Statistical Methods in Medical Research
http://smm.sagepub.com/

A Bayesian network meta-analysis for binary outcome: how to do it

Teresa Greco, Giovanni Landoni, Giuseppe Biondi-Zoccai, Fabrizio D'Ascenzo and Alberto Zangrillo

Stat Methods Med Res published online 22 August 2013
DOI: 10.1177/0962280213500185

Original Article

SMMR
STATISTICAL METHODS IN MEDICAL RESEARCH

Statistical Methods in Medical Research
0(0) 1–17
© The Author(s) 2013
Reprints and permissions:
sagepub.co.uk/journalsPermissions.nav
DOI: 10.1177/0962280213500185
smm.sagepub.com

A Bayesian network meta-analysis for binary outcome: how to do it

Teresa Greco,[1,2] Giovanni Landoni,[1]
Giuseppe Biondi-Zoccai,[3,4] Fabrizio D'Ascenzo[4,5]
and Alberto Zangrillo[1]

Abstract

This study presents an overview of conceptual and practical issues of a network meta-analysis (NMA), particularly focusing on its application to randomised controlled trials with a binary outcome of interest. We start from general considerations on NMA to specifically appraise how to collect study data, structure the analytical network and specify the requirements for different models and parameter interpretations, with the ultimate goal of providing physicians and clinician-investigators a practical tool to understand pros and cons of NMA. Specifically, we outline the key steps, from the literature search to sensitivity analysis, necessary to perform a valid NMA of binomial data, exploiting Markov Chain Monte Carlo approaches. We also apply this analytical approach to a case study on the beneficial effects of volatile agents compared to total intravenous anaesthetics for surgery to further clarify the statistical details of the models, diagnostics and computations. Finally, datasets and models for the freeware WinBUGS package are presented for the anaesthetic agent example.

Keywords

anaesthetic agents, Bayesian, binary outcomes, hierarchical models, mixed treatment comparison, network meta-analysis, WinBUGS

1 Introduction

The search for accurate and reliable sources of evidence represents an ongoing challenge in medicine, as only a comprehensive yet synthetic analytical effort can accurately guide clinical decision making. Any single empirical observation on the apparent relationship between events and exposures or between events and interventions may provide useful information,[1] but systematic reviews and meta-analyses of large high quality randomised controlled trials (RCT) with low heterogeneity represent

[1] Anaesthesia and Intensive Care Department, San Raffaele Scientific Institute, Milan, Italy
[2] Section of Medical Statistics and Biometry Giulio A. Maccacaro, Department of Occupational and Environmental Health, University of Milan, Milan, Italy
[3] Department of Medico-Surgical Sciences and Biotechnologies, Sapienza University of Rome, Italy
[4] Meta-analysis and Evidence based medicine Training in Cardiology (METCARDIO), Italy
[5] Division of Cardiology, Department of Internal Medicine, Città Della Salute e Della Scienza, Turin, Italy

Corresponding author:
Teresa Greco, Anaesthesia and Intensive Care Department, San Raffaele Scientific Institute, Via Olgettina 60, Milan 20132, Italy.
Email: greco.teresa@hotmail.it

the highest degree of evidence, as they offer increased precision and external validity.[2] Moreover, meta-analyses offer a quick and cost-effective method to gather information for clinical decision making. When head-to-head treatment comparisons are not available or conclusive, the limitations of standard (i.e. pairwise) meta-analyses can be overcome by network meta-analysis (NMA, i.e. mixed treatment comparisons (MTC)), which can provide estimates of treatment efficacy or safety of multiple treatment regimens. Different treatment strategies are analysed by statistical inference methods rather than simply summing up trials that evaluated the same intervention compared to another intervention, standard care or placebo. The classic example is the following: if a first trial compares drug A to drug B, showing that drug A is significantly superior to drug B, and a second trial investigates the same or a similar patient population comparing drug B versus drug C (demonstrating that drug B is equivalent to drug C), NMA may allow to infer that drug A is also potentially superior to drug C for this given patient population, even though there was no direct test of drug A against drug C. Specifically, if two particular treatments have never been compared against each other but have been compared to a common comparator, then an adjusted indirect treatment comparison (ITC) can exploit the direct effects of the two treatments versus the common comparator to estimate the indirect treatment effect.[3–5] In such perspective, ITC represents the simplest type of NMA or MTC. In addition to this, it is important to understand that both direct and indirect information provide data for evidence synthesis, and thus any NMA is inherently more efficient and accurate. Nonetheless, NMA, as pairwise meta-analysis, can be accurate and clinically useful only when it combines studies that are similar enough to be grouped, with the aim to explore and limit as much as possible the sources of variability while concomitantly maximising the statistical precision. The results obtained from the combination of direct and indirect estimates may also strengthen the validity within comparison.[6] Even when the results of the direct evidence are conclusive, merging them with the results of indirect estimates in MTC may give a more accurate estimate optimising the existing information of the network.[7,8]

The aim of this work is to outline the principal steps required to perform an efficient and valid NMA. We focus our attention on data which can be analysed with a binomial model applying the Bayesian hierarchical approach proposed by Smith et al.[9] and using Markov Chain Monte Carlo (MCMC) approaches. We first review and examine the methodology underlying NMA. The principal references on this topic are harvested and discussed. Then, we exploit an example of anaesthetic drugs, applying the NMA methods to quantify the clinical effect of these agents on 30-day mortality. Finally, the very codes used for the analysis in WinBUGS 1.4 (freely available in the BUGS project website) are given in the online only appendix.[10]

2 Methods

2.1 Literature search

The network of evidence must include all randomised clinical trials of relevant treatments (interventions, drugs or procedures) that have been compared directly in a reasonably similar patient and diagnostic setting. The inclusion of all relevant evidence in systematic reviews is crucial to avoid bias and maximise precision.[11] The literature search for a NMA applies the same basic standards exploited for pairwise comparisons. Indeed, a standard meta-analysis involves a single search for any trial that compares the treatment of interest with any other therapy that may be theoretically exploited also in a NMA focusing on differences in treatment effects.

However, the choice of treatments to include in a NMA is on the contrary more challenging. Since the literature search is time consuming and requires resources, one may decide that it is not worthwhile to search for all the possible indirect evidences. Hawkins et al.[12] suggest an efficient search strategy to identify clinical trials that may provide indirect evidence when comparing different treatment comparators: a series of iterative searches where the set of comparators included in each search is dependent on the results of the previous one. This iterative process continues or stops considering the marginal cost of searching for higher order indirect data and the marginal benefit of progressively less informative data. If the search is stopped before finding all the entire evidence, the missed treatments are assumed as missing at random,[13] but it is important to pay attention to the applicability of results.

2.2 Configuration of network

Before starting a NMA, it is important to have a complete view of the distribution of included studies. The network diagram allows an intuitive approach to symbolically represent all the direct comparisons among treatments. This graph consists of a set of nodes representing the interventions linked by lines that depict how many RCTs have been included. Two important properties of network configuration are geometry and asymmetry.[14,15] Geometry refers to the overall structure of treatment contrasts, while asymmetry summarises the amount of data for a specific comparison. The network structure must be carefully built and examined so that each pattern of data may be used to reveal particular characteristics that may assist in the choice of the analytical method.[5] For example, Figure 1(a) (star-shaped) allows an ITC analysis of treatments B, C and D all linked to the common comparator A. The graph in Figure 1(b) comprises three nodes representing three

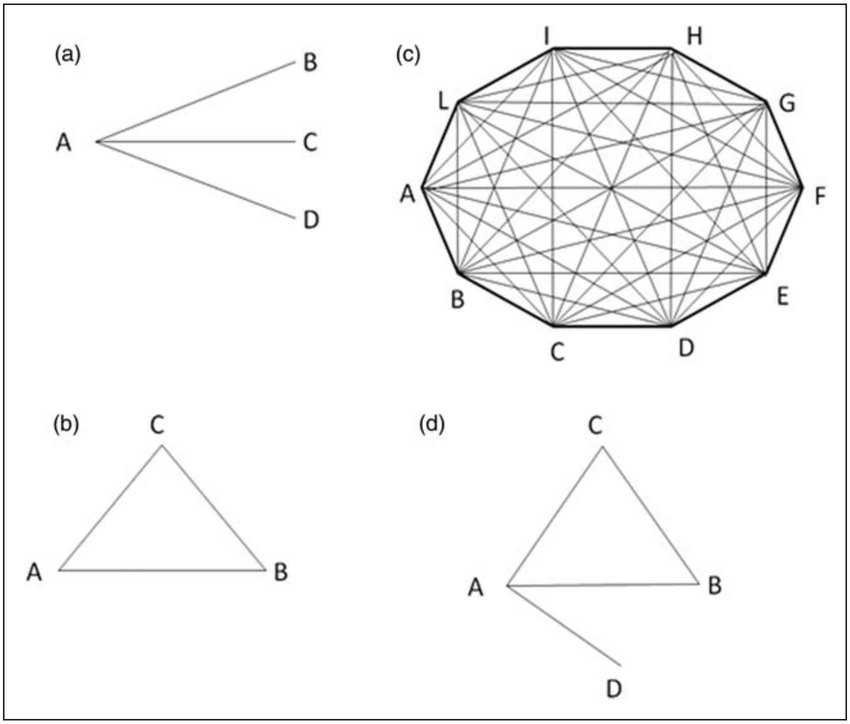

Figure 1. Examples of network configurations.

interventions (A, B and C) and three edges (arrows). An important property of this network is that each contrast has both direct and indirect evidence (closed loops). For example, the BC comparison obtains direct information from trials that compare BC and indirect evidence from trials that analyse AB and AC treatment differences. Moreover, the structure of the network can become extremely complex as in Figure 1(c) (multiple loops). A network where some pairwise contrasts have both direct and indirect evidence can be analysed performing an MTC analysis.[6,8] Actually, all connected networks can be examined using NMA, as even pairwise meta-analysis is only a special case of NMA, and can be analysed using the exact same model. Network configuration can help to establish which treatment can be defined as reference (usually the one that appears more frequently than others, or the one which is most commonly used in clinical practice) or to show the head-to-head comparative relations. In the network structure, the effect estimates may be included, with the corresponding 95% confidence, credibility or credible intervals, the number of studies and the study reference for each pairwise comparison.

Focusing on diagram asymmetry, it is crucial to understand the extent to which different nodes or links are present in the diagram, weighting for the number of trials. Salanti et al.[14,15] propose metrics and tests employed in the ecological literature.[16] Specifically, they suggest to investigate the diversity and the co-occurrence inherent in the network.[17,18] Diversity refers to the number of nodes contained in a network and to different frequencies in treatments. Conversely, co-occurrence is measured with the C-score[19] or other widely cited similarity scores (e.g. Jaccard, Dice or Cosine coefficients),[20] and represents the tendency of a particular comparison to occur more frequently than expected by chance. In other words, it tests for the presence of favourite couples of treatment.

2.3 Assumptions

When there is no or insufficient evidence from direct comparison trials, it may be possible to use results of different studies to obtain the pooled estimates of relative treatment effect. One fundamental assumption is the preservation of the randomisation process within each trial, comparing the estimates of relative effect among treatments. Let us suppose that we have three treatments (A, B and C) compared head-to-head in N trials. Treatment estimates will not be accurate if the researcher only calculates the indirect effect estimate of B versus C, by AB and AC trials, by balancing the observed fraction of respondents on treatment B from AB trials to the observed fraction of respondents on treatment C from AC trials. In fact, in this way, the analysis fails to separate the treatment effects from the other sources of variability.[21,22] However, one can compare the (log) odds ratio (lnOR) of A versus B from the AB trials to the lnOR for A versus C from AC trials.[5] This indirect comparison, adjusted according to the results of their direct comparison with a common comparator, largely preserves the force and validity of the randomised trials.[23] Both homogeneity (no variation in treatment effect between trials within pairwise contrasts) and consistency (no variation in treatment effect between pairwise contrasts) ensure the validity of the analysis. Evaluation of heterogeneity represents another milestone and should be based on the use of a random-effect meta-analysis approach assuming that each individual estimate is different at random and generated from a common distribution. For a large sample, classical inference based on the standard DerSimonian-Laid method is unbiased,[24] even when the distribution of the effects is extremely non-normal.[25,26] However, the performance of the method deteriorates rapidly as the number of studies decreases, especially for meta-analyses of five studies or fewer.[24,25,26] The advantage of Bayesian methods in comparison to such Frequentist approaches is that inference is exact for any sample size, assuming the prior assumptions are valid. In Bayesian statistics, uncertainty is evaluated using the posterior distribution of a parameter, which is the conditional probability distribution of all unknown quantities (i.e. the parameters), given the data, the model and what we knew about these quantities before conducting the analysis.

The between-trial variability can be attributed to specific characteristics (i.e. inclusion criteria, choice of outcomes, differences in follow-up or methods of randomisation) that can be sources of confounding and bias. The true treatment effect would be similar across all trials of network even if these did not include one or both of these two comparative elements. If the included trials have differences that are modifiers of the relative treatment effect, the similarity will be violated and the pooled estimated will be biased.[27–30] The classic measure of heterogeneity is Cochran Q, calculated as the weighted sum of squared differences between individual study effects and the pooled effect across studies. However, the Q test statistic has generally low power, even more so when the data are sparse.[24,31] Another commonly used statistic is I^2 [23,32,33] that describes the percentage of variation across studies that is due to heterogeneity rather than chance. However, use of I^2 can be challenging since it is not independent from the meta-analysis size.[34]

Nevertheless, Bayesian meta-analysis allows the incorporation, into the random effect model, of between-study heterogeneity, including a prior distribution for it as well. Heterogeneity should also be taken into account by performing adjusted analysis, planning appropriate subgroup analyses or using meta-regression techniques to adjust for differences in study-level characteristics.

The exchangeability assumption[35] justifies the fact that the treatment effects may be non-identical but their magnitudes cannot be differentiated a priori. Within the context of NMA, it is important that the indirect estimate is not biased and that there is no divergence between the direct and indirect comparisons. If the AB and AC trials are comparable in effect modifiers (and are thus similar), an indirect estimate ($\tilde{\theta}_{BC}$) for the true difference effect between B versus C can be obtained from the direct estimates of A versus B ($\hat{\theta}_{AB}$) and direct estimates of A versus C ($\hat{\theta}_{AC}$). To perform NMA, it is indispensable that the following *consistency equation* is satisfied:

$$\tilde{\theta}_{BC} = \hat{\theta}_{AC} - \hat{\theta}_{AB} \tag{1}$$

where the effectiveness of each treatment is measured on a scale symmetric to zero such as lnOR, log-hazard ratio or difference in mean.[5,8,14,36,37]

Consistency regards a loop (closed network) rather than individual comparisons. Indeed, to verify presence of inconsistency, the treatments involved must belong to a loop in the network configuration. Lu and Ades[13] propose a general method for assessing evidence inconsistency in the framework of Bayesian hierarchical models. They suggest to represent evidence consistency as a set of linear relations among basic parameters on the log odds scale. Then, these relations will be complicated by introducing some random terms, called inconsistency factors (ICF), and finally this model which incorporates ICF will be compared with the standard one without ICF. Dias et al.[38] also propose an extension of Bucher methods[3] to carry out tests for inconsistency in a network with multiple loops and with only two-arm trials. In this work, we compared the goodness of fit of the consistency model (that obtains the indirect treatment effects by means of the consistency equation) with the inconsistency model (which estimates all relative effects for all treatment contrasts).

Health decisions should be based on models that are internally coherent and if the data cannot be fitted by a consistent model, some adjustment must be made to correct for possible causes of discrepancy. More details are provided in the following sections, but careful reading of the pioneering work of Lu and Ades[13] and Dias et al.[38] is also recommended.

2.4 Statistical details

2.4.1 Contrast-level and arm-level summary data

The input data in NMA are usually the summary statistics extracted from the published literature (aggregate data [AD] or study-level data), rather than the original data directly collected from trial authors (individual patient data [IPD] or patient-level data). Besides, the aggregate input data are available in two formats: as arm-level summaries, where effect measures are reported for each arm (i.e. odds, absolute risk, hazard or mean) or as contrast-level summaries (i.e. OR, risk ratio, hazard ratios or mean difference), where results are presented as the difference in effect between arms. One advantage of the arm-level approach is that it is possible to adopt the exact likelihood for the data (i.e. binomial for binary data) rather than its normal approximation, as for the contrast-level summary. Both Frequentist and Bayesian approaches can be used to specify models based on either two formats.[39] Hereafter, we discuss the analysis of data by means of arm-level summaries, which enable more flexible and precise analyses.

2.4.2 Fixed and random effect models

Suppose that N RCTs make mixed comparisons among K treatments. The number of events on treatment k in the trial i is denoted with r_{ik} and the number of total observations with n_{ik}. Let p_{ik} be the probability of event occurrence, then the number of events, r_{ik}, leads a Binomial distribution:

$$r_{ik} \sim Bi(p_{ik}, n_{ik}), \quad i = 1, 2, \ldots N; \ k = 1, 2, \ldots K \tag{2}$$

The probability of event occurrence p_{ik} is modelled on the logit scale as:

$$logit(p_{ib}) = \log\left(\frac{p_{ib}}{1 - p_{ib}}\right) = \mu_i, \quad i = 1, 2, \ldots N; \ k = b = 1, 2, \ldots K \tag{3}$$

$$logit(p_{ik}) = \log\left(\frac{p_{ik}}{1 - p_{ik}}\right) = \mu_i + \delta_{i,1k}, \quad i = 1, 2, \ldots N; \ k = 2, 3, \ldots K; b < k \tag{4}$$

where μ_i are the trial-specific baselines and represent the log odds of event in the referent treatment ($k = b$), while $\delta_{i,bk}$ are the trial-specific lnOR of event occurrence of the treatment group k compared with referent treatment.

The nature of effect $\delta_{i,bk}$ depends of assumption underlying the fitted model: fixed or random effect model. The difference consists in the way variability of the between-trial results is treated.[11] The fixed effect model considers this variability as exclusively due to random variation (assume between-trial variance equal to zero) and individual studies are simply weighted by their precision. Therefore, if all the studies were infinitely large, then they would give identical results. For fixed effect model, the equation (4) will be replaced as follows:

$$logit(p_{ik}) = \mu_i + d_{i,bk}, \quad i = 1, 2, \ldots N; \ b = 1, 2, \ldots K; \ k = 2, 3, \ldots K; \ b < k \tag{5}$$

where μ_i are the trial-specific baselines and $d_{i,bk}$ are the fixed ($\sigma^2_{i,bk} = \sigma^2_i = 0$), trial-specific lnOR of event occurrence of the treatment group k compared with referent treatment.

The random effect model, instead, assumes a different underlying effect for each study and takes this into consideration as an additional source of variation. This model has been advocated if there is heterogeneity in between-trial results. For a random effect model, the trial-specific log odd ratio $\delta_{i,bk}$ is commonly generated from a Normal distribution

$$\delta_{i,bk} \sim N(d_{bk}, \sigma^2) \tag{6}$$

We assumed equal within-trial variance between relative treatment effect ($\sigma^2_{i,bk} = \sigma^2_i$). For more details, Lu and Ades[8] explain the heterogeneous within-trial variance models. The Bayesian structure requires the prior specification for unknown parameter μ_i, $\delta_{i,bk}$ and σ. Dias et al.[40] recommend to give independent weakly priors such as $\mu_i, \delta_{i,bk} \sim N(0, 100^2)$ and $\sigma \sim Uniform(0, 2)$.

From the consistency assumption, the indirect estimate δ_{st} is:

$$\delta_{st} = \delta_{bt} - \delta_{bs}, \quad b = 1, 2, \ldots K; \; s = 2, 3, \ldots K; \; t = 3, 4 \ldots K; \; s < t \tag{7}$$

The $K - 1$ direct treatment effects δ_{bk} (between k and baseline treatment groups) represent the *basic parameters* of the model on which prior distributions of Bayesian approach are placed,[40] while the *functional parameters* δ_{st} are all the remaining contrasts that are function of basic parameters.

2.4.3 Network meta-regression

Network meta-regression represents a useful tool to explain the heterogeneity between the different treatment effects in the studies by regression of aggregate (study-level) covariates or on IPD, if available, exactly like head-to-head comparisons.[7,13,41,42] Nixon et al.[43] develop methods to simultaneously compare several treatments and to adjust for study-level covariates by combining ideas from MTC and meta-regression. In general, the meta-regression model is fitted specifying fixed or random effect models and adjusting the lnOR for study-level prognostic factors. The meta-regression procedure can reduce bias and inconsistency when covariates are distributed uniformly.[44,45]

The meta-regression model with fixed treatment effect is:

$$logit(p_{ib}) = \mu_i + \beta x_i, \quad i = 1, 2, \ldots N; \; b = 1, 2, \ldots K \tag{8}$$

$$logit(p_{ik}) = \mu_i + d_{i,1k} + \beta x_i, \quad i = 1, 2, \ldots N; \; k = 2, 3, \ldots K; \; b < k \tag{9}$$

where x_i is the trial-level covariate for trial i, which can represent a subgroup or a continue variable. In the meta-regression with random treatment effect, equation (9) is replaced with

$$logit(p_{ik}) = \mu_i + \delta_{i,bk} + \beta x_i, \quad i = 1, 2, \ldots N; \; b = 1, 2, \ldots K; \; k = 2, 3, \ldots K; \; b < k \tag{10}$$

where the trial-specific lnORs are generated from a common distribution $\delta_{i,bk} \sim N(d_{bk}, \sigma^2)$. The parameters μ_i, d_{bk},β and σ will be given independent weakly priors such as $\mu_i, d_{bk}, \beta \sim N(0, 100^2)$ and $\sigma \sim Uniform(0, 5)$.[42]

If the number of studies in a network is limited, the validity of incorporating study-level covariates with meta-regression model may be questionable, given the limited statistical power and risk of overfitting.[22,46] Besides, aggregate covariates adjustment might be prone to ecological bias[47] that represents the failure of study-level associations to properly reflect individual-level associations. Network meta-analyses of IPD are considered the gold standard, as they provide the opportunity to explore differences in effects between subgroups. When IPD are available, meta-regression usually has sufficient power to evaluate heterogeneity and to identify effect-modifying factors.[41,45]

2.4.4 Multi-arm trials

Let us suppose that we include in a network one or more multi-arm trials where the number of comparators is three or greater. A single multi-arm trial i which compares a_i treatments produces a vector δ_i of $a_i - 1$ random treatment effect, $\delta_i = (\delta_{i,12}, \ldots, \delta_{i,ba_i})^T$ that is correlated. Over the between-trial variance, it needs to include the random effect covariance.[8,14,40] The specification of the variance–covariance matrix for the random effects vary from constant and equal structure to totally unrestricted positive–definite matrix.[13,14] The assumption of homogeneous between-trial variance means that all σ^2_{bk} are the same and equal to σ^2, and this implies that the covariance between two contrasts in a multi-arm trial is $\sigma^2/2$.[27] The univariate Normal distribution (6) for multi-arm trial i which compares a_i treatments a_i will be a multivariate Normal distribution

$$\delta_i = \begin{pmatrix} \delta_{i,12} \\ \vdots \\ \delta_{i,ba_i} \end{pmatrix} \sim N_{a_i-1}\left(\begin{pmatrix} \delta_{i,12} \\ \vdots \\ \delta_{i,ba_i} \end{pmatrix}, \begin{pmatrix} \sigma^2 & \cdots & \sigma^2/2 \\ \vdots & \ddots & \vdots \\ \sigma^2/2 & \cdots & \sigma^2 \end{pmatrix}\right) \tag{11}$$

Let us imagine we have $K = 4$ trials included in the network and $\binom{K}{2} = \frac{K!}{2!(K-2)!} = 6$ contrasts between A, B, C and D (Figure 1d). One study is a multi-arm trial that compares A, B and C treatments and two studies are two-arm trials that produce AB and AD comparisons. This network will estimate three basic parameters (d_{AB}, d_{AC} and d_{AD}) and three functional parameters (d_{BC}, d_{BD} and d_{CD}) obtained from the consistence equation (1). The following random effect model can be specified:

$$logit\begin{pmatrix} p_{1B} \\ p_{2D} \\ p_{3B} \\ p_{3C} \end{pmatrix} = \begin{pmatrix} \mu_1 \\ \mu_2 \\ \mu_3 \\ \mu_3 \end{pmatrix} + \begin{pmatrix} 1\,0\,0\,0 \\ 0\,1\,0\,0 \\ 0\,0\,1\,0 \\ 0\,0\,0\,1 \end{pmatrix}\begin{pmatrix} \delta_{1,AB} \\ \delta_{2,AD} \\ \delta_{3,AB} \\ \delta_{3,AC} \end{pmatrix} \tag{12}$$

and

$$\begin{pmatrix} \delta_{1,AB} \\ \delta_{2,AD} \\ \delta_{3,AB} \\ \delta_{3,AC} \end{pmatrix} \sim N_4\left(\begin{pmatrix} \delta_{1,AB} \\ \delta_{2,AD} \\ \delta_{3,AB} \\ \delta_{3,AC} \end{pmatrix}, \begin{pmatrix} \sigma^2 & 0 & 0 & 0 \\ 0 & \sigma^2 & 0 & 0 \\ 0 & 0 & \sigma^2 & \frac{\sigma^2}{2} \\ 0 & 0 & \frac{\sigma^2}{2} & \sigma^2 \end{pmatrix}\right) \tag{13}$$

2.4.5 Bayesian framework and WinBUGS

There is a large literature based on Bayesian analysis, hierarchical modelling and implementation of MCMC methods to perform statistical inference.[27,35,48] In contrast with the classical statistical theory, the Bayesian approach can incorporate any available information about the parameters before we observe the data, and hence the parameters are considered as random variables that are characterised by a prior distribution. Priors may be divided into four categories according to the analytical goal and use: (a) informative, based on existing evidence; (b) weakly informative, that provide enough information to avoid results that contradict the previous knowledge; (c) least informative, determined solely by the model and observed data to minimise the amount of subjective acquaintance and (d) non-informative.[49] Van Dongen et al.[50] stated that non-informative priors, especially when derived from small sample sizes, lead to different results from reference non-Bayesian models, and weak priors generate information closer to the referral model. Consequently, the priors should be looked as a distribution that should reflect a biologically plausible parameter space.[50] With a large number of comparisons in a well-defined network configuration and with a large number of trials included, a reasonable choice of prior distributions will have minor effects on posterior inferences. If the data are sparse or there are no events in one or more arms of contrast, the prior distribution becomes more important. In general, if the information is strong the inference is based primarily on prior beliefs and if it is weak the numerical estimation can be unstable. Various weakly informative prior distributions have been suggested for scale parameters in hierarchical models. It has been suggested[40,51] to use vague priors for μ_i and d_{bk} parameters, such as $N(0, \sigma^2)$ with variance equal to 0.001 or 0.0001. A Uniform distribution, $\sigma \sim \mathit{Uniform}(0, A)$, can be used as prior for the standard deviation of Binomial distributions and logit link function. The upper limit of distribution, A, represents a huge range of trial-specific treatment effect.[40] For a finite but sufficient large A, inferences are not sensitive to the choice of A.[51] The approach to set a Gamma prior on precision, $1/\sigma \sim \mathit{Gamma}(\varepsilon, \varepsilon)$, produces a sharply peaked near zero distribution and further distorts posterior inferences, because of the marginal likelihood that σ^2 remains close to zero. Where σ is estimated to be near zero, the resulting inferences will be sensitive to ε. On the other hand, the use of an Inverse-Gamma distribution is suitable when data are sparse, improving stability and convergence. Usually the hyperparameter ε is set to a low value such as 0.001. As priors are part of the model specification, initial values are part of the computing process. Initial values can be derived from the current dataset or may be generated from prior distributions. The evaluation of posterior distributions is dependent on the MCMC chains convergence. Most convergence checking, such as the Gelman and Rubin approach, are graphical, and either compare the results from different chains or divide one chain into sections and compare these sections. If the simulation has not yet converged, the chains or part-chains will look different when plotted.[48] Finally, the Monte Carlo error (an index that reflect the number of simulation and the autocorrelation degree) should be no more than 5% of the posterior standard deviation of the parameters of interest to minimise the bias inherent to the resampling method.[40]

2.4.6 Goodness of fit

Statistical models, in addition to drive the inference process to provide prediction results, allow to describe how well the model itself fits a set of observations and to discriminate between alternative models. The *likelihood ratio test* represents one of the classic ways to compare two nested models.[52,53] Alternatives include the Akaike Information Criterion (AIC) and the Bayesian Information Criterion (BIC). The AIC[54] for a given model is a function of its maximised log-likelihood and the number of estimable parameters p:

$$AIC = -2 \log L\left(\hat{\theta}|y\right) + 2p \tag{14}$$

For a non-hierarchical model with p parameters and n observations, the Bayes (or Schwarz) Information Criterion[55] is given by

$$BIC = -2 \log L\left(\hat{\theta}|y\right) + 2p \cdot \log(n) \tag{15}$$

The advantage of the AIC and BIC statistics is that these can also be used for non-nested models. To compare two competitive models, as the comparison between the fixed and the random effect models, smaller values of these model assessment statistics are better, and efficiency remains paramount. Subsequently, Spiegelhalter et al.[56] developed a model comparison criterion called the Deviance Information Criterion (DIC), which is a generalisation and Bayesian version of AIC and is also related to the BIC, following the original suggestion of Dempster[57] for model choice in the Bayesian framework. Indeed, the Frequentist approach to model assessment is based on *deviance*, which the difference in the log-likelihoods between the fitted and the saturated models (the model with as many parameters as observations, with perfect to the data). Similarly, Dempster suggested to examine the posterior distribution of the classical deviance defined by

$$D(\theta) = -2\log f(y|\theta) + 2\log f(y) \tag{16}$$

for observations y and parameter vector θ. The DIC is thus based on the posterior distribution of $D(\theta)$ and it is defined as the sum of two components. The first component measures the goodness of fit of a model by the posterior expectation of the overall residual deviance:

$$E_{\theta|y}[D] = \bar{D} \tag{17}$$

The second measures the complexity of the model by the effective number of parameters, p_D, defined as the difference between the posterior mean of the overall residual deviance and the deviance evaluated at the posterior mean of the parameter of interest:

$$p_D = E_{\theta|y}[D] - D\big(E_{\theta|y}[\theta]\big) = \bar{D} - D(\hat{\theta}) \tag{18}$$

Ultimately, models may be compared using a DIC[40,56,58–60] defined by the sum of expressions (21) and (22)

$$DIC = \bar{D} + p_D = 2\bar{D} - D\left(\hat{\theta}\right) = D\left(\hat{\theta}\right) + 2p_D \tag{19}$$

The model with the smallest DIC is estimated to be the model that would best and most efficiently predict the observed data. It is difficult to say what would constitute an important difference in DIC, as both subjectivity and experience must be applied. As a rule of thumb, a difference of more than 10 might definitely rule out the model with the higher DIC, differences between 5 and 10 are considerable, but if the difference in DIC is <5 and the models provide very different inferences, care should be taken when referring the model with the lowest DIC.[10] The aforementioned statistics (AIC, BIC and DIC) are easily calculated during an MCMC run by monitoring both θ and $D(\theta)$. The DIC tool of WinBUGS system directly provides the posterior mean of the overall residual deviance (Dbar), the deviance of the posterior means of interested parameter (Dhat), the p_D and the DIC value.

Another promising method for comparing different models, nested or not, is to use only the posterior distribution of the sum of residual deviance $\bar{D}$ of each competing model.[13,49] The sum of residual deviance for a binomial likelihood function is provided by:

$$\bar{D} = \sum_{i=1}^{N} Dev_i = \sum_{i=1}^{N}\sum_{k=1}^{K} 2\left[\log\left(\frac{r_{ik}}{n_{ik}p_{ik}}\right) + (n_{ik} - r_{ik})\log\left(\frac{n_{ik} - r_{ik}}{n_{ik} - n_{ik}p_{ik}}\right)\right] \tag{20}$$

where, as mentioned earlier, r_{ik} denotes the number of events on treatment k in the trial i, n_{ik} represents the number of total observations and p_{ik} is the probability of event occurrence. The posterior distribution of the model deviance difference can be obtained as $\bar{D}_{1,2} = \bar{D}_1 - \bar{D}_2$ and it may be used to calculate the posterior probability

$$P\big[\bar{D}_{1,2} > \beta(\bar{D})\big] \tag{21}$$

as an analytic method for model selection. The choice of the value of β can vary for different purposes.

In order to make an association with the Frequentist approach, the difference between the deviances of two nested models is approximately a chi-squared distribution with df degrees of freedom, where $df = p_2 - p_1$ is the difference between the number of parameters estimated. In this case, one can choose $\beta = \chi^2_{1-\alpha;df}$. The higher this probability the stronger is the evidence in favour of model 2 against model 1. In addition, it is possible to calculate the value of β that gives:

$$P\big[\bar{D}_{1,2} > \beta(\bar{D})\big] = 0.5 \tag{22}$$

and use, for example, the table of Kass and Raftery[61] (Table 1) to quantify the evidence against model 1. It is worth noting that these numbers are driven more from intuition, rather than a scientific justification.[49] The posterior probability check is performed in WinBUGS using the step function.

2.4.7 Rank probability estimate

An advantage of Bayesian approach is that the posterior distribution of estimate, with its credible interval, can be interpreted in terms of probability which allows an intuitive and direct interpretation of which treatment is the best or the subsequent. In each MCMC run, every treatment is ranked according to its estimated magnitude. Then, the proportion of MCMC cycles in which the treatment k ranks first gives the probability that such specific treatment *is the best* among all K treatments. Other probabilities are calculated for being the second best, the third best and so on for each treatment. Salanti et al.[62] propose some graphical and numerical summaries of rank probabilities (rankograms). These authors also suggest a simple method to show the cumulative rank probabilities for each treatment estimating the surface under the cumulative rank curve (SUCRA). For each treatment k and for each rank w ($k, w = 1, 2, \dots K$), it is possible to calculate the vector of cumulative probabilities $cum_{k,w}$ and the SUCRA index will be:

$$SUCRA_k = \frac{\sum_{w=1}^{K-1} cum_{k,w}}{K-1} \tag{23}$$

Table 1. Scale of evidence proposed by Kass and Raftery.[61]

β	Evidence in favour of model 2
0–2	Not worth more than a bare mention
2–6	Positive
6–10	Strong
>10	Very strong

The SUCRA index simplifies the entire information about treatment ranking into a single number. SUCRA is equal to 1 if the treatment is surely the best, and equal to 0 if the treatment is surely the worst.

2.4.8 Sensitivity analysis

Various techniques may be used to check whether the assumptions of the model are valid and whether the fit of the model is adequate. In the Bayesian setting, it is important to pay attention to the robustness of the posterior distribution. One can assess how posterior distribution changes over different prior distributions.[48] When prior information is available, sensitivity analysis focuses on the structure of the prior distribution. When weak priors are used, it focuses on how different choices of prior parameters may influence the posterior inference. Besides, sensitivity analysis can be performed discussing the different findings from competing models (fixed or random effect models, consistency or inconsistency model) or executing the NMA on a subgroup of RCTs (high quality RCTs only or specific stratification by other baseline covariates).

3 Example: case study on anaesthetic drugs

3.1 Background

A case study on the beneficial effects on a 30-day mortality of volatile agents in comparison to total intravenous anaesthetics is hereby presented to clarify the statistical features of the models, the necessary diagnostics and computation details. We considered four anaesthetic agents used in cardiac surgery (called here A, B, C and D) and investigated through NMA. The scientific community agrees that there is initial evidence suggesting that different anaesthetic drugs could lead to apparent differences in survival rate in patients undergoing cardiac surgery,[63] with volatile agents having beneficial effects (or total intravenous anaesthetic having detrimental effects). At the same time, there are few (if any) direct comparisons between different anaesthetic agents to define which treatment is the best. The WinBUGS code to analyse data on anaesthetic drugs is shown in the online appendix using the indications of Ades et al.[64] and Dias et al.[38,40,42] The online Appendix A presents the MTC data structure of the four different anaesthetic drugs for the primary endpoint (30-day mortality after cardiac surgery).

3.2 Search strategy and network configuration

Pertinent studies were independently searched in BioMedCentral, MEDLINE/PubMed, Embase, and the Cochrane Central Register of clinical trials by two expert investigators. The search included all potentially relevant studies characterised by random allocation to treatment, comparison between at least two of the prespecified anaesthetic drugs (A, B, C and D) and performed in cardiac surgical patients with no restriction in dose and time of administration. We excluded duplicate publications (in this case the article reporting the longest follow-up was abstracted), nonhuman experiments and studies with lack of outcome data on the 30-day mortality. The final analysis included four treatment groups (A, B, C and D), 3140 patients and 30 RCT.

Figure 2 shows the network configuration of the anaesthetic agent example. It represents the odds ratio (OR) estimates, corresponding 95% confidence intervals, the number of studies and the study reference for each pairwise comparisons. Treatment A is chosen as reference because it is the most frequently used.

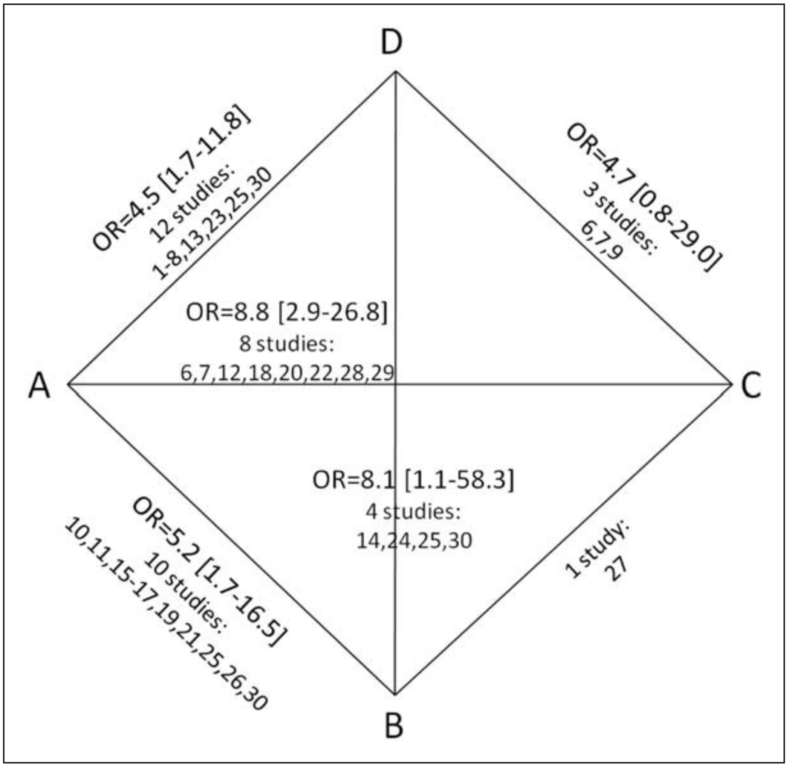

Figure 2. Network configuration of the anaesthetic agent example.
OR: odds ratios.

3.3 Data analysis: fixed and random effect models

The evaluation of the treatment effect for the dichotomous outcome (dead/alive at 30 days) is done by comparing the different effect estimates, by means of natural lnOR with its 95% credible intervals. The online Appendix B (B.1.1 and B.2.1) presents the WinBUGS code to perform both fixed effect and random effect models. The nodes *d* and *tau* (the latter only for the random effect model) must be monitored to obtain the posterior summaries of the parameters of interest for inference, while the posterior mean of DIC (from WinBUGS DIC tool) is needed to assess model fit. The results from the two models (running three chains: 20,000 iterations after a burn-in phase of 20,000 iterations for the fixed effect model and 100,000 iterations after a burn-in of 100,000 iterations for the random effect model[40]) are shown in Table 2. For each model, convergence and lack of auto-correlation were checked and confirmed graphically, by exploring the corresponding WinBUGs tools.[10] The OR estimates were calculated from the corresponding posterior mean of lnOR obtained from WinBUGS (see the online Appendix B, B.1.2 and B.2.2), while the indirect estimates were obtained from the consistency equation. The corresponding 95% credible intervals were calculated from the actual posterior distribution exploiting the normal approximation. For example, for the BC indirect estimate we have: $lnOR_{BC} = lnOR_{AC} - lnOR_{AB}$, where the treatment A is there reference. The standard error of the logarithm of indirect estimate $SE(lnOR_{BC})$ is obtained by $\sqrt{SE(lnOR_{AC})^2 + SE(lnOR_{AB})^2}$ and the 95% credible interval is equal to

Table 2. Anaesthetic agent example.

	Fixed effect model		Random effect model	
	Odds ratio	95% credible interval	Odds ratio	95% credible interval
d_{AB}	0.46	0.12–1.72	0.39	0.05–1.81
d_{AC}	0.39	0.12–1.20	0.34	0.06–1.25
d_{AD}	0.42	0.13–1.29	0.41	0.10–1.53
d_{BC}	0.84[a]	0.14–4.94	0.86[a]	0.08–8.92
d_{BD}	0.91[a]	0.16–5.34	1.04[a]	0.11–9.96
d_{CD}	1.09[a]	0.21–5.56	1.20[a]	0.15–9.52
Tau			128.8	
DIC	105.35		105.84	

Note: Posterior distribution of odds ratios and 95% credible intervals for both the fixed and random effect consistency model, comparing treatments A, B, C and D.

DIC: deviance information criterion.

[a] Indirect treatment effect calculated from consistency equation.

$lnOR_{BC} \pm SE(lnOR_{BC}).1,96$. The corresponding odds ratio OR_{BC} is calculated as $e^{lnOR_{BC}}$ and it 95% credible interval as $e^{lnOR_{BC} \pm SE(lnOR_{BC}).1,96}$. To compare the random effect and fixed effect model, we use the DIC tool of WinBUGS (shown in the online Appendix B, B.1.2 and B.2.2). Let $Dhat = -2\log L(\hat{\theta}|y)$ and $Dbar = \bar{D}$, we can use the DIC as a diagnostic test, in order to choose between fixed and the random effects models. The DIC statistic is given by $Dbar + p_D$ or directly from the WinBUGS tool: $DIC_{FE} = 105.35$ and $DIC_{RE} = 105.83$. We can conclude that the two models have the same goodness of fit. However, we choose the fixed model because it leads to more precise estimates than the random effect model (that is more conservative), and because it is by definition more parsimonious (Table 2). The analysis didn't reveal any significant association in the 30-day mortality.

3.4 Data analysis: consistency checking

The codes to fit both the fixed and random inconsistency models and the output from the DIC tool to compare the two nested models, for a network with a binomial likelihood and log-it link function, are displayed in the online Appendix C. The online Appendix D shows the posterior probability check method to compare the residual deviance difference between the consistency and inconsistency fixed effect models. As presented in the output, the posterior of $P[\bar{D}_{1,2} > \beta(\bar{D})]$, where $\bar{D}_{1,2} = 1.8$ and $\beta(\bar{D})$ that corresponds to a critical value of the chi-squared distribution with three degrees of freedom ($\chi_3^2 = 7.82$) is equal to 0.1. Hence, there is little evidence favouring model 2 under the assumption of equal prior probabilities on each model. This result is confirmed by the second analytical procedure. The value of β that gives $P[\bar{D}_{1,2} > \beta(\bar{D})] = 0.5$ is 2, which corresponds to 'not worth more than a bare mention' evidence in favour of model 2 (Table 1 of Kass and Raftery). In conclusion, we cannot reject the hypothesis of non-significance of the additional inconsistency regression coefficients. The consistency model is appropriate even if we cannot totally exclude the presence of inconsistency in the network configuration.

3.5 Data analysis: posterior rank probabilities

The Table 3 displays the probability that each treatment is best or the subsequent to prevent a 30-day mortality (i.e. 42% probability of being ranked 1st for C), and the cumulative rank

Table 3. Anaesthetic agent example.

	Treatment j			
Rank b	A	B	C	D
1 (best)	0.01 (0.01)	0.27 (0.27)	0.42 (0.42)	0.31 (0.31)
2	0.03 (0.04)	0.30 (0.57)	0.30 (0.72)	0.37 (0.68)
3	0.16 (0.20)	0.32 (0.89)	0.25 (0.97)	0.28 (0.96)
4 (worst)	0.80 (1)	0.11 (1)	0.03 (1)	0.04 (1)
SUCRA (%)	8.3	57.7	70.3	65

Note: Posterior distribution of the ranking probability, and cumulative probability in brackets, for each treatment.

probabilities for each treatment to estimate the SUCRA index calculated by equation (23). It can be seen from Table 3 that the SUCRA index for treatment C is obtained as $(0.42 + 0.72 + 0.97)/3$, thus is 0.703. Hence, drug C emerges as the best treatment, followed by drug D, drug B and finally drug A. The code in the online Appendix E includes the commands to calculate the posterior rank probability to be the best, the second, the third and the worst treatment.

4 Discussion

We have provided a comprehensive and detailed overview of the conceptual and practical issues involved in performing and interpreting NMA on binomial data while applying a Bayesian hierarchical model. We have discussed the general topics related to NMA, including how to collect study data, structure the network and set assumptions about the network that lead to different models and interpretations of model parameters. Many papers have been published on these topics and other will follow suite, describing Bayesian methods for NMA with binary data in a concise way and in quite some detail.[5–10,12–15,21,22,29,30,32,36,38–45] We have strived to put together the most important topics (making available the major references) and we offer, for the first time, a thorough yet manageable guideline to conduct (from literature search to results interpretation) a rigorous NMA on binomial data, applying the Bayesian hierarchical model. In addition, we applied the posterior probability check method[49] to compare the posterior distribution of the sum of residual deviance of consistency and inconsistency models. The presented case study on the beneficial effects of anaesthetic agents, in comparison to total intravenous anaesthetics, and the practical guide with the actual WinBUGS codes, available in the online appendix, will allow transparency and ease of replication of all steps that are required when carrying out such quantitative syntheses.

Funding

This research received no specific grant from any funding agency in the public, commercial or not-for-profit sectors.

Conflict of interest

None declared.

References

1. Guyatt GH, Sackett DL, Sinclair JC, et al. Users' guides to the medical literature. IX. A method for grading health care recommendations. Evidence-based Medicine Working Group. *J Am Med Assoc* 1995; **274**: 1800–1804.
2. Sackett DL, Rosenberg WM, Gray JA, et al. Evidence based medicine: what it is and what it isn't. *BMJ* 1996; **312**: 71–72.
3. Bucher HC, Guyatt GH, Griffith LE, et al. The results of direct and indirect treatment comparisons in meta-analysis of randomized controlled trials. *J Clin Epidemiol* 1997; **50**: 683–691.
4. Song F, Altman DG, Glenny AM, et al. Validity of indirect comparison for estimating efficacy of competing interventions: empirical evidence from published meta-analyses. *BMJ* 2003; **326**: 472.
5. Jansen JP, Fleurence R, Devine B, 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.
6. Caldwell DM, Ades AE and Higgins JPT. Simultaneous comparison of multiple treatments: combining direct and indirect evidence. *BMJ* 2005; **331**: 897–900.
7. Lumley T. Network meta-analysis for indirect treatment comparisons. *Stat Med* 2002; **21**: 2313–2324.
8. Lu G and Ades AE. Combination of direct and indirect evidence in mixed treatment comparisons. *Stat Med* 2004; **23**: 3105–3124.
9. Smith TC, Spiegelhalter DJ and Thomas A. Bayesian approaches to random-effects meta-analysis: a comparative study. *Stat Med* 1995; **14**: 2685–2699.
10. Spiegelhalter D, Thomas A, Best N, et al. WinBUGS User Manual. Version 1.4, http://www.mrc-bsu.cam.ac.uk/bugs (2003, accessed April 2012).
11. Egger M, Smith GD and Altman DG. *Systematic reviews in health care: meta-analysis in context*. London: BMJ Publishing Group, 2001, pp.69–86.
12. Hawkins N, Scott DA and Woods B. How far do you go? Efficient searching for indirect evidence. *Med Decis Making* 2009; **29**: 273–281.
13. Lu G and Ades AE. Assessing evidence inconsistency in mixed treatment comparisons. *J Am Stat Assoc* 2006; **101**: 447–459.
14. Salanti G, Higgins JP, Ades AE, et al. Evaluation of networks of randomized trials. *Stat Meth Med Res* 2008; **17**: 279–301.
15. Salanti G, Kavvoura FK and Ioannidis JP. Exploring the geometry of treatment networks. *Ann Intern Med* 2008; **148**: 544–553.
16. Magurran AE. *Ecological diversity and its measurement*. Princeton: Princeton University Press, 1998.
17. Hamilton AJ. Species diversity or biodiversity? *J Environ Manage* 2005; **75**: 89–92.
18. Tiho S and Josensb G. Co-occurrence of earthworms in urban surroundings: a null model analysis of community structure. *Eur J Soil Biol* 2007; **43**: 84–90.
19. Stone L and Robert A. The checkerboard score and species distributions. *Oncologia* 1990; **85**: 74–79.
20. Salton G. *Introduction to modern information retrieval*. New York, NY: McGraw-Hill, 1983.
21. Sutton A, Ades AE, Cooper N, et al. Use of indirect and mixed treatment comparisons for technology assessment. *Pharmacoeconomics* 2008; **26**: 753–767.
22. Jansen JP, Crawford B, Bergman G, et al. Bayesian meta-analysis of multiple treatment comparisons: an introduction to mixed treatment comparisons. *Value Health* 2008; **11**: 956–964.
23. Higgins JPT and Thompson SG. Quantifying heterogeneity in a meta-analysis. *Stat Med* 2002; **21**: 1539–1558.
24. Brockwell SE and Gordon IR. A comparison of statistical methods for meta-analysis. *Stat Med* 2001; **20**: 825–840.
25. Kontopantelis E and Reeves D. Performance of statistical methods for meta-analysis when true study effects are non-normally distributed: a simulation study. *Stat Meth Med Res* 2012; **21**: 409–426.
26. Kontopantelis E and Reeves D. Performance of statistical methods for meta-analysis when true study effects are non-normally distributed: a comparison between DerSimonian-Laird and restricted maximum likelihood. *Stat Meth Med Res* 2012; **21**: 657–659.
27. Higgins JP and Whitehead A. Borrowing strength from external trials in a meta-analysis. *Stat Med* 1996; **15**: 2733–2749.
28. Turner RM, Davey J, Clarke MJ, et al. Predicting the extent of heterogeneity in meta-analysis, using empirical data from the Cochrane database of systematic reviews. *Int J Epidemiol* 2012; **41**: 818–827.
29. Song F, Loke YK, Walsh T, et al. Methodological problems in the use of indirect comparisons for evaluating healthcare interventions: survey of published systematic reviews. *BMJ* 2009; **3**: b1147.
30. Donegan S, Williamson P, Gamble C, et al. Indirect comparisons: a review of reporting and methodological quality. *PLoS One* 2010; **5**: e11054.
31. Hardy RJ and Thompson SG. Detecting and describing heterogeneity in meta-analysis. *Stat Med* 1998; **17**: 841–856.
32. Higgins JPT, Thompson SG, Deeks JJ, et al. Measuring inconsistency in meta-analyses. *BMJ* 2003; **327**: 557–560.
33. Higgins JPT and Green S. *Cochrane handbook for systematic reviews of interventions*. Version 5.1.0. The Cochrane collaboration, http://www.cochrane-handbook.org/ (2011, accessed April 2012)
34. Mittlböck M and Heinzl H. A simulation study comparing properties of heterogeneity measures in meta-analyses. *Stat Med* 2006; **25**: 4321–4333.
35. Spiegelhalter DJ, Abrams KR and Myles JP. *Bayesian approaches to clinical trials and health-care evaluation*. New York, NY: Wiley, 2004.
36. Cooper NJ, Sutton AJ, Morris D, et al. Addressing between-study heterogeneity and inconsistency in mixed treatment comparisons: application to stroke prevention treatments in individuals with non-rheumatic atrial fibrillation. *Stat Med* 2009; **28**: 1861–1881.
37. Caldwell DM, Welton NJ and 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.
38. Dias S, Welton WJ, Sutton AJ, et al. NICE DSU Technical Support Document 4: inconsistency in networks of evidence based on randomised controlled trial, http://www.nicedsu.org.uk (2011, accessed March 2012).
39. Franchini AJ, Dias S, Ades AE, et al. Accounting for correlation in network meta-analysis with multi-arm trials. *Res Synth Meth* 2012; **3**: 142–160.
40. Dias S, Welton NJ, Sutton AJ, et al. NICE DSU Technical Support Document 2: a generalised linear modelling framework for pairwise and network meta-analysis of randomised controlled trials, http://www.nicedsu.org.uk (2011, accessed March 2012).
41. Hoaglin DC, Hawkins N, Jansen JP, 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.

42. Dias S, Sutton AJ, Welton WJ, et al. NICE DSU technical support document 3: heterogeneity: subgroups, meta-regression, bias and bias-adjustment, http://www.nicedsu.org.uk (accessed March 2012).
43. Nixon RM, Bansback N and Brennan A. Using mixed treatment comparisons and meta-regression to perform indirect comparisons to estimate the efficacy of biologic treatments in rheumatoid arthritis. *Stat Med* 2007; **26**: 1237–1254.
44. Lu G and Ades A. Modeling between-trial variance structure in mixed treatment comparisons. *Biostatistics* 2009; **10**: 792–805.
45. Lambert PC, Sutton AJ, Abrams KR, et al. A comparison of summary patient-level covariates in meta-regression with individual patient data meta-analysis. *J Clin Epidemiol* 2002; **55**: 86–94.
46. Berlin JA, Santanna J, Schmid CH, et al. Anti-Lymphocyte Antibody Induction Therapy Study Group. Individual patient- versus group-level data meta-regressions for the investigation of treatment effect modifiers: ecological bias rears its ugly head. *Stat Med* 2002; **21**: 371–387.
47. Greenland S and Morgenstern H. Ecological bias, confounding, and effect modification. *Int J Epidemiol* 1989; **18**: 269–274.
48. Ntzoufras I. *Bayesian modeling using WinBUGS*. New York, NY: Wiley, 2009.
49. Adamakis S, Raftery CL, Walsh RW, et al. A Bayesian approach to comparing theoretic models to observational data: a case study from solar flare physics. astro-ph.SR 2012; arXiv:1102.0242v3, http://arxiv.org/abs/1102.0242 (accessed November 2012).
50. Van Dongen S. Prior specification in Bayesian statistics: three cautionary tales. *J Theor Biol* 2006; **242**: 90–100.
51. Gelman A. Prior distributions for variance parameters in hierarchical models. *Bayesian Anal* 2006; **1**: 515–533.
52. Crowder MJ. Maximum likelihood estimation for dependent observations. *J Royal Stat Soc* 1976; **B38**: 45–53.
53. Gouriéroux C, Holly A and Monfort A. Likelihood Ratio Test, Wald Test, and Kuhn-Tucker Test in linear models with inequality constraints on the regression parameters. *Econometrica* 1982; **50**: 63–80.
54. Akaike H. Information theory as an extension of the maximum likelihood principle. In: Petrov BN and Casaki (eds) *Second international symposium on information*. Budapest: Akademiai Kiado, 1973, pp.267–281.
55. Schwarz G. Estimating the dimension of a model. *Ann Stat* 1978; **6**: 461–464.
56. Spiegelhalter DJ, Best NG and Carlin BP. Bayesian deviance, the effective number of parameters, and the comparison of arbitrarily complex models. Technical Report, MRC Biostatistics Unit, Cambridge, UK, 1998, http://yaroslavvb.com/papers/spiegelhalter-bayesian.pdf (accessed March 2012).
57. Dempster AP. The direct use of likelihood for significance testing. *Stat Comput* 1997; **7**: 247–252.
58. Berg A, Meyer R and Yu J. Deviance information criterion for comparing stochastic volatility models. *J Bus Econ Stat* 2004; **22**: 107–120.
59. Spiegelhalter DJ, Best ng, Carlin BP, et al. Bayesian measures of model complexity and fit. *J Royal Stat Soc Ser B* 2002; **64**: 583–639.
60. Yuan Y and Johnson VE. Goodness-of-fit diagnostics for Bayesian hierarchical models. *Biometrics* 2012; **68**: 156–164.
61. Kass RE and Raftery AE. Bayes factors. *J Am Stat Assoc* 1995; **90**: 773–795.
62. Salanti G, Ades AE and 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.
63. Landoni G, Rodseth R, Santini F, et al. Randomized evidence for reduction in perioperative mortality. *JCVA* 2012; **26**: 764–772.
64. Ades AE, Sculpher M, Sutton A, et al. Bayesian methods for evidence synthesis in cost-effectiveness analysis. *Pharmacoeconomics* 2006; **24**: 1–19.