statement for network meta-analysis (Liberati *et al*., 2009).

5. Setting up a body of evidence for network meta-analysis

The assumptions of the methodology discussed so far should be taken into account at the early stages of the review, where decisions about collecting and evaluating the body of evidence are made. Researchers may decide to include in the review all treatments for the same condition or a subset of them that is of particular interest, but it is important that only treatments with the same indication are compared. Therefore, when deciding upon the set of competing treatments, reviewers need to make sure that they are 'jointly randomizable': it is possible to imagine that a single randomized trial could have been designed to compare all these treatments (see Section 3.1 for Transitivity interpretation 4).

Once reviewers have decided which treatments they want to compare, two critical issues need to be addressed. The first one is the definition of the included treatments. Each node in a network represents a treatment which, however, is rarely defined in an identical way across trials; more often, a node refers to a collection of arguably similar treatments. The breadth of the definition of a node is an important decision that requires substantial input from clinicians and methodologists. The second issue is the potential inclusion of further 'anchor' treatments that are not of direct clinical interest but provide indirect evidence.

Both decisions relate to the ever present dilemma of lumping and splitting: broader definitions of each node and inclusion of more 'anchor' treatments allow synthesis of more data, increase precision, and enhance the chances of generalizable conclusions. At the same time, the transitivity assumption may seem less defensible when a node embodies many treatment definitions, and there is the risk of introducing heterogeneity and inconsistency into the meta-analysis model. Hawkins *et al*. (2009b) support the use of a wider evidence body and show the benefits when expanding the inclusion criteria for treatments. They suggest a search algorithm to identify indirect evidence via 'anchor' treatments when a particular set of treatments is on focus (Hawkins *et al*., 2009a). On the other hand, Cooper *et al*. (2011) showed the potential pitfalls when the evidence base is extended too much. In one of their examples, the increase in precision gained by expanding the network is negated by increased heterogeneity.

Whereas selective use of direct and indirect information should be avoided when possible, the clinical and methodological characteristics of each research field should guide this decision. If, for example, direct evidence is believed to be biased because of, say, optimism or sponsorship bias, possible sources of indirect evidence may be deemed appropriate. In summary, issues of precision, the plausibility of the transitivity assumption, and the presence of inconsistency should be combined with the possibility of bias in the direct and indirect comparison when setting up the body of evidence. In order to avoid selective use of the evidence, such decisions should be ideally made at the planning stage and described in the protocol of the review.

Presentation of the evidence network can be useful, and most published applications of network meta-analysis do include such information in the form of a table or a graph. Presentation of the network plot, possibly enhanced to show more information about the data (e.g. by plotting nodes and edges proportional to the amount of information they carry), can be informative; an R routine for this purpose can be found in www.mtm.uoi.gr.

Graphical display of the direct and indirect comparison in restricted and extended versions of the network may be helpful to explore how information enters the evidence base and influences the results, see for example, Hawkins *et al*. (2009b).

6. Moving forward toward establishing network meta-analysis as an integral part of comparing multiple intervention reviews

Many researchers and clinicians working in the field of producing evidence for decision-making are at this exciting point of moving a level up in the evidence pyramid and 'upgrading' from pairwise to network meta-analysis. Conventional systematic reviews and meta-analyses are long recognized as very useful tools in providing evidence for the effectiveness of interventions. When first appeared, meta-analysis was greeted with skepticism. The conditions under which meta-analysis yields valid results are now established, although debate about particularly difficult issues, such as the random versus fixed effects or publication bias, is on-going. Systematic reviews that compare multiple interventions have been carried out for many years now, but their statistical component, network meta-analysis, is still debated. The assumption underlying network meta-analysis as outlined in this paper is perceived as an important barrier to the wider uptake of the method. What is important to realize though is that multiple indirect comparisons, often implicit or explicit in reviews, require the same assumptions with network meta-analysis, and they suffer from lack of transparency. Network meta-analysis will be eventually adopted in practice via a process similar to the one that established meta-analysis as a desirable part of a systematic review: via many well-described applications and methodological developments that will show when network meta-analysis gives valid results and how these should be interpreted.

All systematic reviews should consider registration of protocols (Booth *et al*., 2012; Van der Wees *et al*., 2012); comparing multiple interventions reviews even more so. The protocol safeguards against selective use of the indirect evidence and 'anchor' treatments. In the protocol, reviewers shall describe the strategy they plan to follow to evaluate transitivity and consistency, list factors that could introduce heterogeneity and inconsistency, describe the conditions under which they will employ network meta-analysis to synthesize the results, and specify how they plan to deal with inconsistency if present. If the review is carried out as a Cochrane review, the registration of the protocol is compulsory; otherwise, reviewers can use an online database (such as PROSPERO (Booth *et al*., 2012; Palepu *et al*., 2011) www.crd.york.ac.uk/prospero/). Extending the PRISMA statement (Liberati *et al*., 2009) for comparing multiple interventions, reviews will greatly improve reporting, whereas extensions of AMSTAR (Shea *et al*., 2007) will help the end users of research in evaluating the evidence provided by this type of review.

Acknowledgements

The work presented in this paper is part of the IMMA 260559 project funded by the European Research Council. I would like to thank Chris Schmid, Julia Littell and Toshi Fukurawa for helpful comments.

References

Ades AE. 2004. Commentary: evidence synthesis and evidence consistency. *Int J Epidemiol* **33**(2): 426–427.

Ades AE, Caldwell DM, Reken S, Welton NJ, Sutton AJ, Dias S. 2012. Technical support document 7: evidence synthesis of treatment efficacy in decision making: a reviewer's checklist. *Technical Support Document Series No.2*, NICE Decision Support Unit.

Ades AE, Mavranezouli I, Dias S, Welton NJ, Whittington C, Kendall T. 2010. Network meta-analysis with competing risk outcomes. *Value Health* **13**(8): 976–983.

Baker SG, Kramer BS. 2002. The transitive fallacy for randomized trials: if A bests B and B bests C in separate trials, is A better than C? *BMC Med Res Methodol* **2**(1): 13.

Barbui C, Cipriani A, Furukawa TA, Salanti G, Higgins JP, Churchill R, Watanabe N, Nakagawa A, Omori IM, Geddes JR. 2009. Making the best use of available evidence: the case of new generation antidepressants: a response to: are all antidepressants equal? *Evid Based Ment Health* **12**(4): 101–104.

Booth A, Clarke M, Dooley G, Ghersi D, Moher D, Petticrew M, Stewart LA. 2012. The nuts and bolts of PROSPERO: an international prospective register of systematic reviews. *Systematic Reviews* **1**(2): DOI: 10.1186/2046-4053-1-1

Bucher HC, Guyatt GH, Griffith LE, Walter, SD. 1997. The results of direct and indirect treatment comparisons in meta-analysis of randomized controlled trials. *J Clin Epidemiol* **50**(6): 683–691.

Caldwell DM, Ades AE, Higgins JP. 2005. Simultaneous comparison of multiple treatments: combining direct and indirect evidence. *BMJ* **331**(7521): 897–900.

Caldwell DM, Welton NJ, Ades AE. 2010. Mixed treatment comparison analysis provides internally coherent treatment effect estimates based on overviews of reviews and can reveal inconsistency. *J Clin Epidemiol* **63**(8): 875–882.

Casetta I, Iuliano G, Filippini G. 2007. Azathioprine for multiple sclerosis. *Cochrane Database Syst Rev* **4**: CD003982.

Cheung MWL. 2011. metaSEM: meta-analysis: a structural equation modeling approach. *R package version 0.5-3* available from: http://courses.nus.edu.sg/course/psycwlm/Internet/metaSEM/ [Accessed 3 January 2012].
Cipriani A, Barbui C, Salanti G, Rendell J, Brown R, Stockton S, Purgato M, Spineli LM, Goodwin GM, Geddes JR. 2011. Comparative efficacy and acceptability of antimanic drugs in acute mania: a multiple-treatments meta-analysis. *Lancet* **378**(9799): 1306–1315.
Cipriani A, Furukawa TA, Salanti G, Geddes JR, Higgins JP, Churchill R, Watanabe N, Nakagawa A, Omori IM, McGuire H, Tansella M, Barbui C. 2009. Comparative efficacy and acceptability of 12 new-generation antidepressants: a multiple-treatments meta-analysis. *Lancet* **373**(9665): 746–758.
Cooper NJ, Peters J, Lai MC, Juni P, Wandel S, Palmer S, Paulden M, Conti S, Welton NJ, Abrams KR, Bujkiewicz S, Spiegelhalter D, Sutton AJ. 2011. How valuable are multiple treatment comparison methods in evidence-based health-care evaluation? *Value Health* **14**(2): 371–380.
Cooper NJ, Sutton AJ, Morris D, Ades AE, Welton NJ. 2009. Addressing between-study heterogeneity and inconsistency in mixed treatment comparisons: Application to stroke prevention treatments in individuals with non-rheumatic atrial fibrillation. *Stat Med* **28**(14): 1861–1881.
Deeks JJ. 2002. Issues in the selection of a summary statistic for meta-analysis of clinical trials with binary outcomes. *Stat Med* **21**(11): 1575–1600.
DerSimonian R, Laird N. 1986. Meta-analysis in clinical trials. *Control Clin Trials* **7**(3): 177–188.
Dias S, Welton N, Ades A. 2010a. Study designs to detect sponsorship and other biases in systematic reviews. *J Clin Epidemiol* **63**(6): 587–588.
Dias S, Welton NJ, Caldwell DM, Ades AE. 2010b. Checking consistency in mixed treatment comparison meta-analysis. *Stat Med* **29**(7–8): 932–944.
Dias S, Welton N, Marinho V, Salanti G, Ades A. 2010c. Estimation and adjustment of bias in randomised evidence using mixed treatment comparison meta-analysis. *JRSS (A)* **173**(4).
Djulbegovic B, Kumar A, Magazin A, Schroen AT, Soares H, Hozo I, Clarke M, Sargent D, Schell MJ. 2011. Optimism bias leads to inconclusive results-an empirical study. *J Clin Epidemiol* **64**(6): 583–593.
Donegan S, Williamson P, Gamble C, Tudur-Smith C. 2010. Indirect comparisons: a review of reporting and methodological quality. *PLoS One* **5**(11): e11054.
Edwards SJ, Clarke MJ, Wordsworth S, Borrill J. 2009. Indirect comparisons of treatments based on systematic reviews of randomised controlled trials. *Int J Clin Pract* **63**(6): 841–854.
Fadda V, Maratea D, Trippoli S, Messori A. 2011. Network meta-analysis. Results can be summarised in a simple figure. *BMJ* **342**: d1555.
Friedrich JO, Adhikari NK, Beyene J. 2011. Ratio of means for analyzing continuous outcomes in meta-analysis performed as well as mean difference methods. *J Clin Epidemiol* **64**(5): 556–564.
Gasparrini A. 2012. Package 'mvmeta'. Multivariate meta-analysis and meta-regression. *R package* available from: http://cran.r-project.org/web/packages/mvmeta/mvmeta.pdf [Accessed 3 January 2012].
Gotzsche PC. 2000. Why we need a broad perspective on meta-analysis. It may be crucially important for patients. *BMJ* **321**(7261): 585–586.
Gotzsche PC, Hrobjartsson A, Maric K, Tendal B. 2007. Data extraction errors in meta-analyses that use standardized mean differences. *JAMA* **298**(4): 430–437.
Gibbs LE. 2003. *Evidenced-based practice for the helping professions: A practical guide with integrated multimedia*. Pacific Grove, CA: Brooks/Cole Thompson Learning.
Greenland S, Rothman KJ. 2008. Introduction to stratified analysis. *In Modern Epidemiology*. Lippencott-Raven Publishers, Philadelphia, 258–281.
Grimshaw J, McAuley LM, Bero LA, Grilli R, Oxman AD, Ramsay C, Vale L, Zwarenstein M. 2003. Systematic reviews of the effectiveness of quality improvement strategies and programmes. *Qual Saf Health Care* **12**(4): 298–303.
Hawkins N, Scott DA, Woods B. 2009a. How far do you go? Efficient searching for indirect evidence. *Medical Decision Making* **29**(3): 273–281.
Hawkins N, Scott DA, Woods BS, Thatcher N. 2009b. No study left behind: a network meta-analysis in non-small-cell lung cancer demonstrating the importance of considering all relevant data. *Value Health* **12**(6): 996–1003.
Heres S, Davis J, Maino K, Jetzinger E, Kissling W, Leucht S. 2006. Why olanzapine beats risperidone, risperidone beats quetiapine, and quetiapine beats olanzapine: an exploratory analysis of head-to-head comparison studies of second-generation antipsychotics. *Am J Psychiatry* **163**(2): 185–194.
Higgins JPT, Green S. 2008. Cochrane handbook for systematic reviews of interventions, 5.0.1 ed. The Cochrane Collaboration.
Higgins JP, Thompson SG. 2004. Controlling the risk of spurious findings from meta-regression. *Stat Med* **23**(11): 1663–1682.
Higgins JP, Thompson SG, Deeks JJ, Altman DG. 2003. Measuring inconsistency in meta-analyses. *BMJ* **327**(7414): 557–560.
Higgins JP, Thompson SG, Spiegelhalter DJ. 2009. A re-evaluation of random-effects meta-analysis. *JRSS (A)* **172**(1): 137–159.
Hoaglin DC, Hawkins N, Jansen JP, Scott DA, Itzler R, Cappelleri JC, Boersma C, Thompson D, Larholt KM, Diaz M, Barrett A. 2011. conducting indirect-treatment-comparison and network-meta-analysis studies: report of the

ISPOR Task Force on indirect treatment comparisons good research practices—part 2. *Value Health* **14**(4): 429–437.
Horton J, Vandermeer B, Hartling L, Tjosvold L, Klassen TP, Buscemi N. 2010. Systematic review data extraction: cross-sectional study showed that experience did not increase accuracy. *J Clin Epidemiol* **63**(3): 289–298.
Hunt M. 1997. *How Science Takes Stock*. Russel Sage Found, New York.
Ioannidis, JP. 2006. Indirect comparisons: the mesh and mess of clinical trials. *Lancet* **368**(9546): 1470–1472.
Jackson D, Riley R, White IR. 2011. Multivariate meta-analysis: potential and promise. *Stat Med*.
Jansen JP, Crawford B, Bergman G, Stam W. 2008. Bayesian meta-analysis of multiple treatment comparisons: an introduction to mixed treatment comparisons. *Value Health* **11**(5): 956–964.
Jansen JP. 2011. Network meta-analysis of survival data with fractional polynomials. *BMC Med Res Methodol* **11**: 61.
Jansen JP, Fleurence R, Devine B, Itzler R, Barrett A, Hawkins N, Lee K, Boersma C, Annemans L, Cappelleri JC. 2011. 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. *Value Health* **14**(4): 417–428.
Kyrgiou M, Salanti G, Pavlidis N, Paraskevaidis E, Ioannidis JP. 2006. Survival benefits with diverse chemotherapy regimens for ovarian cancer: meta-analysis of multiple treatments. *J Natl Cancer Inst* **98**(22): 1655–1663.
Lambert PC, Sutton AJ, Burton PR, Abrams KR, Jones DR. 2005. How vague is vague? A simulation study of the impact of the use of vague prior distributions in MCMC using WinBUGS. *Stat Med* **24**(15): 2401–2428 available from: PM:16015676
Liberati A, Altman DG, Tetzlaff J, Mulrow C, Gotzsche PC, Ioannidis JP, Clarke M, Devereaux PJ, Kleijnen J, Moher D. 2009. The PRISMA statement for reporting systematic reviews and meta-analyses of studies that evaluate healthcare interventions: explanation and elaboration. *BMJ* **339**: b2700.
Longworth L, Sculpher MJ, Bojke L, Tosh JC. 2011. Bridging the gap between methods research and the needs of policy makers: a review of the research priorities of the National Institute for Health and Clinical Excellence. *Int J Technol Assess Health Care* **27**(2): 180–187.
Lu G, Ades AE. 2006. Assessing evidence inconsistency in mixed treatment comparisons. *JASA* **101**(474): 447–459.
Lu G, Ades A. 2009. Modeling between-trial variance structure in mixed treatment comparisons. *Biostatistics* **10**(4): 792–805.
Lu G, Ades AE. 2004. Combination of direct and indirect evidence in mixed treatment comparisons. *Stat Med* **23**(20): 3105–3124.
Lu G, Ades AE, Sutton AJ, Cooper NJ, Briggs AH, Caldwell DM. 2007. Meta-analysis of mixed treatment comparisons at multiple follow-up times. *Stat Med* **26**(20): 3681–3699.
Lunn DJ, Thomas A, Best N, Spiegelhalter DJ. 2000. WinBUGS—a Bayesian modelling framework: concepts, structure, and extensibility. *Statistics and Computing* **10**: 325–339.
Manzoli L, Salanti G, De VC, Boccia A, Ioannidis JP, Villari P. 2009. Immunogenicity and adverse events of avian influenza A H5N1 vaccine in healthy adults: multiple-treatments meta-analysis. *Lancet Infect Dis* **9**(8): 482–492.
Mavridis D, Salanti G. 2012. A practical introduction to multivariate meta-analysis. *Stat Methods Med Res* [Epub ahead of print]
Mills EJ, Ghement I, O'Regan C, Thorlund K. 2011. Estimating the power of indirect comparisons: a simulation study. *PLoS One* **6**(1): e16237.
Mills EJ, Perri D, Cooper C, Nachega JB, Wu P, Tleyjeh I, Phillips P. 2009. Antifungal treatment for invasive Candida infections: a mixed treatment comparison meta-analysis. *Ann Clin Microbiol Antimicrob* **8**: 23 available from: PM:19558681
Mitka M. 2010. US government kicks off program for comparative effectiveness research. *JAMA* **304**(20): 2230–2231.
Moreno SG, Sutton AJ, Ades AE, Cooper NJ, Abrams KR. 2011. Adjusting for publication biases across similar interventions performed well when compared with gold standard data. *J Clin Epidemiol* **64**(11): 1230–1241.
NICE. Evidence synthesis. http://www.nicedsu.org.uk/Evidence-Synthesis-TSD-series%282391675%29.htm. 2008. National Institute of Health and Clinical Excellence. Guide to the Methods of Technology Appraisal.
Nixon RM, Bansback N, Brennan A. 2006. Using mixed treatment comparisons and meta-regression to perform indirect comparisons to estimate the efficacy of biologic treatments in rheumatoid arthritis. *Stat Med* **26**(6): 1237–1254.
O'Regan C, Ghement I, Eyawo O, Guyatt GH, Mills EJ. 2009. Incorporating multiple interventions in meta-analysis: an evaluation of the mixed treatment comparison with the adjusted indirect comparison. *Trials* **10**: 86.
Padwal R, Klarenbach S, Wiebe N, Birch D, Karmali S, Manns B, Hazel M, Sharma AM, Tonelli M. 2011. Bariatric surgery: a systematic review and network meta-analysis of randomized trials. *Obes Rev* **12**(8): 602–621.
Palepu A, Kendall C, Moher D. 2011. Open Medicine endorses PROSPERO. *Open. Med.* **5**(1): e65–e66.
Patsopoulos NA, Analatos AA, Ioannidis JP. 2005. Relative citation impact of various study designs in the health sciences. *JAMA* **293**(19): 2362–2366.
Piccini JP, Kong DF. 2011. Mixed treatment comparisons for atrial fibrillation: evidence network or bewildering entanglement? *Europace* **13**(3): 295–296.
Price MJ, Welton NJ, Ades AE. 2010. Parameterization of treatment effects for meta-analysis in multi-state Markov models. *Stat Med* **30**(2): 140–151.

Riley RD, Sutton AJ, Abrams KR, Lambert PC. 2004. Sensitivity analyses allowed more appropriate and reliable meta-analysis conclusions for multiple outcomes when missing data was present. *J Clin Epidemiol* **57**(9): 911–924.

Rizos EC, Salanti G, Kontoyiannis DP, Ioannidis JP. 2011. Homophily and co-occurrence patterns shape randomized trials agendas: illustration in antifungal agents. *J Clin Epidemiol* **64**(8): 830–842.

Rothstein HR, Sutton AJ, Borestein ME. 2005. *Publication Bias in Meta-analysis—Prevention, Assessment and Andjustment*. Wiley, Chichester.

Sackett D, Straus S, Richardson W. 2000. *Evidence-based Medicine: How to Practice and Teach EBM*, 2nd ed. Churchill Livingstone, Edinburgh.

Salanti G, Ades AE, Ioannidis JP. 2010a. Graphical methods and numerical summaries for presenting results from multiple-treatment meta-analysis: an overview and tutorial. *J Clin Epidemiol* **64**(2): 163–171.

Salanti G, Dias S, Welton NJ, Ades AE, Golfinopoulos V, Kyrgiou M, Mauri D, Ioannidis JP. 2010b. Evaluating novel agent effects in multiple-treatments meta-regression. *Stat Med* **29**(23): 2369–2383.

Salanti G, Higgins JP, Ades AE, Ioannidis JP. 2008a. Evaluation of networks of randomized trials. *SMMR* **17**(3): 279–301.

Salanti G, Kavvoura FK, Ioannidis JP. 2008b. Exploring the geometry of treatment networks. *Ann Intern Med* **148**(7): 544–553.

Salanti G, Marinho V, Higgins JP. 2009. A case study of multiple-treatments meta-analysis demonstrates that covariates should be considered. *J Clin Epidemiol* **62**(8): 857–864.

Sciarretta S, Palano F, Tocci G, Baldini R, Volpe M. 2011. Antihypertensive treatment and development of heart failure in hypertension: a Bayesian network meta-analysis of studies in patients with hypertension and high cardiovascular risk. *Arch Intern Med* **171**(5): 384–394.

Shea BJ, Grimshaw JM, Wells GA, Boers M, Andersson N, Hamel C, Porter AC, Tugwell P, Moher D, Bouter LM. 2007. Development of AMSTAR: a measurement tool to assess the methodological quality of systematic reviews. *BMC Med Res Methodol* **7**: 10.

Soares HP, Kumar A, Daniels S, Swann S, Cantor A, Hozo I, Clark M, Serdarevic F, Gwede C, Trotti A, Djulbegovic B. 2005. Evaluation of new treatments in radiation oncology: are they better than standard treatments? *JAMA* **293**(8): 970–978.

Song F, Altman DG, Glenny AM, Deeks JJ. 2003. Validity of indirect comparison for estimating efficacy of competing interventions: empirical evidence from published meta-analyses. *BMJ* **326**(7387): 472.

Song F, Harvey I, Lilford R. 2008. Adjusted indirect comparison may be less biased than direct comparison for evaluating new pharmaceutical interventions. *J Clin Epidemiol* **61**(5): 455–463.

Song F, Loke YK, Walsh T, Glenny AM, Eastwood AJ, Altman DG. 2009. Methodological problems in the use of indirect comparisons for evaluating healthcare interventions: survey of published systematic reviews. *BMJ* **338**: b1147.

Song F, Xiong T, Parekh-Bhurke S, Loke YK, Sutton AJ, Eastwood AJ, Holland R, Chen YF, Glenny AM, Deeks JJ, Altman DG. 2011. Inconsistency between direct and indirect comparisons of competing interventions: meta-epidemiological study. *BMJ* **343**: d4909.

Spiegelhalter DJ, Best NG, Bradley PC, van der Linde A. 2002. Bayesian measures of model complexity and fit. *JRSS (B)* **64**: 583–639.

StataCorp LP. STATA. College Station. 2011. Stata Statistical Software: Release 12.

Sterne JA, Juni P, Schulz KF, Altman DG, Bartlett C, Egger M. 2002. Statistical methods for assessing the influence of study characteristics on treatment effects in 'meta-epidemiological' research. *Stat Med* **21**(11): 1513–1524.

Sudlow CL, Counsell CE. 2003. Problems with UK government's risk sharing scheme for assessing drugs for multiple sclerosis. *BMJ* **326**(7385): 388–392 available from: PM:12586677.

Sutton A, Ades AE, Cooper N, Abrams K. 2008. Use of indirect and mixed treatment comparisons for technology assessment. *Pharmacoeconomics* **26**(9): 753–767.

Sutton AJ, Higgins JP. 2008. Recent developments in meta-analysis. *Stat Med* **27**(5): 625–650.

Thijs V, Lemmens R, Fieuws S. 2008. Network meta-analysis: simultaneous meta-analysis of common antiplatelet regimens after transient ischaemic attack or stroke. *Eur Heart J* **29**(9): 1086–1092.

Thompson SG, Higgins JP. 2002. How should meta-regression analyses be undertaken and interpreted? *Stat Med* **21**(11): 1559–1573.

Trelle S, Reichenbach S, Wandel S, Hildebrand P, Tschannen B, Villiger PM, Egger M, Juni P. 2011. Cardiovascular safety of non-steroidal anti-inflammatory drugs: network meta-analysis. *BMJ* **342**: c7086.

Van der Wees P, Qaseem A, Kaila M, Ollenschlaeger G, Rosenfeld R, Board of Trustees of the Guidelines International Network (g-i-n). 2012. Prospective systematic review registration: perspective from the Guidelines International Network (G-I-N). *Systematic Reviews* **1**(3) DOI: 10.1186/2046-4053-1-3

Welton NJ, Cooper NJ, Ades AE, Lu G, Sutton AJ. 2008. Mixed treatment comparison with multiple outcomes reported inconsistently across trials: Evaluation of antivirals for treatment of influenza A and B. *Stat Med* **29**(27): 5620–5639.

White I. 2011. Multivariate random-effects meta-analysis. *The Stata Journal* **9**(1): 40–56.

White IR, Barret JK, Jackson D, Higgins JPT. 2012. Consistency and inconsistency in multiple treatments meta-analysis: model estimation using multivariate meta-regression.

Woods BS, Hawkins N, Scott DA. 2010. Network meta-analysis on the log-hazard scale, combining count and hazard ratio statistics accounting for multi-arm trials: a tutorial. *BMC Med Res Methodol* **10**: 54.