Tutorial

Research Synthesis Methods

Received 21 November 2014, Revised 30 September 2015, Accepted 06 November 2015 Published online in Wiley Online Library

(wileyonlinelibrary.com) DOI: 10.1002/jrsm.1195

GetReal in network meta-analysis: a review of the methodology

Orestis Efthimiou,[a*] Thomas P. A. Debray,[b,c] Gert van Valkenhoef,[d] Sven Trelle,[e,f] Klea Panayidou,[e] Karel G. M. Moons,[b,c] Johannes B. Reitsma,[b,c] Aijing Shang[g] and Georgia Salanti[a,†] on behalf of GetReal Methods Review Group[‡]

Pairwise meta-analysis is an established statistical tool for synthesizing evidence from multiple trials, but it is informative only about the relative efficacy of two specific interventions. The usefulness of pairwise meta-analysis is thus limited in real-life medical practice, where many competing interventions may be available for a certain condition and studies informing some of the pairwise comparisons may be lacking. This commonly encountered scenario has led to the development of network meta-analysis (NMA). In the last decade, several applications, methodological developments, and empirical studies in NMA have been published, and the area is thriving as its relevance to public health is increasingly recognized. This article presents a review of the relevant literature on NMA methodology aiming to pinpoint the developments that have appeared in the field. Copyright © 2016 John Wiley & Sons, Ltd.

Keywords: mixed-treatment comparison; multiple-treatment meta-analysis; indirect treatment comparison; comparing multiple interventions

1. Introduction

When developing a new treatment for a particular disease, a series of clinical trials are conducted where its efficacy is compared with a placebo and/or with a competing treatment that is already implemented in regular care. As a consequence, when reviewing estimates of treatment efficacy, researchers often encounter clinical trials comparing different subsets of all the available treatments for the disease, forming a 'network' of evidence (Salanti *et al*., 2008a). In this situation, a standard meta-analysis restricted to head-to-head comparisons cannot give a definite answer as to which treatment works best for a specific condition. Network meta-analysis (NMA) can then be used to synthesize the totality of the available information and provide clinically relevant estimates to better support decision making (Caldwell *et al*., 2005; Coleman *et al*., 2012; Higgins and Whitehead, 1996; Lu and Ades, 2004; Lumley, 2002).

In the past few years, NMA has become increasingly popular (Coleman *et al*., 2012; D'Ascenzo and Biondi-Zoccai, 2014; Donegan *et al*., 2010; Glenny *et al*., 2005; Nikolakopoulou *et al*., 2014; Sobieraj *et al*., 2013), and several published articles have discussed its advantages and limitations (Cipriani *et al*., 2012; Cooper *et al*., 2011; Cucherat and Izard,

[a]*Department of Hygiene and Epidemiology, University of Ioannina School of Medicine, Ioannina, Greece*
[b]*Julius Center for Health Sciences and Primary Care, University Medical Center Utrecht, Utrecht, The Netherlands*
[c]*The Dutch Cochrane Centre, Julius Center for Health Sciences and Primary Care, University Medical Center Utrecht, Utrecht, The Netherlands*
[d]*Department of Epidemiology, University of Groningen, University Medical Center Groningen, Groningen, The Netherlands*
[e]*Institute of Social and Preventive Medicine, University of Bern, Bern, Switzerland*
[f]*CTU Bern, Department of Clinical Research, University of Bern, Bern, Switzerland*
[g]*F. Hoffmann-La Roche AG, Basel, Switzerland*
**Correspondence to: Orestis Efthimiou, Department of Hygiene and Epidemiology, University of Ioannina School of Medicine, Ioannina, Greece. E-mail: oremiou@gmail.com*
[†]*Present address: Berner Institut für Hausarztmedizin (BIHAM), University of Bern, Bern, Switzerland.*
[‡]*A list of the members of the GetReal methods' review group is given in Appendix A.*

2009; EUnetHTA, 2013; Hoaglin *et al*., 2011; Ioannidis, 2009; Jansen *et al*., 2008, 2011a; Malone, 2007; Mills *et al*., 2011a, 2013b; Naci and Fleurence, 2011; Salanti, 2012; Sutton *et al*., 2008, 2009; Wells *et al*., 2009; Welton *et al*., 2012), which have also been explored in empirical assessments (Jonas *et al*., 2013; Madan *et al*., 2011; O'Regan *et al*., 2009; Song *et al*., 2003, 2008) and simulation studies (Glenny *et al*., 2005; Jonas *et al*., 2013; Song *et al*., 2012). In particular, NMA allows an increase in the precision of estimates of relative effects (as compared with estimates based on direct evidence alone), and it can be used to compare interventions that have never been compared directly, head-to-head. This is especially valuable when active agents are only compared with placebo or standard care for regulatory purposes but not to each other (Falissard *et al*., 2009). In addition, NMA can support (policy-related) decision making by explicitly providing a ranking of all competing interventions (Salanti *et al*., 2011) and to reduce the uncertainty in input parameters for subsequent formal cost-effectiveness models (Thorlund *et al*., 2014).

Despite the aforementioned advantages, the implementation of NMA may be hindered because of several reasons. First, the methodology of NMA rests on the assumption that different sources of evidence (direct and indirect evidence for the same treatment comparison) are in agreement. This assumption is often viewed as an important limitation of the method because it may be difficult to justify in practice. Moreover, the field of NMA is swiftly evolving; during the last few years, there has been an abundance of published articles presenting alternative approaches to deal with issues related to NMA. Hence, an updated review of the methodology is deemed timely to ensure that interested researchers use state-of-the-art methods for practical applications and also when conducting further methodological research. The present article aims to give a comprehensive account of the currently available methods for NMA and discuss in depth conceptual and statistical ways of evaluating the underlying assumptions of the model while providing guidance for researchers that set out to perform an NMA.

This review is part of a work undertaken within the auspices of *GetReal: Incorporating real-life data into drug development*, a project funded by the Innovative Medicines Initiative – a European public–private initiative aiming to speed up the development of better and safer medicines for patients. The aim of the GetReal project is to explore how drug development can become more efficient by better incorporating evidence of relative effectiveness in the process, and to propose ways for enriching the decision making by regulatory authorities and Health Technology Assessment organizations. This article is part of a series of three literature reviews (Debray *et al*., 2015; Panayidou *et al*., 2015).

2. Literature review methods and results

2.1. Search strategy

For the purposes of this review, we searched for published articles that presented new methodology for NMA or articles that evaluated existing methods. We based our search on a previous review of the literature in NMA performed by the *Comparing Multiple Interventions Methods Group* of the Cochrane Collaboration, which resulted in a publicly available database (www.zotero.org/groups/network_meta-analysis_methods/items/). This database has been shared with experts in the field to identify missing relevant articles. We also used the results from a recent literature review performed by Donegan *et al*. (2013b). In addition, we searched the MEDLINE database and also hand-searched key journals. Details on the search strategy we employed can be found in Appendix B.

2.2. Inclusion criteria

We included articles that contributed to the methodology of NMA by introducing new methods and models, articles that reviewed the existing methodology, and articles that provided recommendations or gave guidance on how to perform an NMA. We also included papers that discussed the conceptual issues and the assumptions behind NMA and articles that provided some sort of empirical assessment for the conduct of NMAs in general.

2.3. Exclusion criteria

We excluded publications for which one of the following criteria was met:

- full text of the publication was not available
- published in a language other than English
- conference posters
- applications of NMA without a methodological focus

2.4. Results

A total of 179 articles were included in our database. The identified articles were organized according to their context and are discussed in the relevant sections of this review. The included papers were classified using a number of tags. Each article was assigned one or more tags according to the type of research presented, one or more tags according to the methodological topics addressed, and one or more tags according to the software

it used to implement the methods it presented. An online database of all included articles tagged by topic can be found at https://www.zotero.org/groups/wp4_-_network_meta-analysis/items. More details on the results and the categorization we used are discussed in Appendix B.

In the next sections, we present a review of the methods for NMA as identified in our literature search. In Section 3, we present key concepts and the basic methodology for NMA. We present the underlying assumptions of the model in Section 3.1, and we discuss popular approaches for fitting NMA in Section 3.2. We present methods for detecting inconsistency in the network in Section 4. In Section 5, we give a list of the software options available for fitting NMA models and evaluating inconsistency. In Section 6, we present special issues of the NMA methodology. We discuss the use of different effect sizes (Section 6.1) and how to model time-to-event data (Section 6.2). In Section 6.3, we give methods to allow the inclusion of covariates in the analysis. In Section 6.4, we explore ways for adjusting for possible sources of bias, and in Section 6.5, we discuss the reporting of the results of an NMA. We review available methods for analyzing multiple outcomes and repeated measures in Section 6.6, and we discuss the definition of nodes in the network of treatments in Section 6.7. In Section 6.8, we review the use of individual participant data (IPD) in an NMA, and in Section 6.9, we discuss the use of data from non-randomized studies in the analysis. Finally, in Section 6.10, we discuss how NMA can inform the planning of future studies.

3. Basic methodology in network meta-analysis

3.1. Conceptual issues and underlying assumptions

The key feature of NMA is that it allows the synthesis of direct and indirect estimates for the relative effects of many competing treatments for the same health condition. Two treatments A and B may have been *directly* compared in head-to-head studies (in 'AB' studies). An *indirect* estimate may also be obtained from studies comparing these two treatments with a common comparator treatment C, that is, AC and BC studies (Bucher *et al*., 1997). If both direct and indirect estimates are available, they can be combined in order to estimate a *mixed*-treatment effect, as shown in the left panel of Figure 1. In practice, for most health conditions, there is a plethora of interventions being compared across various randomized trials, forming a network of evidence. For a given treatment comparison within such a network, there may be direct and many different indirect estimates, obtained via many different comparators as shown in the example in the right panel of Figure 1. Using NMA, one can synthesize all these different pieces of information to produce an internally consistent overall estimate of all treatments' relative effects.

Despite the benefits of NMA discussed in Section 1, there is still controversy among researchers about the validity of using indirect treatment comparisons (*indirect evidence*) for decision making. The use of such evidence is particularly challenged when direct treatment comparisons (*direct evidence*) are also available (Edwards *et al*., 2009; Gartlehner and Moore, 2008; Ioannidis, 2006). One focal point of criticism is the nature of evidence NMA provides. Even though patients within a randomized clinical trial (RCT) are randomized to each of the treatments being compared, the treatments are not randomized across the included trials. Therefore, indirect comparisons are non-randomized comparisons and in fact provide observational, rather than randomized, evidence. As a consequence, indirect treatment comparisons may be more subject to biased treatment effect estimates, owing, for example, to confounding (e.g., when randomized AB and AC studies are systematically different than BC; Caldwell *et al*., 2005) and selection bias (e.g., when the choice of comparator in a study is dependent on the relative treatment effect; Salanti *et al*., 2008b). Such considerations are also closely related to the underlying assumptions of NMA; in what follows, we discuss these assumptions in detail.

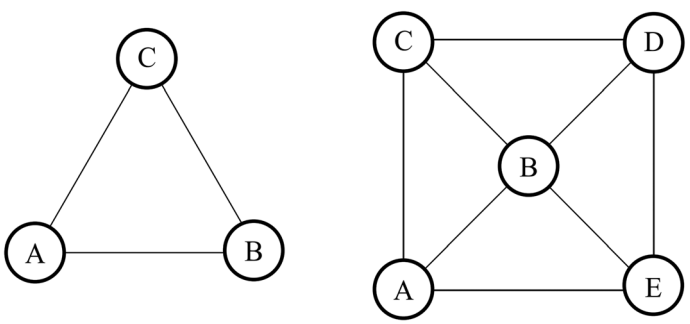

Figure 1. Each circle represents an intervention, and lines represent direct comparisons. Left panel: Three interventions A, B, and C form a simple network. The indirect AB comparison is estimated via C, that is, using the direct AC and BC comparisons. The mixed relative treatment effect for AB is estimated by combining the direct comparison and the indirect comparison. Right panel: A more complex network with five interventions and eight direct comparisons. Overall, one direct comparison and four indirect comparisons contribute evidence to A versus B (indirect comparisons are via C, via E, via C and D, and via E and D).

3.1.1. Transitivity. The objective of an NMA is to enhance the decision-making process regarding the choice among alternative treatments for a certain disorder and a target population. Thus, the estimands one aims to estimate in an NMA are the average relative treatment effect sizes between the competing treatments, as they are expected to exist in the targeted population. The estimates produced by an NMA model for these parameters will be unbiased and consistent provided that the studies included in the dataset yield unbiased estimates and form a representative sample of the population of interest, and when the assumptions employed by the NMA model are valid. NMA adopts the same set of assumptions as a pairwise meta-analysis (Dias *et al*., 2010b), but it also employs an additional assumption that can be hard to assess (Song *et al*., 2009) termed transitivity (Baker and Kramer, 2002) (also called similarity (Donegan *et al*., 2010; Song *et al*., 2003) or exchangeability (Dias *et al*., 2013d)). Transitivity implies that information for the comparison between treatments A and B can be obtained via another treatment C, using the comparisons A versus C and B versus C. This assumption cannot be tested statistically, but its validity can be evaluated in a conceptual and epidemiological manner (Salanti, 2012).

The transitivity assumption implies that we can combine the direct evidence from AC and BC studies to learn (indirectly) about the comparison AB. This, however, will be questionable if there are important differences in the distribution of the effect modifiers (variables or characteristics that modify the observed relative effects, e.g., mean age of the participants and treatment dosage) across the AC and BC trials, which inform the indirect comparison (Baker and Kramer, 2002; Jansen and Naci, 2013). An effect modifier might differ across studies of the same comparison (e.g., mean participant age might be different across the AC trials), but if it has a similar distribution across comparisons (AC and BC), the transitivity assumption may still be valid (Salanti, 2012). Consequently, the plausibility of the transitivity assumption can be evaluated by putting the collection of studies under scrutiny for important differences in the distribution of effect modifiers. If the studies are deemed to be similar, then the transitivity assumption might be realistic, provided that there are no unknown modifiers of the relative treatment effect (Donegan *et al*., 2013b). Obviously, such an evaluation of transitivity may be impossible when the effect modifiers are not reported or when there are few studies per treatment comparison (Cipriani *et al*., 2013). If important differences are identified and there are enough data available, a network meta-regression can be used to improve the transitivity of the network (see also Section 6.3).

This implies, for example, that the common comparator treatment C must be similar in the AC and BC studies in terms of dose, modes of administration, duration, and so on. In an NMA of studies comparing fluoride treatments for the prevention of dental carries, the definition of placebo was different between studies of fluoride toothpaste and studies of fluoride rinse (Salanti *et al*., 2009), casting doubt about the plausibility of the transitivity assumption and thereby challenging the reliability of NMA results. In another example, Julious and Wang (2008) discussed how the use of placebo as an intermediate comparator might bias the results of indirect comparisons because of changes in the placebo response of the population over the years; for example, when studies comparing treatment A to placebo are older than studies comparing B with placebo the indirect estimate for A versus B via placebo may be biased.

Other ways of formulating the transitivity assumption is to assume that regardless of the treatments being compared in each study the true relative effect of A versus B is the same in a fixed-effects model or exchangeable across studies in a random-effects model (Dias *et al*., 2013d; Lu and Ades, 2009), that the 'missing' treatments in each trial are missing at random (Lu and Ades, 2006), or, equivalently, that the choice of treatment comparisons in the trials is not associated either directly or indirectly with the relative effectiveness of the interventions (Salanti, 2012). Finally, an alternative way of postulating this assumption is to state that the included patients could in principle be randomized to any of the treatments included in the network (Salanti, 2012).

3.1.2. Consistency. The statistical manifestation of transitivity is called consistency (Cipriani *et al*., 2013). Checking the network for consistency constitutes an additional method of inferring indirectly about the plausibility of the transitivity assumption. Consistency refers to the statistical agreement between the observed direct and (possibly many) indirect sources of evidence. A simple network may only include three treatments A, B, and C. The transitivity assumption then implies that $\mu_{AB} = \mu_{AC} - \mu_{BC}$ (also termed consistency equation), where μ_{AB} denotes the true relative effect of treatment B over C, likewise for μ_{AC} and μ_{BC}. When this equation does not hold for the (direct) estimates, the network is said to be inconsistent (Lu and Ades, 2006) or incoherent (Lumley, 2002). If this is the case, results from an NMA will be more difficult to interpret and become less reliable. In a following section, we review various statistical methods and models that have been suggested for identifying inconsistency and thus assessing the transitivity assumption in NMA.

Statistical inconsistency can be thought of as another form of heterogeneity: heterogeneity results from the variation of effect modifiers within a treatment comparison, while inconsistency results from the variation of effect modifiers across treatment comparisons (Jansen and Naci, 2013). Researchers should keep in mind, though, that the consistency of a network can only be assessed statistically when there is both direct and indirect evidence for one or more treatment comparisons. This situation only occurs when there are *closed loops* in the network (i.e., when three or more interventions are connected by a polygon, the edges of which represent head-to-head comparisons between the corresponding treatments). When there are no closed loops present in the network, a statistical assessment of inconsistency will not be possible. In these situations, there cannot be inconsistency

by definition. This, however, does not imply that the transitivity assumption will necessarily hold. It should also be noted that the absence of statistical inconsistency does not provide proof for the validity of the transitivity assumption, which, as discussed in the previous section, is essentially an untestable assumption. Thus, next to statistical tests for inconsistency, a conceptual/theoretical assessment of the transitivity assumption should always take place before an NMA is conducted (Cipriani *et al*., 2013), and the studies included in an NMA should always be scanned for important differences in terms of patients, interventions, outcomes, study design, methodological characteristics, and reporting biases (Bucher *et al*., 1997; Caldwell *et al*., 2005; Coleman *et al*., 2012; Coory and Jordan, 2010; Cucherat and Izard, 2009; Dewilde and Hawkins, 2012; Donegan *et al*., 2013b; Glenny *et al*., 2005; Griffin *et al*., 2006; Hoaglin *et al*., 2011; Jansen *et al*., 2008; Madan *et al*., 2011; van Valkenhoef *et al*., 2012c; Wells *et al*., 2009; Xiong *et al*., 2013).

3.2. Statistical methods for network meta-analysis

A simple network may include three treatments of interest, A, B, and C. An estimate of the indirect treatment effect of A versus B can then be obtained by utilizing the direct observations A versus C and B versus C as $\widehat{\mu}_{AB}^{Ind} = \widehat{\mu}_{AC}^{Dir} - \widehat{\mu}_{BC}^{Dir}$ (Bucher *et al*., 1997). This result is sometimes also referred to as the 'adjusted indirect comparison'. The variance of the indirect estimate is the sum of the variances of the two direct ones. When direct evidence is also available for the A-versus-B comparison, it can be combined with the indirect estimate using the inverse variance method to produce a mixed estimate. Note that this approach for obtaining indirect estimates is only valid for 'triangular networks', where three treatments have been compared in a number of two-arm trials (and it can be readily extended for 'polygonal networks' with no diagonals); it can also be used for 'star-shaped' networks, where all treatments are compared with a common comparator (e.g., placebo) but not with each other. For complex networks, like the one depicted in the right panel of Figure 1, there will be multiple sources of indirect information, and thus, more advanced models need to be used.

Popular implementations of NMA models adopt meta-regression (Section 3.2.1), hierarchical modeling (Section 3.2.2), or a multivariate meta-analysis approach (Section 3.2.3). A common feature of all of these approaches is that the use of the consistency equations minimizes the number of parameters that need to be estimated. The minimum set of parameters needed to model the relative treatment effects is usually termed as the set of 'basic parameters' or 'basic contrasts'; these parameters are (assuming that the network is connected) in number equal to the number of treatments minus one and can be used to generate estimates for all possible treatment comparisons via the consistency equations (i.e., $\mu_{XY} = \mu_{XZ} - \mu_{YZ}$ for treatments *X*, *Y*, and *Z*). The basic parameters can be chosen arbitrarily as long as they form a 'spanning tree' of the evidence; that is, they form a connected subgraph of the network containing all vertices but no cycles (Lu and Ades, 2006); if this condition is satisfied, the actual choice of basic parameters does not affect the NMA results. These parameters are commonly taken to be the relative effects of each treatment versus a reference (e.g., the placebo, if present in the network). For example, for a network of four treatments A, B, C, and D, three basic parameters are needed. These can be chosen to correspond to the relative treatment effects of all other treatments versus A, that is, AB, AC, and AD. All other treatment effects can be generated from these three parameters; for example, the relative treatment effect for BD can be estimated using the AB and AD parameters. Choosing instead BA, BC, and BD as the basic parameters would have no impact on the NMA results.

In what follows, we describe some of the most common approaches for fitting an NMA model. These approaches should be considered equivalent; the choice between them should be primarily dictated by the availability of software packages for implementing them and by the technical expertise of the researchers. Different methods may give slightly different estimates; for example, fitting an NMA in a Bayesian framework as a hierarchical model may lead to slightly different results than fitting it in a frequentist setting using multivariate meta-analysis techniques. These differences, however, should not be seen as a limitation of NMA models in general; they are due to the different approaches employed for making statistical inferences, in a similar way as it applies to a pairwise meta-analysis. We discuss the currently available software options for fitting all the models and implementing all the methods in Section 5.

3.2.1. Network meta-analysis as a meta-regression. In the meta-regression approach, first proposed by Lumley (2002), the various treatment comparisons are treated as covariates in a meta-regression model (Salanti *et al*., 2008a). The usual NMA meta-regression model can be summarized in the following equation: $\boldsymbol{y} = \boldsymbol{X\mu} + \boldsymbol{\varepsilon} + \boldsymbol{\delta}$, with $\boldsymbol{y}$ being the vector of observed relative treatment effects, $\boldsymbol{\mu}$ the vector of basic parameters, $\boldsymbol{\varepsilon}$ the vector of random errors, and $\boldsymbol{\delta}$ the vector of random effects. Note that for a study *i* comparing T_i different treatments, only $T_i - 1$ observations on treatment comparisons need to enter the model. For a parallel randomized three-arm ABC trial, for example, we only need to include two of the three comparisons, for example, AB and AC; the BC comparison is just a linear function of the other two. This means that $\boldsymbol{y}$, $\boldsymbol{\varepsilon}$, and $\boldsymbol{\delta}$ have a length equal to $\sum(T_i - 1)$. Random errors are assumed to follow a multivariate normal distribution, $\boldsymbol{\varepsilon} \sim N(0, \boldsymbol{S})$, with $\boldsymbol{S}$ being the (block-diagonal), within-study variance covariance matrix. A study *i* with T_i treatment arms will contribute a $(T_i - 1) \times (T_i - 1)$ matrix to $\boldsymbol{S}$; a two-arm AB study, for example, will only contribute to $\boldsymbol{S}$ the variance of the relative treatment effect of A versus B. A three-arm trial ABC will contribute to $\boldsymbol{S}$ a 2×2 matrix with the variances and the covariance of the two relative treatment effects chosen to be included in $\boldsymbol{y}$, for example, AB and AC. Similarly, $\boldsymbol{\delta} \sim N(0, \Delta)$ for the random effects,

with Δ being the heterogeneity variance–covariance matrix. Matrix $\boldsymbol{X}$, the *design matrix*, has as elements 1, –1, and 0 and describes the structure of the network, providing information on which comparison is being performed in each study (Salanti *et al*., 2008a). If for example the network is built by an AB study (study 1), an AC study (study 2), and a BC study (study 3), the model would be written as

$$\begin{pmatrix} y_{1\mathrm{AB}} \\ y_{2\mathrm{AC}} \\ y_{3\mathrm{BC}} \end{pmatrix} = \begin{pmatrix} 1 & 0 \\ 0 & 1 \\ -1 & 1 \end{pmatrix} \begin{pmatrix} \mu_{AB} \\ \mu_{\mathrm{AC}} \end{pmatrix} + \begin{pmatrix} \varepsilon_{1AB} \\ \varepsilon_{2\mathrm{AC}} \\ \varepsilon_{3\mathrm{BC}} \end{pmatrix} + \begin{pmatrix} \delta_{1\mathrm{AB}} \\ \delta_{2\mathrm{AC}} \\ \delta_{3\mathrm{BC}} \end{pmatrix}.$$

The basic parameters can be estimated as $\widehat{\boldsymbol{\mu}} = (\boldsymbol{X}^{\mathrm{T}}\boldsymbol{W}\boldsymbol{X})^{-1}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{W}\boldsymbol{y}$, with variance $\mathrm{var}(\widehat{\boldsymbol{\mu}}) = (\boldsymbol{X}^{\mathrm{T}}\boldsymbol{W}\boldsymbol{X})^{-1}$, where $\boldsymbol{W}$ is the weight matrix, $\boldsymbol{W} = (\boldsymbol{S} + \Delta)^{-1}$. The within-study variance–covariance matrix $\boldsymbol{S}$ can be estimated from the observed data (Franchini *et al*., 2012; Higgins and Whitehead, 1996), while for the between-study variance–covariance matrix Δ, one can use various ways of estimation including likelihood methods or the methods of moments (Berkey *et al*., 1998; Jackson *et al*., 2010; van Houwelingen *et al*., 2002). Estimating Δ may be difficult especially when the data are sparse or in the presence of multi-arm studies. For this reason, it is common to introduce additional assumptions to reduce the number of parameters in Δ and simplify the estimation. The most common approach is to assume equal heterogeneity variances across comparisons; that is, the between-study heterogeneity of the relative treatment effects is the same for all treatment comparisons (Higgins and Whitehead, 1996; Lumley, 2002). This assumption is, however, quite strong and may often be unrealistic. Lu and Ades (2009) discussed how the consistency equations impose restrictions on the heterogeneity of each comparison, based on the (different) heterogeneity variances of each of the basic parameters. Thorlund *et al*. (2013) presented models for exchangeable heterogeneity variances and also discussed the use of informative prior distributions in the context of a Bayesian analysis.

In a different approach, Lu *et al*. (2011) proposed a two-stage method for performing an NMA as a meta-regression. At the first stage, a meta-analysis is performed in each group of trials comparing the same treatments, for example, all two-arm trials that compare A versus B and all three-arm trials that compare A versus B versus C. This provides the direct estimates on treatment comparisons. At the second stage of the meta-analysis, a weighted linear regression is performed with the direct estimates as dependent variables. This two-stage method can be used to investigate how the first-stage (direct) evidence influences the network estimates and may therefore help to assess the consistency of the network (Section 4).

3.2.2. Network meta-analysis as a hierarchical model. Hierarchical NMA models (Lu and Ades, 2004; Salanti *et al*., 2008a) appear to be implemented most often (Coleman *et al*., 2012; Nikolakopoulou *et al*., 2014). An important advantage of this approach is that if arm-level data are available, their exact likelihood can be used (Dias *et al*., 2013a).

The likelihood of the arm-level data is defined in terms of a set of unknown parameters γ and a link function, $g(\gamma)$, which is used to map these parameters in the $(-\infty, \infty)$ range. For a study i comparing treatments A and B, we set

$$g(\gamma_{i\mathrm{A}}) = u_i,$$
$$g(\gamma_{i\mathrm{B}}) = u_i + \theta_{i\mathrm{AB}}.$$

For the case of binary data, for example, we can choose g to be the logit function and γ the probability of observing an event. We set

$$\mathrm{logit}(p_{i\mathrm{A}}) = u_i,$$
$$\mathrm{logit}(p_{i\mathrm{B}}) = u_i + \theta_{i\mathrm{AB}}.$$

Here, u represents the log-odds of the outcome for treatment A and $\theta_{i\mathrm{AB}}$ the log-odds ratio (OR) of A versus B; the event probabilities for each arm parameterize the binomial likelihood, $r_{iT} \sim \mathrm{Bin}(p_{iT}, n_{iT})$, with r_{iT} denoting the events and n_{iT} the total number of randomized patients in each treatment arm (T = A, B). We then allow $\theta_{i\mathrm{AB}} \sim \mathrm{N}(\mu_{\mathrm{AB}}, \tau^2_{\mathrm{AB}})$ for a random-effects meta-analysis. If two non-reference treatments are compared in a study, for example, treatments B and C, we utilize the consistency equations by setting $\theta_{i\mathrm{BC}} \sim \mathrm{N}(\mu_{\mathrm{AC}} - \mu_{\mathrm{AB}}, \tau^2_{\mathrm{BC}})$. In the presence of multi-arm studies, multivariate normal distributions should be used instead, where the within-study and between-study variances are replaced by the corresponding variance–covariance matrices $\boldsymbol{S}_i$ and Δ_i. Details on how to model other types of data can be found in Dias *et al*. (2013a). Note that the issues discussed in the previous section regarding the estimation of the between-trial heterogeneity hold for the hierarchical models as well.

Network meta-analysis can be fitted as a hierarchical model also if only contrast-level data are available from the studies (i.e., when the reported data are on the relative treatment effects of the treatments being compared, but not on the specific arms). For a two-arm study i comparing A (reference treatment) and B, the model is written as $y_{i\mathrm{AB}} \sim \mathrm{N}(\theta_{i\mathrm{AB}}, s_i^2)$. Note here that the normality assumption can be justified even if the underlying patient-level distributions are skewed, owing to the central limit theorem (Dias *et al*., 2013a).

Hierarchical models can also be fitted when a combination of arm-level and contrast-level data are available, using the exact likelihood for the arm-level data and the normal approximation for the contrast-level data in a so-called shared parameter model (Dias *et al*., 2013a).

3.2.3. Network meta-analysis as a multivariate meta-analysis model. White *et al*. (2012) suggested a method for performing NMA as a multivariate meta-analysis by treating the basic comparisons as different outcomes and by employing standard multiple-outcome meta-analytical techniques (Mavridis and Salanti, 2013). For this model to work, all studies need to report on the reference treatment; if this is not the case for some studies, a data-augmentation technique is required to impute a minimally informative reference treatment arm. The model is written as $\boldsymbol{y}=\boldsymbol{X}^{*}\boldsymbol{\mu}+\varepsilon+\boldsymbol{\delta}$, with $\boldsymbol{X}^{*}$ being a matrix with all elements being either 0 or 1, depending on which 'outcomes' are reported in each study.

Assume for example that treatments A, B, and C are compared in a number of studies, and also assume that treatment A is taken to be the reference treatment. In this approach, $\boldsymbol{\mu}$ will be a 2×1 vector of the basic parameters, AB and AC. A study comparing A versus B will contribute an element 1 in the first column of $\boldsymbol{X}^{*}$ and 0 in the second, as in this study, only the first 'outcome' is reported. An A-versus-C study will report the second outcome only; thus, the relevant elements in $\boldsymbol{X}^{*}$ will be 0 and 1. For a B-versus-C study, however, an A arm must be imputed; this study becomes a three-arm one and reports on both 'outcomes'. The model for these three studies is as follows:

$$\begin{pmatrix} y_{1\mathrm{AB}} \\ y_{2\mathrm{AC}} \\ y_{3\mathrm{AB}} \\ y_{3\mathrm{AC}} \end{pmatrix} = \begin{pmatrix} 1 & 0 \\ 0 & 1 \\ 1 & 0 \\ 0 & 1 \end{pmatrix} \begin{pmatrix} \mu_{AB} \\ \mu_{\mathrm{AC}} \end{pmatrix} + \begin{pmatrix} \varepsilon_{1\mathrm{AB}} \\ \varepsilon_{2\mathrm{AC}} \\ \varepsilon_{3\mathrm{AB}} \\ \varepsilon_{3\mathrm{AC}} \end{pmatrix} + \begin{pmatrix} \delta_{1\mathrm{AB}} \\ \delta_{2\mathrm{AC}} \\ \delta_{3\mathrm{AB}} \\ \delta_{3\mathrm{AC}} \end{pmatrix}.$$

Note that the two random errors and also the two random effects that were included for the third study will be correlated. Also note that, in this approach, the vector of observations, $\boldsymbol{y}$, has been modified to account for the imputed arms. Standard methods for multiple-outcome meta-analysis can now be used to fit the model.

Alternative models for performing an NMA have also been recently proposed in the literature. Rücker (2012) described the analogy between NMA and electrical networks and applied graph-theoretical methods to perform a fixed-effects NMA. This method was extended to also accommodate for random effects and was shown to be equivalent to the meta-regression method described in Section 3.2.1 (Rücker and Schwarzer, 2014). Also, Yang *et al*. (2014) introduced a confidence distribution approach for performing an NMA. In this approach, instead of combining point estimates from each study, the authors combine confidence distributions.

For more in-depth reviews of the methodology that include the statistical details of the models we presented, we refer the reader to Cucherat and Izard (2009), Greco *et al*. (2013), and Ohlssen *et al*. (2013).

4. Statistical methods for evaluating inconsistency in network meta-analysis

As we have previously discussed (Section 3.1), transitivity is a central assumption of NMA. A statistical assessment of the plausibility of this assumption can be made by checking whether the various sources of evidence fit together in a coherent way. This assessment is vital for ensuring that the NMA results are valid and interpretable for clinical decision making (Donegan *et al*., 2012) but may be difficult to do in practice, especially in the case of complex networks or when multi-arm studies are included in the network.

Statistical consistency can be assessed only in closed loops of evidence, and there are two general approaches to do this: either locally (by focusing on the inconsistency of a specific treatment comparison) or globally (by checking for inconsistency in the entire network of evidence). In what follows, we discuss methods and models that have been proposed for both of these approaches.

4.1. Local methods to detect inconsistency

4.1.1. Loop-specific approach. A straightforward approach for evaluating the presence of inconsistency in a network is to apply a loop-specific approach (Bucher *et al*., 1997); in this approach, we examine each loop of the network in isolation from the rest of the network. For an ABC loop in the network, for example, we choose one of the comparisons (e.g., B versus C) and compute the direct ($\widehat{\mu}_{\mathrm{BC}}^{\mathrm{Dir}}$) and indirect estimates ($\widehat{\mu}_{\mathrm{BC}}^{\mathrm{Ind}}$). Their difference measures inconsistency and is usually termed inconsistency factor (Lu and Ades, 2006): $\widehat{w}_{\mathrm{ABC}}=\widehat{\mu}_{\mathrm{BC}}^{\mathrm{Dir}}-\widehat{\mu}_{\mathrm{BC}}^{\mathrm{Ind}}$, with variance $\mathrm{var}(\widehat{w}_{\mathrm{ABC}})=\mathrm{var}(\widehat{\mu}_{\mathrm{BC}}^{\mathrm{Ind}})+\mathrm{var}(\widehat{\mu}_{\mathrm{BC}}^{\mathrm{Dir}})$. A 95% confidence interval can be obtained as $\widehat{w}_{\mathrm{ABC}}\pm1.96\sqrt{\mathrm{var}(\widehat{w}_{\mathrm{ABC}})}$, and a *Z*-statistic for the null hypothesis of consistency, that is, $\hat{w}_{\mathrm{ABC}}=0$, can be constructed as $z_{\mathrm{ABC}}=\widehat{w}_{\mathrm{ABC}}/\sqrt{\mathrm{var}(\widehat{w}_{\mathrm{ABC}})}$; this can be compared with the standard normal distribution to obtain a *p*-value (Bucher *et al*., 1997). Inconsistency is a property of the loop, in the sense that choosing a different treatment

comparison of the loop and repeating the computations would give the exact same results (Dias *et al*., 2013d; Lu and Ades, 2006); thus, we denote the inconsistency factor with an ABC subscript.

The loop-specific approach can be applied for each loop in a network to point out hot spots for inconsistency. The major advantage of this approach is that it is easy to implement; it suffers however from important limitations: when a treatment comparison is part of more than one loop, this method does not compare direct evidence for this comparison with all available indirect information, but with evidence from only one loop at a time, also in this case the tests for different loops sharing this comparison will not be independent. In addition, for networks with many loops, there are multiple testing issues.

4.1.2. Composite test. It is possible to extend the loop-specific approach by accounting for more than one indirect estimates for a treatment comparison using a 'composite test' for inconsistency (Caldwell *et al*., 2010; Dias *et al*., 2013d). Suppose that there are L loops that provide independent indirect information for the A-versus-B comparison; these can be combined with the direct information using the usual inverse variance method to obtain a pooled, overall estimate of the relative treatment effect of A versus B. Under the null hypothesis that the $L+1$ different estimates are in agreement, a test statistic following a chi-squared distribution with L degrees of freedom can be constructed to check for inconsistency. One should keep in mind, however, that the presence of multi-arm studies induces correlations among the estimates for the treatment effects, which this method, as well as the loop-specific approach, fails to account for.

4.1.3. Back-calculation. Dias *et al*. (2010b) also proposed two additional methods for locally checking inconsistency. The first method ('back-calculation') can be applied when the only available data are the pooled summaries of the pairwise meta-analyses. In the first step, the data are used to obtain a network estimate for each pairwise comparison in the network. It is then assumed that this estimate is a weighted average of the direct and indirect evidence, coming from the rest of the network. This allows a back-calculation of the indirect estimate and its variance, which in turn can be used to construct a Z-test for the difference of direct and indirect evidence. Note that this method is problematic for a random-effects meta-analysis, as the posterior distribution of the heterogeneity variance will in general be different between the NMA model and the model for the pairwise meta-analysis.

4.1.4. Node splitting. The second method proposed by Dias *et al*. (2010b), the node-splitting approach, can be used when trial-level data are available. In this method, the direct evidence for a specific treatment comparison is excluded from the rest of the network and is used to obtain a direct estimate. The remaining information in the network is used to obtain an indirect estimate for this comparison, after fitting an NMA model. The two estimates, direct and indirect, are then used to evaluate inconsistency with a Z-test. The main drawbacks of this approach (as well as the back-calculation approach) are that they might be computationally intensive, especially for large networks with many treatment comparisons, and that they cannot properly handle multi-arm studies.

4.2. Global methods to detect inconsistency

4.2.1. The Lu and Ades model. Lu and Ades (2006) introduced a model in which the consistency equations are 'bent' by including extra terms, the inconsistency factors. For an ABC loop, for example, the consistency equation is written as $\mu_{BC} = \mu_{AC} - \mu_{AB} + w_{ABC}$, where the w parameter measures the discrepancy of direct and indirect evidence. For networks comprising many loops, a different inconsistency factor needs to be included in each loop. When the network only includes two-arm studies, the number of independent inconsistency factors (the 'inconsistency degrees of freedom', *ICDF*) is $ICDF = C - T - 1$, with C being the number of available pairwise comparisons in the data and T the number of different treatments. The inconsistency factors can be assumed to follow a common distribution in order to increase the power in their estimation. A χ^2 test can be used to assess the inconsistency of the whole network, under the null assumption that all inconsistency factors are zero. In the presence of multi-arm studies, however, this model is problematic (van Valkenhoef *et al*., 2012b). Higgins *et al*. (2012) showed that different parameterizations of the model (choosing different reference treatments) may lead to different results. Thus, when multi-arm studies are present, the use of the Lu and Ades model should be avoided.

4.2.2. The 'design-by-treatment' inconsistency model. Higgins *et al*. (2012) and White *et al*. (2012) introduced an alternative inconsistency model, the 'design-by-treatment' interaction model, which encompasses both loop and design inconsistencies. The latter corresponds to the possible discrepancies in the treatment effects across designs, where 'design' refers to the treatments being compared in a study. For example, a study comparing treatments A and B is considered to be an AB design. The A-versus-B estimate coming from such a study may be different than the A-versus-B estimate coming from a three-arm study comparing treatments A, B, and C (an ABC design); this difference is referred to as design inconsistency. In the absence of multi-arm studies, the Lu and Ades model is equivalent to the design-by-treatment model. Similarly to the Lu and Ades model, when multi-arm studies are present, the estimates of the inconsistency factors depend on the parameterization. Unlike

the Lu and Ades model, however, the global statistic for inconsistency in the design-by-treatment interaction model is invariant under re-parameterization. One limitation of the model is the increased complexity and the accompanying loss of power for the statistical test (Higgins *et al*., 2012). Another point of criticism for this approach is that it adopts a somewhat arbitrary definition of design inconsistency, as it is mainly dictated by methodological rather than clinical considerations.

Jackson *et al*. (2014) presented an implementation of the design-by-treatment inconsistency model where they assumed random effects on the inconsistency factors. In the same paper, the authors also derived R and I^2 statistics to quantify heterogeneity and inconsistency. In addition, an approach based on the notion of design-by-treatment inconsistency was also proposed by Piepho and colleagues (Piepho *et al*., 2012; Piepho, 2014; Piepho *et al*., 2015).

4.2.3. Unrelated mean effects model. An alternative method was introduced by Dias *et al*. (2013d). In this model, the consistency equations are completely removed, and the NMA model is equivalent to a series of separate, independent meta-analyses for each pairwise contrast, sharing, however, a common heterogeneity variance (Caldwell *et al*., 2012). The fit of the model is then compared with the standard consistency NMA model using the posterior deviance and the deviance information criterion (Spiegelhalter *et al*., 2002). In addition, estimating the contribution to posterior mean deviance for each data point can help identify possibly 'problematic' studies, that is, studies not fitting well with the rest of the evidence. Each data point is expected to have a contribution of about 1 to the posterior mean deviance. A larger value will suggest a poor fit to the model, pointing out possibly inconsistent pieces of evidence. The use of leverage plots has been also suggested as a diagnostic tool for identifying studies with poor fit (Dias *et al*., 2010b). A limitation of the unrelated mean effects model is that, in the presence of multi-arm studies, a re-parameterization will affect the results of a random-effects meta-analysis.

4.2.4. Q statistic for inconsistency. A different approach for globally assessing inconsistency is by using the Q statistic for inconsistency (König *et al*., 2013; Krahn *et al*., 2013), which is analogous to the Q statistic for heterogeneity in simple meta-analysis. This approach is based on the two-stage method for fitting NMA (Lu *et al*., 2011). On the first stage, we perform a pairwise meta-analysis for the studies of each design available in the dataset and obtain the direct relative treatment effect. From the AB studies, for example, we estimate $\widehat{\mu}_{AB}$; this can be used to compute the Q_{AB}^{het} statistic for heterogeneity as $Q_{AB}^{het} = \sum_i \left(1/s_i^2\right)\left(y_i - \widehat{\mu}_{AB}\right)^2$, where i runs through all AB studies. Similarly, on the second stage of the analysis, we obtain network estimates for all pairwise comparisons; using these estimates and the direct estimates of the first stage, a Q statistic for the inconsistency of the whole network can be obtained, and the null hypothesis of consistency in the network can be tested using a χ^2 distribution with $C - T - 1$ degrees of freedom. This approach can be used in the presence of multi-arm studies. The authors also propose the use of a graphical tool, the *net-heat* plot, which provides a display of the flow of evidence in the network and can help identify potential hotspots for inconsistency.

Also, Rücker and Schwarzer suggested a Q statistic for inconsistency using a graph-theoretical approach and showed that it is identical to the Q statistic described earlier (Rücker, 2012; Rücker and Schwarzer, 2014).

4.2.5. Multidimensional scaling. In another graphical approach proposed by Chung and Lumley (2008), the multidimensional scaling method is used to infer about inconsistency in a network. For each pairwise comparison, a usual inverse variance meta-analysis is performed; the magnitude of the relative treatment effects is considered to be a measure of the observed pairwise 'dissimilarity' of the treatments. The pairwise estimates are summarized in a dissimilarity matrix, to which a weighted multidimensional scaling is applied in order to obtain the 'fitted dissimilarities'. Important differences between observed and fitted dissimilarities are an indicator of possible inconsistencies. Note that this method cannot properly handle multi-arm studies.

4.3. Empirical studies and simulations on inconsistency

Empirical studies show that the prevalence of inconsistency in published networks is non-trivial. Song *et al*. (2011) performed a meta-epidemiological study that included 112 published triangular networks, 16 of which were found to be statistically inconsistent. Veroniki *et al*. (2013) evaluated inconsistency in 40 published networks including a total of 303 loops. They found that 2–9% of the loops were inconsistent, depending on the effect measure used and the assumptions for heterogeneity; also, approximately one-eighth of the networks were found to be inconsistent using the design-by-treatment method.

The various methods for assessing inconsistency have been rarely and poorly applied in published NMAs (Donegan *et al*., 2010). In a meta-epidemiological study by Nikolakopoulou *et al*. (2014), it was found that, in 24% of the published NMAs, the authors did not use appropriate methods to evaluate inconsistency, while in 44%, the authors did not report using any method at all.

Song *et al*. (2012) performed simulations to evaluate the statistical properties of various methods for inferring about network inconsistency. They explored the use of the loop-specific approach, the node-splitting technique, and the Lu and Ades model. They found that even though all methods are unbiased, they have little power in

detecting inconsistency. It is also important to note that inferences on inconsistency heavily depend on the extent of heterogeneity and the method used to evaluate it (Veroniki *et al*., 2013). Thus, analysts should keep in mind that a statistically non-significant estimate for inconsistency should not be interpreted as proof of consistency. In addition, even when statistically significant inconsistency is found, its magnitude should be interpreted in terms of clinical relevance; thus, a statistically significant inconsistency in a certain loop might be clinically unimportant.

4.4. Choosing between the methods for evaluating inconsistency

If the network structure allows it, i.e. if there are closed loops in the network, a statistical assessment of inconsistency should always take place after fitting the NMA model. In the previous paragraphs we presented a variety of methods and models currently available for statistically checking the network for consistency and we discussed the advantages and limitations of each approach. In Table 1 we provide an overview of these approaches, including a brief summary of the limitations of each one.

An assessment of inconsistency may start with the loop-specific approach, which, despite its shortcomings, is the easiest one to implement and can pinpoint possibly problematic loops. Afterwards, if all studies in the network are two-armed, all presented strategies are valid choices for checking for inconsistencies. We generally recommend the application of both local and global methods to gain a better understanding of the source of possible discrepancies between direct and indirect evidence and the plausibility of the consistency assumption in the network as a whole. If the network includes multi-arm studies, only the design-by-treatment model and the Q statistic approach will lead to results that are independent of the parameterization of the model (i.e., the choice of the basic parameters). Researchers may still choose to implement some of the other methods as exploratory analyses; they should bear in mind, however, that their results might not be robust.

Approaches for evaluating inconsistency can also be selected based on the available technical expertise and/or software packages. In Table 2, we provide a summary of the currently available software solutions for implementing the various approaches.

If statistically significant inconsistency is detected, researchers are advised to explore potential sources of it and try to explain it. Local methods for assessing inconsistency can indicate outlying studies, which should be checked for data extraction errors, important differences in the distribution of effect modifiers, or other possible biases. In Section 6.7, we present various models for adding covariates and adjusting for suspected biases in the analysis. If sufficient studies are available, such models can be applied to explain and possibly eliminate inconsistencies, while, if inconsistency persists, researchers can consider splitting up the network (see discussion in Section 6.3). Finally, in the case of unexplained inconsistency, researchers may choose not to synthesize the evidence in an NMA at all or to present the results from the appropriate inconsistency model (the Lu and Ades model when all studies are two-armed; the design-by-treatment model when multi-arm trials are present) along with the direct evidence and a warning to the readers of the limitations of the analysis.

5. Software options for fitting a network meta-analysis model and evaluating inconsistency

Our literature search showed that BUGS software is a popular choice for implementing new methods in NMA, the majority of the articles included in our database reported using WinBUGS, OpenBUGS (Lunn *et al*., 2000, 2009), or JAGS (Plummer, 2003): Achana *et al*., (2013), Ades (2003), Ades *et al*. (2010), Caldwell *et al*. (2005, 2010, 2012), Carlin *et al*. (2013), Chaimani *et al*. (2013b), Chaimani and Salanti (2012), Cooper *et al*. (2009, 2011), Cope and Jansen (2013), Dakin *et al*. (2011), Del Giovane *et al*. (2013), Dewilde and Hawkins (2012), Dias *et al*. (2010a, 2010b, 2010c, 2013a, 2013b, 2013c, 2013d), Ding and Fu (2012), Donegan *et al*. (2013a, 2013b), Efthimiou *et al*. (2014a, 2014b), Franchini *et al*. (2012), Govan *et al*. (2010), Griffin *et al*. (2006), Higgins and Whitehead (1996), Hoaglin *et al*. (2011), Hong *et al*. (2013a, 2013b), Jansen (2011), Jansen *et al*. (2008), Jansen and Cope (2012), Jonas *et al*, (2013), Jones *et al*. (2011), Lu *et al*. (2007), Lu and Ades (2004, 2006, 2009), Madan *et al*. (2011, 2014), Mavridis *et al*. (2013), Mills *et al*. (2013a), Nixon *et al*. (2007), Ohlssen *et al*. (2013), O'Regan *et al*. (2009), Ouwens *et al*. (2010), Price *et al*. (2011a, 2011b), Salanti *et al*. (2009, 2010, 2011), Saramago *et al*. (2012), Schmid *et al*. (2013), Schmidli *et al*. (2013), Schmitz *et al*. (2012, 2013), Soares *et al*. (2014), Song *et al*. (2012), Spineli *et al*. (2013), Sturtz and Bender (2012), Thorlund *et al*. (2013, 2014), Thorlund and Mills (2012a), Trinquart *et al*. (2012a, 2012b), Tu and Faggion (2012), van Valkenhoef *et al*. (2012a), Warren *et al*, (2014), Wells *et al*. (2009), Welton *et al*. (2008, 2009, 2010, 2012), White *et al*. (2012), Woods *et al*. (2010), and Zhang *et al*. (2013). An alternative option for implementing Bayesian statistical inference is STAN, a recently developed programming language (Stan Development Team, 2014). However, we did not identify any articles using STAN.

Also, there were many articles that reported the use of R (R Core Team, 2014): Chung and Lumley (2008), Dias *et al*. (2010b), Donegan *et al*. (2012, 2013a), Jansen (2012), Krahn *et al*. (2013), Lu *et al*. (2011), Lu and Ades (2009), Lumley (2002), Mills *et al*. (2011b, 2013a), Naudet *et al*. (2013), Rücker (2012), Rücker and Schwarzer (2014), Salanti *et al*. (2009), Schmid *et al*. (2013), Song *et al*. (2012), Sturtz and Bender (2012), Tan *et al*. (2014), Thorlund *et al*. (2014), Thorlund and Mills (2012a), Trinquart *et al*. (2012b), van Valkenhoef *et al*. (2012a, 2012c), Veroniki *et al*.

(2013), and Zhang *et al*. (2013); some papers used STATA (StataCorp., 2013): Ballesteros (2005), Chaimani *et al*. (2013a), Franchini *et al*. (2012), Higgins *et al*. (2012), Jonas *et al*. (2013), Madan *et al*. (2011), Norton *et al*. (2012), Tu and Faggion (2012), Veroniki *et al*. (2013), Warren *et al*. (2014), and White *et al*. (2012); and a few papers reported using SAS software (SAS Institute, 2011): Carlin *et al*. (2013), Hong *et al*. (2013b), Jonas *et al*. (2013), Jones *et al*. (2011), Piepho *et al*. (2012), Piepho (2014), Piepho *et al*. (2015), and Signorovitch *et al*. (2012). Finally, Van Valkenhoef *et al*. (2012a) presented GEMTC, a freely available, open-source program for performing NMA.

Neupane *et al*. (2014) performed a review of the available automated packages for performing an NMA in R, aiming to summarize the key features and functionality of each package. In Table 2, we describe possible software solutions for some of the models presented in this review.

6. Other issues in network meta-analysis

6.1. The use of different measures of effect size

There is a wide choice of summary effect measures that can be used for the meta-analysis of evidence on a binary outcome. The most common choices are the OR, risk ratio for harmful or beneficial outcomes (RR_H and RR_B), risk difference (RD), and hazard ratio for time-to-event data. Donegan *et al*. (2010) reported that the majority of the published NMAs of dichotomous data used OR and RR (50% and 40%, respectively), with RD being used in only 10% of the analyses. Veroniki *et al*. (2013) analyzed 40 published networks and showed that the choice of effect measure may have an impact on the inferences about the statistical inconsistency. This was also discussed by Coory and Jordan (2010); using information from published networks, they concluded that the use of OR and RR is preferable over that of RD. In addition, it has been demonstrated that the choice of the scale may have an impact on the results of an NMA (Gleser and Olkin, 2009). In particular, Eckermann *et al*. (2009) showed that the use of RR may lead to inferential fallacies and advocated the use of OR. Norton *et al*. (2012) discussed that different choices of scale may lead to differences in the ranking of the treatments in an NMA. They propose that researchers should explore how sensitive the NMA results are in the choice of effect measure. Van Valkenhoef and Ades (2013) on the other hand discuss that a rank reversal is unlikely to take place unless the assumptions underlying NMA do not hold or the data are very sparse; thus, if a rank reversal is observed when using a certain effect measure, then the transitivity assumption is unlikely to hold irrespective of the choice of effect measure.

In principle, the choice of the summary statistic in any meta-analysis should be dictated (if possible) by biological justifications regarding the disease (Walter, 2000) and also by empirical evidence (Deeks, 2002). Alternatively, it is possible to fit models using alternative effect measures and make a choice by comparing model fit. Caldwell *et al*. (2012) proposed the use of the posterior mean deviance and the deviance information criterion to evaluate the model fit of the different effect measures in an NMA. The choice of the effect measure can also be guided by comparing the estimates of between-study heterogeneity, with lower values being preferable. This, however, may be impractical when there are few data available, in which case, the choice of scale may be driven

Table 1. An overview of the methods for assessing statistical inconsistency along with the limitations of each method.

Approach to inconsistency	Method	Limitations
Local methods	Loop specific	Does not compare direct with all indirect evidence for each comparison Different loops sharing a comparison are not independent Multiple testing issues
	Composite test	Fails to account for correlations induce by multi-arm studies
	Back-calculation	Problematic for a random-effects meta-analysis Cannot properly handle multi-arm studies
	Node splitting	Computationally intensive Cannot properly handle multi-arm studies
Global methods	Lu and Ades model	Depends on parameterization when multi-arm studies are included
	Design-by-treatment model	Non-intuitive definition of inconsistency
	Unrelated mean effects model	Depends on parameterization when multi-arm studies are included
	Q statistic for inconsistency	Based on the notion of design-by-treatment inconsistency model (non-intuitive definition of inconsistency)
	Multidimensional scaling	Cannot properly handle multi-arm studies

Table 2. Available software solutions for network meta-analysis (NMA).

Method	Software
Bucher method	STATA (indirect command) ITC (Microsoft Windows application, available at https://www.cadth.ca/resources/itc-user-guide/introduction)
NMA as meta-regression	STATA (metareg command: when only two-armed trials are included in the network; mvmeta and network commands: in the presence of multi-armed studies in the network; lincom command can be used to estimate the functional parameters in both cases.) R (rma command from metafor package: when only two-armed studies are included) (Viechtbauer, 2010)
Hierarchical NMA model	BUGS codes available at: http://www.mtm.uoi.gr and http://www.bris.ac.uk/social-community-medicine/projects/mpes/ R (gemtc package (van Valkenhoef and Kuiper, 2015), information available at https://drugis.org/software/r-packages/gemtc; pcnetmeta package (Lin *et al*., 2015) available at http://cran.r-project.org/web/packages/pcnetmeta/index.html)
NMA as a multivariate meta-analysis	STATA (mvmeta; network) R (mvmeta) (Gasparrini *et al*., 2012)
Graph-theoretical approach to NMA	R (netmeta package: accounts for multi-armed studies available at http://cran.r-project.org/web/packages/netmeta/index.html) (Rücker *et al*., 2015)
Loop-specific approach for inconsistency	R (routines available at http://mtm.uoi.gr/index.php/how-to-do-an-mtm) STATA (network graphs, available at http://mtm.uoi.gr/index.php/stata-routines-for-network-meta-analysis)
Node splitting approach	BUGS (codes available at http://www.bristol.ac.uk/cobm/research/mpes) GeMTC software STATA (network package)
Lu and Ades model	BUGS (codes available at http://www.bristol.ac.uk/social-community-medicine/projects/mpes) STATA (mvmeta) (Gasparrini *et al*., 2012) GeMTC software
Design-by-treatment model	STATA (mvmeta; network)
Q-statistics in NMA	R (netmeta) STATA (mvmeta; network)
Graphical presentation	STATA (network graphs, available at http://mtm.uoi.gr/index.php/stata-routines-for-network-meta-analysis) R (netmeta)

by the ease of interpretation and understanding of the disease process (Caldwell *et al*., 2012; Dias *et al*., 2013a). Analysts should also keep in mind that the choice of the effect measure for modeling may be different than the scale used for decision making (Egger *et al*., 1997); for example, OR may be preferred for data analysis, while RR or RD is easier to interpret and can be used to better communicate results with patients or clinicians (Walter, 2000).

Moreover, hazard ratios should always be considered as a suitable choice of scale for the case where there is an underlying time-to-event process and the proportional hazards assumption is deemed plausible (Caldwell *et al*., 2012; Soares *et al*., 2014) (see Section 6.3 for additional details on the methodology). Also note that the discussion of this paragraph pertains to the analysis of a binary outcome. When continuous data are available, the analysts should avoid dichotomization because it translates into a loss of power and also because the choice of a cutoff point may impact on the inferences of NMA (Schmitz *et al*., 2012).

6.2. Modeling time-to-event data

In many RCTs, the outcome measured is the time to the occurrence of an event (e.g., death and disease progression). Welton *et al*. (2010) described a method for simultaneously synthesizing survival and disease progression outcomes in a single NMA analysis; also, Woods *et al*. (2010) provided guidance on how to perform an NMA on the log-hazard scale when studies report different survival statistics.

Analysts should keep in mind that the synthesis of time-to-event data in terms of hazard ratios relies on the proportional hazards assumption. Treatment effects, however, may vary over time, and this might threaten the validity of the meta-analysis results. For NMA, this might have an extra impact on the consistency of the results. Ouwens *et al*. (2010) and Jansen (2011) modeled the hazard functions using parametric survival curves and fractional polynomials of time, respectively; in these models, the hazard ratio is allowed to vary over time. Jansen

and Cope (2012) discussed methods for extending these models by including study-level covariates (such as study date) that can reduce possible inconsistencies and bias. In another paper by the same authors (Cope and Jansen, 2013), various alternative summaries were presented for summarizing the estimates of the relative treatment effects obtained from an NMA of survival data.

6.3. Extension of network meta-analysis models to account for effect modifiers

In a pairwise meta-analysis, a meta-regression on important effect modifiers can explain the presence of between-study heterogeneity, which may hinder the interpretation of the results and may have important implications in decision making (Dias *et al*., 2013b). In NMA, interpreting results will be even more problematic in the presence of evidence inconsistency; meta-regression techniques in NMA adjust the treatment effects for possible effect modifiers and can reduce heterogeneity and inconsistency in the results that may be present when these covariates are distributed unevenly among studies (Achana *et al*., 2013; Cooper *et al*., 2009; Dias *et al*., 2013b; Greco *et al*., 2013; Hoaglin *et al*., 2011; Jonas *et al*., 2013; Nixon *et al*., 2007; Salanti *et al*., 2009). The effect modifiers can be either categorical or continuous variables and may represent either patient-level or trial-level characteristics.

6.3.1. General model for including effect modifiers in the analysis. Nixon *et al*. (2007) first combined NMA and meta-regression techniques to develop models that allow the simultaneous comparison of multiple competing treatments while adjusting for study-level covariates in an attempt to investigate and explain heterogeneity. Salanti *et al*. (2009) and Cooper *et al*. (2009) proposed the use of meta-regression as a tool for eliminating inconsistency as well as heterogeneity in NMA. As an example of adding covariates in NMA, Salanti *et al*. (2009) considered the year of randomization in each trial as a covariate in an NMA for topical fluoride treatments for the prevention of dental carries. The covariate adjusted for possible time trends in the placebo-controlled comparisons, and relative treatment effects were estimated for a predefined year of randomization (the year of the most recent study).

In general, there are three main approaches in the meta-regression of study-level covariates for NMA (Cooper *et al*., 2009; Dias *et al*., 2013b): using different and unrelated interaction terms (coefficients), using exchangeable interaction terms, and using a common interaction term.

6.3.1.1. Unrelated interaction terms. In this approach, there are a number of interaction terms for each covariate equal to the number of the basic parameters of the model. Let us assume for simplicity that we are only interested in one study-level covariate x_i. We can augment the hierarchical random-effects model previously presented as follows: for a study i, comparing treatments B versus C, we allow $\theta_{i\mathrm{BC}} \sim \mathrm{N}(\mu_{\mathrm{BC}} + x_i\beta_{\mathrm{BC}}, \tau^2)$, assuming a common heterogeneity variance τ^2 for the treatment effects. If treatment A is chosen to be the reference treatment, we can utilize the consistency equations to write $\theta_{i\mathrm{BC}} \sim \mathrm{N}(\mu_{\mathrm{AC}} - \mu_{\mathrm{AB}} + x_i(\beta_{\mathrm{AB}} - \beta_{\mathrm{AC}}), \tau^2)$; in a Bayesian analysis, the β_{AT} 'basic' coefficients (where $\mathrm{T} \neq \mathrm{A}$) can be assigned unrelated vague prior distributions.

6.3.1.2. Exchangeable interaction terms. The model has the same structure as the model for unrelated interaction terms, but now the basic coefficients are drawn from a common distribution, $\beta_{\mathrm{AT}} \sim \mathrm{N}(b, \tau_b^2)$, where index T runs through all treatments except reference treatment A. The mean b and the variance τ_b^2 of the common distributions can be assigned vague priors.

6.3.1.3. Common interaction term. The common interaction term model is the same as the exchangeable interaction model described in the previous paragraph, but now all basic interaction terms are assumed equal, $\beta_{\mathrm{AT}} = \beta$, for all treatments $\mathrm{T} \neq \mathrm{A}$. A vague prior is then assigned to β. This model implies that the relative treatment effects between the non-reference treatments are independent of the covariate, as the interaction terms cancel out. In this case, the choice of the reference treatment becomes important, as it might change the meta-regression results (Dias *et al*., 2013b).

6.3.2. Network meta-analysis meta-regression for baseline risk. The underlying risk of the disease, usually termed as 'baseline risk', is a proxy for important patient characteristics that may be possible modifiers of the treatment effects, and it can be included as a covariate in an NMA; however, care should be taken to account for its correlation with the treatment effects (Achana *et al*., 2013; Dias *et al*., 2013b). Achana *et al*. (2013) proposed a random-effects meta-regression model in which the effect of the reference treatment was used as a measure of the baseline risk. In order to include studies not reporting the reference treatment, the authors proposed three alternative distributional assumptions for the 'true' unobserved baseline risk. Following Cooper *et al*. (2009), the interaction terms were taken to be independent, exchangeable, or common. The authors recommended that the goodness of fit of the various alternative configurations can be based on residual deviance.

6.3.3. Limitations of network meta-analysis meta-regression models. Analysts should keep in mind that network meta-regression inherits all the interpretation difficulties of meta-regression for the case of pairwise meta-analysis, most importantly, the inability to directly infer on causal relationships (Dias *et al*., 2013b), and the risk of ecological bias if study-level covariates are used to infer about individuals. Another drawback of meta-regression models for

decision making in general is that in order to assess the magnitude of the relative treatment effects, the analyst must choose a value of the covariate at which to make the comparison (Hoaglin *et al*., 2011).

Regarding the choice between the three meta-regression models we presented, the unrelated interaction terms approach requires a sufficient number of studies to inform each pairwise comparison. In the common case where some of the comparisons in the network are performed in only few studies (or even worse in only one study), the coefficients cannot be properly estimated. Moreover, Dias *et al*. (2013b) advocate that even though the use of models with exchangeable coefficients seems attractive, they are likely to lead to situations where decision making on the optimal treatment will depend on the value of the covariate. In addition, when there are comparisons in the network informed by a low number of studies, the ranking of the treatments might be based in non-robust results. Therefore, even though the exchangeable coefficient model – or even more complex models – can be fitted, the authors suggest that their use should be limited to exploratory analyses. For these reasons, the common interaction approach might be the safest choice for most networks, but analysts should keep in mind the rather strong assumption it employs (i.e., all interaction terms are equal) and that it heavily depends on the choice of reference treatment.

6.4. Investigating potential sources of bias in network meta-analysis

When combining results from different studies in a (pairwise or network) meta-analysis, researchers always run the risk of obtaining biased estimates of the relative treatment effects for the target population. This may be the case when some of the studies provide a biased estimate of the treatment effect for their population of interest (which may or may not be the same as the target population of the meta-analysis). For example, treatment effects may be overestimated in studies of low methodological quality. This relates to the so-called internal validity of the included studies. The pooled result of a (network) meta-analysis may, however, still be biased even if the included studies have no internal biases. This might be the case, for example, when the included studies refer to populations different than the one targeted by the meta-analysis or when studies without 'statistically significant' results have not been published (Turner *et al*., 2009). This rather relates to the generalizability (sometimes referred to as 'external validity') of the included studies with respect to the targeted question or population.

Dias *et al*. (2013b) discussed that, when confronted with studies of mixed quality, researchers have three options: they can choose to analyze only the high-quality studies, thus ignoring a possibly important amount of information; they can choose to analyze all evidence, thus risking a bias in the pooled estimates; or they can include all studies after taking into account and adjusting for possible biases in the studies. In what follows, we focus on the last choice, presenting various available approaches for adjusting for suspected internal biases in the included studies, and also for adjusting for various forms of external bias.

6.4.1. Accounting for study limitations in network meta-analysis. Several approaches can be used to account for the potential presence of biases in meta-analysis. Spiegelhalter and Best (2003) proposed an additive bias model to adjust for both external and internal biases. More advanced methods account for multiple sources of bias, where the contribution of each source is modeled separately (see, e.g., Eddy and Shachter, 1992; Greenland, 2003, 2005; Wolpert and Mengersen, 2004). A straightforward but more resource-intensive approach to adjust for bias is by eliciting bias distributions (Turner *et al*., 2009). This involves assigning a number of independent experts to evaluate each study in terms of some predefined criteria that may indicate the presence of bias. This information can be used to construct an overall bias distribution. The parameters of this distribution are combined with the estimates of the studies in order to produce a bias-adjusted estimate of the treatment effect in each study. These estimates can then be synthesized using standard NMA techniques.

Models that account for bias usually assume that biased studies yield inaccurate treatment effects of size $\mu + \beta$, where μ is the unbiased treatment effect and β is a bias parameter. For the case of NMA, if the study-specific bias parameters are assumed to be exchangeable across studies, the unbiased treatment effects and the mean bias can be estimated from the network (Dias *et al*., 2013b, 2013c). Dias *et al*. (2010c) presented a model where exchangeable bias parameter with non-zero mean was included in studies that compared active versus inactive treatments and were considered to be of a high risk of bias (according to some predefined measure such as allocation concealment, blinding, or other trial characteristics). They also explored the use of two different bias parameters, one for active-versus-inactive comparisons and one for active-versus-active comparisons; note that in this approach some assumption on the direction of bias in the active-versus-active trials is necessary to be made. Salanti *et al*. (2010) considered a similar model in which the newest treatments were favored, thus adjusting the treatment effects for possible 'optimism' or 'novelty' bias (Dias *et al*., 2010a). Study size can be a proxy for the study's risk of bias, and Chaimani and Salanti (2012) presented a method for adjusting for the 'small study effects', the exaggeration of treatment effects in smaller trials. This exaggeration might be due to methodological differences between smaller and larger trials that affect treatment effectiveness, owing to publication bias or owing to reporting bias. They proposed a network meta-regression model, where the bias parameter is multiplied with the observed variance of the treatment effects in each study; the standard error or the precision (inverse variance) can be used alternatively. Their model can also adjust for suspected 'sponsorship' bias, for the case when interventions are sponsored in some of the studies. A similar model was also presented by Trinquart *et al*. (2012b).

6.4.2. Selection model to account for publication bias. Mavridis *et al*. (2013) proposed a Bayesian implementation of the Copas (1999) selection model for addressing possible publication bias in NMA. The idea behind selection models is that the observed set of published studies is considered to be a 'biased' sample, owing to the nature of the publication process. This is addressed by introducing a latent variable for each study, the 'propensity of publication', which is assumed to be correlated with the study's effect size. Mavridis *et al*. (2013) modeled propensity via a regression model, where it was assumed to be inversely proportional to the standard error of the effect size. They considered alternative scenarios of how the selection model parameters depend on the treatments being compared in each study. Trinquart *et al*. (2012b) also presented a selection model that modeled the propensity score of a trial as a linear function of the standard error. The effect sizes of the studies were weighted according to their propensity. Their model was shown to yield similar results to the model by Mavridis *et al*. (2013).

6.4.3. Accounting for ecological bias. The meta-analysis of aggregated data (AD) can lead to *ecological bias*. This refers to a bias that may arise when using AD in order to make inferences about patient-level interactions. Govan *et al*. (2010) proposed an NMA model to control for ecological bias by specifically modeling the effects of the covariates. Their model allows the inclusion of studies that provide information on all covariates, studies that report marginal data on some of the covariates, and also studies not providing any covariate information at all. The model allows the joint estimation of both the treatment and the covariate effects.

6.4.4. Graphical approaches to bias. In a different approach, Salanti *et al*. (2008b) discussed how the geometry of the network can offer insight on the presence of a 'comparator preference' bias, that is, when head-to-head comparisons between effective treatments are deliberately avoided, which in turn would imply that the treatment effects versus the reference treatments might be exaggerated. The authors utilized two indices from ecological literature: diversity, which is a measure of the number of treatments present in the network and how often they were tested, and co-occurrence, which measures whether specific treatment comparisons were preferred in the network while others were avoided. Limited diversity and significant co-occurrence in a network are indicators of possible preference bias in the network (Ioannidis, 2009).

Jansen *et al*. (2012) discussed the use of directed acyclic graphs (DAGs) as a graphical tool for conceptually evaluating the consistency assumption and also identifying possible sources of bias in indirect and mixed-treatment estimates. By means of DAGs, they showed that NMA estimates can be biased when relative treatment effect modifiers vary across comparisons and are not adjusted for in the analysis. They also showed that adjusting for covariates that are not effect modifiers is not only unnecessary but can introduce bias.

6.4.5. Empirical assessments of the impact of bias in network meta-analysis. Chaimani *et al*. (2013b) performed a network meta-epidemiological study to explore the effect of trial characteristics and study precision in NMA. They analyzed 32 networks and found evidence that imprecise studies (studies reporting broader confidence intervals for their estimates) tend to report larger effects compared with more precise studies, thus altering the results of the NMA. However, they found no evidence of association between effect size and previously identified indicators of bias, including blinding, allocation concealment, and random sequence generation. Trinquart *et al*. (2012a) used data from 74 Food and Drug Administration (FDA)-registered placebo-controlled studies on 12 antidepressants along with 51 corresponding publications in order to assess the impact of publication bias. They found that the effect sizes derived from published studies differed from the ones derived from the FDA data by at least 100% for almost half of the pairwise comparisons. They concluded that reporting bias alters NMA estimates and changes the treatments' ranking. They also noted that the impact of reporting bias may be more important in NMA compared with classical meta-analyses, in the sense that reporting bias in one treatment comparison may have an effect in the ranking of all treatments in NMA.

6.5. Reporting results from network meta-analysis

Although the implementation of NMA is increasingly gaining popularity, the quality of reporting has been rather low. Various meta-epidemiological studies of published NMAs showed that the methods used and the assumptions made were not routinely reported (Bafeta *et al*., 2013; Donegan *et al*., 2010; Lee, 2013; Nikolakopoulou *et al*., 2014). Ohlssen *et al*. (2013) presented a checklist of items that should be reported in a drug safety Bayesian NMA while Ades *et al*. (2013) and Mills *et al*. (2012a) gave guidelines for those reviewing an NMA for the purposes of decision making.

One possible hurdle in the reporting of an NMA is that presenting all results can be a challenging task, especially for networks with many treatments and multiple outcomes. The literature offers a plethora of graphical and tabular methods for visualizing the evidence base (Chaimani *et al*., 2013a; Chung and Lumley, 2008; Senn *et al*., 2013), the assumptions made (Chaimani *et al*., 2013a; Krahn *et al*., 2013), and the results obtained from an NMA (Chaimani *et al*., 2013a; Chung and Lumley, 2008; Fadda *et al*., 2011; König *et al*., 2013; Salanti *et al*., 2011; Senn *et al*., 2013; Tan *et al*., 2014). In a meta-epidemiological study on the presentational approaches used,

Tan *et al*. (2013) examined NMAs published in the UK and found that there is no standardized presentational approach for reporting the results of NMA. The authors concluded that a standardization of reporting is required.

6.6. Modeling multiple outcomes and repeated measures

Randomized clinical trials commonly report on more than one outcome. These outcomes may be correlated within a study (because of the fact that observations come from the same set of patients), and in addition, the true treatment effects on the outcomes can be correlated across studies (reflecting the way outcomes are related when measured in different settings). The usual meta-analytical approach on multiple outcomes is to analyze each one separately, ignoring all possible correlations. On the other hand, a joint meta-analysis of all outcomes, which incorporates possible correlations, can increase precision by 'borrowing of strength' across outcomes and may reduce the impact of outcome reporting bias (Jackson *et al*., 2011; Kirkham *et al*., 2012).

Welton *et al*. (2008) described a method for performing an NMA of two correlated outcomes for the case when all studies are two armed. Hong *et al*. (2013a) presented a model for multiple outcomes that does not take into account within-study correlations. Schmid *et al*. (2013) proposed a model for unordered categorical data that also allows the inclusion of studies with partially observed data. Efthimiou *et al*. (2014a) developed a method for multiple binary outcomes, where information on the correlations is elicited from expert clinicians. Also, Efthimiou *et al*. (2014b) presented two models that can be used to perform a joint NMA for correlated dichotomous, continuous, time-to-event, or mixed outcomes: the first of these models can be used when within-study correlations are available, the second when these correlations are unavailable.

Competing risks is a special case of multiple-outcome structure where the outcomes are mutually exclusive; Ades *et al*. (2010) presented methods for performing a competing-risks NMA. Price *et al*. (2011a) discussed methods for an NMA in multi-state Markov models; a model averaging technique was also proposed (Price *et al*., 2011b) for combining estimates from alternative multi-state models.

In some cases, studies may report on a single outcome for multiple time points, which leads to a series of correlated observations. Lu *et al*. (2007) proposed a hierarchical NMA model for synthesizing repeated measures of a discrete outcome. Dakin *et al*. (2011) suggested a model that can handle a continuous outcome but did not include in the analysis the correlations between the observations. Ding and Fu (2012) also presented a model for a continuous outcome that automatically modeled the correlations between the observations at different time points. Madan *et al*. (2014) presented methods for analyzing two dichotomous outcomes reported on multiple time points, for studies comparing complex interventions.

Using NMA results to decide which of all available treatments is optimal for a specific condition might be a non-trivial issue, when the treatments are compared for more than one outcome. In order to facilitate decision making in the presence of multiple outcomes, van Valkenhoef *et al*. (2012c) proposed a method for multiple-criteria benefit–risk assessment of all competing treatments in an NMA; also, Hong *et al*. (2013b) described a similar method for producing an overall ranking of the treatments in the network using a scoring system for combining efficacy and safety outcomes.

6.7. Definition of nodes in the treatment network

One important decision that analysts must make at the onset of an NMA regards the number of nodes (treatments) to be included in the network, with a simple choice being to include all relevant treatments. As discussed in Section 3.1.1, the parameters of interest in an NMA are the relative treatment effects between the various competing treatments for a target disorder and a target population; however, a decision maker may not be interested in all pairwise comparisons. The relative treatment effect between two active treatments might be more important to decision making than the relative effects of active treatments versus placebo. When this is the case, researchers may simplify the network by focusing on just a subset of the available treatments, that is, the ones that are deemed to be clinically relevant (e.g., newer/more effective treatments). This, however, poses a dilemma: including in the evidence base studies that compare treatments that are not clinically interesting may provide additional indirect evidence for the clinically interesting ones, which in turn may increase the precision of the estimates (Li *et al*., 2011; Salanti, 2012); on the other hand, obtaining all relevant evidence in order to include clinically uninteresting treatments in the network may be very time–consuming, and it may also lead to additional inconsistencies (Sturtz and Bender, 2012).

Hawkins *et al*. (2009b) performed an empirical study that supported the use of all potentially relevant data; in another empirical study, Mills *et al*. (2013a) concluded that the exclusion of treatments in an NMA might have an important effect on the results and might limit its usefulness, if important comparisons are excluded. To address this issue Hawkins *et al*. (2009a) presented two alternative iterative search strategies for identifying an efficient set of evidence, where the comparators included in each search is determined by the results of the previous iteration. In addition, Cooper *et al*. (2011) showed that extending the network to include more treatments might lead to increased heterogeneity, which in turn will increase the uncertainty in the results despite the inclusion of additional information.

Another challenging issue of NMA relates to the definition of nodes in the network. It is not uncommon for treatments – although defined similarly – to be administered in different ways across the trials; for instance, trials that involve the same drug may apply different doses. Alternatively, it might be the case that some of the drugs belong to the same drug class. In such scenarios, it is unclear whether and how similar treatments should be grouped into nodes. Also, as the included studies adopt increasingly different definitions of the nodes, transitivity becomes a less plausible assumption. This, in turn, may lead to increased inconsistency and/or heterogeneity in the results (Dewilde and Hawkins, 2012; Salanti, 2012). Decisions about whether or not to group similar treatments in a node should be therefore guided by clinical considerations regarding their similarity and the research question at hand. In addition, Del Giovane *et al*. (2013) and also Warren *et al*. (2014) discussed several alternative approaches to account for variability in treatment definition. The choice of the network structure can then be guided by statistical considerations, after comparing goodness-of-fit measures between alternative plausible models.

Another frequently encountered scenario in which the definition of nodes can be of importance is when interventions are administered as a combination of more than one treatment; the simplest approach would be to analyze each combination as a different node in the NMA. Welton *et al*. (2009) and Mills *et al*. (2012b) proposed possible scenarios for modeling how interventions interact with each other when combined into a complex intervention, with one of the approaches being the assumption of additive treatment effects. Thorlund and Mills (2012a) performed a simulation study that showed that when the treatment effects are truly additive, the conventional NMA model performs poorly in comparison with the additive effects model.

In summary, even though there is no exact recipe available for setting up the network and defining the nodes, the choice should be guided by considerations on the transitivity assumption, the presence of statistical inconsistency, the possibility of bias, and also practical constraints on the resources available for setting up the database. Ideally, whenever possible, decisions about the network structure should be described *a priori* in the protocol in order to avoid selective use of data (Salanti, 2012).

6.8. Incorporating individual patient data in network meta-analysis

The NMA models we have discussed so far can be used only for the analysis of AD, while the meta-regression approaches presented in Section 6.3.1 allow the exploration of the effects of only study-level covariates to the relative treatment effects. On the other hand, the use of IPD in an NMA (either exclusively or in combination with AD) is expected to increase precision and also allows the distinction between within-study and across-study associations to be made, so as to avoid possible ecological bias (Jansen, 2012). We extensively discuss the statistical methodology and the potential advantages of an IPD-MA when pooling head-to-head trials in a companion paper (Debray *et al*., 2015). These advantages also apply to NMA, and access to IPD is particularly relevant when the number of included studies is small and the validity of using meta-regression of study-level covariates becomes increasingly questionable. The use of patient-level covariates will allow a better evaluation of the heterogeneity and inconsistency in the network (Greco *et al*., 2013; Hoaglin *et al*., 2011; Jansen *et al*., 2008, 2011b).

A few models for including IPD in an NMA have been recently proposed. Saramago *et al*. (2012) developed a series of NMA models set in a Bayesian background that can be used for the simultaneous synthesis of IPD and AD while incorporating both study-level and individual-level covariates. Their models also allow the inclusion of studies with different designs (cluster and individual allocation). The authors found that the incorporation of IPD in the network resulted in an increase in the precision compared with an AD-only analysis, even when IPD are available only for a fraction of the studies. Donegan *et al*. (2012, 2013a) presented a model for combining IPD and AD in a single analysis with three alternative specifications (unrelated, exchangeable, and common interactions; see also Section 6.3.1). The inclusion of both IPD and AD in the analysis was shown to lead to an increased precision of the estimates of the regression coefficients and a better assessment of the consistency assumption. A similar model was proposed by Jansen (2012). In the same paper, a second, alternative model was also suggested for the case of a binary covariate. The author performed a simulation study indicating that the second model is less affected by bias at the cost of larger uncertainty in the results. Finally, Ali *et al*. (2013) discussed the use of IPD in order to identify possible interactions between treatment effects and potential effect modifiers; when such modifiers are found to be unevenly distributed among studies, the authors suggest that NMA models need to account for these differences.

6.9. Utilizing data from non-randomized and observational studies

Ades and Sutton (2006) discuss that results obtained from RCTs may not be necessarily generalizable to a wide population and that randomized studies' results could be combined with information from observational studies or patient registries, after adjusting for potential biases. Randomized and non-randomized evidence can be regarded as being complimentary, in the sense that observational studies can be considered to be reliable sources of information regarding the population baseline, while RCTs regarding intervention affects data. Dias *et al*. (2013c) describe how non-randomized studies can be used to inform a 'baseline natural history model'. Evidence from such studies can be used to estimate the absolute effect for a reference treatment. This can in turn be

combined with NMA results for the relative effects of active treatments, in order to obtain an estimate of the absolute treatment effects.

Schmitz *et al*. (2013) proposed three alternative models for jointly synthesizing information from RCTs as well as non-randomized studies: the simplest approach presented was to perform a naïve pooling, disregarding differences in study design; the second approach was to utilize non-randomized studies as prior information, while adjusting for bias due to study design; the third was a three-level hierarchical model that accounts for bias and for heterogeneity between trial designs. The first of the models (naïve pooling) should only be used as the first step of the analysis, as it disregards potential biases in non-randomized trials. The second model (using observational evidence as prior information) allows adjusting for biases, but between-trial design heterogeneity is not taken into account, and it is not possible to include more than two different trial designs. The third model (three-level hierarchical model) addresses these issues and should be preferred.

Finally, Soares *et al*. (2014) discussed the use of observational data for the case that there are sparse and few data in an NMA. In their approach, such data were used to inform the baseline effects but did not directly contribute to the relative treatment effects.

6.10. Planning future studies

The issue of planning future studies based on the results of an existing NMA has received little attention in the literature. Thorlund and Mills (2012b) and also Snapinn and Jiang (2011) provided sample size considerations for determining the statistical power of indirect evidence, and Mills *et al*. (2011b) performed a simulation study to estimate the power of indirect comparisons; however, there is currently little guidance on the design (i.e., treatments to be compared) and the sample size needed for updating an existing NMA in an optimal manner.

Naci and O' Conor (2013) alternatively suggest the design and conduct of prospective NMAs; this would go against the current practice of retrospective NMA, where each individual study is planned in isolation from others. They also suggest that the regulatory agencies should have an active role in the design of future trials, especially in the selection of comparators and in ensuring that the patient populations are comparable in terms of treatment effect modifiers.

7. Concluding remarks

In this article, we have presented the currently available methodology on NMA. We have described major developments in the field, discussed all known relevant issues, and highlighted key assumptions while providing guidance for the interested researchers.

Our review has several limitations. Pragmatic decisions needed to be made given the lack of a widely accepted terminology referring to NMA, the abundance of recently published articles, and the complexity of new methods in order to ensure a timely publication of this review. Thus, there may have been articles that presented methodological advances for NMA that we failed to identify by not including in our search more online databases and by not hand-searching additional journals. We believe, however, that even if the identified set of articles might not be complete, it is representative of the currently available methods for NMA and that the most important methodological aspects, challenges, and solutions of NMA are covered. Moreover, although we present some of the mathematical features of the various models and methods, we do it in a descriptive manner, and we do not provide all relevant details. Hence, this review serves as a road map for researchers: the keen reader should refer to the original articles for details, keeping also in mind that NMA is still an active, rapidly developing research field.

Appendix A

The following are the members of the GetReal Methods Review Group:

- Mark Belger, Eli Lilly, UK
- Silvie Bozzi, Sanofi, France
- Maximo Carreras, F. Hoffmann-La Roche AG, Switzerland
- Thomas P. A. Debray, UMC Utrecht, The Netherlands
- Orestis Efthimiou, University of Ioannina, Greece
- Matthias Egger, ISPM Bern, Switzerland
- Christine Fletcher, Amgen Ltd., UK
- Sandro Gsteiger, ISPM Bern, Switzerland
- Noemi Hummel, ISPM Bern, Switzerland
- Gablu Kilcher, ISPM Bern, Switzerland
- Brice Kitio-Dschassi, Sanofi, France
- Amr Makady, Zorginstituut Nederland, The Netherlands

- Karel G.M. Moons, UMC Utrecht, The Netherlands
- Rolf H. H. Groenwold, UMC Utrecht, The Netherlands
- Klea Panayidou, ISPM Bern, Switzerland
- Johannes B. Reitsma, UMC Utrecht, The Netherlands
- Georgia Salanti, University of Ioannina, Greece
- Aijing Shang, F. Hoffmann-La Roche AG, Switzerland
- Sven Trelle, CTU Bern, Switzerland
- Gert van Valkenhoef, University of Groningen, The Netherlands
- Brigitta Monz, GSK, Germany

Appendix B

We based our search on a previous review of the literature in NMA performed by the 'Comparing Multiple Interventions Methods Group' of the Cochrane Collaboration, which resulted in a publicly available database (www.zotero.org/groups/network_meta-analysis_methods/items/) that included a total of 186 papers. This database has been shared with experts in the field to identify missing relevant articles. We also used the results from a recent literature review performed by Donegan *et al*. (2013b), where 116 papers on methods for assessing the homogeneity and consistency assumptions of NMA were identified (referred to as 'key paper' in Figure B1).

In addition, we searched the MEDLINE database for relevant hits using the following algorithm:

(network OR mixed treatment* OR multiple treatment* OR mixed comparison* OR indirect comparison* OR umbrella OR simultaneous comparison*) AND (meta-analysis).

This query produced 1789 hits (14 March 2014), 88 of which were deemed relevant.

Articles that have appeared in two methodological journals, namely, *Journal of Research Synthesis Methods (RSM)* and *Journal of the Royal Statistical Society (JRSS)*, series A, B and C, are not indexed by the MEDLINE database. We performed a search using the following algorithm for RSM: 'network meta-analysis' in Abstract OR 'mixed treatment' in Abstract OR 'multiple treatment' in Abstract OR 'indirect comparison' in Abstract OR 'umbrella' in Abstract OR 'simultaneous comparison' in Abstract in *Research Synthesis Methods*. This provided 17 articles, four of which were deemed relevant. In order to identify relevant articles published in *JRSS*, we performed a hand-search, which identified two relevant papers.

In Figure B1, we present the flow chart of the papers identified in our search. Eligible papers were classified using a number of tags. Each article was assigned one or more tags according to the type of research presented, one or more tags according to the methodological topics addressed, and one or more tags according to the software it used to implement the methods it presented.

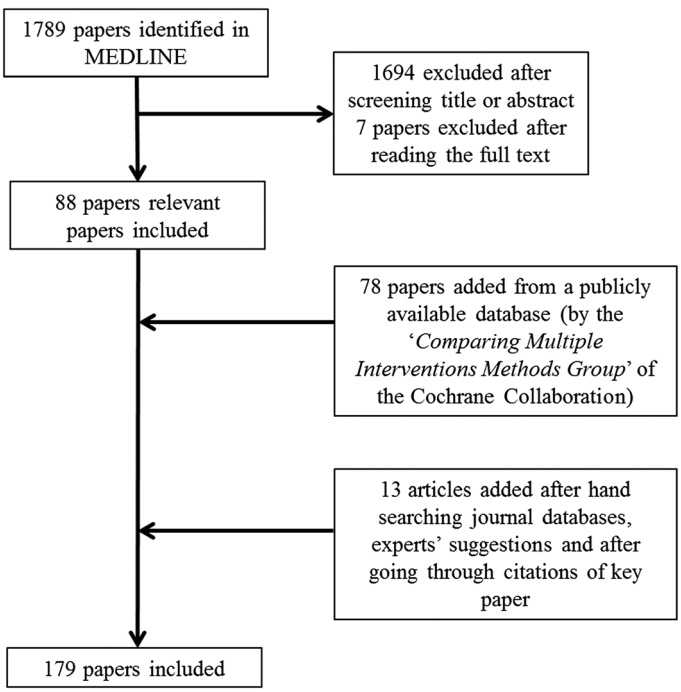

Figure B1. Flow chart of included and excluded methodological papers for the network meta-analysis review.

We used five *contribution* tags:

- *Methodology development*: assigned to papers presenting a novel methodology
- *Didactical/good practice/recommendations*: for papers giving guidance or advice
- *Methodology overview*: assigned to papers presenting a summary of the existing methodology for NMA
- *Simulation*: for papers using a simulated dataset to make assessments
- *Empirical assessment*: for papers presenting an assessment based on published NMAs

There were 14 *methodology* tags:

- *Basic methodology*: for papers presenting novel methodology for addressing fundamental issues of NMA
- *Definition of nodes*: assigned to articles presenting methodology regarding the definition of treatment nodes included in an NMA
- *Effect sizes*: for papers addressing issues on the different effect sizes that can be used in an NMA
- *Conceptual issues/assumptions underlying NMA*: assigned to papers that elucidate the conceptual issues of an NMA and discuss the assumptions that need to hold in order for an NMA to give valid results
- *Statistical inconsistency*: for papers discussing methods for quantifying statistical inconsistency in NMA, for papers presenting ways for addressing inconsistency, or for papers examining the prevalence of inconsistency in published networks
- *Risk of bias*: for papers presenting methods for addressing the risk of bias in an NMA
- *Non-randomized and observational studies*: assigned to papers suggesting ways to include data from non-randomized and observational studies in an NMA
- *Publication bias*: for papers presenting methods for addressing the risk of publication bias in an NMA
- *Multiple outcomes/repeated measures/survival analysis*: for papers presenting methods for the joint analysis of multiple correlated outcomes, repeated measures, and analysis of survival data
- *NMA meta-regression:* for papers discussing the use of covariates in an NMA
- *IPD in NMA studies*: for papers presenting ways to include evidence from studies reporting IPD in an NMA
- *Sensitivity analyses*: for papers presenting some form of sensitivity analysis
- *Planning future studies*: for papers discussing methods for planning future trials
- *Reporting NMA*: for papers discussing methods for reporting the results of an NMA

Finally we used 4 *software implementation* tags:

- *BUGS*: for papers using WinBUGS, OpenBUGS, or JAGS
- *R*: assigned to papers using the R programming language
- *STATA*: for papers using the STATA software package
- *SAS:* for papers using the SAS software package

Acknowledgements

The research leading to these results was conducted as part of the GetReal consortium. For further information, please refer to www.imi-getreal.eu

Funding

This work was supported by the Innovative Medicines Initiative Joint Undertaking under grant agreement no. 115546, resources of which are composed of financial contribution from the European Union's Seventh Framework Programme (FP7/2007-2013) and EFPIA companies in kind contribution.

References

Achana FA, Cooper NJ, Dias S, Lu G, Rice SJ, Kendrick D, Sutton AJ. 2013. Extending methods for investigating the relationship between treatment effect and baseline risk from pairwise meta-analysis to network meta-analysis. *Statistics in Medicine* **32**: 752–771. DOI:10.1002/sim.5539.

Ades AE. 2003. A chain of evidence with mixed comparisons: models for multi-parameter synthesis and consistency of evidence. *Statistics in Medicine* **22**: 2995–3016. DOI:10.1002/sim.1566.

Ades AE, Caldwell DM, Reken S, Welton NJ, Sutton AJ, Dias S. 2013. Evidence synthesis for decision making 7: a reviewer's checklist. *Medical Decision Making* **33**: 679–691. DOI:10.1177/0272989X13485156.

Ades AE, Mavranezouli I, Dias S, Welton NJ, Whittington C, Kendall T. 2010. Network meta-analysis with competing risk outcomes. *Value in Health* **13**: 976–983. DOI:10.1111/j.1524-4733.2010.00784.x.

Ades AE, Sutton AJ. 2006. Multiparameter evidence synthesis in epidemiology and medical-decision making: current approaches. *Journal of the Royal Statistical Society A* **169**: 5–35.

Ali S, Mealing S, Hawkins N, Lescrauwaet B, Bjork S, Mantovani L, Lampertico P. 2013. The use of individual patient-level data (IPD) to quantify the impact of pretreatment predictors of response to treatment in chronic hepatitis B patients. *BMJ Open* **3**: . DOI:10.1136/bmjopen-2012-001309.

Bafeta A, Trinquart L, Seror R, Ravaud P. 2013. Analysis of the systematic reviews process in reports of network meta-analyses: methodological systematic review. *BMJ* **347**: f3675.

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 Medical Research Methodology* **2**: 13.

Ballesteros J. 2005. Orphan comparisons and indirect meta-analysis: a case study on antidepressant efficacy in dysthymia comparing tricyclic antidepressants, selective serotonin reuptake inhibitors, and monoamine oxidase inhibitors by using general linear models. *Journal of Clinical Psychopharmacology* **25**: 127–131.

Berkey CS, Hoaglin DC, Antczak-Bouckoms A, Mosteller F, Colditz GA. 1998. Meta-analysis of multiple outcomes by regression with random effects. *Statistics in Medicine* **17**: 2537–2550. DOI:10.1002/(SICI)1097-0258(19981130)17:22<2537::AID-SIM953>3.0.CO;2-C.

Bucher HC, Guyatt GH, Griffith LE, Walter SD. 1997. The results of direct and indirect treatment comparisons in meta-analysis of randomized controlled trials. *Journal of Clinical Epidemiology* **50**: 683–691.

Caldwell DM, Ades AE, Higgins JPT. 2005. Simultaneous comparison of multiple treatments: combining direct and indirect evidence. *BMJ* **331**: 897–900. DOI:10.1136/bmj.331.7521.897.

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. *Journal of Clinical Epidemiology* **63**: 875–882. DOI:10.1016/j.jclinepi.2009.08.025.

Caldwell DM, Welton NJ, Dias S, Ades AE. 2012. Selecting the best scale for measuring treatment effect in a network meta-analysis: a case study in childhood nocturnal enuresis. *Research Synthesis Methods* **3**: 126–141.

Carlin BP, Hong H, Shamliyan TA, Sainfort F, Kane RL. 2013. Case Study Comparing Bayesian and Frequentist Approaches for Multiple Treatment Comparisons. (Prepared by the Minnesota Evidence-Based Practice Center under Contract No. 290-2007-10064-I2). AHRQ Publication No. 12(13)-EHC103-EF. Rockville, MD: Agency for Healthcare Research and Quality.

Chaimani A, Higgins JPT, Mavridis D, Spyridonos P, Salanti G. 2013a. Graphical tools for network meta-analysis in STATA. *PLoS One* **8**: e76654. DOI:10.1371/journal.pone.0076654.

Chaimani A, Salanti G. 2012. Using network meta-analysis to evaluate the existence of small-study effects in a network of interventions. *Research Synthesis Methods* **3**: 161–176.

Chaimani A, Vasiliadis HS, Pandis N, Schmid CH, Welton NJ, Salanti G. 2013b. Effects of study precision and risk of bias in networks of interventions: a network meta-epidemiological study. *International Journal of Epidemiology* **42**: 1120–1131. DOI:10.1093/ije/dyt074.

Chung H, Lumley T. 2008. Graphical exploration of network meta-analysis data: the use of multidimensional scaling. *Clinical Trials* **5**: 301–307. DOI:10.1177/1740774508093614.

Cipriani A, Barbui C, Rizzo C, Salanti G. 2012. What is a multiple treatments meta-analysis? *Epidemiology and Psychiatric Sciences* **21**: 151–153. DOI:10.1017/S2045796011000837.

Cipriani A, Higgins JPT, Geddes JR, Salanti G. 2013. Conceptual and technical challenges in network meta-analysis. *Annals of Internal Medicine* **159**: 130–137. DOI:10.7326/0003-4819-159-2-201307160-00008.

Coleman CI, Phung OJ, Cappelleri JC, Baker WL, Kluger J, White CM, Sobieraj DM. 2012. Use of Mixed Treatment Comparison in Systematic Reviews. (Prepared by the university of Connecticut/Harford Hospital Evidence-Based Practice Center under Contract No. 290-2007-10067-I). AHRQ Publication No. 12-EHC119-EF. Rockville, MD: Agency for Healthcare Research and Quality.

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 in Health* **14**: 371–380. DOI:10.1016/j.jval.2010.09.001.

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. *Statistics in Medicine* **28**: 1861–1881. DOI:10.1002/sim.3594.

Coory M, Jordan S. 2010. Frequency of treatment-effect modification affecting indirect comparisons: a systematic review. *PharmacoEconomics* **28**: 723–732. DOI:10.2165/11535670-000000000-00000.

Copas J. 1999. What works?: selectivity models and meta-analysis. *Journal of the Royal Statistical Society: Series A Statistic Society* **162**: 95–109. DOI:10.1111/1467-985X.00123.

Cope S, Jansen JP. 2013. Quantitative summaries of treatment effect estimates obtained with network meta-analysis of survival curves to inform decision-making. *BMC Medical Research Methodology* **13**: 147. DOI:10.1186/1471-2288-13-147.

Cucherat M, Izard V. 2009. Summary report: Indirect comparisons: methods and validity. HAS.

Dakin HA, Welton NJ, Ades AE, Collins S, Orme M, Kelly S. 2011. Mixed treatment comparison of repeated measurements of a continuous endpoint: an example using topical treatments for primary open-angle glaucoma and ocular hypertension. *Statistics in Medicine*. DOI:10.1002/sim.4284.

D'Ascenzo F, Biondi-Zoccai G. 2014. Network meta-analyses: the "white whale" for cardiovascular specialists. *Journal of Cardiothoracic and Vascular Anesthesia* **28**: 169–173. DOI:10.1053/j.jvca.2013.01.004.

Debray TPA, Moons KGM, van Valkenhoef G, Efthimiou O, Hummel N, Groenwold RHH, Reitsma JB, on behalf of GetReal Methods Review Group. 2015. Get real in individual participant data (IPD) meta-analysis: a review of the methodology. *Research Synthesis Methods*. DOI:10.1002/jrsm.1160.

Deeks JJ. 2002. Issues in the selection of a summary statistic for meta-analysis of clinical trials with binary outcomes. *Statistics in Medicine* **21**: 1575–1600. DOI:10.1002/sim.1188.

Del Giovane C, Vacchi L, Mavridis D, Filippini G, Salanti G. 2013. Network meta-analysis models to account for variability in treatment definitions: application to dose effects. *Statistics in Medicine* **32**: 25–39. DOI:10.1002/sim.5512.

Dewilde S, Hawkins N. 2012. Investigating incoherence gives insight: clopidogrel is equivalent to extended-release dipyridamole plus aspirin in secondary stroke prevention. *Journal of Clinical Epidemiology* **65**: 835–845. DOI:10.1016/j.jclinepi.2012.01.019.

Dias S, Sutton AJ, Ades AE, Welton NJ. 2013a. Evidence synthesis for decision making 2: a generalized linear modeling framework for pairwise and network meta-analysis of randomized controlled trials. *Medical Decision Making International Journal of Society Medical and Decision Making* **33**: 607–617. DOI:10.1177/0272989X12458724.

Dias S, Sutton AJ, Welton NJ, Ades AE. 2013b. Evidence synthesis for decision making 3: heterogeneity – subgroups, meta-regression, bias, and bias-adjustment. *Medical Decision Making* **33**: 618–640. DOI:10.1177/0272989X13485157.

Dias S, Welton NJ, Ades AE. 2010a. Study designs to detect sponsorship and other biases in systematic reviews. *Journal of Clinical Epidemiology* **63**: 587–588. DOI:10.1016/j.jclinepi.2010.01.005.

Dias S, Welton NJ, Caldwell DM, Ades AE. 2010b. Checking consistency in mixed treatment comparison meta-analysis. *Statistics in Medicine* **29**: 932–944. DOI:10.1002/sim.3767.

Dias S, Welton NJ, Marinho VCC, Salanti G, Higgins JPT, Ades AE. 2010c. Estimation and adjustment of bias in randomized evidence by using mixed treatment comparison meta-analysis. *Journal of the Royal Statistical Society A* **173**: 613–629.

Dias S, Welton NJ, Sutton AJ, Ades AE. 2013c. Evidence synthesis for decision making 5: the baseline natural history model. *Medical Decision Making* **33**: 657–670. DOI:10.1177/0272989X13485155.

Dias S, Welton NJ, Sutton AJ, Caldwell DM, Lu G, Ades AE. 2013d. Evidence synthesis for decision making 4: inconsistency in networks of evidence based on randomized controlled trials. *Medical Decision Making* **33**: 641–656. DOI:10.1177/0272989X12455847.

Ding Y, Fu H. 2012. Bayesian indirect and mixed treatment comparisons across longitudinal time points. *Statistics in Medicine*. DOI:10.1002/sim.5688.

Donegan S, Williamson P, D'Alessandro U, Garner P, Tudur Smith C. 2013a. Combining individual patient data and aggregate data in mixed treatment comparison meta-analysis: Individual patient data may be beneficial if only for a subset of trials. *Statistics in Medicine* **32**: 914–930. DOI:10.1002/sim.5584.

Donegan S, Williamson P, D'Alessandro U, Tudur Smith C. 2013b. Assessing key assumptions of network meta-analysis: a review of methods. *Research Synthesis Methods* **4**: 291–323. DOI:10.1002/jrsm.1085.

Donegan S, Williamson P, D'Alessandro U, Tudur Smith C. 2012. Assessing the consistency assumption by exploring treatment by covariate interactions in mixed treatment comparison meta-analysis: individual patient-level covariates versus aggregate trial-level covariates. *Statistics in Medicine* **31**: 3840–3857. DOI:10.1002/sim.5470.

Donegan S, Williamson P, Gamble C, Tudur SC. 2010. Indirect comparisons: a review of reporting and methodological quality. *PLoS One* **5**: e11054. DOI:10.1371/journal.pone.0011054.

Eckermann S, Coory M, Willan AR. 2009. Indirect comparison: relative risk fallacies and odds solution. *Journal of Clinical Epidemiology* **62**: 1031–1036. DOI:10.1016/j.jclinepi.2008.10.013.

Eddy DM, Shachter R. 1992. Meta-analysis by the Confidence Profile Method. Har/Dskt edn.Boston: Academic Press.

Edwards SJ, Clarke MJ, Wordsworth S, Borrill J. 2009. Indirect comparisons of treatments based on systematic reviews of randomised controlled trials. *International Journal of Clinical Practice* **63**: 841–854. DOI:10.1111/j.1742-1241.2009.02072.x.

Efthimiou O, Mavridis D, Cipriani A, Leucht S, Bagos P, Salanti G. 2014a. An approach for modelling multiple correlated outcomes in a network of interventions using odds ratios. *Statistics in Medicine* **33**: 2275–2287. DOI:10.1002/sim.6117.

Efthimiou O, Mavridis D, Riley RD, Cipriani A, Salanti G. 2014b. Joint synthesis of multiple correlated outcomes in networks of interventions. *Biostatistics (Oxford, England)* . DOI:10.1093/biostatistics/kxu030.

Egger M, Smith GD, Phillips AN. 1997. Meta-analysis: principles and procedures. *BMJ* **315**: 1533–1537. DOI:10.1136/bmj.315.7121.1533.

EUnetHTA – European network for Health Technology Assessment (last).. 2013. Comparators and comparisons: Direct and indirect comparisons. Available http://www.eunethta.eu/sites/5026.fedimbo.belgium.be/files/Direct%20and%20indirect%20comparisons.pdf

Fadda V, Maratea D, Trippoli S, Messori A. 2011. Results can be summarised in a simple figure. *BMJ* **342**: d1555–d1555. DOI:10.1136/bmj.d1555.

Falissard B, Zylberman M, Cucherat M, Izard V, Meyer F. 2009. Real medical benefit assessed by indirect comparison. *Thérapie* **64**: 225–232. DOI:10.2515/therapie/2009032.

Franchini AJ, Dias S, Ades AE, Jansen JP, Welton NJ. 2012. Accounting for correlation in network meta-analysis with multi-arm trials. *Research Synthesis Methods* **3**: 142–160.

Gartlehner G, Moore CG. 2008. Direct versus indirect comparisons: a summary of the evidence. *The International Journal of Technology Assessment in Health Care* **24**: 170–177. DOI:10.1017/S0266462308080240.

Gasparrini A, Armstrong B, Kenward MG. 2012. Multivariate meta-analysis for non-linear and other multi-parameter associations. *Statistics in Medicine* **31**: 3821–3839. DOI:10.1002/sim.5471.

Glenny AM, Altman DG, Song F, Sakarovitch C, Deeks JJ, D'Amico R, Bradburn M, Eastwood AJ. 2005. Indirect comparisons of competing interventions. *Health TechnolAssess* **9**: 1–iv.

Gleser L, Olkin LV. 2009. Stochastically dependent effect sizes. In Cooper H, Hedges LV (eds.) The Handbook of Research Synthesis and Meta-analysis. Russell Sage Foundation.

Govan L, Ades AE, Weir CJ, Welton NJ, Langhorne P. 2010. Controlling ecological bias in evidence synthesis of trials reporting on collapsed and overlapping covariate categories. *Statistics in Medicine* **29**: 1340–1356. DOI:10.1002/sim.3869.

Greco T, Landoni G, Biondi-Zoccai G, D'Ascenzo F, Zangrillo A. 2013. A Bayesian network meta-analysis for binary outcome: how to do it. *Statistical Methods in Medical Research* . DOI:10.1177/0962280213500185.

Greenland S. 2005. Multiple-bias modelling for analysis of observational data. *Journal of the Royal Statistical Society: Series A Statistic Society* **168**: 267–306. DOI:10.1111/j.1467-985X.2004.00349.x.

Greenland S. 2003. The impact of prior distributions for uncontrolled confounding and response bias: a case study of the relation of wire codes and magnetic fields to childhood leukemia. *Journal of the American Statistical Association* **98**: 47–54.

Griffin S, Bojke L, Main C, Palmer S. 2006. Incorporating direct and indirect evidence using Bayesian methods: an applied case study in ovarian cancer. *Value in Health* **9**: 123–131. DOI:10.1111/j.1524-4733.2006.00090.x.

Hawkins N, Scott DA, Woods B. 2009a. How far do you go? Efficient searching for indirect evidence. *Medical Decision Making* **29**: 273–281. DOI:10.1177/0272989X08330120.

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 in Health* **12**: 996–1003. DOI:10.1111/j.1524-4733.2009.00541.x.

Higgins JPT, Jackson D, Barrett JK, Lu G, Ades AE, White IR. 2012. Consistency and inconsistency in network meta-analysis: concepts and models for multi-arm studies. *Research Synthesis Methods* **3**: 98–110.

Higgins JPT, Whitehead A. 1996. Borrowing strength from external trials in a meta-analysis. *Statistics in Medicine* **15**: 2733–2749. DOI:10.1002/(SICI)1097-0258(19961230)15:24<2733::AID-SIM562>3.0.CO;2-0.

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 in Health* **14**: 429–437. DOI:10.1016/j.jval.2011.01.011.

Hong H, Carlin BP, Chu H, Shamliyan TA, Wang S, Kane RL. 2013a. A Bayesian missing data framework for multiple continuous outcome mixed treatment comparisons. In AHRQ Methods for Effective Health Care. Rockville (MD): Agency for Healthcare Research and Quality (US).

Hong H, Carlin BP, Shamliyan TA, Wyman JF, Ramakrishnan R, Sainfort F, Kane RL. 2013b. Comparing Bayesian and frequentist approaches for multiple outcome mixed treatment comparisons. *Medical Decision Making*. DOI:10.1177/0272989X13481110.

Ioannidis JP. 2009. Integration of evidence from multiple meta-analyses: a primer on umbrella reviews, treatment networks and multiple treatments meta-analyses. *CMAJ* **181**: 488–493. DOI:10.1503/cmaj.081086.

Ioannidis JP. 2006. Indirect comparisons: the mesh and mess of clinical trials. *Lancet* **368**: 1470–1472. DOI:10.1016/S0140-6736(06)69615-3.

Jackson D, Barrett JK, Rice S, White IR, Higgins JPT. 2014. A design-by-treatment interaction model for network meta-analysis with random inconsistency effects. *Statistics in Medicine* **33**: 3639–3654. DOI:10.1002/sim.6188.

Jackson D, Riley R, White IR. 2011. Multivariate meta-analysis: potential and promise. *Statistics in Medicine* **30**: 2481–2498. DOI:10.1002/sim.4172.

Jackson D, White IR, Thompson SG. 2010. Extending DerSimonian and Laird's methodology to perform multivariate random effects meta-analyses. *Statistics in Medicine* **29**: 1282–1297. DOI:10.1002/sim.3602.

Jansen JP. 2012. Network meta-analysis of individual and aggregate level data. *Research Synthesis Methods* **3**: 177–190.

Jansen JP. 2011. Network meta-analysis of survival data with fractional polynomials. *BMC Medical Research Methodology* **11**: 61. DOI:10.1186/1471-2288-11-61.

Jansen JP, Cope S. 2012. Meta-regression models to address heterogeneity and inconsistency in network meta-analysis of survival outcomes. *BMC Medical Research Methodology* **12**: 152. DOI:10.1186/1471-2288-12-152.

Jansen JP, Crawford B, Bergman G, Stam W. 2008. Bayesian meta-analysis of multiple treatment comparisons: an introduction to mixed treatment comparisons. *Value in Health* **11**: 956–964. DOI:10.1111/j.1524-4733.2008.00347.x.

Jansen JP, Fleurence R, Devine B, Itzler R, Barrett A, Hawkins N, Lee K, Boersma C, Annemans L, Cappelleri JC. 2011a. 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 International Society For Pharmacoeconomics and Outcomes Research* **14**: 417–428. DOI:10.1016/j.jval.2011.04.002.

Jansen JP, Fleurence R, Devine B, Itzler R, Barrett A, Hawkins N, Lee K, Boersma C, Annemans L, Cappelleri JC. 2011b. 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 in Health* **14**: 417–428. DOI:10.1016/j.jval.2011.04.002.

Jansen JP, Naci H. 2013. Is network meta-analysis as valid as standard pairwise meta-analysis? It all depends on the distribution of effect modifiers. *BMC Medicine* **11**: 159. DOI:10.1186/1741-7015-11-159.

Jansen JP, Schmid CH, Salanti G. 2012. Directed acyclic graphs can help understand bias in indirect and mixed treatment comparisons. *Journal of Clinical Epidemiology* **65**: 798–807. DOI:10.1016/j.jclinepi.2012.01.002.

Jonas DE, Wilkins TM, Bangdiwala S, Bann CM, Morgan LC, Thaler KJ, Amick HR, Gartlehner G. 2013. Findings of Bayesian Mixed Treatment Comparison Meta-Analysis: Comparison and Exploration Using Real-word Trial Data and Simulation. (Prepared by RTI-UNC Evidence-Based Practice Center under Contract No. 290-2007-10056-I). AHRQ Publication No. 13-EHC039-EF. Rockville, MD: Agency for Healthcare Research and Quality.

Jones B, Roger J, Lane PW, Lawton A, Fletcher C, Cappelleri JC, Tate H, Moneuse P. 2011. Statistical approaches for conducting network meta-analysis in drug development. *Pharmaceutical Statistics* **10**: 523–531. DOI:10.1002/pst.533.

Julious SA, Wang SJ. 2008. How biased are indirect comparisons, particularly when comparisons are made over time in controlled trials. *Drug Information Journal* **42**: 625.

Kirkham JJ, Riley RD, Williamson PR. 2012. A multivariate meta-analysis approach for reducing the impact of outcome reporting bias in systematic reviews. *Statistics in Medicine* **31**: 2179–2195. DOI:10.1002/sim.5356.

König J, Krahn U, Binder H. 2013. Visualizing the Flow of Evidence in Network Meta-analysis and Characterizing Mixed Treatment Comparisons. Statistics in Medicine n/a–n/a. DOI: 10.1002/sim.6001

Krahn U, Binder H, Konig J. 2013. A graphical tool for locating inconsistency in network meta-analyses. *BMC Medical Research Methodology* **13**: 35. DOI:10.1186/1471-2288-13-35.

Lee AW. 2013. Review of mixed treatment comparisons in published systematic reviews shows marked increase since 2009. *Journal of Clinical Epidemiology*. DOI:10.1016/j.jclinepi.2013.07.014.

Lin L, Zhang J, Chu H. 2015. pcnetmeta: patient-centered network meta-analysis. Available at: http://www.r-project.org/

Li T, Puhan MA, Vedula SS, Singh S, Dickersin K, The Ad Hoc Network Meta-analysis Methods Meeting Working Group. 2011. Network meta-analysis-highly attractive but more methodological research is needed. *BMC Medicine* **9**: 79. DOI:10.1186/1741-7015-9-79.

Lu G, Ades A. 2009. Modeling between-trial variance structure in mixed treatment comparisons. *Biostatistics* **10**: 792–805. DOI:10.1093/biostatistics/kxp032.

Lu G, Ades AE. 2006. Assessing evidence inconsistency in mixed treatment comparisons. *Journal of the American Statistical Association* **101**: 447–459.

Lu G, Ades AE. 2004. Combination of direct and indirect evidence in mixed treatment comparisons. *Statistics in Medicine* **23**: 3105–3124. DOI:10.1002/sim.1875.

Lu G, Ades AE, Sutton AJ, Cooper NJ, Briggs AH, Caldwell DM. 2007. Meta-analysis of mixed treatment comparisons at multiple follow-up times. *Statistics in Medicine* **26**: 3681–3699. DOI:10.1002/sim.2831.

Lu G, Welton NJ, Higgins JPT, White IR, Ades AE. 2011. Linear inference for mixed treatment comparison meta-analysis: a two-stage approach. *Research Synthesis Methods* **2**: 43–60.

Lumley T. 2002. Network meta-analysis for indirect treatment comparisons. *Statistics in Medicine* **21**: 2313–2324. DOI:10.1002/sim.1201.

Lunn DJ, Thomas A, Best N, Spiegelhalter D. 2000. WinBUGS – a Bayesian modelling framework: concepts, structure, and extensibility. *Statistics and Computing* **10**: 325–337. DOI:10.1023/A:1008929526011.

Lunn D, Spiegelhalter D, Thomas A, Best N. 2009. The BUGS project: evolution, critique and future directions. *Statistics in Medicine* **28**: 3049–3067. DOI:10.1002/sim.3680.

Madan J, Chen Y-F, Aveyard P, Wang D, Yahaya I, Munafo M, Bauld L, Welton N. 2014. Synthesis of evidence on heterogeneous interventions with multiple outcomes recorded over multiple follow-up times reported inconsistently: a smoking cessation case-study. *Journal of the Royal Statistical Society: Series A Statistic Society* **177**: 295–314. DOI:10.1111/rssa.12018.

Madan J, Stevenson MD, Cooper KL, Ades AE, Whyte S, Akehurst R. 2011. Consistency between direct and indirect trial evidence: is direct evidence always more reliable? *Value in Health* **14**: 953–960. DOI:10.1016/j.jval.2011.05.042.

Malone DC. 2007. Using indirect comparisons in pharmacoeconomic studies – time for implementation. *Clinical Therapeutics* **29**: 2454–2455. DOI:10.1016/j.clinthera.2007.11.011.

Mavridis D, Salanti G. 2013. A practical introduction to multivariate meta-analysis. *Statistical Methods in Medical Research* **22**: 133–158. DOI:10.1177/0962280211432219.

Mavridis D, Sutton A, Cipriani A, Salanti G. 2013. A fully Bayesian application of the Copas selection model for publication bias extended to network meta-analysis. *Statistics in Medicine* **32**: 51–66. DOI:10.1002/sim.5494.

Mills EJ, Bansback N, Ghement I, Thorlund K, Kelly S, Puhan MA, Wright J. 2011a. Multiple treatment comparison meta-analyses: a step forward into complexity. *Journal of Clinical Epidemiology* **3**: 193–202. DOI:10.2147/CLEP.S16526.

Mills EJ, Ghement I, O'Regan C, Thorlund K. 2011b. Estimating the power of indirect comparisons: a simulation study. *PLoS One* **6**: e16237. DOI:10.1371/journal.pone.0016237.

Mills EJ, Ioannidis JP, Thorlund K, Schunemann HJ, Puhan MA, Guyatt GH. 2012a. How to use an article reporting a multiple treatment comparison meta-analysis. *JAMA* **308**: 1246–1253. DOI:10.1001/2012.jama.11228.

Mills EJ, Kanters S, Thorlund K, Chaimani A, Veroniki AA, Ioannidis JP. 2013a. The effects of excluding treatments from network meta-analyses: survey. *BMJ* **347**: f5195.

Mills EJ, Thorlund K, Ioannidis JP. 2013b. Demystifying trial networks and network meta-analysis. *BMJ* **346**: f2914.

Mills EJ, Thorlund K, Ioannidis JP. 2012b. Calculating additive treatment effects from multiple randomized trials provides useful estimates of combination therapies. *Journal of Clinical Epidemiology* **65**: 1282–1288. DOI:10.1016/j.jclinepi.2012.07.012.

Naci H, Fleurence R. 2011. Using indirect evidence to determine the comparative effectiveness of prescription drugs: do benefits outweigh risks? *Health Outcomes Research medical* **2**: 241–249.

Naci H, O'Connor AB. 2013. Assessing comparative effectiveness of new drugs before approval using prospective network meta-analyses. *Journal of Clinical Epidemiology* **66**: 812–816. DOI:10.1016/j.jclinepi.2013.04.002.

Naudet F, Millet B, Charlier P, Reymann JM, Maria AS, Falissard B. 2013. Which placebo to cure depression? A thought-provoking network meta-analysis. *BMC Medicine* **11**: 230. DOI:10.1186/1741-7015-11-230.

Neupane B, Richer D, Bonner AJ, Kibret T, Beyene J. 2014. Network meta-analysis using R: a review of currently available automated packages. *PLoS ONE* **9**: e115065. DOI:10.1371/journal.pone.0115065.

Nikolakopoulou A, Chaimani A, Veroniki AA, Vasiliadis HS, Schmid CH, Salanti G. 2014. Characteristics of networks of interventions: a description of a database of 186 published networks. *PLoS One* **9**: e86754. DOI:10.1371/journal.pone.0086754.

Nixon RM, Bansback N, Brennan A. 2007. Using mixed treatment comparisons and meta-regression to perform indirect comparisons to estimate the efficacy of biologic treatments in rheumatoid arthritis. *Statistics in Medicine* **26**: 1237–1254. DOI:10.1002/sim.2624.

Norton EC, Miller MM, Wang JJ, Coyne K, Kleinman LC. 2012. Rank reversal in indirect comparisons. *Value in Health* **15**: 1137–1140. DOI:10.1016/j.jval.2012.06.001.

Ohlssen D, Price KL, Amy XH, Hong H, Kerman J, Fu H, Quartey G, Heilmann CR, Ma H, Carlin BP. 2013. Guidance on the implementation and reporting of a drug safety Bayesian network meta-analysis. *Pharmaceutical Statistics* . DOI:10.1002/pst.1592.

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. DOI:10.1186/1745-6215-10-86.

Ouwens MJNM, Philips Z, Jansen JP. 2010. Network meta-analysis of parametric survival curves. *Research Synthesis Methods* **1**: 258–271.

Panayidou K, Gsteiger S, Kilcher G, Egger M, Carreras M, Efthimiou O, Debray TPA, Trelle S, Hummel N, on behalf of GetReal Methods Review Group. 2015. Get Real in mathematical modelling: A review of methods predicting drug effectiveness from randomized trial data. *Research Synthesis Methods* Submitted.

Piepho H-P. 2014. Network-meta analysis made easy: detection of inconsistency using factorial analysis-of-variance models. *BMC Medical Research Methodology* **14**: 61. DOI:10.1186/1471-2288-14-61.

Piepho H-P, Madden LV, Williams ER. 2015. Multiplicative interaction in network meta-analysis. *Statistics in Medicine* **34**: 582–594. DOI:10.1002/sim.6372.

Piepho HP, Williams ER, Madden LV. 2012. The use of two-way linear mixed models in multitreatment meta-analysis. *Biometrics* **68**: 1269–1277.

Plummer M. 2003. JAGS: a program for analysis of Bayesian graphical models using Gibbs sampling.

Price MJ, Welton NJ, Ades AE. 2011a. Parameterization of treatment effects for meta-analysis in multi-state Markov models. *Statistics in Medicine* **30**: 140–151. DOI:10.1002/sim.4059.

Price MJ, Welton NJ, Briggs AH, Ades AE. 2011b. Model averaging in the presence of structural uncertainty about treatment effects: influence on treatment decision and expected value of information. *Value in Health* **14**: 205–218. DOI:10.1016/j.jval.2010.08.001.

R Core Team. 2014. R: A Language and Environment for Statistical Computing. Vienna, Austria: R Foundation for Statistical Computing. http://www.R-project.org

Rücker G. 2012. Network meta-analysis, electrical networks and graph theory. *Research Synthesis Methods* **3**: 312–324.

Rücker G, Schwarzer G. 2014. Reduce dimension or reduce weights? Comparing two approaches to multi-arm studies in network meta-analysis. *Statistics in Medicine* **33**: 4353–4369. DOI:10.1002/sim.6236.

Rücker G, Schwarzer G, Krahn U, König J. 2015. netmeta: network meta-analysis using frequentist methods. Available at: http://www.r-project.org/

Salanti G. 2012. Indirect and mixed-treatment comparison, network, or multiple-treatments meta-analysis: many names, many benefits, many concerns for the next generation evidence synthesis tool. *Research Synthesis Methods* **3**: 80–97. DOI:10.1002/jrsm.1037.

Salanti G, Ades AE, Ioannidis JPA. 2011. Graphical methods and numerical summaries for presenting results from multiple-treatment meta-analysis: an overview and tutorial. *Journal of Clinical Epidemiology* **64**: 163–171. DOI:10.1016/j.jclinepi.2010.03.016.

Salanti G, Dias S, Welton NJ, Ades AE, Golfinopoulos V, Kyrgiou M, Mauri D, Ioannidis JP. 2010. Evaluating novel agent effects in multiple-treatments meta-regression. *Statistics in Medicine* **29**: 2369–2383. DOI:10.1002/sim.4001.

Salanti G, Higgins JP, Ades AE, Ioannidis JP. 2008a. Evaluation of networks of randomized trials. *Statistical Methods in Medical Research* **17**: 279–301. DOI:10.1177/0962280207080643.

Salanti G, Kavvoura FK, Ioannidis JP. 2008b. Exploring the geometry of treatment networks. *Annals of Internal Medicine* **148**: 544–553.

Salanti G, Marinho V, Higgins JP. 2009. A case study of multiple-treatments meta-analysis demonstrates that covariates should be considered. *Journal of Clinical Epidemiology* **62**: 857–864. DOI:10.1016/j.jclinepi.2008.10.001.

Saramago P, Sutton AJ, Cooper NJ, Manca A. 2012. Mixed treatment comparisons using aggregate and individual participant level data. *Statistics in Medicine* **31**: 3516–3536. DOI:10.1002/sim.5442.

SAS Institute. 2011. The SAS System for Windows. Release 9.2. Cary, NC: SAS Inst..

Schmid CH, Trikalinos TA, Olkin I. 2013. Bayesian network meta-analysis for unordered categorical outcomes with incomplete data. *Research Synthesis Methods*n/a–n/a. DOI . DOI:10.1002/jrsm.1103.

Schmidli H, Wandel S, Neuenschwander B. 2013. The network meta-analytic-predictive approach to non-inferiority trials. *Statistical Methods in Medical Research* **22**: 219–240. DOI:10.1177/0962280211432512.

Schmitz S, Adams R, Walsh C. 2013. Incorporating data from various trial designs into a mixed treatment comparison model. *Statistics in Medicine*. DOI:10.1002/sim.5764.

Schmitz S, Adams R, Walsh C. 2012. The use of continuous data versus binary data in MTC models: a case study in rheumatoid arthritis. *BMC Medical Research Methodology* **12**: 167. DOI:10.1186/1471-2288-12-167.

Senn S, Gavini F, Magrez D, Scheen A. 2013. Issues in performing a network meta-analysis. *Statistical Methods in Medical Research* **22**: 169–189. DOI:10.1177/0962280211432220.

Signorovitch J, Erder MH, Xie J, Sikirica V, Lu M, Hodgkins PS, Wu EQ. 2012. Comparative effectiveness research using matching-adjusted indirect comparison: an application to treatment with guanfacine extended release or atomoxetine in children with attention-deficit/hyperactivity disorder and comorbid oppositional defiant disorder. *Pharmacoepidemiol. Saf* **21**(Suppl 2): 130–137. DOI:10.1002/pds.3246.

Snapinn S, Jiang Q. 2011. Indirect comparisons in the comparative efficacy and non-inferiority settings. *Pharmaceutical Statistics* **10**: 420–426. DOI:10.1002/pst.509.

Soares MO, Dumville JC, Ades AE, Welton NJ. 2014. Treatment comparisons for decision making: facing the problems of sparse and few data. *Journal of the Royal Statistical Society: Series A (Statistics in Society)* **177**: 259–279. DOI:10.1111/rssa.12010.

Sobieraj DM, Cappelleri JC, Baker WL, Phung OJ, White CM, Coleman CI. 2013. Methods used to conduct and report Bayesian mixed treatment comparisons published in the medical literature: a systematic review. *BMJ Open* **3**. DOI:10.1136/bmjopen-2013-003111.

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**: 472. DOI:10.1136/bmj.326.7387.472.

Song F, Clark A, Bachmann MO, Maas J. 2012. Simulation evaluation of statistical properties of methods for indirect and mixed treatment comparisons. *BMC Medical Research Methodology* **12**: 138. DOI:10.1186/1471-2288-12-138.

Song F, Harvey I, Lilford R. 2008. Adjusted indirect comparison may be less biased than direct comparison for evaluating new pharmaceutical interventions. *Journal of Clinical Epidemiology* **61**: 455–463. DOI:10.1016/j.jclinepi.2007.06.006.

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. 2003. Bayesian approaches to multiple sources of evidence and uncertainty in complex cost-effectiveness modelling. *Statistics in Medicine* **22**: 3687–3709. DOI:10.1002/sim.1586.

Spiegelhalter DJ, Best NG, Carlin BP, Van Der Linde A. 2002. Bayesian measures of model complexity and fit. *Journal of the Royal Statistical Society, Series B: Statistical Methodology* **64**: 583–639. DOI:10.1111/1467-9868.00353.

Spineli LM, Higgins JP, Cipriani A, Leucht S, Salanti G. 2013. Evaluating the impact of imputations for missing participant outcome data in a network meta-analysis. *Clinical Trials London England* **10**: 378–388. DOI:10.1177/1740774512470317.

Stan Development Team. 2014. Stan modeling language users guide and reference manual, version 2.5.0.

StataCorp. 2013. Stata Statistical Software: Release 13. College Station, TX: StataCorp LP.

Sturtz S, Bender R. 2012. Unsolved issues of mixed treatment comparison meta-analysis: network size and inconsistency. *Research Synthesis Methods* **3**: 300–311.

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

Sutton AJ, Cooper NJ, Jones DR. 2009. Evidence synthesis as the key to more coherent and efficient research. *BMC Medical Research Methodology* **9**: 29. DOI:10.1186/1471-2288-9-29.

Tan SH, Bujkiewicz S, Sutton A, Dequen P, Cooper N. 2013. Presentational approaches used in the UK for reporting evidence synthesis using indirect and mixed treatment comparisons. *Journal of Health Services Research & Policy* **18**: 224–232. DOI:10.1177/1355819613498379.

Tan SH, Cooper NJ, Bujkiewicz S, Welton NJ, Caldwell DM, Sutton AJ. 2014. Novel presentational approaches were developed for reporting network meta-analysis. *Journal of Clinical Epidemiology*. DOI:10.1016/j.jclinepi.2013.11.006.

Thorlund K, Mills E. 2012a. Stability of additive treatment effects in multiple treatment comparison meta-analysis: a simulation study. *Journal of Clinical Epidemiology* **4**: 75–85. DOI:10.2147/CLEP.S29470.

Thorlund K, Mills EJ. 2012b. Sample size and power considerations in network meta-analysis. *Systematic Reviews* **1**: 41. DOI:10.1186/2046-4053-1-41.

Thorlund K, Thabane L, Mills EJ. 2013. Modelling heterogeneity variances in multiple treatment comparison meta-analysis-are informative priors the better solution? *BMC Medical Research Methodology* **13**: 2. DOI:10.1186/1471-2288-13-2.

Thorlund K, Zafari Z, Druyts E, Mills EJ, Sadatsafavi M. 2014. The impact of incorporating Bayesian network meta-analysis in cost-effectiveness analysis – a case study of pharmacotherapies for moderate to severe COPD. *Cost Effectiveness and Resource Allocation CE* **12**: 8. DOI:10.1186/1478-7547-12-8.

Trinquart L, Abbe A, Ravaud P. 2012a. Impact of reporting bias in network meta-analysis of antidepressant placebo-controlled trials. *PLoS One* **7**: e35219. DOI:10.1371/journal.pone.0035219.

Trinquart L, Chatellier G, Ravaud P. 2012b. Adjustment for reporting bias in network meta-analysis of antidepressant trials. *BMC Medical Research Methodology* **12**: 150. DOI:10.1186/1471-2288-12-150.

Turner RM, Spiegelhalter DJ, Smith GCS, Thompson SG. 2009. Bias modelling in evidence synthesis. *Journal of the Royal Statistical Society: Series A Statistic Society* **172**: 21–47. DOI:10.1111/j.1467-985X.2008.00547.x.

Tu YK, Faggion CM. 2012. A primer on network meta-analysis for dental research. *ISRN Dentistry* **2012**: 276520. DOI:10.5402/2012/276520.

van Houwelingen HC, Arends LR, Stijnen T. 2002. Advanced methods in meta-analysis: multivariate approach and meta-regression. *Statistics in Medicine* **21**: 589–624.

van Valkenhoef G, Ades AE. 2013. Evidence synthesis assumes additivity on the scale of measurement: response to 'Rank reversal in indirect comparisons' by Norton et al. *Value in Health* **16**: 449–451. DOI:10.1016/j.jval.2012.11.012.

van Valkenhoef G, Kuiper J. 2015. gemtc: network meta-analysis using Bayesian. Available at: http://www.r-project.org/

van Valkenhoef G, Lu G, de Brock B, Hillege H, Ades AE, Welton NJ. 2012a. Automating network meta-analysis. *Research Synthesis Methods* **3**: 285–299.

van Valkenhoef G, Tervonen T, de Brock B, Hillege H. 2012b. Algorithm parameterization of mixed treatment comparisons. *Statistics and Computing* **22**: 1099–1111.

van Valkenhoef G, Tervonen T, Zhao J, de Brock B, Hillege HL, Postmus D. 2012c. Multicriteria benefit–risk assessment using network meta-analysis. *Journal of Clinical Epidemiology* **65**: 394–403. DOI:10.1016/j.jclinepi.2011.09.005.

Veroniki AA, Vasiliadis HS, Higgins JP, Salanti G. 2013. Evaluation of inconsistency in networks of interventions. *International Journal of Epidemiology* **42**: 332–345. DOI:10.1093/ije/dys222.

Viechtbauer W. 2010. Conducting meta-analyses in R with the metafor package. *Journal of Statistical Software* **36**: 1–48. DOI:10.18637/jss.v036.i03.

Walter SD. 2000. Choice of effect measure for epidemiological data. *Journal of Clinical Epidemiology* **53**: 931–939. DOI:10.1016/S0895-4356(00)00210-9.

Warren FC, Abrams KR, Sutton AJ. 2014. Hierarchical network meta-analysis models to address sparsity of events and differing treatment classifications with regard to adverse outcomes. *Statistics in Medicine*. DOI:10.1002/sim.6131.

Wells GA, Sultan SA, Chen L, Khan M, Coyle D. 2009. Indirect evidence: Indirect treatment comparisons in meta-analysis. *Ottawa Canadian Agency for Drugs and Technologies in Health*. https://www.cadth.ca/media/pdf/H0462_itc_tr_e.pdf

Welton NJ, Caldwell DM, Adamopoulos E, Vedhara K. 2009. Mixed treatment comparison meta-analysis of complex interventions: psychological interventions in coronary heart disease. *American Journal of Epidemiology* **169**: 1158–1165. DOI:10.1093/aje/kwp014.

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. *Statistics in Medicine* **27**: 5620–5639. DOI:10.1002/sim.3377.

Welton NJ, Sutton AJ, Cooper N, Abrams KR, Ades AE. 2012. Mixed and indirect treatment comparisons. In Evidence Synthesis for Decision Making in Healthcare. New York: John Wiley & Sons.

Welton NJ, Willis SR, Ades AE. 2010. Synthesis of survival and disease progression outcomes for health technology assessment of cancer therapies. *Research Synthesis Methods* **1**: 239–257.

White IR, Barrett JK, Jackson D, Higgins JPT. 2012. Consistency and inconsistency in network meta-analysis: model estimation using multivariate meta-regression. *Research Synthesis Methods* **3**: 111–125.

Wolpert RL, Mengersen KL. 2004. Adjusted likelihoods for synthesizing empirical evidence from studies that differ in quality and design: effects of environmental tobacco smoke. *Statistical Science* **19**: 450–471.

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 Medical Research Methodology* **10**: 54. DOI:10.1186/1471-2288-10-54.

Xiong T, Parekh-Bhurke S, Loke YK, Abdelhamid A, Sutton AJ, Eastwood AJ, Holland R, Chen YF, Walsh T, Glenny AM, Song F. 2013. Overall similarity and consistency assessment scores are not sufficiently accurate for predicting discrepancy between direct and indirect comparison estimates. *Journal of Clinical Epidemiology* **66**: 184–191. DOI:10.1016/j.jclinepi.2012.06.022.

Yang G, Liu D, Liu RY, Xie M, Hoaglin DC. 2014. Efficient network meta-analysis: a confidence distribution approach. *Statistical Methodology*. DOI:10.1016/j.stamet.2014.01.003.

Zhang J, Carlin BP, Neaton JD, Soon GG, Nie L, Kane R, Virnig BA, Chu H. 2013. Network meta-analysis of randomized clinical trials: reporting the proper summaries. *Clinical Trials*. DOI:10.1177/1740774513498322.