Background document

Stream 2 (Statistical issues in NMA)

CMIMG Meeting

16 & 17 July, Bristol, UK

Prepared by Anna Chaimani and Georgia Salanti for CMIMG WG2

Table of Contents

1 Introduction to the meeting

The Cochrane Comparing Multiple Interventions Methods group (CMIMG) received funding from the Cochrane Methodological Innovation project to produce guidance for Cochrane reviewers undertaking systematic reviews of multiple interventions. The work is divided into three streams.

Stream 2 is concerned with statistical issues and presentation of results and aims to provide guidance about the statistical methods associated with Network Meta-Analysis (NMA).The present document provides background and discussion topics for the Stream 2 meeting on 16-17 July 2013.

There is a vast literature on NMA methods and Stream 2 will produce guidance considering established and recently developed methods. We have collected the majority of the methodological papers on this topic and we have summarized the presented methodology (see Section 0). As NMA-related terminology varies considerably in the literature, CMIMG has put together a glossary of terms (see Section **Error! Reference source not found.**) that we use throughout this document.

2 Objectives of the meeting and anticipated output

It is important to highlight that the objectives of this stream do not include the production of new methods. We aim to establish the Cochrane NMA-related guidance drawing from the existing technical literature. At the end of the meeting we would hopefully:

a) Decide which methods are to be recommended for Cochrane reviews

The evaluation of the existing methodologies (to estimate effects, evaluate assumptions, present results etc.) should consider not only their statistical integrity but also (if not primarily!) their relevance for Cochrane reviews. *Simple and efficient methods, easily understood by clinicians are to be preferred over sophisticated ones without however compromising on quality.* Methods should also be placed in the context of the existing statistical guidance as outlined in the Cochrane Handbook. The various NMA-related methods are presented in Sections 3, 4, and 5.

b) Address some important uncertainties and reach consensus

A few issues have been identified in the literature for which there is no wide agreement. These are outlined in Section 6.

c) How to present data and results from NMA

A collection of the various graphical options in the literature is presented in Section 7. RevMan could possibly accommodate some of the options used to present the data.

d) Discuss what do we expect to see in the methods section of a protocol and Cochrane review

There are a few published papers make suggestions about how an NMA should be presented. An initiative to extend the PRISMA statement to NMA is ongoing. More information is in Section 8.

3 Models for NMA and fitting options

The various models can be seen from different angles. To our understanding, these different model presentations are equivalent and could be fit, in principal, in any Bayesian and frequentist software. However, ease of implementation and the availability of suitable software routines put practical constraints on the use of each method. For simplicity we focus on networks with two-arm studies only and we discuss implications for multi-arm studies.

3.1 Notation & basic concepts of network meta-analysis models

Let us consider a network with S studies and T treatments informing N direct pairwise comparisons.
The aim of NMA is to estimate all possible relative treatment effects (such as OR, MD etc.) between pairs of the T treatments, e.g. μ_{kc} the relative effect of treatment k vs. treatment c $(k, c = 1, \dots, T)$. In reality we do not need to estimate all μ_{kc}, but only a few of them, called basic parameters. At the start of every network meta-analysis we need to choose a set of basic parameters μ_t $(t = 1, \dots, T-1)$ representing the summary treatment effects of $T-1$ independent treatment comparisons. An easy way to define the basic parameters is to choose one of the T interventions as the reference (denoted with A) and each μ_t would represent the comparison of treatment t vs. A $(t = 1, \dots, T$ with $t \neq A)$, hence $\mu_t \equiv \mu_{tA}$. NMA will estimate all $\mu_t s$ and then, the network summary effects of all other (functional) comparisons can be derived using the consistency equations

$$\mu_{kc} = \mu_{kA} - \mu_{cA} = \mu_k - \mu_c \qquad (1)$$

We can choose any set of basic contrasts with the only constraint that all the competing treatments in the network should be included in at least one basic contrast.
Each study i $(i = 1, \dots, S)$ reports at least one treatment effect y_{ikc} that corresponds to the comparison of treatments *k* and *c*. In case that a network includes only two-arm trials (or all trials are analyzed as two-arm studies) we suppress the subscript 'kc' and we denote each observed effect size as y_i and its variance s_i^2.
We also consider that studies estimate the underlying 'true' treatment effects with random errors ε_i, which are assumed normally distributed $\varepsilon_i \sim N(0, s_i^2)$. We assume that there is heterogeneity in each pairwise comparison which is presented with the variance τ_{kc}^2. We can assume a common heterogeneity for all the comparisons in the network, for which we suppress the subscript 'kc' and we write τ^2.
In all meta-regression models the covariates representing effect modifiers are denoted with z, while a covariate x (present in some of the models that follow) represents a dummy variable corresponding to a basic comparison.
When we need to distinguish between direct and indirect estimates we use the superscripts 'dir', 'ind' respectively. We finally denote with w the difference between direct and indirect estimates in a specific loop (or between designs), which is often called inconsistency factor.

Summary of notation

Description	Symbol	Description	Symbol
index for studies	i	reference treatment in the network	A
total number of studies	S	network summary effects for the functional contrasts	μ_{kc}
total number of treatments	T	direct summary estimates for a specific comparison	μ_{kc}^{dir}
number of (direct) comparisons with available data	N	indirect summary estimates for a specific comparison	μ_{kc}^{ind}
index for basic contrasts	t	variance of the summary network treatment effects (of μ_{kc})	v_{kc}
index for any comparison in the network	kc	between-study variance (heterogeneity) for a specific comparison. This is the variance of θ_{ikc} or δ_{ikc}.	τ_{kc}^2
observed effect size in a study for a specific comparison (in a study in networks with only two-arm trials)	y_{ikc} (y_i)	common between-study variance (heterogeneity) for all comparisons.	τ^2
		random term in a study	$\delta_{ikc} = \mu_{kc} - \theta_{ikc}$ $\delta_i = \mu_{kc} - \theta_i$
variance of the observed effect size in a study for a specific comparison (in a study in networks with only two-arm trials)	s_{ikc}^2 (s_i^2)	number of independent loops in the network	L
		inconsistency factor for a specific loop	w_{kch}
underlying treatment effect in a study for a specific comparison	θ_{ikc} (θ_i)	covariate denoting a basic contrast (takes values 0,1,-1)	x_t
random error in a study	ε_i	covariate corresponding to an effect modifier	z
network summary effects for the basic contrasts	μ_t	network estimates adjusted for the effects of covariates	μ_{kc}^{adj}

3.2 Network meta-analysis as a hierarchical model

Hierarchical models for NMA offer increased flexibility in modeling the underlying assumptions and are easy to extend (see section 5) but they are also more difficult to fit using frequentist software.

We assume a normal distribution for the contrast-based study-specific outcome data

$$y_i \sim N(\theta_i, s_i^2)$$

The underlying study-specific treatment effects for each comparison are assumed either fixed and equal to the common summary effect, $\theta_i = \mu_{kc}$, or random, coming from a common distribution (usually a normal distribution), hence

$$\theta_i \sim N(\mu_{kc}, \tau_{kc}^2) \qquad (2)$$

$$\text{or} \quad \theta_i = \mu_{kc} + \delta_i$$

$$\text{with} \quad \delta_i \sim N(0, \tau_{kc}^2) \qquad (3)$$

Then, the consistency equations (1) link the summary mean effects μ_{kc} to the $T-1$ basic parameters μ_t. In the presence of multi-arm studies multivariate distributions can be used in the likelihood of (2) or (3).

An advantage of hierarchical models for NMA is that they easily enable modeling separately the outcome in in each study arm. Thus, they can incorporate several assumptions for the baseline response if needed [1].

Implementation: This model has been routinely fitted in a Bayesian framework using WinBUGS or OpenBUGS and codes for any type of data are readily available [2,3]. The new software GeMTC also uses this model. Recently, it has been implemented in SAS using the *genmod* procedure for fixed effect and the *glimmix* procedure for random effects model and codes have been provided for dichotomous data [4-6]. In a hierarchical model it is easy to account for the correlation in the observed and underlying effect sizes induced by multi-arm trials via multivariate (instead of univariate) normal distributions and to use the appropriate likelihood of the data (e.g. binomial for dichotomous data).

3.3 Network meta-analysis as a meta-regression model

This approach, presented by Lumley [7], treats the different treatment comparisons as covariates in a meta-regression model. The model without an intercept is easier to understand as the estimated regression coefficients are directly interpreted as the NMA summary effects for comparisons representing the basic contrasts. The model for a network with T treatments and two-arm studies is

$$y_i = \mu_1 x_{i1} + \cdots + \mu_{(T-1)} x_{i(T-1)} + \delta_i + \varepsilon_i$$

The covariates x_{it} can take the values -1, 0 or 1. More specifically, if study i compares treatments *A* and $t, t = 1, \ldots T-1$ then $x_{it} = 1$ and all other covariates are set equal to 0. For studies that do not include treatment *A* we rely on the consistency equations in (1). Thus, if study i compares treatment k to c, we would have $x_{ic} = 1$, $x_{ik} = -1$ and all other covariates equal 0. The values of the t covariates for all S studies in the network form what we call the design matrix $\boldsymbol{X}$**,** which is a matrix with S rows and $T-1$ columns.

Under the consistency assumption we can derive the summary treatment effects for all other non-basic (functional) contrasts using again the equation (1). For multi-arm studies, estimates for pairwise comparison are included in the model and hence the components of δ_i, ε_i that refer to the same study are correlated.

Implementation: The model can be fitted in any software able to perform meta-regression (e.g. STATA) but it requires the construction of the design matrix $\boldsymbol{X}$. The main drawback of the implementation of this model in frequentist software is that it does not easily account for the inherent correlation in multi-arm trials (i.e. it is not easy to impose the correct multivariate form on the distributions of δ_i, ε_i). Also, this approach is usually implemented after transforming the study data to obtain effect sizes that are assumed to be normally distributed, i.e. the exact likelihood for the data is not often employed.

3.4 Network meta-analysis as a multivariate meta-analysis model

This approach, first described by White et al. [8], treats the different treatment comparisons as different outcomes. The $T-1$ basic parameters μ_t represent the "outcomes" (comparisons) tA, a subset of which is presented in every study. In case there are studies not including the reference treatment A, data augmentation techniques can be used to 'impute' an A arm with minimal information. This reflects the consistency assumption which implies that the missing arm is missing at random and all comparisons in the networks can be expressed via the basic contrasts. Then, each study reports on one or more outcomes $y_{itA} := y_{it}$ and the model can be written as a multivariate meta-analysis model:

$$(y_{i1} \dots y_{i(T-1)}) = (\mu_1 \dots \mu_{(T-1)}) + (\delta_{i1} \dots \delta_{i(T-1)}) + (\varepsilon_{i1} \dots \varepsilon_{i(T-1)})$$

As in any multivariate meta-analysis model it is not necessary to have all 'outcomes' (basic contrasts) reported in all studies.

For trials comparing treatment k to c that do not involve A, data is augmented so that the study becomes a three-arm study and hence multivariate normal distributions on $(\varepsilon_{i1} \dots \varepsilon_{i(T-1)})$ and $(\delta_{i1} \dots \delta_{i(T-1)})$ are employed to incorporate the covariances between the different outcomes within each study.

Implementation: This model can be fitted in STATA using the *mvmeta* command or in WinBUGS. The main disadvantage of this approach, when is fitted in STATA, is that it does not use the exact likelihood of the data (e.g. binomial, poisson, etc.). The R function *mvmeta* and the SAS procedure *proc mixed* can also be used with some caution (the between studies variance-covariance matrix is only unrestricted, see later considerations about heterogeneity).

3.5 Network meta-analysis as a linear model using a two-stage approach

Lu et al. [9] suggested that network meta-analysis can be performed as a linear model using a two-stage approach. According to this approach, at the first stage we use pairwise meta-analysis to derive the direct pooled estimates for all comparisons with available data and at the second stage we perform meta-regression on the direct pooled estimates to derive the network estimates assuming consistency via the design matrix. The approach is described below for a network including only two-arm trials.

Let us consider that each comparison kc with available data in the network includes S_{kc} number of studies. At the first stage we perform N pairwise meta-analyses for each direct comparison to derive the direct pooled pairwise treatment effects. Hence, at the end of the first stage we have derived for each comparison $\hat{\mu}_{kc}^{dir}$ and its variance $\hat{v}_{kc}^{dir}$. To account for the between-study variance within each comparison we can use a random effects meta-analysis.

The pairwise direct estimates are then used as 'data' in a meta-regression model at the second stage of the analysis, which is similar to the model described in the Section 3.3:

$$\hat{\mu}_{kc}^{dir} = \mu_1 x_1 + \cdots + \mu_{(T-1)} x_{(T-1)} + \varepsilon_{kc}^*$$

with the only difference that here the data on the left hand side of the equation are not the observed treatment effects in the individual studies but the direct summary effects. Consequently, the random errors ε_{kc}^* relate to $\hat{v}_{kc}^{dir}$. The covariates x_t are the entries of the design matrix and represent the linear relationships between the available N direct comparisons and the basic contrasts based on the consistency equations (1).

The idea can be employed for datasets with multi-arm studies as well; in that case we would pool separately the XY and the XYZ studies to derive two different direct XY estimates. In that way the effect sizes in multi-arm studies are meta-analysed simultaneously in a model that incorporates their covariances. Then, the direct summary effects from all studies (two- and multi-arm) are pooled at the second stage to derive the network summary estimates accounting again for the correlated estimates. For more details on this approach see [9,10]. The two-stage approach is easily associated with the estimation of the Q-statistics to test the presence of inconsistency and heterogeneity in NMA (see section 4.8).

In the absence of multi-arm studies this approach can be performed in any software that performs meta-analysis and meta-regression. For a network with multi-arm studies and R routine can be used (which can be found at http://www.unimedizin-mainz.de/fileadmin/kliniken/imbei/Dokumente/Biometrie/Software/netheat.R) or by using twice the mvmeta in STATA.

3.6 Considerations for heterogeneity

Several assumptions for the comparison-specific heterogeneities τ_{kc}^2 can be employed. We review these below. In general, stronger assumptions about similarity of heterogeneity parameters make estimation more efficient; in most networks we have only a few studies per comparison so comparison-specific heterogeneity parameters are not estimated well.

We often assume that all pairwise comparisons share the same heterogeneity parameter, as was first proposed by Higgins et al. [11], hence $\tau_{kc}^2 = \tau^2$ for any treatment comparison k vs. c. The advantage of this approach is that we decrease the number of parameters to be estimated and hence we increase precision in the estimation of heterogeneity (and the treatment effects). However, some investigators claim that this might be a strong assumption. This assumption also implies that the correlation between all pairs of comparisons is 0.5. This has important implications for the structure of the between-studies variance-covariance matrix of δ_i. Available software to fit this model: WinBUGS, STATA (metareg and mvmeta), R (any meta-regression routine), SAS and GeMTC but only WinBUGS, STATA, SAS and GeMTC account properly for the multi-arm structure in the between-studies variance-covariance matrix.

We can allow for different and completely independent comparison-specific heterogeneity parameters putting no structural restrictions on the variance-covariance matrix of the random effects ($\tau_{kc}^2 \neq \tau_{hl}^2$). In the presence of comparisons with very few studies, though, this approach might not be efficient leading to very imprecise heterogeneity estimates. Moreover, it has been suggested that the consistency equations impose restrictions to the heterogeneities, so that they cannot be completely independent [12]. Available software to fit this model: WinBUGS and STATA (mvmeta). However, note that STATA can incorporate this model only in networks with available data for all possible pairwise comparisons in the network.

Lu & Ades [12] suggested that the assumption of consistency restricts the heterogeneity parameters for different comparisons. More specifically, the heterogeneity for a comparison kc is bounded by the heterogeneities of the basic parameters

$$|\tau_{kA}^2 - \tau_{cA}^2| \leq \tau_{kc}^2 \leq |\tau_{kA}^2 + \tau_{cA}^2|$$

Accounting for these constraints requires further re-parameterization of the between studies variance-covariance matrix, which makes the approach rather cumbersome. This model can be incorporated only in the hierarchical model approach (section 3.2) and fitted in WinBUGS.

3.7 Summary of the methods

	Software available	**Assumptions about heterogeneity that that can be accommodated (by software)**	**Possibility to use the exact likelihood (by software)**	**Method to estimate heterogeneity**	**Assumption of consistency…**
Hierarchical	WinBUGS or OpenBUGS, GeMTC, SAS (proc genmod & proc glimmix)	WinBUGS or OpenBUGS: any assumption GeMTC, SAS: only common heterogeneity for all comparisons	All software options	MCMC, ML	links the comparison-specific summary effects and puts constrains into their estimation
Meta-regression	WinBUGS or OpenBUGS, STATA (metareg, without multi-arm studies), R (rma from metaphor, without multi-arm studies), SAS (proc mixed)	WinBUGS or OpenBUGS: any assumption STATA, R, SAS: only common heterogeneity for all comparisons	Only WinBUGS or OpenBUGS	MCMC, REML, ML, DL, Hunter-Schmidt, Hedges, Sidik-Jonkman, Empirical Bayes	is used to derive the design matrix

Multivariate meta-analysis	WinBUGS or OpenBUGS and STATA (mvmeta), R (mvmeta), SAS (proc mixed)	WinBUGS or OpenBUGS: any assumption STATA: common or different heterogeneities across comparisons R: Only different heterogeneities SAS: only common heterogeneity for all comparisons	Only in WinBUGS or OpenBUGS	MCMC, REML, ML,	is employed to impute the 'missing' reference arm in studies
Two-stage approach	WinBUGS or OpenBUGS, STATA (metareg or mvmeta), SAS (proc mixed), self-programmed available routine in R	WinBUGS or OpenBUGS: any assumption STATA, R, SAS: only common heterogeneity for all comparisons	only WinBUGS or OpenBUGS	MCMC, REML, ML, DL, Hunter-Schmidt, Hedges, Sidik-Jonkman, Empirical Bayes but in the R routine possibly only DL	is used to derive the design matrix at the second stage of the analysis

4 Methods to evaluate statistically the consistency in a network

Some of the methods described below are related or even identical under circumstances (e.g. the Lu&Ades model with the design-by-treatment model or the Q test for heterogeneity and the X^2 test in the design-by-treatment model). Note also that little is known about the power of the tests for inconsistency.

4.1 Loop-specific approach

This is perhaps the simplest approach to estimate inconsistency in a closed loop. For example, in the triangular loop ABC, that would be $\widehat{w}_{ABC} = |\hat{\mu}_{CB}^{dir} - \hat{\mu}_{CB}^{ind}|$ with variance $var(\widehat{w}_{ABC}) = \hat{v}_{CB}^{dir} + \hat{v}_{CB}^{ind}$ where $\hat{\mu}_{CB}^{ind} = \hat{\mu}_{CA}^{dir} - \hat{\mu}_{BA}^{dir}$. Then, to infer about the statistical significance of the $\widehat{w}_{ABC}$, we can use a z-test. The method can be implemented in any software and there are readily available routines in R and STATA (see http://www.mtm.uoi.gr) that identify all triangular and quadratic (non-overlapping) closed loops of a network and estimate the ws and their variances.

The main drawbacks of the method is that indirect evidence is defined as the evidence in the loop rather than this coming from the entire network and that provides multiple correlated tests. Also, the correlation in loops including multi-arm trials is not properly accounted for. However, such correlation can be minimized by including in the estimations only two of the three available comparisons in a loop from multi-arm studies. Finally, the method can point into problematic loops but cannot infer about consistency in the entire network.

4.2 Composite test for inconsistency

The composite test for inconsistency, suggested by Caldwell et al. [13], is similar to the previous approach but considers that differences can be present also between the various indirect estimates derived from all possible independent loops in a network. The method estimates for a specific comparison kc the inverse variance weighted average $\hat{\mu}_{kc}$ of the direct summary effect $\hat{\mu}_{kc}^{dir}$ and all independent (from non-overlapping L available loops in the network) summary indirect estimates $\hat{\mu}_{kc}^{ind_1}, \dots, \hat{\mu}_{kc}^{ind_L}$:

$$\hat{\mu}_{kc} = \frac{\frac{1}{\hat{v}_{kc}^{dir}}\hat{\mu}_{kc}^{dir} + \sum_{j=1}^{L}\frac{1}{\hat{v}_{kc}^{ind_i}}\hat{\mu}_{kc}^{ind_j}}{\frac{1}{\hat{v}_{kc}^{dir}} + \sum_{j=1}^{L}\frac{1}{\hat{v}_{kc}^{ind_j}}}$$

Then the approximate χ^2-statistic $T_{kc} = \frac{1}{\hat{v}_{kc}^{dir}}\left(\hat{\mu}_{kc}^{dir} - \hat{\mu}_{kc}\right)^2 + \sum_{j=1}^{L}\frac{1}{\hat{v}_{kc}^{ind_i}}\left(\hat{\mu}_{kc}^{ind_j} - \hat{\mu}_{kc}\right)^2$ follows a χ_L^2 distribution under the null hypothesis of consistency between all independent estimates of the pairwise comparison. Using a χ^2-test we can infer about the statistical significance of inconsistency.

The method can be implemented in any software; however there are no readily available routines. This approach also fails to account for the correlation between the different estimates in the presence of multi-arm trials and can be applied as an omnibus test only for a collection of independent loops sharing a common comparison; that is in special cases of networks.

4.3 Node-splitting approach

The 'node-splitting' approach, suggested by Dias et al. [14], compares direct and indirect evidence, the latter coming from the entire network. The method excludes one direct comparison at a time and estimates the indirect treatment effect for the excluded comparison using standard NMA methods.
An equivalent approach is the 'back-calculation' method [14]. This approach estimates the direct and network treatment effect estimates using the complete dataset. Then considering that the network estimates are a weighted average of the direct and all indirect estimates extracts the indirect estimates which are estimated as

$$\hat{\mu}_{kc}^{ind} = \left(\frac{\hat{\mu}_{kc}}{\hat{v}_{kc}} - \frac{\hat{\mu}_{kc}^{dir}}{\hat{v}_{kc}^{dir}}\right)\hat{v}_{kc}^{ind} \quad \text{and} \quad \frac{1}{\hat{v}_{kc}^{ind}} = \frac{1}{\hat{v}_{kc}} - \frac{1}{\hat{v}_{kc}^{dir}}$$

Then we obtain the difference between the direct and indirect estimates (w) and using a z-test we can infer about the statistical significance of inconsistency (or we can compare the posterior distributions of the direct and indirect effect If the method is implemented in a Bayesian setting).

The methods can be fitted in WinBUGS (code can be found online from Bristol University) and the 'node-splitting' approach has been implemented also in the GeMTC software.
As with the two previous approaches, the proper modeling of multi-arm trials is an unsolved issue. The methods can point out into comparisons that are associated with inconsistency but do not infer about the entire network.

4.4 Lu & Ades model

Lu & Ades [15] suggested a NMA model that accounts for the possible inconsistency by 'relaxing' the consistency equations. This approach adds an extra term in the usual consistency equations to reflect the possible disagreement between the parameters of the estimated summary effects (μ) in a loop. For example, for the ABC the consistency equation would be:

$$\mu_{CB} = \mu_{CA} - \mu_{BA} + w_{ABC}$$

The parameters w_{kch} can be assumed independent across loops. To increase power they can also be assumed exchangeable following a normal distribution $w_{kch} \sim N(0, \sigma^2)$ with the variance σ^2 being called the 'inconsistency variance'. The rest of the model is the same as in standard NMA. A large w suggests important inconsistency in the respective loop and comparing σ^2 with τ^2 can help inference about the 'extra' variability due to inconsistency.

The main drawback of this approach is that the results are sensitive to the parameterization of multi-arm trials. The model can be fit in STATA and WinBUGS and has been implemented also in the GeMTC software. In STATA and WinBUGS the addition of the inconsistency factors to the consistency equations should be made by hand, whereas in GeMTC is done automatically.

4.5 Design-by-treatment interaction model

An extension of this model is the design-by-treatment interaction model, suggested by Higgins et al. [16] This approach accounts also for a different source of inconsistency (known as design inconsistency) that can arise between estimates of the same comparison coming from studies with different designs (e.g. two-arm vs. three-arm trials). The previous model (by Lu & Ades) is a special case of the design-by-treatment model and the two models are identical for networks that do not have multi-arm studies. Inference about the presence of inconsistency in the entire network is made using a χ^2 test to assess jointly all w parameters. The w factors can be assumed fixed and independent or random as above and drown from $N(0, \sigma^2)$. The random inconsistency model has an elegant analogy to the random-effects model.
This model can be fitted in STATA (fixed *w*) or WinBUGS (both fixed and random *w*).

In the presence of multi-arm trials, parameterization for the inconsistency factors can be difficult particularly for complex and large networks, as discrimination between loop and design inconsistency is not always straightforward. However, a STATA routine is being prepared to automate the process.

The advantage of the approach is that the different possible parameterizations of multi-arm studies do not affect the inference about the presence of inconsistency in the entire network (based on the χ^2-test).

4.6 Comparison of model fit and parsimony between consistency and inconsistency model

This approach is specific to models fitted within a Bayesian framework and employs the measures of model fit and parsimony (e.g. DIC, residual deviance) to check the plausibility of the consistency assumption. We first apply both the consistency and the inconsistency models (the model in which the consistency equations have been omitted). Then, we compare the model fit and parsimony between the two models and if the inconsistency model appears to fit the data better and be more parsimonious this is an indication that the consistency assumption might not hold.

Note that this is a tool to infer only about the inconsistency in the entire network and not in each specific loop.

4.7 Net-Heat matrix

Krahn et al. [10] extended the two-stage NMA model associated with the second stage of the analysis to evaluate inconsistency. This extended model includes an additional parameter per comparison that represents the difference between direct and indirect estimates for that comparison. This model can be seen as a version of the design-by treatment model. Krahn et al suggested comparing the residual inconsistency of the extended model with the residual inconsistency of the initial (non-extended model) to reveal whether a specific comparison is a 'hot-spot' of inconsistency.

4.8 Q-statistics in NMA

The two-stage approach to fit NMA enables the derivation of Q-statistics for heterogeneity and inconsistency analogous to the Q-statistic for heterogeneity in simple meta-analysis. More specifically, at the first stage of the analysis we estimate for each comparison kc the usual Q-statistic from all S_{kc} studies that report it

$$Q_{kc} = \sum_{i=1}^{S_{kc}} \frac{1}{s_i^2} \left(y_i - \hat{\mu}_{kc}^{dir}\right)^2 \sim \chi^2_{S_{kc}-1}$$

which represents the distances between the observations in studies y_i and the pooled direct estimates $\hat{\mu}_{kc}^{dir}$ for every direct comparison kc. According to Krahn et al. [10] the sum of all the within-comparisons Q_{kc}s is the Q for heterogeneity in a network meta-analysis (Q^{het}) and follows a χ^2 distribution with $S - N$ degrees of freedom.

A similar Q-statistic can be estimated for the meta-regression model at the second stage of the analysis as

$$Q^{inc} = \left[\left(\hat{\mu}_1^{dir} \cdots \hat{\mu}_N^{dir}\right) - \left(\hat{\mu}_1 \cdots \hat{\mu}_N\right)\right] \begin{pmatrix} 1/\hat{v}_1^{dir} & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & 1/\hat{v}_N^{dir} \end{pmatrix} \left[\begin{pmatrix} \hat{\mu}_1^{dir} \\ \vdots \\ \hat{\mu}_N^{dir} \end{pmatrix} - \begin{pmatrix} \hat{\mu}_1 \\ \vdots \\ \hat{\mu}_N \end{pmatrix}\right]$$

and represents the distances between the direct summary estimates $\hat{\mu}_{kc}^{dir}$ and the network summary estimates $\hat{\mu}_{kc}$ for each comparison kc. Under the null hypothesis of consistency the Q^{inc} follows a χ^2 distribution with $N - T + 1$ degrees of freedom.

Krahn et al. [10] further suggested that the total Q-statistic for both stages of the analysis is the sum of the Q-statisics for heterogeneity and inconsistency ($Q^{tot} = Q^{het} + Q^{inc}$) and under the null hypothesis of consistency and homogeneity in the network follows a χ^2 distribution with $S - T$ degrees of freedom.

Note that all the above Q-statistics have been extended to account for the inclusion of multi-arm trials. For details on these extensions see [9,10]. These Q-statistics can be estimated using a routine in R available from http://www.unimedizin-mainz.de/fileadmin/kliniken/imbei/Dokumente/Biometrie/Software/netheat.R

4.9 I^2 measure for NMA

Jackson et al. [17] suggest in their paper (under revision) the use of the I^2 measure for heterogeneity and inconsistency in NMA using a definition previously suggested for multivariate meta-analysis [18].

4.10 Multidimensional scaling

Chung & Lumley [19] proposed a graphical method and the use of multidimensional scaling for the evaluation of inconsistency. Their approach considers that the direct pairwise summary effects are the 'observed dissimilarities' between the treatments and uses them to construct a dissimilarity matrix. Then, it employs weighted multidimensional scaling techniques to estimate the 'fitted dissimilarities'. Comparing the observed and the fitted dissimilarities we can infer about the presence of inconsistency; discrepancies imply possible important inconsistency.

4.11 Comparison of methods

	Identifies spots of inconsistency	**Can infer about the entire network**	**Software**	**Sensitive to parameterization of 3-arm studies**
Loop-specific approach	Yes	No	WinBUGS or OpenBUGS, SAS, readily available routines in STATA and R	Yes
Composite test	Yes	No	WinBUGS or OpenBUGS, STATA, SAS, R (no routines readily available)	Yes
Node-splitting & back-calculation	Yes	No	available code in WinBUGS or OpenBUGS, but also STATA, SAS, R, GeMTC (node-splitting)	Yes
Lu & Ades model	Yes	Yes	WinBUGS or OpenBUGS, STATA, GeMTC, possibly SAS	Yes
Design-by-treatment	Yes	Yes	WinBUGS or OpenBUGS and STATA	No
Comparison of model fit and parsimony	No	Yes	WinBUGS or OpenBUGS, STATA, SAS, R	Yes
Net-heat matrix	Yes	No	Could be done in WinBUGS or OpenBUGS, STATA, SAS, but needs self-programming, there is an available routine in R	No
Q	No	Yes	WinBUGS or OpenBUGS, STATA, SAS but needs self-programming, available routine	No

			in R	
I2	No	Yes	WinBUGS or OpenBUGS, STATA, R, SAS but needs self-programming	No
Multidimensional scaling	Yes	No	STATA, R, SAS but needs self-programming	No

5 Network meta-regression

Extensions of NMA models to account for important effect modifiers have been presented in the literature.

The meta-regression model used to perform NMA can be extended to incorporate additional covariates (on the top of the basic contrasts) that might impact on the treatment effect estimates. Such an extended model would be

$$y_i = \mu_1^{adj} x_{i1} + \cdots + \mu_{(T-1)}^{adj} x_{i(T-1)} + \beta_1 z_{i1} + \cdots + \beta_p z_{ip} + \delta_i + \varepsilon_i$$

where the covariates $z_{i1}, \dots, z_{ip}$ are the possible effect modifiers and the coefficients $\beta_1, \dots, \beta_p$ show how much these covariates differentiate the estimated treatment effects. In this model the estimated effects $\hat{\mu}_1^{adj}, \dots, \hat{\mu}_{(T-1)}^{adj}$ are 'adjusted' for the effect of the p covariates.

Similarly, the hierarchical NMA model can be extended to account for the impact of possible effect modifiers and can be fitted in WinBUGS and possibly in SAS using the *glimmix* procedure but no codes have been provided in SAS. Covariates can be included also when NMA is fitted as a multivariate meta-analysis model and can be fitted in STATA using the *mvmeta* command and in WinBUGS.

Note that the model above assumes that the effect modifiers have the same effect on all comparisons. We can assume comparison-specific coefficients or consistent coefficients within comparisons (see Cooper et al. [20] for a detailed description) . Note however that only models fitted within WinBUGS offer maximum flexibility regarding assumptions for the coefficients.

6 Challenging issues in network meta-analysis

6.1 Choice of the appropriate effect measure

The measure of analysis is not necessarily the same as the measure of presentation of results from meta-analysis. The different effect measures have different properties and their interpretability varies e.g. risk difference is well understood but has poor mathematical properties. Therefore, reviewers are encouraged to undertake the analysis using an appropriate measure and then present the result transformed into a scale that is easy to understand.

A recent publication by Norton et al. [21] shows that the three different effect measures, commonly used for dichotomous data (OR, RR, RD), can result in different rankings for the treatments in an indirect comparison if the baseline risk is not constant across studies. If there is no particular reason to choose OR, the authors recommend to use either the RR or the RD and then to perform a sensitivity analysis. However, they don't give a clear justification for this preference.

In their response to this paper, van Valkenhoef and Ades [22] highlight the fact that the phenomenon of rank reversal is a consequence of the underlying assumptions in any indirect comparison. The exchangeability assumption and the additivity of treatment effects cannot hold simultaneously for all three measures. Thus, the choice of the analysis measure should be based not on convenience and interpretation criteria but on the appropriateness of each measure for the respective data. Extreme variation in the baseline risk across studies that would lead to the rank reversal phenomenon is a sign that the exchangeability assumption might not be plausible.

The ideas in van Valkenhoef and Ades were first described in Caldwell et al. [23] where it was shown that the choice of scale can impact strongly on the results and should be based on scientific grounds such as heterogeneity and goodness-of-fit measures (e.g. DIC, residual diagnostics, etc.). They also suggest that OR occasionally gives larger treatment effects and can be misinterpreted while for time-to-event data HR should be given greater consideration. Finally, they note that the scale of analysis is a different issue than the scale of reporting.

An empirical study by Veroniki et al. [24] showed that there are *a priori* no important differences between the different measures in terms of inconsistency when estimated using the loop-based approach or the design-by-treatment model.

Earlier studies [25,26] had suggested that OR might be preferable for network meta-analysis compared to RR. It is known that RRb (beneficial) and RRh (harmful) can result in different effect estimates in magnitude and precision but consistent in direction. It was shown that in indirect comparisons RRb and RRh can also give relative treatment effects with different direction (favoring different treatments).

6.2 Stability of ranking probabilities

There is agreement that the ranking measures are useful but they should be presented only in the context of the estimated effect sizes. It has been previously suggested that ranking should be based on cumulative probabilities rather than the 'probability of being the best' as the latter ignores uncertainty in ranking [27]. Mills et al. [28] explored the effect of excluding treatments in a collection of 18 network meta-analyses and found that the effect sizes and ranking are conditional on the number of treatments and trials of the network.

Jonas et al. [29] also used two different networks and compared the ranking probabilities (derived from Bayesian NMA) in the complete datasets and after excluding parts of the datasets to create sub-networks with specific patterns (e.g. star network, ladder, etc.). This study also found differences in the treatment order across the different network patterns.

7 Presentation options in network meta-analysis

7.1 Presentation of the evidence base

Diagram showing the comparisons in the individual studies of a network.
Example in Hoaglin et al.[30] (Supplementary material)
The vertical axis of the graph shows a list of the studies and the horizontal axis presents the three competing treatments in the network (BMS, PES, SES). Each horizontal line connects the (two) treatments compared in each study. The parentheses report the total number of patients for each trial arm.

Network graph showing all the available comparisons in the network.
Example in Hoaglin et al.[30] (Supplementary material)
Each node in the graph represents an intervention and each edge a direct comparison between two treatments.

Network graph showing all the available comparisons in the network and the available number of trials for each comparison as well as the presence of multi-arm trials.
Example in Lu et al.[9] (Figure 2)
This network graph separates the direct comparisons derived from two-arm studies (solid lines) from those derived from multi-arm studies (dashed lines). The number next to the edges show the total number of trials (two- and multi-arm) in each comparison.

Network graph showing all the available comparisons in the network and the available number of trials for each comparison using weighted nodes and edges.
Example in Salanti et al.[31] (Figure 1)
In this network graph the size of the nodes and edges is proportional to the number of studies including each treatment and comparison respectively.

Network graph showing all the available comparisons in the network using weighted nodes and edges and the risk of bias level (for a specific item) for each comparison using colored edges.
Example in Chaimani et al. [32] (Figure 2)
In this network graph the size of the nodes and edges is proportional to the number of studies including each treatment and comparison respectively. The green, yellow and red (not present in this graph) edges represent

those comparisons in which most trials are of low, unclear and high risk of bias respectively with respect to the adequacy of allocation concealment.

Table showing the network structure; the available study designs in the network.
Example in Lu et al.[9] (Figure 2)
The table below the number of studies including a specific group of treatment arms (denoted with X)

Matrix showing the available direct comparisons in the network.
Example in Ioannidis [33]
The matrix shows the number of available direct comparisons between two treatments (drugs with different doses are grouped together)

Graph showing the data provided by the individual studies of the network.
Example in Lu & Ades [15] (Figure 1)
In this graph the horizontal axis presents the four competing treatments in the network and the vertical axis includes the observed log-odds values in studies for each trial arm. Each line represents a study and connects the treatments that compares. The two blue thick lines correspond to three-arm trials. Since this network has a beneficial outcome, lines with positive slope favor the intervention on the right and lines with negative slope favor the intervention on the left.

Graph showing the contribution of each direct comparison in the network estimates.
Example in Chaimani et al.[32] (Figure 3)
The graph shows the percentage contribution of each direct comparison in each network meta-analysis estimate and in the entire network along with the number of included studies in each comparison. More specifically, the network presented in this graph includes 4 treatments and studies form 4 direct comparisons. Each value in the upper part of the graph (with title 'Network meta-analysis estimates' in the vertical axis) is the percentage contribution of the direct estimate (in column) to the network estimate (in row). For example, the comparison BMS vs. MT is informed 55.7% by its own direct estimate and 22.2% by each of the comparisons BMS vs. PTCA and MT vs. PTCA. The second part of the graph ('Entire network') shows the percentage contribution of the 4 direct comparisons in the entire network. The squares are proportional to the percentage contributions.

7.2 Presentation of results

Forest plot with the treatment effects for all pairwise comparisons including the network estimates, the direct estimates and the individual studies' estimates.
Example in Hawkins et al.[34] (Figure 3)

Forest plot with the network estimates for all treatments compared to the reference.
Example in Hoaglin et al.[30] (Supplementary material)

'Hsu mean-mean plot' showing all the available pairwise comparisons in the network and the respective network estimates with the 95% CI.
Example in Senn et al.[35] (Figure 6)
In this graph placebo is taken as a reference treatment and arbitrarily given the value 0. All other treatments are expressed as differences to placebo and plotted twice: once on the X axis and once on the Y axis. Each of the 45 possible pairwise contrasts is shown as a point in the X,Y plane with its 95% confidence limits which are joined to the point estimates by diagonal lines. The magnitude of a relative treatment effect is the distance of each point from the dotted diagonal line rising from bottom left to top right. Solid lines represent the significant contrasts and dashed lines the non-significant contrasts.

Shade plot showing the level of significance of the treatment effects for all available pairwise comparisons in the network.
Example in Senn et al.[35] (Figure 4)

'*Bubble-plot' including the ranking probabilities for all treatments.*
Example in Hawkins et al.[34] (Figure 4)
The shaded circles in the graph are proportional to the probability for each treatment to be at a specific rank in treatment hierarchy.

'Rankograms' showing the probability for each treatment to be at a specific rank in the treatment hierarchy.
Example in Salanti et al.[36] (Figure 3)
The horizontal axis in the graphs includes the five possible ranks and the vertical axis the probability of a treatment to achieve a rank.

Bar plots showing the probability for each treatment to be at a specific rank in the treatment hierarchy.
Example in van Valkenhoef et al. [37] (Figure 4)
The horizontal axis in the graph includes the five competing treatments in the network. The size of each bar corresponds to the probability of each treatment to be at a specific rank in treatment hierarchy.

'SUCRA plots' showing the cumulative probability for each treatment to be up to a specific rank in the treatment hierarchy.
Example in Salanti et al.[27] (Figure 5)
The horizontal axis in the graphs includes the five possible ranks and the vertical axis the cumulative probability of a treatment to be up to a specific rank. The larger the surface under the cumulative ranking curve, the more effective/safer the treatment.

'SUCRA plots' comparing the ranking results from different models.

Example in Salanti et al.[38] (Figure 3)

Each line pattern in the graphs corresponds to a specific model.

Scatterplot combining the ranking results for two different outcomes that have been analyzed separately.

Example in Chaimani et al. [32] (Figure 9)

The horizontal axis includes the SUCRA values for efficacy and the vertical axis the SUCRA values for acceptability. The 14 competing treatments have been separated into 5 groups using clustering methods according to their performance on both outcomes. The treatments that perform well on both outcomes are those plotted in the upper right corner.

Network graph presenting the relative treatment effects for each pairwise comparison in the network.

Example in Fadda et al. [39]

In the graph the solid lines represent the direct comparisons and the dotted lines the indirect comparisons. The symbols '+' and '−' show the favored and non-favored treatment respectively in each comparison according to the relative treatment effects. For comparisons with no statistically significant treatment effects we replace these symbols with '=t' and '=' respectively.

Table showing all the pairwise relative treatment effects with their 95% CI for one or two outcomes.

Example in Cipriani et al. [40] (Figure 4)

Each cell in the table includes the OR (95% CI) for the comparison of the treatment in the respective column vs. the treatment in the respective row. The gray cells correspond to the dropout rate outcome and the light blue cells to the efficacy outcome. The diagonal shows the names of the competing treatments.

7.3 Presentation of inconsistency in the network

Forest plot with inconsistency factors estimated by the loop-specific approach for inconsistency

Example in Salanti et al. [31] (Figure 3)

The squares in the graph represent the inconsistency factors for every closed loop and the horizontal lines the 95% CI. If the 95% CI line does not cross the zero line for a specific loop, there might be important inconsistency in the this loop. For example, this graph implies that there is no loop in the network with statistically significant inconsistency. However, the loops NGR and NRV show a quite large difference between direct and indirect estimates. This implies that these two loops might be possible sources of important inconsistency that needs exploration.

'Net-heat plot': a plot showing the contribution of the direct estimates to the network estimates and the change in inconsistency when detaching one design from the network

Example in Krahn et al.[10] (Figure 5)

The colors are associated with the change in inconsistency between direct and indirect evidence in the comparison shown in the row after detaching the effect of the comparison shown in the column. Blue colors indicate an increase and warm colors indicate a decrease (the stronger the intensity of the color, the stronger the change). In this way comparisons corresponding to warm colors (mostly orange and red) in the off-diagonal elements are possible important sources of inconsistency (for example the comparison metl vs. SUal). The comparison-specific inconsistencies are the summands in the Q^{inc} (see section 4.8) corresponding to each comparison. (see also the '*Graph showing the contribution of each direct comparison in the network estimates*').

Density plot comparing the posterior densities of the treatment effect for a specific comparison between network, direct and indirect estimates.

Example in Dias et al. [14] (Figure 4)

The dotted line is the density of the log-odds ratio for a specific treatment comparison derived from the full NMA model (including all available data), the solid line is the density for the same comparison derived only from direct comparisons and the dashed line represents the density derived from the indirect evidence (the comparison is excluded from the network). There is an obvious disagreement between the sources of evidence.

7.4 Presentation of heterogeneity in the network

Graph showing the network summary estimates with their CI/CrI and the predictive interval for all comparisons.

Example in Chaimani et al.[32] (Figure 6)

The black horizontal lines represent the CI of the network summary estimates in odds ratio scale and the red lines the predictive intervals. In this example a common heterogeneity for all comparisons has been assumed.

8 What we expect to see in a Cochrane protocol and review

Several papers have been published that make recommendations regarding the reporting of NMA [30,41-47]. According to these publications, any NMA should follow the conventional reporting guidelines for systematic reviews and pairwise meta-analyses plus a set of additional items specific to the context of NMA. The various publications appear to agree in a number of items that need to be reported. Researchers should clearly define the research question and the search strategy including the specification of the target population, inclusion criteria for the individual trials, the competing treatments and the outcomes of interest. A discussion on the similarity of patient populations and study designs within and across treatment

comparisons is important to infer about the plausibility of the homogeneity and transitivity assumptions, while the evaluation of the validity of results requires assessment of the level of credibility of the individual studies. Researchers are also encouraged to describe the methods they used to evaluate the statistical heterogeneity and inconsistency. Presentation of the analysis should include a detailed specification of the model, a statement about modeling multi-arm trials, measures of model fit, the assumption employed for the heterogeneity (see section 3.6), the software used and the methods they employed to explain important heterogeneity and inconsistency, if present. Finally, it is advisable to present both the direct and network treatment effect estimates for the pairwise comparisons of interest, the relative ranking results and results from subgroup analyses or network meta-regression, if such analyses have been applied.

Two initiatives are currently working towards this direction and their output is based on (narrow or wider) consensus:

1. An attempt to extend the PRISMA (Preferred Reporting Items for Systematic reviews and Meta-Analysis) statement [48] is in progress and preliminary results of this initiative will be distributed to the meeting attendants.
2. The AMCP/NPC/ISPOR Comparative Effectiveness Research Collaborative Initiative has developed an instrument to assess the relevance and credibility of an indirect treatment comparison or network meta-analysis (unpublished work). The developed assessment tool consists of 6 categories with a total of 26 questions related to the relevance and credibility of an indirect comparison or network meta-analysis. Although this is not a reporting guideline, it highlights the methodological components associated with the credibility of the NMA results. The tool will be discussed in the meeting.

9 List of Appendices

Appendix 1: References to methodological papers

Appendix 2: Glossary

Reference List

1. Dias S, Welton NJ, Sutton AJ, Ades AE: **NICE DSU Technical Support Document 5: Evidence synthesis in the baseline natural history model.** *available from http://www.nicedsu.org.uk* 2011.

2. Dias S, Sutton AJ, Ades AE, Welton NJ: **A Generalized Linear Modeling Framework for Pairwise and Network Meta-analysis of Randomized Controlled Trials.** *Med Decis Making* 2012.

3. Dias S, Sutton AJ, Ades AE, Welton NJ: Evidence Synthesis for Decision Making 2: **A Generalized Linear Modeling Framework for Pairwise and Network Meta-analysis of Randomized Controlled Trials.** *Med Decis Making* 2013, **33**: 607-617.

4. Carlin BP, Hong H, Shamliyan TA, Sainfort F, Kane RL: **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* 2013, **Publication No. 12(13)-EHC103-EF. Rockville, MD: Agency for Healthcare Research and Quality**.

5. Hong H, Carlin BP, Shamliyan TA, Wyman JF, Ramakrishnan R, Sainfort F *et al.*: **Comparing Bayesian and Frequentist Approaches for Multiple Outcome Mixed Treatment Comparisons.** *Med Decis Making* 2013, **33**: 702-714.

6. Jones B, Roger J, Lane PW, Lawton A, Fletcher C, Cappelleri JC *et al.*: **Statistical approaches for conducting network meta-analysis in drug development.** *Pharm Stat* 2011, **10:** 523-531.

7. Lumley T: **Network meta-analysis for indirect treatment comparisons.** *Stat Med* 2002, **21:** 2313-2324.

8. White IR, Barrett JK, Jackson D, Higgins JPT: **Consistency and inconsistency in network meta-analysis: model estimation using multivariate meta-regression.** *Res Synth Meth* 2012, **3:** 111-125.

9. Lu G, Welton NJ, Higgins JPT, White IR, Ades AE: **Linear inference for mixed treatment comparison meta-analysis: A two-stage approach.** *Res Synth Meth* 2011, **2:** 43-60.

10. Krahn U, Binder H, Konig J: **A graphical tool for locating inconsistency in network meta-analyses.** *BMC Med Res Methodol* 2013, **13:** 35.

11. Higgins JP, Whitehead A: **Borrowing strength from external trials in a meta-analysis.** *Stat Med* 1996, **15:** 2733-2749.

12. Lu G, Ades A: **Modeling between-trial variance structure in mixed treatment comparisons.** *Biostatistics* 2009, **10:** 792-805.

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

14. Dias S, Welton NJ, Caldwell DM, Ades AE: **Checking consistency in mixed treatment comparison meta-analysis.** *Stat Med* 2010, **29:** 932-944.

15. Lu G, Ades AE: **Assessing evidence inconsistency in mixed treatment comparisons.** *J Amer Statist Assoc* 2006, **101:** 447-459.

16. Higgins JPT, Jackson D, Barrett JK, Lu G, Ades AE, White IR: **Consistency and insconsistency in network meta-analysis: concepts and models for multi-arm studies .** *Res Synth Meth* 2012, **3:** 98-110.

17. Jackson D, Barrett JK, Stephen R, White IR, Higgins JPT: **A design-by-treatment interaction model for network meta-analysis with random inconsistency effects.** [under revision].

18. Jackson D, White IR, Riley RD: **Quantifying the impact of between-study heterogeneity in multivariate meta-analyses.** *Stat Med* 2012,**31**: 3805-3820.

19. Chung H, Lumley T: **Graphical exploration of network meta-analysis data: the use of multidimensional scaling.** *Clin Trials* 2008, **5:** 301-307.

20. Cooper NJ, Sutton AJ, Morris D, Ades AE, Welton NJ: **Addressing between-study heterogeneity and inconsistency in mixed treatment comparisons: Application to stroke prevention treatments in individuals with non-rheumatic atrial fibrillation.** *Stat Med* 2009, **28:** 1861-1881.

21. Norton EC, Miller MM, Wang JJ, Coyne K, Kleinman LC: **Rank reversal in indirect comparisons.** *Value Health* 2012, **15:** 1137-1140.

22. van VG, Ades AE: **Evidence synthesis assumes additivity on the scale of measurement: response to "Rank reversal in indirect comparisons" by Norton et al.** *Value Health* 2013, **16:** 449-451.

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

24. Veroniki AA, Vasiliadis HS, Higgins JP, Salanti G: **Evaluation of inconsistency in networks of interventions.** *Int J Epidemiol* 2013, **42:** 332-345.

25. Coory M, Jordan S: **Frequency of treatment-effect modification affecting indirect comparisons: a systematic review.** *Pharmacoeconomics* 2010, **28:** 723-732.

26. Eckermann S, Coory M, Willan AR: **Indirect comparison: relative risk fallacies and odds solution.** *J Clin Epidemiol* 2009, **62:** 1031-1036.

27. Salanti G, Ades AE, Ioannidis JP: **Graphical methods and numerical summaries for presenting results from multiple-treatment meta-analysis: an overview and tutorial.** *J Clin Epidemiol* 2011, **64:** 163-171.

28. Mills EJ, Katers S, Thorlund K, Chaimani A, Veroniki AA, Ioannidis JPA: **The effects of excluding treatments from network meta-analysis.** *BMJ* 2013 [to appear].

29. Jonas DE, Wilkins TM, Bangdiwala S, Bann CM, Morgan LC, Thaler KJ *et al*.: **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* 2013, **Publication No. 13-EHC039-EF. Rockville, MD: Agency for Healthcare Research and Quality**.

30. Hoaglin DC, Hawkins N, Jansen JP, Scott DA, Itzler R, Cappelleri JC *et al*.: **Conducting indirect-treatment-comparison and network-meta-analysis studies: report of the ISPOR Task Force on**

Indirect Treatment Comparisons Good Research Practices: part 2. *Value Health* 2011, **14:** 429-437.

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

32. Chaimani A, Higgins JPT, Mavridis D, Spyridonos P, Salanti G: **Graphical tools for network meta-analysis in STATA.** *Plos One* 2013 [to appear].

33. Ioannidis JP: **Indirect comparisons: the mesh and mess of clinical trials.** *Lancet* 2006, **368:** 1470-1472.

34. Hawkins N, Scott DA, Woods BS, Thatcher N: **No study left behind: a network meta-analysis in non-small-cell lung cancer demonstrating the importance of considering all relevant data.** *Value Health* 2009, **12:** 996-1003.

35. Senn S, Gavini F, Magrez D, Scheen A: **Issues in performing a network meta-analysis.** *Stat Methods Med Res* 2013, **22:** 169-189.

36. Salanti G, Ades AE, Ioannidis JP: **Graphical methods and numerical summaries for presenting results from multiple-treatment meta-analysis: an overview and tutorial.** *J Clin Epidemiol* 2011, **64:** 163-171.

37. van VG, Tervonen T, Zhao J, de BB, Hillege HL, Postmus D: **Multicriteria benefit-risk assessment using network meta-analysis.** *J Clin Epidemiol* 2012, **65:** 394-403.

38. Salanti G, Dias S, Welton NJ, Ades AE, Golfinopoulos V, Kyrgiou M *et al*.: **Evaluating novel agent effects in multiple-treatments meta-regression.** *Stat Med* 2010, **29:** 2369-2383.

39. Fadda V, Maratea D, Trippoli S, Messori A: **Network meta-analysis: Results can be summarised in a simple figure.** *BMJ* 2011, **342:** d1555.

40. Cipriani A, Barbui C, Salanti G, Rendell J, Brown R, Stockton S *et al*.: **Comparative efficacy and acceptability of antimanic drugs in acute mania: a multiple-treatments meta-analysis.** *Lancet* 2011, **378:** 1306-1315.

41. Jansen JP, Fleurence R, Devine B, Itzler R, Barrett A, Hawkins N *et al*.: **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* 2011, **14:** 417-428.

42. Ades AE, Caldwell DM, Reken S, Welton NJ, Sutton AJ, Dias S: **NICE DSU Technical Support Document 7: Evidence synthesis of treatment efficacy in decision making: A reviewer's checklist.** *available from http://www.nicedsu.org.uk,* 2012.

43. Salanti G: **Indirect and mixed-treatment comparison, network, or multiple-treatments meta-analysis: many names, many benefits, many concerns for the next generation evidence synthesis tool.** *Res Synth Meth* 2012, **3:** 80-97.

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

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

46. Jansen JP, Crawford B, Bergman G, Stam W: **Bayesian meta-analysis of multiple treatment comparisons: an introduction to mixed treatment comparisons.** *Value Health* 2008, **11:** 956-964.

47. Li T, Puhan MA, Vedula SS, Singh S, Dickersin K: **Network meta-analysis-highly attractive but more methodological research is needed.** *BMC Med* 2011, **9:** 79.

48. Moher D, Liberati A, Tetzlaff J, Altman DG: **Preferred reporting items for systematic reviews and meta-analyses: the PRISMA statement.** *Int J Surg* 2010, **8:** 336-341.