NICE DSU TECHNICAL SUPPORT DOCUMENT 7: EVIDENCE SYNTHESIS OF TREATMENT EFFICACY IN DECISION MAKING: A REVIEWER'S CHECKLIST

REPORT BY THE DECISION SUPPORT UNIT

January 2012

AE Ades[1], Deborah M Caldwell[1], Stefanie Reken[2], Nicky J Welton[1], Alex J Sutton[3], Sofia Dias[1]

[1]School of Social and Community Medicine, University of Bristol, Canynge Hall, 39 Whatley Road, Bristol BS8 2PS, UK
[2]National Institute for Health and Clinical Excellence, MidCity Place, 71 High Holborn, London, WC1V 6NA, UK
[3]Department of Health Sciences, University of Leicester, 2nd Floor Adrian Building, University Road, Leicester LE1 7RH, UK

Decision Support Unit, ScHARR, University of Sheffield, Regent Court, 30 Regent Street Sheffield, S1 4DA;

Tel (+44) (0)114 222 0734
E-mail dsuadmin@sheffield.ac.uk

ABOUT THE DECISION SUPPORT UNIT

The Decision Support Unit (DSU) is a collaboration between the Universities of Sheffield, York and Leicester. We also have members at the University of Bristol, London School of Hygiene and Tropical Medicine and Brunel University. The DSU is commissioned by The National Institute for Health and Clinical Excellence (NICE) to provide a research and training resource to support the Institute's Technology Appraisal Programme. Please see our website for further information www.nicedsu.org.uk

ABOUT THE TECHNICAL SUPPORT DOCUMENT SERIES

The NICE Guide to the Methods of Technology Appraisal[i] is a regularly updated document that provides an overview of the key principles and methods of health technology assessment and appraisal for use in NICE appraisals. The Methods Guide does not provide detailed advice on how to implement and apply the methods it describes. This DSU series of Technical Support Documents (TSDs) is intended to complement the Methods Guide by providing detailed information on how to implement specific methods.

The TSDs provide a review of the current state of the art in each topic area, and make clear recommendations on the implementation of methods and reporting standards where it is appropriate to do so. They aim to provide assistance to all those involved in submitting or critiquing evidence as part of NICE Technology Appraisals, whether manufacturers, assessment groups or any other stakeholder type.

We recognise that there are areas of uncertainty, controversy and rapid development. It is our intention that such areas are indicated in the TSDs. All TSDs are extensively peer reviewed prior to publication (the names of peer reviewers appear in the acknowledgements for each document). Nevertheless, the responsibility for each TSD lies with the authors and we welcome any constructive feedback on the content or suggestions for further guides.

Please be aware that whilst the DSU is funded by NICE, these documents do not constitute formal NICE guidance or policy.

Dr Allan Wailoo

Director of DSU and TSD series editor.

[i] National Institute for Health and Clinical Excellence. Guide to the methods of technology appraisal, 2008 (updated June 2008), London.

Acknowledgements

The authors thank Phil Alderson and David Wonderling for their comments on earlier drafts of this document.

The DSU thanks Julian Higgins, Chris Hyde, Steve Palmer, Paul Tappenden and the team at NICE, led by Zoe Garrett, for reviewing this document. The editor for the TSD series is Allan Wailoo.

The production of this document was funded by the National Institute for Health and Clinical Excellence (NICE) through its Decision Support Unit. The views, and any errors or omissions, expressed in this document are of the author only. NICE may take account of part or all of this document if it considers it appropriate, but it is not bound to do so.

This report should be referenced as follows:

Ades, A.E., Caldwell, D.M., Reken, S., Welton, N.J., Sutton, A.J., Dias, S. NICE DSU Technical Support Document 7: Evidence synthesis of treatment efficacy in decision making: a reviewer's checklist. 2012; available from http://www.nicedsu.org.uk

EXECUTIVE SUMMARY

This checklist is for the review of evidence syntheses for treatment efficacy used in decision making based on either efficacy or cost-effectiveness. It is intended to be used for pair-wise meta-analysis, indirect comparisons, and network meta-analysis, without distinction. It does not generate a quality rating, and is not prescriptive. Instead it focuses on a series of questions aimed at revealing the assumptions that the authors of the synthesis are expecting reviewers to accept, on the adequacy of the arguments authors advance in support of their position, and the need for further analyses or sensitivity analyses. The checklist is intended primarily for those who review evidence syntheses, including indirect comparisons and network meta-analyses, in the context of decision making but will also be of value to those submitting syntheses for review, whether to decision making bodies or journals.

The checklist has four main headings:

A. Definition of the decision problem, including: target population for decision; comparators; trial inclusion criteria; treatment definition; trial outcomes; scale of measurement for synthesis; patient population in the trials in relation to the target population; trial quality; presentation of the data.
B. Methods of analysis and presentation of results, including: description of meta-analytic methods; heterogeneity in relative treatment effects; baseline models for trial outcomes; presentations of results.
C. Issues specific to network synthesis, including: model specification; multi-arm trials; connected and disconnected networks; inconsistency.
D. Embedding the synthesis in a probabilistic cost-effectiveness model, including the propagation of parameter uncertainty and correlations.

The headings and implicit advice follow directly from the other Technical Support Documents (TSD) in this series.[1-6] A simple table is provided that could serve as a *pro forma* checklist.

CONTENTS

1. INTRODUCTION 7

2. RELATION TO OTHER CHECKLISTS **9**

3. HOW TO INTERPRET AND USE THE CHECK-LIST **10**

4. THE CHECKLIST **11**

5. REFERENCES **31**

TABLES

Table 1 Checklist Table. Mark ✓ to indicate that the issue has been addressed satisfactorily, and ✗ if there is any cause for concern on the item. The Comments column should be used to answer the question (YES, NO, NA: not applicable) and/or to spell out the reasons for any concerns, the need for sensitivity analyses etc. 24

Abbreviations and Definitions

CEA	cost-effectiveness analysis
DIC	Deviance Information Criterion
MCMC	Markov chain Monte Carlo
RE	Random effects
TSD	Technical Support Document

1. INTRODUCTION

This Technical Support Document (TSD) sets out a practical checklist intended primarily for those who review evidence syntheses, including indirect comparisons and network meta-analyses, in the context of decision making. Of course, the information in such a checklist can also be used by those preparing such evidence syntheses. The checklist is a set of systematic criteria by which an independent reviewer can assess whether or not the synthesis meets the requirements set out in the NICE Methods Guide[7] and further elaborated in the other TSDs in this series.[1-6]

The proposed checklist has a relatively limited scope. Our assumption is that the purpose of the synthesis is to obtain a comparison, for purposes of efficacy and/or cost-effectiveness, of a *pre-specified set of treatments* in patients with *a pre-specified set of characteristics*. The purpose of this restriction is to tie the checklist firmly to the "decision making" context addressed in TSDs 1-6,[1-6] in which a clinician or the policy maker has a particular set of patients and precisely defined treatments in mind. This is the level at which re-imbursement authorities typically operate and in which clinicians are interested. As noted in TSD1,[1] not all evidence synthesis is conceived in precisely this way. A number of systematic reviews and meta-analyses are carried out with a primary objective of summarising literature on a particular treatment comparison or set of comparisons, often in a broader range of patient groups. The proposed checklist is not primarily intended for that broader form of review, although it may be highly relevant to it.

This document reads somewhat differently from other checklists that have been proposed.[8-11] For example, we make no attempt to produce a summary quality rating of the synthesis. A numerical or qualitative rating does not provide the information that decision makers require to determine whether a submitted synthesis represents an adequate basis for the decision they are charged with making. Similarly, a numerical or quality rating does not help an editorial board decide whether to accept a paper based on an evidence synthesis. Instead, we see the checklist as a way of assessing whether the synthesis and the conclusions drawn from it are a fair reflection of what can be concluded from the existing evidence, whatever its quality. The completed checklist would form the basis of a reviewer report to the decision maker or journal editor. Looked at from the other side, the checklist tells those submitting a synthesis precisely what are the critical issues they will be expected to clarify, and puts them on notice as to the arguments and evidence they may be called upon to marshal, and the sensitivity analyses they may be asked to undertake.

Our objective is to provide a framework for open discussion of whether a convincing argument has been made, albeit based on data that may be limited and imperfect. Similarly, there is no attempt to quantify the "strength of evidence", and still less to quantify or suggest a "strength of recommendation".[12,13] A convincing argument can be developed from poor evidence, and the strength of evidence should be fully reflected in the credible interval attached to it, which should incorporate not only sampling error but uncertainty due to bias adjustment, or due to the uncertain relevance of the available data (see TSD3[3]). If decisions are based on cost-effectiveness, the strength of recommendation is better expressed through the commonly used metrics such as incremental cost-effectiveness ratios, cost-effectiveness acceptability curves, and probability that a strategy is optimal, given the model and a threshold willingness-to-pay.[14]

The proposed checklist runs systematically over all the issues raised in this TSD series, including:

A. Definition of the decision problem, including: target population for decision; comparators; trial inclusion criteria; treatment definition; trial outcomes; scale of measurement for synthesis; patient population in the trials in relation to the target population; trial quality; presentation of the data.
B. Methods of analysis and presentation of results, including: description of meta-analytic methods; heterogeneity in relative treatment effects; baseline models for trial outcomes; presentations of results.
C. Issues specific to network synthesis, including: model specification; multi-arm trials; connected and disconnected networks; inconsistency.
D. Embedding the synthesis in a probabilistic cost-effectiveness model, including the propagation of uncertainty and correlations.

We begin with a brief review of the relationship between the checklist we are proposing and earlier checklists that have been oriented to pairwise, indirect comparisons, and network meta-analyses.[11,15,16] Among the key issues are: in what ways is a checklist intended for re-imbursement decisions, such as those made by NICE, different from checklists intended for systematic review and meta-analysis more generally? Another question is whether, given a reviewers' checklist for pair-wise meta-analysis, what additional checklist item are needed for indirect comparisons and network meta-analysis?

2. RELATION TO OTHER CHECKLISTS

Throughout the documents in this series, we have defined network meta-analysis as an extension of pairwise meta-analysis.[2] The statistical model for network meta-analysis has been published previously,[17-19] and software for implementing this model is described in TSD2.[2] A key assumption is that for any pair of treatments under consideration, the true relative treatment effects are either identical (fixed effect model) or exchangeable (random effect, RE, model), across *all* the trials in the set. This identity or exchangeability requirement is present for any pair of treatments X and Y. It is therefore not strictly correct to claim that network meta-analysis requires *extra* assumptions of "trial similarity" and "consistency", *additional* to assumptions that are required in pair-wise meta-analysis, as has been occasionally claimed.[16,20,21] But this is not to say that these properties are unimportant. On the contrary, the fact that pair-wise and network meta-analysis are so very close in their underlying assumptions only serves to emphasise that all the "good-practice" advice that is incorporated in existing guidance[22] and checklists[9-11,23,24] available for pair-wise meta-analysis, are also the essential guarantors of adequacy in network meta-analysis. Equally, however, it highlights that these assumptions deserve scrutiny in the context of pair-wise synthesis, particularly as there is even less possibility of checking them within the data at hand.

For this reason, we have not sought to duplicate existing guidelines for conducting or reporting systematic reviews.[11,22,25] Instead, we assume that these have been followed and that they apply equally to network and pair-wise meta-analysis. Similarly, most items in the proposed checklist apply to both pairwise and network meta-analyses. The only issues that come exclusively under the heading of network synthesis are connectedness of networks, inconsistency, and software implementation.

Setting aside any special issues raised in network synthesis, if we compare the proposals in the six TSDs to those in previous guidelines and checklists, it seems that the TSDs are more restrictive in some areas, and less restrictive in others. They tend to be *more* restrictive in their handling of effect modifiers and potential effect modifiers. For example, we suggest that attention in focussed on a clinically well-defined and homogeneous target population (TSD3[3], TSD4[4]), not *because* this reduces heterogeneity and the risk of inconsistency, but because this is likely to be inherent in the decision question. By contrast, it is easy to find published meta-analyses which have combined trials on patients who could not be considered homogeneous,

in some cases including trials on patients who have failed, *and* trials on patients who have not failed, on some of the treatments in question.

In other respects, our approach is less restrictive, in that we would encourage syntheses of multiple outcomes within a single coherent model,[26-29] rather than a separate synthesis for each outcome suggested in the "PICO" formula.

Although this checklist covers similar items to that of the ISPOR taskforce,[23,24] it is more detailed and has been designed to be more suited to inform an actual decision making process rather than guide academic paper submissions.

3. HOW TO INTERPRET AND USE THE CHECK-LIST

Our objective in providing a check-list is to provide guidance on what questions should be asked by a reviewer of an evidence synthesis. We assume that reviewers can ask for clarification, alternate analyses, sensitivity analyses, details of search algorithms, computer code, and so on. Whether or not these are forthcoming, the reviewer will have to form a series of judgements, which will be passed on to those making the final decision. The objective of the checklist, therefore, is to help reviewers form a precise understanding of exactly what assumptions those submitting the synthesis want them to accept, exactly what empirical evidence or reasoning they have advanced in support of their case, and what steps can be taken to ensure that issues are resolved, perhaps in a revised submission. The suggested checklist, therefore, expresses a record of "fact" about a synthesis, its conduct and assumptions, but also provides room for comments. These may include expressions of doubt about assumptions or interpretations of evidence, and may point to the need for further analyses, or sensitivity analyses.

It is acknowledged that, in certain cases, relatively strong assumptions may be necessary due to the lack of evidence. Furthermore, empirical approaches to testing those assumptions may be limited by the data available. A thorough and transparent discussion of all assumptions and their implications to the results should be provided in the report. The checklist is designed to allow for the reviewer to comment on whether the assumptions are reasonable and adequately justified, and to indicate that the issue in question has been adequately addressed. For example, in reply to a question on whether additional modelling assumptions were made, the reviewer may answer with a "tick" adding the comment "no additional assumptions"; or put a "cross" next to this item with a comment indicating that the

assumptions are questionable; or the reviewer may answer with a "tick" adding the comment "additional assumptions justified".

The checklist comes in four sections. It begins with a set of considerations relating to the definition of the decision problem, including the target patient population or populations, criteria for comparators and for trials, but also what is known, *before* examination of the data, about the potential role of known or unknown effect modifiers. The underlying question is: "Is this set of trials a reasonable basis for a synthesis?", and if not "What further assumptions or adjustments are required to make it a reasonable basis?" The second section turns to the data analysis methods and to the results. Heterogeneity in baseline or relative treatment effects is again a major issue. The third section examines issues that are uniquely concerned with network meta-analysis: connectedness and inconsistency. A final section touches on uncertainty propagation in the cost-effectiveness analysis. We refer to the other TSDs in this series throughout for further details.

4. THE CHECKLIST

A. DEFINITION OF THE DECISION PROBLEM

A1. Target population for decision

A1.1. Has the target patient population for decision been clearly defined?

Reviewers should note whether or not the target population is clearly defined, and whether there is more than one population, and therefore more than one decision, involved. It is assumed that each decision would require its own cost-effectiveness analysis (CEA).

A2. Comparators

A2.1. Decision Comparator Set: Have all the appropriate treatments in the decision been identified?

The *decision comparator set* of treatments includes all the treatments to be compared, as identified in the scoping exercise (TSD1[1]). Ideally, this should include all the candidate treatments for the target population in question.

A2.2. Synthesis Comparator Set: Are there additional treatments in the Synthesis Comparator Set, which are not in the Decision Comparator Set? If so, is this adequately justified?

The *synthesis comparator set* consists of all the treatments in the decision set plus any other treatments which will be used in the synthesis but which are not of interest for the decision (TSD1[1]). The question here is whether additional treatments have been added to the set of comparators, and why. One reason for adding treatments to the synthesis set might be in order to make a connected network.[1,7] It is sometimes possible to extend the comparison set still further,[30] although this would not be regarded as the base-case analysis within the terms of the 2008 Methods Guide.[7] The advantages of this extension are the increased potential to check consistency, the potential to reduce uncertainty by including more evidence, and the fact that the final results will be more robust and less sensitive to the inclusion of any individual trial. This must be weighed against potential disadvantages: as more treatments are included the possibility of heterogeneity in patient populations may increase. If expansion of the network leads to increased heterogeneity this may result in *increased* uncertainty in estimates from RE models.[31] Even so, the increased uncertainty may be an appropriate reflection of the true state of affairs, and the increased robustness conferred by a larger ensemble of data may be seen as outweighing this.

Another reason for extending the set of comparators is to be able to include trials that provide additional information on the relationship between outcomes (see TSD1[1] and TSD5[5]). An example would be in treatments for influenza where trials may report one or more of time to end of fever, time to end of symptoms, time to return to work, and so on.[26,32]

A3. Trial inclusion / exclusion

A3.1. Is the search strategy technically adequate and appropriately reported?

To minimise bias in the systematic review a thorough search of the literature should be conducted. This should be reported in sufficient detail so that it can be judged and reproduced, if required.[22] Standard methods for review protocols and reporting should be adopted according to current best practice.[11,22,25]

A3.2. Have all trials involving at least two of the treatments in the Synthesis Comparator set been included?

If some have been excluded, which are they, and have adequate reasons been given? Should sensitivity to inclusion/exclusion of these studies individually and/or together be provided?

In particular, there is no specific reason to rule out trials on the basis of their size. Nor should they be ruled out on the basis of their design, for example because they were non-inferiority trials. All things being equal, these design features should have no impact on the validity of the estimates obtained, only their variance (TSD1[1]). Possibly, a case could be made for ruling out smaller trials if there was reason to suspect publication bias or small-study bias, but this should be based on a formal analysis, with examination of funnel plots or other methods.[3,33] Cross-over or cluster-randomised trials should also be included, provided that results from a correct analysis (accounting for within patient comparisons or clustering) have been reported. Trials that were stopped early (under a protocol with pre-specified early stopping rules) should also be included, without adjustment for early stopping.[34,35] Multi-arm trials involving at least two of the treatments in the synthesis comparator set should also be included, although it is reasonable to exclude arms of treatments outside the synthesis comparator set, as they contribute nothing to the analysis. Single arm studies cannot be included in a relative efficacy analysis.

A3.3. Have all trials reporting relevant outcomes been included?

If different trials report different, but clearly related, outcomes or the same outcome has been reported in different ways (for example as hazard ratios or median time to an event) or at different time points, a synthesis incorporating different reporting formats or test instruments within a single coherent model should be undertaken, rather than separate syntheses for different formats and instruments. Methods for combining data reported in different formats, such as shared parameter models (see TSD2[2]), should be considered.[26,28]

A3.4. Have additional trials been included? If so, is this adequately justified?

Trials that would not fall within the strict target definition of patients or treatments may be included, if the trial population, treatment protocol or dosing are "similar" to those within the decision problem. The key assumption, that the relative treatment effects are identical or exchangeable with those in the target population, must be explicitly addressed (TSD3[3]), and sensitivity analyses excluding these studies should be considered. If further trials *have* been included, it needs to be established that there has been no arbitrary selection from among a set of eligible trials.

A4. Treatment Definition

A4.1. Are all the treatment options restricted to specific doses and co-treatments, or have different doses and co-treatments been "lumped" together? If the latter, is it adequately justified?

In a decision making context the doses and treatment regimes being considered for every treatment in the decision set are almost always tightly defined, if only to ensure that a treatment can be costed in a consistent way.[1,7] The practice of "lumping"[36] doses or co-treatments together generally makes no sense in decision making, unless the variations in dose or co-treatment are so small that clinicians would agree that the variation has no material effect on efficacy.[4] In the rare cases where "inconsistency" has been found to occur in empirical studies, lumping over different doses or co-treatments appears to be the reason.[17,37-40] If different doses or different co-treatments are considered to have the same efficacy this should be explicitly addressed and justified (TSD3[3]).

A4.2. Are there any additional modelling assumptions?

It is open to investigators to fit, for example, dose-response models,[41] or to fit models in which the effect of a complex intervention can be derived from the effects of the subcomponents.[42] Evidence in the literature that bears on the validity of such models in the current context should be reviewed, and their *a priori* clinical or scientific plausibility discussed. Evidence in the form of goodness of fit of alternative models should be presented (TSD2[2]).

A5. Trial outcomes and scale of measurement chosen for the synthesis

A5.1. Where alternative outcomes are available, has the choice of outcome measure used in the synthesis been justified?

Several different outcomes may be reported in a set of trials, and at more than one follow-up time. In cases where alternatives exist, some justification should be given for excluding the others. Where there are options to carry out synthesis on a single outcome and alternative options to carrying out a combined coherent synthesis on multiple outcomes, this should be discussed and justified. A coherent synthesis of several outcomes should give more robust results (for example, probit or logit models for ordered categorical outcomes such as PASI or ACR – see TSD2[2]), but the validity of such models should be established by citing previous literature, and/or by examining their validity in the synthesis dataset. For example, models for ordered categorical outcomes can be checked by examining the relative treatment effects at

different cut-offs in each trial and seeing if they are approximately the same. The residual deviance check of model fit (TSD2[2]) is also a useful guide.

A5.2. Have the assumptions behind the choice of scale been justified?

The choice of outcome measure that forms the basis for the synthesis, for example log odds ratio, log relative risk, log hazard ratio, risk difference, should be justified, as there is a strong assumption that the true effects are linear on the chosen scale (TSD2[2]). Analysis of rate outcomes in most cases assumes constant hazards over time in each trial arm, and a proportional hazards treatment effect. The plausibility of constant hazards, particularly when trial follow-up times vary greatly needs to be discussed. Conversely, the use of logit models for probability outcomes in studies with different follow-up times implies very different assumptions. One option is to assume that all outcome events that are going to occur, will have occurred before the observation period in the trial has ended, regardless of variation between studies in follow-up time. Another is to assume a proportional odds model, which implies a complex form for the hazard rates.[43] The clinical plausibility of these assumptions should be discussed and supported either by citing relevant literature, or by examination of evidence on changes in outcome rate over the period of follow-up.

A6. Patient population: trials with patients outside the target population

A6.1. Do some trials include patients outside the target population? If so, is this adequately justified?

A6.2. What assumptions are made about the impact, or lack of impact this may have on the relative treatment effects? Are they adequately justified?

A6.3. Has an adjustment been made to account for these differences? If so, comment on the adequacy of the evidence presented in support of this adjustment, and on the need for a sensitivity analysis.

Some trials have a patient population which differs somewhat from the target population for decision. The issue then arises: can these trials be included? They can be included, but investigators must then be explicit about what they are assuming, and give a reasoned argument justifying their approach. There are really only two alternatives, one is to say that the patients may have different characteristics, but that would not be expected to impact on the treatment effects. The other is to include some form of *adjustment* in the analysis, in order to obtain an adjusted estimate that would represent the treatment effect expected in the target population. This adjustment could be based on data, perhaps from another trial or cohort

study. Alternatively, such adjustment could be based on expert elicitation,[44] or on a meta-regression (see TSD3[3]).

A7. Patient population: heterogeneity within the target population

A7.1. Have potential modifiers of treatment effect been considered?

This may be based on clinical opinion or on a separate review of the literature.

A7.2. Are there apparent or potential differences between trials in their patient populations, albeit within the target population? If so, has this been adequately taken into account?

A different possibility is that, although the patient population of every trial appears to lie *within* the definition of the target population, there is still heterogeneity between the trial populations – perhaps, for example, based on age, referral pattern, previous treatment, or disease severity. As before, one option for the investigator is to consider that neither the relative treatment nor the baseline treatment effects are influenced by the patient heterogeneity. A second option is that the relative effects remain unchanged, but baseline effects are different. This would lead to a form of sub-group analysis on baselines, and potentially to different decisions being taken for different patient groups. This is covered in TSD5[5] and taken up in section B3 of this checklist. A final possibility would be that the relative effects vary. This would lead, potentially, to a subgroup analysis based on a covariate that modified the treatment effect (TSD3[3]). This could trigger a consideration of whether or not there was any *a priori* clinical rationale for suspecting such a subgroup effect, whether there was statistical evidence for an interaction in the trial data, and whether this had been examined and reported in previous analyses.

A8. Risk of Bias

A8.1. Is there a discussion of the biases to which these trials, or this ensemble of trials, are vulnerable?

A8.2. If a bias risk was identified, was any adjustment made to the analysis and was this adequately justified?

An account should be given of the characteristics of each of the individual trials that could be associated with bias, and also the possibility of publication or small-study biases attaching to the ensemble of trials. More importantly, there should also be an account of the potential impact trial quality could have on the synthesis results.[45] Biases associated with indicators of trial quality are a particular concern, as there is good evidence that these may act to increase

treatment effect.[46-51] Methods for adjusting for these biases should be considered (see TSD3[3]).

A9. Presentation of the data

A9.1. Is there a clear table or diagram showing which data have been included in the base-case analysis?

A network diagram is a useful way of showing the structure of the evidence. The actual data used in the base-case analysis (trial first author and date, outcomes, treatments compared and covariates if relevant) should be set out in a table. Examples are given in TSDs 1,[1] 2[2] (Appendix) and 4.[4]

A9.2. Is there a clear table or diagram showing which data have been excluded and why?

Details of all trials and outcomes not considered for the analysis should be detailed in a table or diagram, along with reasons.[11] In addition, a note should be made of other potentially relevant data available, such as information on related outcomes or outcomes reported at more than one time point. These data can be useful to inform more complex models or to test some of the key modelling assumptions. For example, if the survival curves are presented in the publication of a relevant trial, but no relevant data can be extracted, this should be noted. Similarly, if the main outcome is reported at 6 weeks, but is also available at other time points, these should be noted.

B. METHODS OF ANALYSIS AND PRESENTATION OF RESULTS

B1. Meta-analytic methods

B1.1. Is the statistical model clearly described?

Reviewers should be provided with a precise description of the meta-analytic method used. The model should either be presented in algebraic form, or a citation should be provided to the statistical model being assumed. If a Bayesian analysis is used, details on priors, convergence and number of iterations should also be given.[15]

Reviewers should check that the meta-analysis method used is statistically sound for the dataset at hand. For example, the addition of 0.5 to zero cells counts can materially bias the estimated treatment effects. If the treatment effects are strong and the event is common or there is large sample size imbalance between the groups, the Peto method should be avoided,[22] although it performs well when events are very rare.[52] Fixed effect estimators,

such as Mantel-Haenzel, should not be used without considering possible heterogeneity. Further guidance is provided in standard texts on meta-analysis.[52-54]

B1.2. Has the software implementation been documented?

The name of the software module and package used for statistical analysis should be given, and any additional computer code should be provided. It is essential that enough information is provided so that the exact same analysis can be replicated. If confidentiality issues exist, then a dummy dataset can be provided.

B2. Heterogeneity in the relative treatment effects

B2.1. Have numerical estimates been provided of the degree of heterogeneity in the relative treatment effects?

An assessment should be made of the degree of heterogeneity in relative treatment effects for each set of pair-wise comparisons. Tests of the null hypothesis of homogeneity,[22] the I^2 statistic,[55] or estimates of the between-trial variation in a RE model can all be useful. The latter are particularly valuable as they can be compared with the estimated treatment effects (see TSD3[3]).

B2.2. Has a justification been given for choice of random or fixed effect models? Should sensitivity analyses be considered?

The results of such analyses can be used, in part, to justify the choice of RE models. In a Bayesian context deviance information criterion (DIC) statistics can also be used for this (see TSD2[2]).

B2.3. Has there been an adequate response to heterogeneity?

If there is substantial heterogeneity in relative treatment effects, the role of known or unknown covariates and potential for random biases, and the possible role of bias adjustment (see section A8) or control for variation by covariate adjustment should be discussed (see TSD3[3]). Covariate adjustment will usually have implications for the decision question as it raises the possibility of different treatment effects in different patient groups. In addition acceptance of subgroup effects requires *a priori* clinical and scientific plausibility and/or demonstration of a statistically significant interaction effect (TSD3[3]).

B2.4. Does the extent of unexplained variation in relative treatment effects threaten the robustness of conclusions?

As the between-studies standard deviation approaches the average treatment effect in magnitude, it is legitimate to ask how this impacts on the validity of conclusions. One might be confident that the mean treatment effect in a RE model is greater than zero, while still being quite uncertain about whether the treatment effect will be positive in a future instance. How can such heterogeneity be interpreted in a decision context? One suggestion (see TSD3[3]), is that the predictive distribution of the treatment effect in a new trial is the appropriate input in a decision analysis, rather than the mean effect.[56-58] This is not a recommendation that is made in the NICE Methods Guide,[7] but it could be considered that it better represents the uncertainty in the treatment effect, without materially changing the expected treatment effect.

B2.5. Has the statistical heterogeneity between baseline arms been discussed?

The extent of heterogeneity in the baseline arms should be discussed, as it may provide information on the heterogeneity of the patient populations. Heterogeneity in baselines should lead to re-examination of trial inclusion criteria and the risk of heterogeneous treatment effects.

B3. Baseline model for trial outcomes

B3.1. Are baseline effects and relative effects estimated in the same model? If so, has this been justified?

TSD5[5] strongly recommends that the model for the relative treatment effects is independent of the model for the baseline model. The intention is to avoid biasing the relative effect model by choosing a baseline model whose assumptions are not correct. The usual approach to this is to model the empirical relative effect measures (log odds ratios, log hazard ratios, etc). The Bayesian approach in TSD2[2] models the trial arms rather than the relative effects, but places vague unrelated priors on the "baseline" arm of each trial. Simultaneous modelling of baseline and relative effects should generally be avoided unless a clear reason can be given. One reason would be that it is necessary to obtain numerically stable results, for example when a number of trials have zero cells (TSD5[5]).

B3.2. Has the choice of studies used to inform the baseline model been explained?

The source of data used for the baseline model should be explained and justified (TSD5[5]). Use of the placebo arms from the trials is one option, but external data could also be considered, or the placebo arms from a suitable subset of the included studies. The question that needs answering is: "what source or sources of data best represent the outcome that would be obtained with the standard treatment in the target population?" If several sources of data are available, methods for averaging them should be justified. TSD5[5] (Section 2.2) discusses the use of the predictive distribution where heterogeneous data is used.

B4. Presentation of results of analyses of trial data

B4.1. Are the relative treatment effects (relative to a placebo or "standard" comparator) tabulated, alongside measures of between-study heterogeneity if a RE model is used?

B4.2. Are the absolute effects on each treatment, as they are used in the cost-effectiveness analysis (CEA), reported?

Guidance on what results should be presented is available in several of the TSDs in this series: TSD1[1] and TSD2[2] provide examples of results tables from pair-wise and network meta-analysis. TSD3[3] illustrates reporting of results from subgroup and meta-regression analyses. A table with the results based only on direct evidence and the full network is very informative[17] as are other graphical and tabular displays[59,60] such as rank-o-grams:[29,61] plots which show the probabilities that each treatment is the best, second best etc and can incorporate multiple outcomes in one display.

B5 Synthesis in other parts of the natural history model

The relative treatment effect model and the baseline model are both based on the short term outcomes that are reported in trials. However, in most CEA models there is a need to project this "downstream" so that the natural history reflects post-trial outcomes.

B5.1. Is the choice of data sources to inform the other parameters in the natural history model adequately described and justified?

B5.2. In the natural history model, can the longer-term differences between treatments be explained by their differences on randomised trial outcomes?

Construction and interpretation of natural history models is greatly facilitated when the values of parameters "downstream" from the trial outcomes are independent of treatment. For example, can it be assumed that the side effects/mortality profile for Biologics in Rheumatoid

arthritis is independent of the initial treatment effect? When these parameters do depend on treatment they will often be informed from observational evidence. The use of observational evidence to drive differences in relative treatment effects needs to be carefully justified and explained. Potential sources of bias should be discussed.

C. ISSUES SPECIFIC TO NETWORK SYNTHESIS

The need for a detailed description of the methods and software implementation applies equally to indirect comparisons and network meta-analysis.

C1. Adequacy of information on model specification and software implementation

For network meta-analysis and indirect comparisons, the WinBUGS code for Bayesian evidence synthesis set out in TSD2[2] is a recommended option. The STATA package mvmeta[62] and implementation in SAS[63] are also recommended.

Technical note: parameterisation of treatment effects

There is a wide variety of alternative software platforms that investigators may wish to use. These range from implementations in well known statistical packages, such as SAS, STATA, S-PLUS, or R, or variants of the coding suggested in TSD2[2] for WinBUGS models. However, the model parameterisation requires care, as a number of apparently innocuous variations may give very different results, or be wrong. The reviewer faced with un-cited models or software devised by the investigator may need to ask for further information. See TSD6[6] for further details.

C2. Multi-arm trials

C2.1 If there are multi-arm trials, have the correlations between the relative treatment effects been taken into account?

When the empirical treatment differences are used as data (e.g. Log Odds Ratios, Log Hazard ratios, etc), the differences in multi-arm trials are correlated and this must be taken into account. This is done in the Bayesian models in TSD2[2], and in STATA's package mvmeta.[62] Most frequentist methods are based on treatment differences. There are a number of software tools under development within a frequentist framework, all of which are based on the treatment differences, and it remains to be seen whether the appropriate adjustments will be made.

C3. Connected and disconnected networks

C3.1. Is the network of evidence based on randomised trials connected?

It is easy to check that a network is "connected" and this should be clear from a network diagram. The approach to network synthesis described in TSD2[2] is intended only for connected networks. However, TSD1[1] sets out an approach by which disconnected networks can be connected. Approaches used to re-connect networks require a number of strong assumptions and these must be explained and, if appropriate, justified.

C4. Inconsistency

C4.1. How many inconsistencies could there be in the network?

The network structure should be set out in a diagram (see TSDs 1,[1] 2[2] and 4[4]), and the number of possible inconsistencies set out.

C4.2. Are there any a priori *reasons for concern that inconsistency might exist, due to systematic clinical differences between the patients in trials comparing treatments A and B, and the patients in trials comparing treatments A and C, etc?*

If the AB trials tend to have been carried out on systematically different patient populations to the AC trials or the BC trials, there is a high risk that indirect or mixed treatment comparisons will be unreliable.

C4.3. Have adequate checks for inconsistency been made?

TSD4[4] sets out a range of possible approaches to assessing the degree of inconsistency, and also detection of trials that might be contributing to this inconsistency. Different methods to check for inconsistent should be used, depending on the structure of the network. TSD3[3] suggests a Bayesian cross-validation approach to detect the presence of outliers.

C4.4 If inconsistency was detected, what adjustments were made to the analysis, and how was this justified?

If there is good evidence for the presence of inconsistency in a network, it is difficult to justify why it should be used as a basis for choosing which treatment is most effective or cost-effective. A range of options are available, including removing trials from the network or incorporating additional parameters to account for bias. There are, however, likely to be a large number of ways of eliminating inconsistency, which all have quite different implications (TSD4[4]).

D. EMBEDDING THE SYNTHESIS IN A PROBABILISTIC COST EFFECTIVENESS ANALYSIS

D1. Uncertainty propagation

D1.1. Has the uncertainty in parameter estimates been propagated through the CEA model?

Probabilistic models are the base-case requirements in submissions to NICE.[7] Failure to take account of the uncertainty in *any* parameter should be explained and justified.

D2. Correlations

D2.1 Are there correlations between parameters? If so, have the correlations been propagated through the CEA model?

Correlations between parameters are induced when they are estimated from the same dataset. Relative treatment effects from networks with loops are always correlated. Absolute effects of treatments based on differences from a common baseline are also correlated. TSD6[6] sets out alternative ways of insuring that correlations are propagated through the decision model, within both Bayesian Markov chain Monte Carlo (MCMC) and Frequentist frameworks.

Table 1 Checklist Table. Mark ✓ to indicate that the issue has been addressed satisfactorily, and ✗ if there is any cause for concern on the item. The Comments column should be used to answer the question (YES, NO, NA: not applicable) and/or to spell out the reasons for any concerns, the need for sensitivity analyses etc.

		Item satisfactory?	Comments
A. DEFINITION OF THE DECISION PROBLEM ***A1. Target population for decision***			
A1.1	*Has the target patient population for decision been clearly defined?*		
A2. Comparators			
A2.1	*Decision Comparator Set: Have all the appropriate treatments in the decision been identified?*		
A2.2	*Synthesis Comparator Set: Are there additional treatments in the Synthesis Comparator Set, which are not in the Decision Comparator Set? If so, is this adequately justified?*		
A3 Trial inclusion / exclusion			
A3.1	*Is the search strategy technically adequate and appropriately reported?*		
A3.2	*Have all trials involving at least two of the treatments in the Synthesis Comparator Set been included?*		
A3.3	*Have all trials reporting relevant outcomes been included?*		

		Item satisfactory?	Comments
A3.4	*Have additional trials been included? If so, is this adequately justified?*		
A4 Treatment Definition			
A4.1	*Are all the treatment options restricted to specific doses and co-treatments, or have different doses and co-treatments been "lumped" together? If the latter, is it adequately justified?*		
A4.2	*Are there any additional modelling assumptions?*		
A5 Trial outcomes and scale of measurement chosen for the synthesis			
A5.1	*Where alternative outcomes are available, has the choice of outcome measure used in the synthesis been justified?*		
A5.2	*Have the assumptions behind the choice of scale been justified?*		
A6 Patient population: trials with patients outside the target population			
A6.1	*Do some trials include patients outside the target population? If so, is this adequately justified?*		

		Item satisfactory?	Comments
A6.2	*What assumptions are made about the impact, or lack of impact this may have on the relative treatment effects? Are they adequately justified?*		
A6.3	*Has an adjustment been made to account for these differences? If so, comment on the adequacy of the evidence presented in support of this adjustment, and on the need for a sensitivity analysis.*		
A7 Patient population: heterogeneity within the target population			
A7.1	*Has there been a review of the literature concerning potential modifiers of treatment effect?*		
A7.2	*Are there apparent or potential differences between trials in their patient populations, albeit within the target population? If so, has this been adequately taken into account?*		
A8 Risk of Bias			
A8.1	*Is there a discussion of the biases to which these trials, or this ensemble of trials, are vulnerable?*		
A8.2	*If a bias risk was identified, was any adjustment made to the analysis and was this adequately justified?*		

		Item satisfactory?	Comments
A9. Presentation of the data			
A9.1	*Is there a clear table or diagram showing which data have been included in the base-case analysis?*		
A9.2	*Is there a clear table or diagram showing which data have been excluded and why?*		
B. METHODS OF ANALYSIS AND PRESENTATION OF RESULTS ***B1 Meta-analytic methods***			
B1.1	*Is the statistical model clearly described?*		
B1.2	*Has the software implementation been documented?*		
B2. Heterogeneity in the relative treatment effects			
B2.1	*Have numerical estimates been provided of the degree of heterogeneity in the relative treatment effects?*		
B2.2	*Has a justification been given for choice of random or fixed effect models? Should sensitivity analyses be considered?*		
B2.3	*Has there been adequate response to heterogeneity?*		

		Item satisfactory?	Comments
B2.4	*Does the extent of unexplained variation in relative treatment effects threaten the robustness of conclusions?*		
B2.5	*Has the statistical heterogeneity between baseline arms been discussed?*		
B3 Baseline model for trial outcomes			
B3.1	*Are baseline effects and relative effects estimated in the same model? If so, has this been justified?*		
B3.2	*Has the choice of studies to inform the baseline model been explained?*		
B4 Presentation of results of analyses of trial data			
B4.1	*Are the relative treatment effects (relative to a placebo or "standard" comparator) tabulated, alongside measures of between-study heterogeneity if a RE model is used?*		
B4.2	*Are the absolute effects on each treatment, as they are used in the CEA, reported?*		
B5 Synthesis in other parts of the natural history model			
B5.1	*Is the choice of data sources to inform the other parameters in the natural history model adequately described and justified?*		

		Item satisfactory?	Comments
B5.2	*In the natural history model, can the longer-term differences between treatments be explained by their differences on randomised trial outcomes?*		
C. ISSUES SPECIFIC TO NETWORK SYNTHESIS ***C1 Adequacy of information on model specification and software implementation***			
C2. Multi-arm trials			
C2.1	*If there are multi-arm trials, have the correlations between the relative treatment effects been taken into account?*		
C3 Connected and disconnected networks			
C3.1	*Is the network of evidence based on randomised trials connected?*		
C4 Inconsistency			
C4.1	*How many inconsistencies could there be in the network?*		
C4.2	*Are there any* a priori *reasons for concern that inconsistency might exist, due to systematic clinical differences between the patients in trials comparing treatments A and B, and the patients in trials comparing treatments A and C, etc?*		

		Item satisfactory?	Comments
C4.3	*Have adequate checks for inconsistency been made?*		
C4.4	*If inconsistency was detected, what adjustments were made to the analysis, and how was this justified?*		
D EMBEDDING THE SYNTHESIS IN A PROBABILISTIC COST EFFECTIVENESS ANALYSIS ***D1. Uncertainty Propagation***			
D1.1	*Has the uncertainty in parameter estimates been propagated through the CEA model?*		
D2 Correlations			
D2.1	*Are there correlations between parameters? If so, have the correlations been propagated through the CEA model?*		

5. REFERENCES

1. Dias, S., Welton, N.J., Sutton, A.J., Ades, A.E. NICE DSU Technical Support Document 1: Introduction to evidence synthesis for decision making. 2011; last updated April 2012; available from http://www.nicedsu.org.uk

2. Dias, S., Welton, N.J., Sutton, A.J., Ades, A.E. NICE DSU Technical Support Document 2: A generalised linear modelling framework for pair-wise and network meta-analysis of randomised controlled trials. 2011; last updated April 2012; available from http://www.nicedsu.org.uk

3. Dias, S., Sutton, A.J., Welton, N.J., Ades, A.E. NICE DSU Technical Support Document 3: Heterogeneity: subgroups, meta-regression, bias and bias-adjustment. 2011; last updated April 2012; available from http://www.nicedsu.org.uk

4. Dias, S., Welton, N.J., Sutton, A.J., Caldwell, D.M., Lu, G., Ades, A.E. NICE DSU Technical Support Document 4: Inconsistency in networks of evidence based on randomised controlled trials. 2011. last updated April 2012; available from http://www.nicedsu.org.uk

5. Dias, S., Welton, N.J., Sutton, A.J., Ades, A.E. NICE DSU Technical Support Document 5: Evidence synthesis in the baseline natural history model. 2011; last updated April 2012; available from http://www.nicedsu.org.uk

6. Dias, S., Sutton, A.J., Welton, N.J., Ades, A.E. NICE DSU Technical Support Document 6: Embedding evidence synthesis in probabilistic cost-effectiveness analysis: software choices. 2011; last updated April 2012; available from http://www.nicedsu.org.uk

7. National Institute for health and Clinical Excellence. Guide to the methods of technology appraisal (updated June 2008). 2008.

8. Oxman, A.D. Systematic Reviews: Checklists for review articles. *British Medical Journal* 1994; 309:648-651.

9. Shea, B.J., Grimshaw, J.M., Wells, G.A., Boers, M., Andersson, N., Hamel, C. et al. Development of AMSTAR: a measurement tool to assess the methodological quality of systematic reviews. *BMC Medical Research Methodology* 2007; 7:10.

10. Oxman, A.D., Guyatt, G.H. Guidelines for reading literature reviews. *Canadian Medical Association Journal* 1988; 138:697-703.

11. Moher, D., Liberati, A., Tetzlaff, J., Altman, D.G., The PRISMA Group. Preferred Reporting Items for Systematic Reviews and Meta-Analyses: The PRISMA Statement. *BMJ* 2009; 339:b2535.

12. Guyatt, G.H., Oxman, A.D., Vist, G. GRADE: an emerging consensus on rating quality of evidence and strength of recommendations. *BMJ* 2008; 336:924-926.

13. Grades of Recommendation Assessment Development and Evaluation (GRADE) Working Group. Grading quality of evidence and strength of recommendations. *BMJ* 2004; 328:1490-1494.

14. Briggs, A., Claxton, K., Sculpher, M. Decision modelling for health economic evaluation. Oxford University Press, Oxford; 2008.

15. Spiegelhalter, D.J., Myles, J.P., Jones, D.R., Abrams, K.R. Bayesian methods in Health Technology Assessment: a review. *Health Technology Assessment* 2000; 4(38).

16. Donegan, S., Williamson, P., Gamble, C., Tudor-Smith, C. Indirect Comparisons: a review of reporting and methodological quality. *PLoS ONE* 2011; 5:e11054.

17. Caldwell, D.M., Ades, A.E., Higgins, J.P.T. Simultaneous comparison of multiple treatments: combining direct and indirect evidence. *BMJ* 2005; 331:897-900.

18. Lu, G., Ades, A. Assessing evidence consistency in mixed treatment comparisons. *Journal Of The American Statistical Association* 2006; 101:447-459.

19. Ades, A.E., Sculpher, M., Sutton, A., Abrams, K., Copper, N., Welton, N.J. et al. Bayesian methods for evidence synthesis in cost-effectiveness analysis. *Pharmacoeconomics* 2006; 24(1):1-19.

20. Song , F., Loke, Y.-K., Walsh, T., Glenny, A.-M., Eastwood, A.J., Altman, D.G. Methodological problems in the use of indirect comparisons for evaluating healthcare interventions: survey of published systematic reviews. *BMJ* 2009; 338(31):b1147.

21. O'Regan, C., Ghement, I., Eyawo, O., Guyatt, G.H., Mills, E.J. Incorporating multiple interventions in meta-analysis: an evaluation of the mixed treatment comparison with the adjusted indirect comparison. *Trials* 2009; 10(86): available from http://www.trialsjournal.com/content/10-1/86.

22. Higgins, J.P.T., Green, S. Cochrane Handbook for Systematic Reviews of Interventions Version 5.0.0 (updated February 2008). The Cochrane Collaboration, Wiley, Chichester; 2008.

23. Jansen, J.P., 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 in Health* 2011; 14:417-428.

24. Jansen, J.P., Fleurence, R., Devine, B., Itzler, R., Barrett, A., Hawkins, N. et al. Conducting Indirect Treatment Comparisons and Network Meta-Analysis Studies: Report of the ISPOR Task Force on Indirect Treatment Comparisons Good Research Practices: Part 2. *Value in Health* 2011; 14:429-437.

25. Centre for Reviews and Dissemination. Systematic Reviews: CRD's Guidance for Undertaking Reviews in Healthcare. Third ed. CRD, University of York, York; 2009.

26. Welton, N.J., Cooper, N.J., Ades, A.E., Lu, G., Sutton, A.J. Mixed treatment comparison with multiple outcomes reported inconsistently across trials: evaluation of

antivirals for treatment of influenza A and B. *Statistics In Medicine* 2008; 27:5620-5639.

27. Welton, N.J., Willis, S.R., Ades, A.E. Synthesis of Survival and Disease Progression Outcomes for Health Technology Assessment of Cancer Therapies. *Res Synth Method* 2010; 1:239-257.

28. Lu, G., Ades, A.E., Sutton, A.J., Cooper, N.J., Briggs, A.H., Caldwell, D.M. Meta-analysis of mixed treatment comparisons at multiple follow-up times. *Statistics In Medicine* 2007; 26(20):3681-3699.

29. Ades, A.E., Mavranezouli, I., Dias, S., Welton, N.J., Whittington, C., Kendall, T. Network meta-analysis with competing risk outcomes. *Value in Health* 2010; 13(8):976-983.

30. Hawkins, N., Scott, D.A., Woods, B. How far do you go? Efficient searching for indirect evidence. *Medical Decision Making* 2009; 29:273-281.

31. Cooper, N.J., Peters, J., Lai, M.C.W., Juni, P., Wandel, S., Palmer, S. et al. How valuable are multiple treatment comparison methods in evidence-based health-care evaluation? *Value in Health* 2011; 14:371-380.

32. Burch, J., Paulden, M., Conti, S., Stock, C., Corbette, M., Welton, N.J. et al. Antiviral drugs for the treatment of influenza: a systematic review and economic evaluation. *Health Technology Assesment* 2010; 13(58):1-290.

33. Sutton, A.J., Song, F., Gilbody, S., Abrams, K.R. Modelling publication bias in meta-analysis: a review. *Statistical Methods in Medical Research* 2000; 9:421-445.

34. Goodman, S., Berry, D., Wittes, J. Bias and Trials Stopped Early for Benefit [Letter]. *Journal of the American Medical Association* 2010; 304:157.

35. Goodman, S.N. Systematic reviews are not biased by results from trials stopped early for benefit [Letter]. *Journal of Clinical Epidemiology* 2008; 61:95-96.

36. Gotzsche, P.C. Why we need a broad perspective on meta-analysis. *British Medical Journal* 2000; 321:585-586.

37. Song, F., Altman, D., Glenny, A.-M., Deeks, J. Validity of indirect comparison for estimating efficacy of competing interventions: evidence from published meta-analyses. *British Medical Journal* 2003; 326:472-476.

38. Song , F., Xiong, T., Parekh-Bhurke, S., Loke, Y.K., Sutton, A.J., Eastwood, A.J. et al. Inconsistency between direct and indirect comparisons of competing interventions: meta-epidemiological study. *BMJ* 2011; 343:d4909.

39. Chou, R., Fu, R., Hoyt Huffman, L., Korthuis, P.T. Initial highly-active antiretroviral therapy with a protease inhibitor versus non-nucleoside reverse transcriptase inhibitor: discrepancies between direct and indirect meta-analyses. *Lancet* 2006; 368:1503-1515.

40. Caldwell, D.M., Gibb, D.M., Ades, A.E. Validity of indirect comparisons in meta-analysis.[Letter]. *Lancet* 2007; 369(9558):270.

41. Greenland, S., Longnecker, M.P. Methods for Trend Estimation from Summarized Dose-Response Data, with Applications to Meta-Analysis. *American Journal Of Epidemiology* 1992; 135:1301-1309.

42. Welton, N.J., Caldwell, D.M., Adamopoulos, E., Vedhara, K. Mixed Treatment Comparison Meta-analysis of Complex Interventions: Psychological interventions in coronary heart disease. *American Journal Of Epidemiology* 2009; 169(9):1158-1165.

43. Collett, D. Modelling survival data in medical research. Chapman & Hall, 1994.

44. Turner, R.M., Spiegelhalter, D.J., Smith, G.C.S., Thompson, S.G. Bias modelling in evidence synthesis. *Journal of the Royal Statistical Society (A)* 2009; 172:21-47.

45. Higgins, J.P.T., Altman, D.G. Assessing risk of bias in included studies. In: Higgins J., Green S., eds. *Cochrane Handbook for systematic reviews of interventions Version 5.0.1 (updated September 2008)*. The Cochrane Collaboration; 2008.

46. Schulz, K.F., Chalmers, I., Hayes, R.J., Altman, D.G. Empirical Evidence of Bias. Dimensions of Methodological Quality Associated With Estimates of Treatment Effects in Controlled Trials. *JAMA, J* 1995; 273(5):408-412.

47. Wood, L., Egger, M., Gluud, L.L., Schulz, K., Juni, P., Altman, D. et al. Empirical evidence of bias in treatment effect estimates in controlled trials with different interventions and outcomes: meta-epidemiological study. *British Medical Journal* 2008; 336:601-605.

48. Juni, P., Holenstein, F., Sterne, J., Bartlett, C., Egger, M. Direction and impact of language bias in meta-analyses of controlled trials: empirical study. *International Journal Of Epidemiology* 2002; 31(1):115-123.

49. Kirkham, J.J., Dwan, K.M., Altman, D., Gamble, C., Dodd, S., Smyth, R. et al. The impact of outcome reporting bias in randomised controlled trials on a cohort of systematic reviews. *BMJ* 2010; 340:c365.

50. Kjaergard, L.L., Villumsen, J., Gluud, C. Reported methodologic quality and discrepancies between large and small randomized trials in meta-analyses. *Annals of Internal Medicine* 2001; 135:982-989.

51. Moher, D., Pham, B., Jones, A., Cook, D.J., Jadad, A., Moher, M. et al. Does quality of reports of randomised trials affect estimates of intervention efficacy reported in meta-analyses? *Lancet* 1998; 352(9128):609-613.

52. Bradburn, M.J., Deeks J.J, Berlin, J.A., Localio, A.R. Much ado about nothing: a comparison of the performance of meta-analysis methods with rare events. *Statistics In Medicine* 2007; 26:53-77.

53. Sutton, A.J., Abrams, K.R., Jones, D.R., Sheldon, T.A., Song , F. Methods for meta-analysis in medical research. Wiley, London; 2000.

54. Sweeting, M.J., Sutton, A.J., Lambert, P.C. What to add to nothing? Use and avoidance of continuity corrections in meta-analysis of sparse data. *Statistics In Medicine* 2004; 23:1351-1375.

55. Higgins, J.P.T., Thompson, S.G., Deeks, J.J., Altman, D.G. Measuring inconsistency in meta-analyses. *BMJ* 2003; 327:557-560.

56. Spiegelhalter, D.J., Abrams, K.R., Myles, J. Bayesian approaches to clinical trials and Health-Care Evaluation. Wiley, New York; 2004.

57. Higgins, J.P.T., Thompson, S.G., Spiegelhalter, D.J. A re-evaluation of random-effects meta-analysis. *Journal of the Royal Statistical Society (A)* 2009; 172:137-159.

58. Ades, A.E., Sutton, A.J. Multiparameter evidence synthesis in epidemiology and medical decision making: current appoaches. *Statistics in Society, JRSS(A)* 2006; 169(1):5-35.

59. Salanti, G., Ades, A.E., Ioannidis, J.P.A. Graphical methods and numerical summaries for presenting results from multiple-treatment meta-analysis: an overview and tutorial. *Journal of Clinical Epidemiology* 2011; 64:163-171.

60. Cooper, N.J., Sutton, A.J., Lu, G., Khunti, K. Mixed comparison of stroke prevention treatments in individuals with nonrheumatic atrial fibrillation. *Archives of Internal Medicine* 2006; 166(12):1269-1275.

61. Cipriani, A., Furukawa, T.A., Salanti, G., Geddes, J.R., Higgins, J.P.T., Churchill, R. et al. Comparative efficacy and acceptability of 12 new generation antidepressants: a multiple-treatments meta-analysis. *Lancet* 2009; 373:746-758.

62. White, I.R. Multivariate random-effects meta-regression: Updates to mvmeta. *The Stata Journal* 2011; 11:255-270.

63. Jones, B., Roger, J., Lane, P.W., Lawton, A., Fletcher, C., Cappellen, J.C. et al. Statistical approaches for conducting network meta-analysis in drug development. *Pharmaceutical Statistics* 2011; 10:523-531.