UNIVERSITY OF IOANNINA
SCHOOL OF HEALTH SCIENCES
FACULTY OF MEDICINE
DEPARTMENT OF HYGIENE AND EPIDEMIOLOGY

STUDY OF THE HETEROGENEITY AND INCONSISTENCY IN NETWORKS OF MULTIPLE INTERVENTIONS

Veroniki Areti Angeliki
Mathematician

Dissertation Prepared for the Degree of

DOCTOR OF PHILOSOHPY

IOANNINA 2014

ΠΑΝΕΠΙΣΤΗΜΙΟ ΙΩΑΝΝΙΝΩΝ
ΣΧΟΛΗ ΕΠΙΣΤΗΜΩΝ ΥΓΕΙΑΣ
ΤΜΗΜΑ ΙΑΤΡΙΚΗΣ
ΕΡΓΑΣΤΗΡΙΟ ΥΓΙΕΙΝΗΣ ΚΑΙ ΕΠΙΔΗΜΙΟΛΟΓΙΑΣ

ΜΕΛΕΤΗ ΤΗΣ ΕΤΕΡΟΓΕΝΕΙΑΣ ΚΑΙ ΤΗΣ ΑΣΥΝΕΠΕΙΑΣ ΣΕ ΔΙΚΤΥΑ ΠΟΛΛΑΠΛΩΝ ΠΑΡΕΜΒΑΣΕΩΝ

Βερονίκη Αρετή Αγγελική
Μαθηματικός

ΔΙΔΑΚΤΟΡΙΚΗ ΔΙΑΤΡΙΒΗ

ΙΟΑΝΝΙΝΑ 2014

Η έγκριση της διδακτορικής διατριβής από την Ιατρική Σχολή του Πανεπιστημίου Ιωαννίνων δεν υποδηλώνει αποδοχή των γνώμων του συγγραφέα Ν. 5343/32, άρθρο 202, παράγραφος 2 (νομική καταχώρηση του Ιατρικού Τμήματος).

Ημερομηνία αίτησης της κ. Βερονίκης Αρετής Αγγελικής: 18-1-2011

Ημερομηνία ορισμού Τριμελούς Συμβουλευτικής Επιτροπής: 708α/8-3-2011

Μέλη Τριμελούς Συμβουλευτικής Επιτροπής:
Επιβλέπουσα
Ντζάνη Ευαγγελία Επίκουρη Καθηγήτρια Υγιεινής με έμφαση στην Επιδημιολογία του Τμήματος Ιατρικής της Σχολής Επιστημών Υγείας του Πανεπιστημίου Ιωαννίνων
Μέλη
Σαλαντή Γεωργία Λέκτορας Επιδημιολογίας του Τμήματος Ιατρικής της Σχολής Επιστημών Υγείας του Πανεπιστημίου Ιωαννίνων
Σπυρίδωνος Παναγιώτα Λέκτορας της Ιατρικής Πληροφορικής του Τμήματος Ιατρικής της Σχολής Επιστημών Υγείας του Πανεπιστημίου Ιωαννίνων

Ημερομηνία ορισμού θέματος: 4-5-2011
«Μελέτη της ετερογένειας και της ασυνέπειας σε δίκτυα πολλαπλών παρεμβάσεων»

Ημερομηνία 1ης ανασυγκρότησης της Τριμελούς Συμβουλευτικής Επιτροπής: 714α/7-6-2011
Επιβλέπουσα
Σαλαντή Γεωργία Επίκουρη Καθηγήτρια Επιδημιολογίας του Τμήματος Ιατρικής της Σχολής Επιστημών Υγείας του Πανεπιστημίου Ιωαννίνων
Μέλη
Ντζάνη Ευαγγελία Επίκουρη Καθηγήτρια Υγιεινής με έμφαση στην Επιδημιολογία του Τμήματος Ιατρικής της Σχολής Επιστημών Υγείας του Πανεπιστημίου Ιωαννίνων
Σπυρίδωνος Παναγιώτα Λέκτορας της Ιατρικής Πληροφορικής του Τμήματος Ιατρικής της Σχολής Επιστημών Υγείας του Πανεπιστημίου Ιωαννίνων

Ημερομηνία 2ης ανασυγκρότησης της Τριμελούς Συμβουλευτικής Επιτροπής: 721α/11-10-2011
Επιβλέπουσα
Σαλαντή Γεωργία Επίκουρη Καθηγήτρια Επιδημιολογίας του Τμήματος Ιατρικής της Σχολής Επιστημών Υγείας του Πανεπιστημίου Ιωαννίνων
Μέλη
Σπυρίδωνος Παναγιώτα Λέκτορας της Ιατρικής Πληροφορικής του Τμήματος Ιατρικής της Σχολής Επιστημών Υγείας του Πανεπιστημίου Ιωαννίνων
Julian Piers Thomas Higgins Senior Epidemiologist MRC, Ινστιτούτο Δημόσιας Υγείας του Cambridge

ΔΙΟΡΙΣΜΟΣ ΕΠΤΑΜΕΛΟΥΣ ΕΞΕΤΑΣΤΙΚΗΣ ΕΠΙΤΡΟΠΗΣ : 763α/10-4-2014

Δημολιάτης Ιωάννης	Επίκουρος Καθηγητής Υγιεινής του Τμήματος Ιατρικής του Πανεπιστημίου Ιωαννίνων
Σαλαντή Γεωργία	Επίκουρη Καθηγήτρια Επιδημιολογίας του Τμήματος Ιατρικής του Πανεπιστημίου Ιωαννίνων
Ευαγγέλου Ευάγγελο	Λέκτορας Υγιεινής με έμφαση στην Κλινική και Μοριακή

Βιολογία του Τμήματος Ιατρικής του Πανεπιστημίου Ιωαννίνων

Μαυρίδης Δημήτρης Λέκτορας του Παιδαγωγικού Τμήματος Δημοτικής Εκπαίδευσης του Πανεπιστημίου Ιωαννίνων

Σπυρίδωνος Παναγιώτα Λέκτορας Ιατρικής Πληροφορικής του Τμήματος Ιατρικής του Πανεπιστημίου Ιωαννίνων

Higgins Julian Piers Thomas Professor of Evidence Synthesis University of Bristol and University of York H.B.

Sharon Straus Professor of Clinical Epidemiology, Department of Medicine, University of Calgary and University of Toronto

Έγκριση Διδακτορικής Διατριβής με βαθμό **«ΑΡΙΣΤΑ»** στις **21-5-2014**

ΠΡΟΕΔΡΟΣ ΤΗΣ ΙΑΤΡΙΚΗΣ ΣΧΟΛΗΣ
Ανδρέας Φωτόπουλος
Καθηγητής Πυρηνικής Ιατρικής

Η Γραμματέας της Σχολής

Καπιτοπο

ΜΑΡΙΑ ΚΑΠΙΤΟΠΟΥΛΟΥ

ΙΑΤΡΙΚΗ · ΣΧΟΛΗ · ΠΑΝΕΠΙΣΤΗΜΙΟ ΙΩΑΝΝΙΝΩΝ
1977

Acknowledgements

I am deeply indebted to my supervisor Dr. Georgia Salanti for her excellent guidance, caring and wise advice through all these years. The support and guidance of my second supervisor, Prof. Julian Higgins, has been invaluable on the academic level, for which I am extremely grateful. I am very thankful to the European Research Council, which has supported my study. Thank you also to all my other colleagues and friends in School of Medicine, Department of Hygiene and Epidemiology at the University of Ioannina in Greece for having made this building very special to me. I would never have been able to finish my dissertation without the help and support from my family and my husband.

This dissertation is dedicated to the memory of my beloved father Charalambos Veronikis (1937-2005).

Table of Contents

1. Introduction

1.1 Background to the research

Meta-analysis is the statistical technique that pools data from several trials in an effort to increase power over the individual studies and in the hope of identifying patterns among trial results or potential sources of disagreement among those results. However, conventional meta-analytic models are restricted to pairwise comparisons, pooling results only from studies that compare two interventions. It is very common in health-care decision making to have more than two competing interventions. When multiple interventions have been developed to address the same problem, their relative effectiveness is a key concern for policy makers and practitioners who need to choose which specific intervention to implement. There might be studies comparing pair or triplets of interventions but it is impossible to find enough studies comparing all possible pairs of all the available interventions referred to the same condition. In comparative effectiveness reviews with more than two interventions, multiple separate and pairwise meta-analyses need to be conducted. This becomes confusing and taking into account the results from all the available comparisons would potentially lead to biased inferences (1). Moreover, pairwise meta-analysis would not answer which treatment is better when there are no studies directly comparing the treatments of interest. Network meta-analysis (NMA) addresses this problem by extending conventional meta-analytic models to enable comparisons between different sets of interventions to be combined in a single analysis.

When interventions are compared to a common treatment, e.g. placebo, it is possible to compare them indirectly via this common comparator. This methodology has been discussed in an early paper by Bucher et *al* (2). Suppose there are three treatments A, B and C and there are no studies comparing directly treatments B and C, but both of them are compared to a common comparator A. The true relative effects of the two treatments versus the common comparator, B *vs*. A and C *vs.* A, may contribute to make inference on the comparison C *vs.* B via the indirect comparison method. Although direct comparisons are better than indirect ones in terms of statistical power and mean squared error (3), studies have demonstrated that under certain circumstances the indirect comparison provides less biased estimates than pairwise meta-analysis (3,4). Mills et *al* (5) showed via a simulation study that indirect comparisons have low power when heterogeneity is moderate to large.

Network meta-analysis is used to combine the results of clinical trials that undertake different comparisons of treatments. When various treatment comparisons are connected in a network these can be presented in a network diagram as long as every pairwise comparison belongs to a chain that connects all treatments. A network of treatments should be connected in the sense that at least one comparison or path exists between two interventions in the network. If the comparisons from primary studies do not form a connected network, then NMA cannot be applied. To visualise the available evidence graphical representations can be used via network diagrams. The NMA methodology synthesizes simultaneously evidence from a network of studies involving multiple treatments. The relative effectiveness of each pair of interventions can then be estimated, regardless of whether the two interventions were directly compared in any of the primary studies. NMA is more advantageous of pairwise meta-analysis and indirect comparison as we can draw inferences for the comparability between interventions not directly studied in an individual clinical trial, and we are given the option to rank the interventions according to their efficacy.

The ever-increasing number of alternative treatment options and the plethora of clinical trials have increased the use of NMA the last fifteen years (6). Despite the advantages of network meta-analysis, it is not yet a widely established approach of evidence synthesis in the literature. Many review authors are sceptical towards the assumptions of the method (7). The statistical expertise required in fitting the model, the presentation of complex results in an understandable format and the evaluation of the risk of bias and the quality of evidence in the network meta-analysis are far more challenging than in conventional meta-analysis.

The potential utility of network meta-analysis and the generalisation of its results rest on the validity of the required assumptions (8). As in the case of conventional meta-analysis, the validity of NMA findings rests on the randomization process of the included RCTs and preserves this within-study randomization (1,9). A possible violation of randomization would arise if participants of treatment arm A of AB trial were directly compared with those in arm C of BC trial in order to estimate the relative effect of A *vs.* C. It is not valid to compare individuals in different studies. The NMA model respects that participants are randomized to interventions within a trial and not across trials (1,10,11).

Even if conventional meta-analysis and NMA use well-conducted randomized trials, threats to the validity of the homogeneity assumption might arise. The assumption of homogeneity is violated when there is excessive discrepancy among the study-specific treatment effects in the sense that the discrepancy is greater than what one would expect due to random error. A certain degree of variability in study estimates is almost always present due to chance. Additional variability might occur due to many reasons such as differences in the way studies are conducted and how the estimates are measured; this additional variability is often termed heterogeneity. The estimation of the heterogeneity is of interest and provides insight in the interpretation of the meta-analysis' results. Several methods have been suggested to quantify or express heterogeneity the properties of which differ under several circumstances.

A key assumption in NMA is that the trials are exchangeable in the sense that they are sufficiently similar regarding particular characteristics of the studies (1,10,12). The ability to learn about a pairwise comparison via an intermediate treatment and make a valid indirect comparison constitutes transitivity. The transitivity assumption is comparable to the homogeneity assumption to a clinical and methodological way. In order that the transitivity assumption holds the similarity of the distribution of the effect modifiers across comparisons is required. The effect modifiers are study-level characteristics that influence the relative effects of the interventions being compared. The ideal evidence would be to use large, multi-arm trials that randomly allocate participants to all eligible interventions. Multi-arm trials are by definition consistent and in case a treatment effect estimate for one comparison is missing then it can be calculated from the remaining estimates. The transitivity assumption implies that participants in the network could theoretically have been randomized to any of the treatments in the network.

Some study-level characteristics vary across studies which is inevitable. These variations can include, for example, the way in which an intervention is defined and delivered or participant characteristics. The transitivity assumption holds when a) the treatments are equivalent in the sense that they are given for the same condition, b) studies are sufficiently similar in the sense that the effect modifiers are equally distributed, and c) the missing arm in a study is missing at random suggesting that these arms are only unobserved. Transitivity assumes that observed and unobserved estimates do not differ beyond what can be explained by heterogeneity.

Lack of transitivity in NMA can question the consistency of the underlying estimates and the reliability of the results. It is therefore crucial to evaluate the consistency assumption between the different sources of evidence before analysing them jointly. Consistency refers to the extension of transitivity in network estimates, where direct and indirect estimates obtained for the same comparison are in agreement (12,13). Disagreement between direct and indirect estimates is called inconsistency. Note that the distinction between transitivity and consistency is analogous to the one between methodological and statistical heterogeneity in pairwise meta-analysis. Similarly to the assumption of homogeneity, the assumption of consistency is violated in the presence of important discrepancy, beyond what is expected by sampling error, between the overall treatment effects of the different sources of evidence.

Consistency is a property of closed loops within networks, i.e. the paths that begin from an intervention node and end to the same node via two or more intermediate interventions, as well as entire networks. Consistency in individual loops can be measured by testing for statistical differences between direct and indirect estimates. The assumption of consistency can be statistically evaluated with several approaches in either certain parts of the network (e.g. separating direct and indirect evidence (SIDE), loop-specific (LS), back-calculation) or in the entire network (e.g. Lumley model, Lu and Ades (LA) model, design-by-treatment interaction (DBT) model, comparing the model fit and parsimony from consistency and inconsistency models) (9,13–18). Consistency should always be statistically assessed and reported when network meta-analysis is used. However, statistical tests are underpowered and high levels of heterogeneity can mask inconsistency (12,13,16,19,20). A large heterogeneity in the treatment effects leads to greater uncertainty in estimates of the mean effect sizes, and statistical inconsistency is less likely to be detected. Finding no statistical evidence of inconsistency does not necessarily imply that a network is consistent or that the transitivity assumption is valid.

The ability to detect inconsistency might depend on the estimation of the amount of heterogeneity which can vary using different methods (e.g. DerSimonian and Laird, restricted maximum likelihood etc.) (21). Similarly, different assumptions for the heterogeneity, being the same or different in different parts of the same loop or the network of evidence, may impact on the detection of inconsistency. Inconsistency can be possibly affected by the use of different effect measures. Empirical evidence suggests that ratio measures are more homogeneous than absolute effect measures (22,23). These differences depend on the extent of variation in baseline risk across studies. If these are substantially different in different parts of a loop, then inconsistency may be greater for some effect measures than others; if baseline risks vary substantially within each comparison, then more or less heterogeneity may be present (22). Although there are strong indications that the presence, magnitude and estimation method of heterogeneity might influence the detection of inconsistency, this association has not been studied extensively. For instance, the impact of two alternative methods to express uncertainty about the pairwise summary effects (Wald type (Wt) and Knapp-Hartung (KH) method (24,25)) remains unclear. It has been shown that the KH method outperforms Wt and that it is insensitive to the estimator of the heterogeneity used (24,25). I anticipate that differences in the properties of the two methods will impact on the estimation of inconsistency.

If the required assumptions for NMA are violated the results of a network meta-analysis can be biased. Despite its importance, investigators commonly combine direct and indirect evidence without evaluating the validity of the consistency assumption. A recent survey showed that only 14% of the authors applying NMA have evaluated the assumption of consistency, the 24% of whom have used inappropriate approaches (e.g. comparison of direct estimates with NMA estimates) to evaluate consistency (6). Several reviews evaluating NMAs and the validity of the prerequisite assumptions, highlighted the importance of assessing and reporting the methods applied (4,6,14,26–29). Thus, there is an urgent need to improve the quality of published NMAs with respect to the uptake, application and reporting of methods to evaluate inconsistency. The poor quality might also highlight that the methods for NMA are in development and there is a lack of agreement on the methods that should be employed.

1.2 Justification for the research

The aim of this thesis is to evaluate the prevalence of inconsistency and the importance of statistical considerations that might influence its detection. I explore the role of factors that may impact on inferences about inconsistency. The factors that I explore are associated with the amount of data available in the loop (e.g. number, size and distribution of trials across comparisons, frequency of events), heterogeneity in the pairwise comparisons (magnitude and estimation method) and the method for inference about the uncertainty in pairwise summary effects. I examine whether the different effect measures for dichotomous outcome data are associated with differences in inconsistency, and I evaluate whether different approaches to evaluate inconsistency impact on inferences on the prevalence and magnitude of inconsistency. I evaluate inconsistency in 40 complex networks of interventions (involving 303 closed loops) with dichotomous outcome data, at least four treatments and at least one closed loop. I also conduct a simulation study considering realistic scenarios and I estimate the properties (type I error, power and coverage probability) for the test of consistency. The simulation scenarios are informed by the previous empirical study with the large collection of 303 loops from published networks of interventions (13), and a study about the empirical distribution of heterogeneity on dichotomous outcomes (30).

2. Heterogeneity and uncertainty in meta-analysis

2.1 Introduction

Meta-analysis is a technique that pools data from several trials and returns an overall estimate of treatment effect size. It requires the studies whose data are pooled to be ‘similar’ in design and to provide sufficient information for computing estimates. A certain degree of variability in study estimates is almost always present due to chance. Additional variability might occur due to many reasons such as differences in the way studies are conducted and how the estimates are measured; this additional variability is often termed heterogeneity. There are three different types of heterogeneity: i) clinical heterogeneity, which is referred to the variability in the participants, interventions, and outcomes, ii) methodological heterogeneity, which reflects the variability in study design and risk of bias, and iii) statistical heterogeneity, which is referred to the variability in the intervention effects. In the next sections, I will refer to the statistical heterogeneity, which is a consequence of clinical or methodological variability, or both, among trials, as heterogeneity.

Heterogeneity refers to the variation across study-findings beyond random error and its quantification is often of interest and improves the interpretation of results of a meta-analysis. One of the most widely statistical methods used for meta-analysis is the inverse variance method which uses the reciprocal of the within-study variance as weight. The magnitude of the heterogeneity impacts on the estimation of the weights assigned to each study and hence to the estimated variance of the overall treatment effect.

Several methods have been proposed to estimate the heterogeneity variance (τ^2) and they vary in popularity and complexity. The estimators for τ^2 are categorised to closed form (non-iterative) and iterative methods. In contrast to closed form estimators, the iterative methods require checking for convergence and run the risk of estimating τ^2 erroneously or failing to converge to a solution. The estimators can be generally categorised to the method of moments approaches (e.g. DerSimonian and Laird (DL) (31), and Paule and Mantel (PM) (32) methods), the maximum likelihood estimators (e.g. maximum likelihood (ML) (33,34), and restricted maximum likelihood (REML) (33) methods), the weighted least squares estimators (e.g. Sidik and Jonkman (SJ) (35) method), and the Bayes estimators (e.g. full Bayes (36) method). Veroniki et *al.* (37) describe in a recent review all the existing methods in detail.

The uncertainty around the summary treatment effect can be estimated using a wide variety of methods. The most popular categories of the confidence intervals (CIs) for the summary treatment effect are the asymptotically normal-based CIs (e.g. Wald type (Wt) (31)), the likelihood-based CIs (e.g. profile likelihood (34)), the CIs based on *t*-distribution (e.g. Knapp and Hartung (KH) (38)), the quantile approximation CIs (39), and the Henmi and Copas CIs (40). For a comparison of the methods see Sánchez-Meca and Marín-Martínez, (25).

In this chapter I start with a short description of the statistical models for combining studies in meta-analysis and the properties a good estimator should be associated with. Then I present the most popular estimation methods for the heterogeneity and the uncertainty of the summary treatment effect, which I also use in the empirical and simulation studies for the evaluation of inconsistency (see chapter 4).

2.2 Models for meta-analysis

The main two models used to pool study results in the meta-analysis are the fixed-effect (FE) and the random-effects (RE) models. The FE model assumes that all studies share the same (fixed) true effect, i.e. there is one 'true effect' size and all differences in observed effects are due to sampling error. In the RE model the effect sizes observed in the studies represent a random sample from a particular distribution of the underlying treatment effects. They are distributed around a mean with the width of the distribution describing the degree of heterogeneity. The CIs around the summary effect obtained from a RE meta-analysis describe the uncertainty in the location of the mean effect and its width depends on the magnitude of the heterogeneity variance, the number of studies and the precision of the individual study estimates (41). A RE model takes into account both within-study (v_i) and between-study (τ^2) variation, in contrast to the FE model that accounts for within-study variation only. It follows that in the presence of heterogeneity ($\tau^2 > 0$) the RE model results in a wider CI compared to the FE model reflecting greater uncertainty around the mean (42). When the heterogeneity equals to zero the RE model simplifies to the FE model.

Let y_i be the observed relative treatment effect (e.g. log-odds ratio (LOR)) in study $i = 1,..k$ with v_i its respective within-study variance, μ_{FE} the common mean under the FE model and ε_i the random error in study i.

$$y_i = \mu_{FE} + \varepsilon_i$$

$$\varepsilon_i \sim N(0, v_i)$$

The summary treatment effect is estimated as

$$\hat{\mu}_{FE} = \frac{\sum_{i=1}^{k} y_i w_{i,FE}}{\sum_{i=1}^{k} w_{i,FE}} \qquad (2.1)$$

where $w_{i,FE} = 1/v_i$ is the weight assigned to each study under the FE model.

Under the RE model μ_{RE} is the mean of the distribution of the underlying effects, δ_i is the difference between the mean μ_{RE} and the underlying study-specific mean θ_i, and τ^2 is the variance of the random effects distribution.

$$y_i = \theta_i + \varepsilon_i$$

$$\theta_i = \mu_{RE} + \delta_i$$

$$\varepsilon_i \sim N(0, v_i)$$

$$\delta_i \sim N(0, \tau^2)$$

The estimated summary treatment effect $\hat{\mu}_{RE}$ is computed as in (2.1) using the weights under the RE model, $w_{i,RE} = 1/(v_i + \tau^2)$. In the next sections I will use the notation $\hat{\mu}_{RE}(\hat{\tau}^2)$ to denote that the overall treatment effect depends on the estimated heterogeneity.

Both models are structured assuming that the within-study variances, v_i are known. Since v_i are only estimated from the observed study data the distributions of the test statistics discussed in the following sections are only approximated for large v_i values.

2.3 Estimating the heterogeneity variance

2.3.1 Properties of a good estimation method

A good estimator should a) be unbiased, and b) have low mean square error (MSE).

Bias is the difference between the expected value of the estimator (or the mean of the estimator) and its true value and is given by

$$Bias(\hat{\tau}^2) = E(\hat{\tau}^2) - \tau^2 = E(\hat{\tau}^2 - \tau^2).$$

Negatively or positively biased estimators lead to an under- or over-estimation of the true heterogeneity variance, respectively. A good estimator should not only be unbiased, but also remain unaffected as much as possible by sampling fluctuation (efficiency). The MSE is commonly used as an efficiency measure and represents the squared distance between the estimator and its true value:

$$MSE(\hat{\tau}^2) = E[(\hat{\tau}^2 - \tau^2)^2] = Var(\hat{\tau}^2) + \left(Bias(\hat{\tau}^2)\right)^2$$

where $$Var(\hat{\tau}^2) = E\left[\left(\hat{\tau}^2 - E(\hat{\tau}^2)\right)^2\right].$$

If $\hat{\tau}^2$ is an unbiased estimator of τ^2 $(E(\hat{\tau}^2) = \tau^2)$, then the variance of this estimator is bounded as $Var(\hat{\tau}^2) \geq \left(I_F(\tau^2)\right)^{-1}$ under the Cramer-Rao inequality, with $I_F(\tau^2)$ the Fisher information. The efficiency of an unbiased estimator is defined by (43–45):

$$e(\hat{\tau}^2) = \frac{\left(I_F(\tau^2)\right)^{-1}}{Var(\hat{\tau}^2)}.$$

A good estimator has $e(\hat{\tau}^2)$ close to unity, with low variance. The efficiency of an estimator is measured relative to other estimators and is called ‘relative efficiency’. Consider for example two estimation methods that yield $\hat{\tau}_1^2$ and $\hat{\tau}_2^2$. If $Var(\hat{\tau}_1^2) < Var(\hat{\tau}_2^2)$ then $\hat{\tau}_1^2$ is said to be more efficient than $\hat{\tau}_2^2$, and the relative efficiency, $e(\hat{\tau}_1^2, \hat{\tau}_2^2)$, of the two unbiased estimators lies within the interval [0,1]. The relative efficiency $e(\hat{\tau}_1^2, \hat{\tau}_2^2)$ is generally defined as:

$$e(\hat{\tau}_1^2, \hat{\tau}_2^2) = \frac{MSE(\hat{\tau}_1^2)}{MSE(\hat{\tau}_2^2)}.$$

For unbiased estimators $MSE(\hat{\tau}^2) = Var(\hat{\tau}^2)$ and the $e(\hat{\tau}_1^2, \hat{\tau}_2^2)$ simplifies to the ratio of the two variances. The MSE of the estimator is minimised relative to other estimators. Good estimators should also have small type I error (rejecting the null hypothesis $H_0\colon \tau^2 = 0$ when it is true) and high power (rejecting the null hypothesis when $\tau^2 \neq 0$).

2.3.2 Estimation methods for the heterogeneity variance

DerSimonian and Laird (DL) method

The DL estimator is possibly the most frequently used as it is a closed form and simple to implement method (31). The DL estimator can be obtained as:

$$\hat{\tau}_{DL}^2 = \mathbf{max}\left\{0, \frac{Q-(k-1)}{\sum_{i=1}^{k} w_{i,FE} - \frac{\sum_{i=1}^{k} w_{i,FE}^2}{\sum_{i=1}^{k} w_{i,FE}}}\right\}$$

where $Q = \sum_{i=1}^{k} w_{i,FE}(y_i - \hat{\mu}_{FE})^2$. The estimated values of the heterogeneity might either be negative setting therefore $\hat{\tau}_{DL}^2$ equal to zero, or might be non-negative keeping the same non-truncated value. While the method before truncation is unbiased under the assumptions of the RE model, it has been shown that after truncating negative values to zero it might produce biased estimators (46,47). Hence, under the assumptions of the RE model the DL estimator might be positively biased over-estimating the true heterogeneity. Bias inflates not only due to truncation, but also because weights are calculated from within-study variances that are assumed fixed and known. As the size and number of studies included in the meta-analysis decreases and the sampling variances increase the DL estimator becomes more variable and truncation is more prevalent increasing bias. On the other hand, as τ^2 deviates zero the truncation bias decreases since the possibility of finding a negative $\hat{\tau}_{DL}^2$ decrease.

Although $\hat{\tau}_{DL}^2$ must be positively biased due to truncation, simulation studies suggest that the DL method performs well for small or close to zero heterogeneity and large k (21,46). It should be noted that the magnitude and direction of bias of the estimator depends on the selection of the simulation scenarios. A general conclusion from the published studies is that the estimation method underestimates the true heterogeneity when it is large and particularly when the size and number of studies is relatively small (21,35,46,48). This results in poor control of type I error and low coverage probability of the underlying CIs (25,39,49). The method is associated with lower MSE than the SJ and PM estimators in meta-analyses where the true heterogeneity is not too large (21). Jackson et *al.* (50) suggested that the DL estimator is inefficient when the studies included in the meta-analysis are of different sizes and particularly for large τ^2.

Paule-Mandel (PM) method

Paule and Mandel (32) proposed this method (PM) to estimate τ^2 by iterating the generalised Q-statistic

$$Q_{gen} = \sum_{i=1}^{k} w_{i,RE}\left(y_i - \hat{\mu}_{RE}(\tau^2)\right)^2 \sim \chi_{k-1}^2 \qquad (2.1)$$

until Q_{gen} equals to its expected value ($E\left(Q_{gen}\right) = k - 1$) (47). The method is also known as empirical Bayes estimator and has been discussed by Morris (51) with $w_{i,RE} = 1/(\tau^2 + v_i)$. The process requires at each iteration step non-negative values, otherwise $\hat{\tau}_{PM}^2$ is set equal to zero, and guaranties one solution of $\hat{\tau}_{PM}^2$ (52). When the normality assumption does not hold it has been shown that the PM method is more robust for the estimation of τ^2 in contrast to the DL estimator that depends on large studies (52). The method mirrors the REML estimation when the normality assumption holds (53). Many authors recommend the use of $\hat{\tau}_{PM}^2$ because of its good properties (37,48,52).

An empirical study (48) showed that as heterogeneity increases $\hat{\tau}_{PM}^2$ becomes greater than $\hat{\tau}_{DL}^2$. It has been suggested that the PM estimator is nearly unbiased for large number (k ≥30) and size (arm sample size larger than 100 participants) of studies, and that performs best in terms of bias among the DL, REML and PM methods (54). Sidik and Jonkman (21) noted the methodological similarity between the SJ and PM estimators, and stated that differences between the two estimates can largely be accounted for by the fact that the SJ estimator is simplified to two-steps and that yields always positive heterogeneity estimates. Generally the PM estimator has similar MSE with the SJ method (21). More specifically when the heterogeneity variance is small the PM estimator has slightly smaller MSE than the SJ method, whereas for large heterogeneity the PM estimator has slightly larger MSE than the SJ method. It has been shown that the PM method upwards bias for small k and τ^2, whereas for large k and τ^2 downwards bias (21,55). Knapp and Hartung (38) in a comparison of the DL, REML and PM methods, found that the PM estimator is less efficient than the DL and REML methods, and that it does not perform well for small k.

Maximum Likelihood (ML) method

The ML method is asymptotically efficient but requires an iterative solution (33,34). Setting the marginal distribution $y_i \sim N(\mu, v_i + \tau^2)$ the estimate $\hat{\tau}_{ML}^2$ is produced by the log–likelihood function

$$lnL(\mu,\tau^2) = -\frac{k}{2}ln(2\pi) - \frac{1}{2}\sum_{\iota=1}^{\kappa} ln(v_i + \tau^2) - \frac{1}{2}\sum_{\iota=1}^{\kappa} \frac{(y_i-\mu)^2}{(v_i+\tau^2)}.$$

Setting partial derivatives with respect to μ and τ^2 equal to zero and solving the likelihood equations for the two parameters to be estimated, the ML estimators for μ and τ^2 can be obtained by

$$\hat{\mu}_{RE}(\hat{\tau}_{ML}^2) = \frac{\sum_{i=1}^{k} w_{i,RE} y_i}{\sum_{i=1}^{k} w_{i,RE}}$$

$$\hat{\tau}_{ML}^2 = \max\left\{0, \frac{\sum_{i=1}^{k} w_{i,RE}^2 \left(\left(y_i - \hat{\mu}_{RE}(\hat{\tau}_{ML}^2)\right)^2 - v_i\right)}{\sum_{i=1}^{k} w_{i,RE}^2}\right\}$$

where $w_{i,RE} = 1/(v_i + \hat{\tau}_{ML}^2)$. An initial estimate of $\hat{\tau}_{ML}^2$ can be decided a priori as a plausible value of the heterogeneity variance or it can be estimated with any other estimator or even it can be set zero. The ML estimates are obtained by iterating both $\hat{\tau}_{ML}^2$ and $\hat{\mu}_{RE}(\hat{\tau}_{ML}^2)$ until convergence. Each iteration step requires non-negativity and sets negative estimated values equal to zero. Instead of the underlying procedure, the maximisation of the likelihood can be performed using various techniques (e.g. Newton-Raphson method, the simplex method etc.) with different convergence properties. It should be noted that likelihood based methods are asymptotically unbiased with variance approaching the Cramer-Rao lower bound. Hence, when k is large the maximum likelihood estimators are fully efficient.

Simulation studies suggest that although the ML estimator has lower MSE across all values of k and τ^2 than the DL and REML methods, the estimator exhibits a large amount of negative bias as τ^2 increases when k and sample size are small to moderate (21,41,46). Within-study variances v_i are assumed to be known which may account for the negative bias (21). The estimator also assumes effect estimates are normally distributed and there is currently little evidence to suggest how the ML method performs under non-normal conditions. It has been shown that the ML method has the smallest MSE in comparison to the REML, SJ, and PM methods, but exhibits the largest amount of bias among them (21,41). It is suggested to avoid applying the ML estimator due to substantial bias (46,54).

Restricted Maximum Likelihood (REML) method

The REML method is a well-known estimation technique and is produced by the log–likelihood function (51)

$$lnL(\tau^2) = -\frac{k}{2}ln(2\pi) - \frac{1}{2}\sum_{\iota=1}^{\kappa} ln(v_i + \tau^2) - \frac{1}{2}\sum_{\iota=1}^{\kappa} \frac{\left(y_i - \hat{\mu}_{RE}(\hat{\tau}_{ML}^2)\right)^2}{(v_i + \tau^2)} - \frac{1}{2}ln\left(\sum_{\iota=1}^{\kappa} \frac{1}{(v_i + \tau^2)}\right),$$

by setting the partial derivative with respect to τ^2 equal to zero. The REML estimator can be obtained by:

$$\hat{\tau}_{REML}^2 = \max\left\{0, \frac{\sum_{i=1}^{k} w_{i,RE}^2\left(\left(y_i - \hat{\mu}_{RE}(\hat{\tau}_{REML}^2)\right)^2 - v_i\right)}{\sum_{i=1}^{k} w_{i,RE}^2} + \frac{1}{\sum_{\iota=1}^{\kappa} w_{i,RE}}\right\},$$

where $w_{i,RE} = 1/(v_i + \hat{\tau}_{REML}^2)$ (21,31). Similarly, to ML method the REML estimator is calculated by an iterative process that requires non-negativity at each iteration step with a closed form initial estimate.

Simulation studies suggest that the REML method underestimates τ^2 especially when the data is sparse and in such cases should probably be avoided (21,46). Is has been shown that the method is less downwardly biased than the DL and ML estimators, but has greater MSE (21,39). Viechtbauer (46) showed that the REML method is the best approach compared to other methods, including DL and ML estimators, when large studies are included in the meta-analysis. Knapp and Hartung (38) in a comparison of the DL, REML and PM methods, found that the REML estimation is more efficient than the DL and PM methids. Jackson et *al.* (50) investigated the asymptotic efficiency of the DL, ML and REML methods and showed that for large τ^2 the DL and REML estimators are more efficient. Although most simulation studies have shown that the REML estimation produces superior results to the DL method, an empirical study including 920 Cochrane reviews with dichotomous outcome data and meta-analyses including at least three studies has shown that the magnitude of the REML estimator can be smaller or larger than the DL method (56).

Sidik-Jonkman (SJ) method

Sidik and Jonkman (35) introduced a non-iterative estimation method of the heterogeneity variance based on weighted least squares. To obtain the SJ estimator (known also as model error variance estimator) we first calculate the values $\hat{q}_i = \hat{r}_i + 1$ with

$\hat{r}_i = v_i/\hat{\tau}_0^2$ (assuming $\hat{\tau}_0^2 \neq 0$) and $\hat{\tau}_0^2 = \sum_{i=1}^{k}(y_i - \bar{y})^2/k$. The SJ estimation method can be derived by setting the quantity $\sum_{i=1}^{k}\hat{q}_i^{-1}\left(y_i - \hat{\mu}_{\hat{q},RE}\right)^2$ equal to its expected value and obtain

$$\hat{\tau}_{SJ}^2 = \frac{1}{k-1}\sum_{i=1}^{k}\hat{q}_i^{-1}\left(y_i - \hat{\mu}_{\hat{q},RE}\right)^2, \qquad (2.2)$$

where $\hat{\tau}_0^2$ is an initial estimate of the heterogeneity variance and $\hat{\mu}_{\hat{q},RE} = \sum_{i=1}^{k}\hat{q}_i^{-1}y_i / \sum_{i=1}^{k}\hat{q}_i^{-1}$ is the weighted random-effects pooled estimate. The method always yields a positive estimate of the true heterogeneity variance.

The SJ estimator as already mentioned above has methodological similarities with the PM method. Weights assigned to each study for $\hat{\tau}_{SJ}^2$ can be re-expressed as $\hat{q}_i = \hat{r}_i + 1 = \hat{\tau}_0^2(v_i + \hat{\tau}_0^2)^{-1}$, i.e. random effect weights as in the PM estimator multiplied by the constant term $\hat{\tau}_0^2$. Thus, if the initial estimate of $\hat{\tau}_{SJ}^2$ was defined as $\hat{\tau}_0^2 = \hat{\tau}_{SJ}^2$ in the above equation for $\hat{\tau}_{SJ}^2$ (2.2), it can be rearranged into the form $E\left(Q_{gen}\right) = k - 1$ which is identical to how the PM estimator is derived. In practice, the SJ method differs from the PM estimator in being always positive and non-iterative.

Simulation studies suggest that the SJ estimation method has smaller MSE and substantially smaller bias than the DL estimator for large values of k and τ^2, whereas the opposite occurs when k and τ^2 are small (35). Simulation studies have also suggested that the SJ estimation method has the largest bias among the DL, ML, REML, and PM methods for relatively small values of τ^2, whereas the magnitude of bias relative to the other estimators tends to decrease as τ^2 increases (21,54). For large τ^2 the SJ and PM methods are the best estimators in terms of bias according to Sidik and Jonkman (21). In agreement to most simulation studies, an empirical study (56) showed that the SJ estimator produces larger values than the DL method.

2.3.3 Assumptions made for the heterogeneity in an NMA model

In a pairwise meta-analysis we can either assume that trials estimate a single underlying effect size (FE model) or that the study-specific underlying effect sizes are different but drawn from the same distribution (RE model) with heterogeneity τ^2. Let τ_{XY}^2 be the heterogeneity in the Y *vs.* X comparison. Consider the network defined by two triangular loops, ABC and BCD, informed by AB, AC, BC, BD and CD comparisons. An assumption is that all studies share the same (fixed) parameter, i.e. there is one true effect size for each comparison and all differences in observed effects are due to sampling error with $\tau^2_{\text{AB}} = \tau^2_{\text{AC}} = \tau^2_{\text{BC}} = \tau^2_{\text{BD}} = \tau^2_{\text{CD}} = \tau^2 = 0$, where τ^2_{AB}, τ^2_{AC}, τ^2_{BC}, τ^2_{BD}, and τ^2_{CD} are the heterogeneity variances in the B *vs.* A , C *vs.* A, C *vs.* B, B *vs.* D, and C *vs.* D comparisons, respectively. Alternatively, we might assume that the true effect sizes differ implying that the parameters of the underlying studies follow some distribution. Under the latter scenario, it is common to assume that heterogeneity is the same for all comparisons being made $\tau^2_{\text{AB}} = \tau^2_{\text{AC}} = \tau^2_{\text{BC}} = \tau^2_{\text{BD}} = \tau^2_{\text{CD}} = \tau^2 = \tau^2_{ntw}$ and call it common within-network heterogeneity (τ^2_{ntw}) assumption. In case each comparison in the network is informed by a single study τ^2_{ntw} is set zero. Another assumption would be to allow all comparisons to have a different amount of the heterogeneity $\tau^2_{\text{AB}} \neq \tau^2_{\text{AC}} \neq \tau^2_{\text{BC}} \neq \tau^2_{\text{BD}} \neq \tau^2_{\text{CD}}$, but the consistency of NMA structure imposes some implicit constraints on the variances and covariances of the random effects. Lu and Ades (57) discuss further these issues and propose an approach to model the heterogeneity structure that is compatible with the consistency assumptions. Finally, a frequent assumption is the common within-loop heterogeneity (τ^2_{loop}) in which all comparisons in a particular loop have the same amount of heterogeneity; ABC loop: $\tau^2_{\text{AB}} = \tau^2_{\text{AC}} = \tau^2_{\text{BC}} = \tau^2_{loop,1}$, BCD loop: $\tau^2_{\text{BC}} = \tau^2_{\text{BD}} = \tau^2_{\text{CD}} = \tau^2_{loop,2}$. Assuming a common within-loop heterogeneity allows comparisons that have been addressed by only one study to 'borrow strength' from the rest of the comparisons included in the loop. When all comparisons involved in a loop are informed by a single study then τ^2_{loop} equals zero. It should be noted that τ^2_{loop} may be different for the same comparison when it is involved in different loops. The common within-loop heterogeneity assumption can be made simultaneously within one analysis and is only reasonable in separate, loop-specific, analyses (see section 3.5.1).

2.4 Estimating the uncertainty for the summary treatment effect

Apart from estimating the true summary treatment effect using a variety of methods it is also important to quantify the uncertainty for the estimate, i.e. the CI for μ. The CIs produced by different methods are often compared in terms of a) coverage probability and b) width.

The coverage probability of a CI for μ is the proportion of times the interval comprises the true overall treatment effect, $P(\mu \in CI)$. A good CI for a certain level of confidence allows a small 'room' for μ to vary. The greater the precision of μ the narrower the CI. The most frequently encountered CIs to quantify the extent of the summary treatment effect are described below.

2.4.1 Confidence Intervals for the summary treatment effect

Wald-type (Wt) Confidence Intervals

The Wt CI (31) is the most popular technique for establishing CIs for μ. Assuming $w_{i,RE} = 1/(v_i + \hat{\tau}^2)$ and $var(\hat{\mu}_{RE}(\hat{\tau}^2)) = 1/\sum w_{i,RE}$, a 95% CI is given by

$$\hat{\mu}_{RE}(\hat{\tau}^2) \pm 1.96\sqrt{var\big(\hat{\mu}_{RE}(\hat{\tau}^2)\big)}.$$

The method has coverage probability considerably below 95%, unless a large number of studies is included in the meta-analysis with a large study size and low or zero heterogeneity (25,39,49). A simulation study examined the performance of the method using a variety of estimators, including the DL, REML and SJ methods, and showed that Wt depends on the estimator employed (25). The method performs poorly for small samples ($k < 16$) (50). Normand (58) suggests the use of the method with the REML estimator so as to take into account the loss in degrees of freedom caused by the estimation of μ.

Knapp-Hartung (KH) Confidence Intervals

The KH method was proposed by Hartung (59) and later discussed by Knapp and Hartung (38), which relies on a *t*-distribution with $k-1$ degrees of freedom and assumes a weighted variance of μ. Sidik and Jonkman (24) independently developed this approach as well. Hartung (59) showed that the approximate distribution of the S statistic is

$$S = \frac{\hat{\mu}_{RE}(\hat{\tau}^2) - \mu}{\sqrt{var\big(\hat{\mu}_{RE}(\hat{\tau}^2)\big)}} \sim t_{k-1}$$

where $$var\big(\hat{\mu}_{RE}(\hat{\tau}^2)\big) = Q_{gen}\frac{1}{(k-1)\sum_{i=1}^{k} w_{i,RE}}$$

with Q_{gen} the generalised Q statistic as defined in (2.1). The approximate 95% CI for μ is given by

$$\hat{\mu}_{RE}(\hat{\tau}^2) \pm t_{k-1,a/2}\sqrt{var\big(\hat{\mu}_{RE}(\hat{\tau}^2)\big)}.$$

The method is easy to implement and no iterative computation is required. This test was proposed by Hartung (59) as it is not influenced by the magnitude of the heterogeneity variance or the heterogeneity estimator in contrast to the standard test. In agreement, Makambi (49) showed that KH is robust against changes in the magnitude of τ^2 and the selection of estimator. Similarly, Sidik and Jonkman (60) employed different estimators and showed that the coverage probability of the method is less affected by the estimator than the Wald-type method. This is in agreement with Knapp and Hartung (38) who showed that the use of different estimators makes little difference in practice. A simulation study suggested (25) that the method has good properties with high coverage in general and that it is insensitive to the number of trials. Higgins and Thompson (11) showed in simulations that the KH method has more appropriate false positive rates than the standard normal test. It has been also shown that the approach provides coverage close to the nominal level (25,60) and that exhibits better control of type I error than the Wt method with the DL estimator (49). Knapp and Hartung (38) suggested the use of the PM estimator for the heterogeneity along with the KH method for obtaining CIs for μ so as to get a cohesive approach based on generalised Q.

2.5 Comparison of methods

The estimation of the heterogeneity using a variety of methods may lead to different conclusions and the selection of an appropriate estimator for τ^2 is crucial. In this chapter I show that no estimator is clearly best under all circumstances in terms of both bias and efficiency. To select the most appropriate estimator one should consider whether a zero value of heterogeneity is possible, the properties of the various estimators in terms of bias and efficiency, and ease of application, which gives preference to closed form methods. It should be also taken into account that the performance of an estimator depends on the number and size of studies included the meta-analysis, as well as on the magnitude of the true heterogeneity. Empirical studies have shown that the majority of the pairwise meta-analyses are informed by less than ten studies (30,61) and that most meta-analyses with dichotomous outcome data yield $\tau^2 \leq 0.4$. In such cases, evidence from simulation studies shows that the SJ method overestimates τ^2 (21,54); the ML method is associated with substantial negative bias (21,54); REML estimation is less downwardly biased than the DL and ML estimators with greater MSE though (21); and the PM estimator is less downwardly biased than the DL or REML methods (54). An empirical study (62) showed that non-negative methods perform well on average, but produce biased results for meta-analyses with few studies where positive heterogeneity methods are to be preferred. Novianti et al (63) compared in a simulation study for sequential meta-analysis among others the DL, PM, REML and SJ estimators when true heterogeneity is zero and showed that all methods overestimate τ^2 with the DL, PM and REML methods having the best approximation of τ^2. Thompson and Sharp (41) as well as Viechtbauer (46) in a comparison of various estimators including the DL, ML, and REML methods concluded that the REML estimation is the most appropriate technique in terms of bias and efficiency. Panityakul er al (54) applied a simulation study and suggested that the PM estimator is less biased than the DL and REML methods. Empirical evidence illustrated that heterogeneity might vary with different effect measures (22,23,64). It is therefore possible that the performance of the estimators might differ according to the outcome data. Novianti et al (63) recommended the use of PM estimators for both dichotomous and continuous outcome data, while stated that REML for continuous data is a valid alternative as well. With respect to CIs for the overall treatment effect the KH approach is recommended as one of the best options. The KH method is the only method that has been suggested that provides good coverage irrespectively to the heterogeneity and the number of studies included in the meta-analysis (25,38). However, before these approaches can be confidently used, a thorough investigation of all available methods for the estimation of τ^2 and CI for μ using realistic scenarios informed by empirical evidence would be necessary for completeness. A summary of the five heterogeneity estimators and the two methods for the estimation of the uncertainty for the summary treatment effect is presented in Table 1.

In the following chapters I drop the subscript RE from $\hat{\mu}_{RE}$ as every μ is estimated in the random effects model. In case this does not hold I will make this clear.

Table 1. Overview of the estimators for the heterogeneity ($\boldsymbol{\tau^2}$) and the methods to estimate the uncertainty of the overall treatment effect according to simulation studies

Method	**Abbre viation**	**Iterative /Closed form**	**Positive/ Non-negative**	**Simulation studies suggest that the method…**
Estimators for the heterogeneity				
Method of moments estimators				
DerSimonian and Laird	DL	Closed form	Non-negative	performs well for small or close to zero τ^2 and large number (≥ 20) and size (≥ 160) of studies (21,46)
Paule and Mandel	PM	Iterative	Non-negative	is generally less downwardly biased than DL, it is easy to calculate, mirrors more computationally intensive methods (e.g. REML), and does not require distributional assumptions (41,47)
Maximum Likelihood estimators				
Maximum likelihood	ML	Iterative	Non-negative	reveals substantial negative bias for large τ^2 ($\tau^2 \geq 0.5$) which decreases as number and size of studies increase (21,41,46)
Restricted maximum likelihood	REML	Iterative	Non-negative	is less downwardly biased than DL and ML, and bias decreases as number and size of studies increase (21,46)
Model error variance estimator				
Sidik and Jonkman	SJ	Closed form	Positive	performs well for large τ^2 ($\tau^2 \geq 0.5$), but not suitable for small sample sizes (< 30) (21,54)
Estimation methods for the uncertainty of the overall treatment effect				
Wald-type	Wt	Closed form	-	performs well for small τ^2 ($\tau^2 < 0.5$) and large number (≥ 30) and size (≥ 40) of studies (25,39,49)
Knapp-Hartung	KH	Closed form	-	performs well irrespectively the magnitude and estimator for the heterogeneity, as well as the number of studies (25)

3. Statistical approaches to evaluate the assumption of consistency

3.1 Introduction

One of the key advantages of NMA is the appropriate modelling of studies with multiple arms as all study arms can be included in contrast to pairwise meta-analysis that forces separate comparisons (see Franchini et *al.* (65)). The estimates obtained from multi-arm studies for different comparisons are correlated. Consider for example the three-arm trial comparing interventions A, B, and C. The AB and AC comparisons use the same data from participants in A and hence modelling AB and AC treatment effects is sufficient, as the third contrast (BC) is calculated by the difference of the AB and AC treatment effects.

The NMA technique combines simultaneously both direct comparisons within trials and indirect comparisons across trials. Before combining the results of direct and indirect comparisons the extent to which they are in agreement with each other should be examined. Inconsistency, the statistical disagreement of the information coming from various sources of evidence, namely direct and indirect, can occur in NMA as the result of intransitivity or by chance. It should be noted that multi-arm studies are inherently consistent in an evidence loop, which might complicate the consistency assessment.

Several methodologies to evaluate consistency have been outlined in the literature (for a review see NICE DSU Technical Support Document 4 (16)). The methods can be broadly categorised into methods that detect local inconsistency at a specific part of the network (e.g. a specific loop of evidence) and methods that evaluate global inconsistency for the entire network (14,15,17,18). Methods in the former category are useful to locate sources of inconsistency whereas methods in the latter category provide global tests. The evaluation of inconsistency can be equivalently evaluated in either a Bayesian or a frequentist setting.

The simplest and most popular statistical approach to evaluate the prevalence of inconsistency is by contrasting direct and indirect information in a loop of evidence (2). Investigators should interpret the results carefully as the method is associated with a number of limitations (3,20). More sophisticated and appropriate approaches have been presented for complex networks to evaluate local and global inconsistency and a recent review highlights their advantages and limitations (14). Dias et *al*. (15) suggested the node-splitting approach to identify inconsistency between the evidence provided from direct studies for a specific treatment comparison and the indirect evidence based on the entire network after the comparison of interest has been removed. Various models have been proposed to evaluate consistency in the entire network and to synthesise evidence so as to reflect the extra variability beyond what is expected by heterogeneity or random error (14,16). Lu and Ades (66) developed a statistical model to account for random inconsistency in each closed loop of evidence in the network. The presence of studies with multiple arms makes the results of the Lu and Ades model sensitive to their parameterisation, and this prompted Higgins et *al.*(17) and White et *al.* (18) to introduce the idea of design inconsistency and develop models that encompass the potential conflict between studies including different sets of treatments, named 'designs'. Krahn et *al.*(67) and Jackson et *al.* (68) have also derived formulae for the Q-statistic and the I^2 metric for inconsistency in the entire network. As several of these developments are new, they haven't been applied to more than a handful of networks.

3.2 Notation

Consider a network of evidence comprising of S treatments in the set $\Omega = \{\mathrm{A}, \mathrm{B}, \mathrm{C}, \dots\}$ and K studies in total. Each study $k = 1, \dots, K$ compares a specific number of treatments S_k. Studies that compare the same treatments belong to the same design d. Design d refers to studies with S_d specific treatments in the set $\Omega_d \subseteq \Omega$ investigated in N_d studies and the network has in total D designs ($d = 1, \dots, D$). Let A be the arbitrarily chosen reference treatment and M an index for any of the S-1 remaining treatments. I set $y_{d,k,\mathrm{AM}}$ the observed effect size (e.g. LOR) of treatment M relative to treatment A in study k and design d. Let also $v_{d,k,\mathrm{AM}}$ be the variance of $y_{d,k,\mathrm{AM}}$ and μ_{AM} the parameter for the 'true' relative treatment effect of M relative to A. Any parameter μ_{AM} that includes treatment A is named *basic* parameter and all other parameters are named *functional.* Under the assumption of consistency a functional parameter associated with treatment comparison MX can be expressed via the consistency equation $\mu_{\mathrm{MX}} = \mu_{\mathrm{AX}} - \mu_{\mathrm{AM}}$. If consistency does not hold the functional parameters are subject to loop inconsistency and parameters that can be estimated in different designs (e.g. μ_{AB} estimated in AB and ABC studies) are subject to design inconsistency. Loop inconsistency refers to the difference between direct (e.g. μ_{MX}) and indirect (e.g. $\mu_{\mathrm{AX}} - \mu_{\mathrm{AM}}$) estimate for the same comparison. Design inconsistency refers to the difference in the relative effect of two treatments when this is estimated in studies with different designs. I will term the amount of disagreement between different sources of evidence inconsistency factor (IF). The IF parameters included in the models below might be treated either as random effects assuming they all come from a common normal distribution $IF \sim N(0, \sigma^2)$ with σ^2 the inconsistency variance, or as fixed effects allowing different sources of evidence to differ by a fixed quantity.

3.3 Consistency model

The consistency (or NMA) model is defined as a multivariate random-effects meta-analysis. The observed effect size $y_{d,k,\mathrm{AM}}$ of treatment M relative to treatment A of study k with design d is modelled under the consistency assumption as:

$$y_{d,k,\mathrm{AM}} = \mu_{\mathrm{AM}} + \delta_{d,k,\mathrm{AM}} + \varepsilon_{d,k,\mathrm{AM}} \tag{3.1}$$

The consistency model relies on the transitivity assumption and that the missing arms are missing at random. Hence, White et *al.* (69) use the data augmentation technique and impute data with a very small amount of information for designs that do not include the reference treatment. The study random errors are normally distributed $\boldsymbol{\varepsilon_{d,k}} \sim N(\mathbf{0}, \boldsymbol{V}_k)$, where $\boldsymbol{V}_k$ is the within-study variance-covariance matrix assumed to be known. Note that for a two-arm study k the within-study variance-covariance matrix $\boldsymbol{V}_k$ reduces to a 1×1 matrix including the sample variance of study k. In the general case that a study has S_k arms the dimension of $\boldsymbol{V}_k$ is $(S_k - 1) \times (S_k - 1)$. The study-specific RE are normally distributed as shown below with $\boldsymbol{T}_k$ being the between studies variance-covariance matrix involving the heterogeneity variance for each treatment comparison:

$$\boldsymbol{\delta_{d,k}} \sim N\left(\mathbf{0}, \boldsymbol{T}_k = \begin{pmatrix} \tau^2 & \dots & \tau^2/2 \\ \vdots & \ddots & \vdots \\ \tau^2/2 & \dots & \tau^2 \end{pmatrix}\right)$$

I discuss the structure of $\boldsymbol{T}_k$ in section 4.2.3 and the assumptions of the heterogeneity in 2.3.2.

3.4 Models to evaluate global inconsistency

3.4.1 Design-by-treatment interaction (DBT)

The DBT method evaluates whether a network as a whole demonstrates inconsistency by employing an extension of multivariate meta-regression that allows for different treatment effects in studies with different designs (the 'design-by-treatment interaction approach') (17,18). To exemplify the idea of the design-by-treatment interaction approach, consider a network of evidence constructed from an ABC three-arm trial and an ABCD four-arm trial. Both ABC and ABCD trials are inherently consistent. However, the two studies are considered to have different designs and design inconsistency reflects the possibility that they might give different estimates for the same comparisons the two studies include (AB, AC and BC).

The inconsistency model is an extension of model (3.1) and is defined as a multivariate random-effects meta-regression with additional covariates for the different designs:

$$y_{d,k,\mathrm{AM}} = \mu_{\mathrm{AM}} + IF_{d,\mathrm{AM}} + \delta_{d,k,\mathrm{AM}} + \varepsilon_{d,k,\mathrm{AM}}. \tag{3.2}$$

where $IF_{d,\mathrm{AM}}$ represents inconsistency in comparison AM for design d, which may correspond to either design or loop inconsistency. As described in detail elsewhere (17,18) not all possible $IF_{d,\mathrm{AM}}$ covariates are required, since otherwise the model is over-parameterised. The number of inconsistency factors depends on both the total number of treatments in the network and the number of treatments in each design, and is defined as $df_{DBT} = \sum_d (S_d - 1) - (S - 1)$. The number of inconsistency terms df_{DBT} is the difference in the number of parameters between the consistency and inconsistency models. If any AM comparison can be estimated only via direct evidence and there are no multi-arm studies involving both A and M treatments then inconsistency cannot be estimated and I set $IF_{d,\mathrm{AM}} = 0$.

I assess the null hypothesis $H_0\colon \boldsymbol{IF} = \boldsymbol{0}$ the χ^2-test with df_{DBT} degrees of freedom:

$$W^{DBT} = \boldsymbol{IF}'\boldsymbol{\Sigma}^{-1}\boldsymbol{IF}$$

where $\boldsymbol{IF}$ is the $df_{DBT} \times 1$ vector comprising the inconsistency terms and $\boldsymbol{\Sigma}$ is the $df_{DBT} \times df_{DBT}$ variance-covariance matrix of $\boldsymbol{IF}$. The model accounts for possible correlations in the likelihood in multi-arm trials and is insensitive to their parameterisation. Note that the

W^{DBT} statistic is equivalent to the Q-statistic for the evaluation of the assumption of consistency as presented elsewhere (67).

3.4.2 Lu and Ades (LA)

Lu and Ades (66) proposed a special case of the DBT model that accounts for inconsistency in each loop of the network as long as this loop is not only informed by multi-arm studies. This is because multi-arm studies are inherently consistent and therefore loops informed by studies with multiple arms are not expected to show inconsistency. Lu and Ades (66) implemented their model (LA) in a Bayesian framework assuming random IF terms. As the LA model does not distinguish between different designs, I drop the respective subscript. Suppose now MX comparison is included in a closed loop AMX, and the study k compares all three treatments, then the observed treatment effect $y_{k,\mathrm{MX}}$ is modelled as:

$$y_{k,\mathrm{MX}} = \mu_{\mathrm{AX}} - \mu_{\mathrm{AM}} + \delta_{k,\mathrm{AM}} + IF_{\mathrm{AMX}} + \varepsilon_{k,\mathrm{AM}}$$

where IF_{AMX} measures the magnitude of inconsistency in the loop that MX belongs to. Let N be the number of total distinct comparisons observed in the network and G the number of functional parameters that indirectly are only estimated by multi-arm trials, then $df_{LA} = N - (S-1) - G$ are the number of inconsistency factors in the network (see Lu and Ades (66) for more details). I assess the null hypothesis $H_0\colon \boldsymbol{IF} = \boldsymbol{0}$ using the χ^2-test with *f* degrees of freedom:

$$W^{LA} = \boldsymbol{IF}'\boldsymbol{\Sigma}^{-1}\boldsymbol{IF}$$

with $\boldsymbol{\Sigma}$ the $df_{LA} \times df_{LA}$ variance-covariance matrix of $\boldsymbol{IF}$. The model provides a global test for loop consistency in the entire network. The presence of multi-arm trials though might complicate the consistency assessment. It is possible that differences in the parameterisation of the multi-arm studies can yield different values for the IF parameters with different W^{LA} values and hence different inference on inconsistency.

3.4.3 $\boldsymbol{Q}$-statistic and I^2 measure for evaluating and measuring inconsistency

As in pairwise meta-analysis, one can employ the Q-statistics in NMA to infer about homogeneity, consistency or both. The total variability in the entire network can be split into the variation within (Q_{Het}) and between (Q_{Inc}) designs; these refer to heterogeneity and inconsistency. Krahn et al. (67) used the decomposition of Cochran's Q-test to evaluate the assumptions of consistency and homogeneity in the network. The decomposition resembles the one used in the study-level subgroups in the context of pairwise meta-analysis (70). The total network Q-statistic (Q_{Net}) is separated into the heterogeneity statistic (Q_{Het}) and the inconsistency statistic (Q_{Inc}):

$$Q_{Net} = Q_{Het} + Q_{Inc}$$

Under the homogeneity and consistency assumptions, the global Cochran's Q-statistic follows a $\chi^2_{df_{Net}}$ distribution with degrees of freedom the number of data points minus the number of basic parameters ($df_{Net} = \sum_{k=1}^{K}(S_k - 1) - (S - 1)$). The Q_{Net} statistic is defined as the weighted sum of squared deviations of the observed treatment effects from the consistent effect estimates. If we stack all observed treatment effects $y_{d,k,AM}$ into the vector $\boldsymbol{y}$ with length $\sum_{k=1}^{K}(S_k - 1)$ and $\boldsymbol{V}$ is a block diagonal variance-covariance matrix with blocks $\boldsymbol{V_i}$ and $\widehat{\boldsymbol{\mu}}$ the $S - 1$ estimates of the basic parameters, then

$$Q_{Net} = (\boldsymbol{y} - \boldsymbol{X}\widehat{\boldsymbol{\mu}})'\boldsymbol{V}^{-1}(\boldsymbol{y} - \boldsymbol{X}\widehat{\boldsymbol{\mu}})$$

where $\boldsymbol{X}$ is a $\sum_{k=1}^{K}(S_k - 1) \times (S - 1)$ design matrix that denotes the comparisons presented in each element of $\boldsymbol{y}$. The associated test examines whether the total variation can be compatible with chance. To evaluate the heterogeneity within designs we use the heterogeneity statistic defined as the sum of the within-design Q-statistics (Q^d_{Het}):

$$Q_{Het} = \sum_{d=1}^{D} Q^d_{Het} = \sum_{d=1}^{D} (\boldsymbol{y}_d - \boldsymbol{X}_d\widehat{\boldsymbol{\mu}}_d)'\boldsymbol{V}_d^{-1}(\boldsymbol{y}_d - \boldsymbol{X}_d\widehat{\boldsymbol{\mu}}_d).$$

Using the fixed effect consistency model $\widehat{\boldsymbol{\mu}}_d$ represents the $(S_d - 1) \times 1$ vector of the treatment effects estimated in design d with S_d treatments. Suppose design d includes K_d studies, then $\boldsymbol{y}_d$ is a vector of length $[K_d(S_d - 1)]$ and $\boldsymbol{X}_d$ is the $[K_d(S_d - 1)] \times (S_d - 1)$ design matrix. The variance-covariance matrix $\boldsymbol{V}_d$ is a $[K_d(S_d - 1)] \times [K_d(S_d - 1)]$ block diagonal matrix containing the within-study variances and covariances of the observed treatment effects included in design d. The Q_{Het} statistic has approximately a $\chi^2_{df_{Het}}$ distribution with $df_{Het} = \sum_{d=1}^{D}(N_d - 1)\ (S_d - 1)$. The between-designs Q-statistic for inconsistency is a likelihood-ratio test statistic as shown by Lu et *al.*(71) and is defined as

$$Q_{Inc} = \boldsymbol{IF}'\boldsymbol{\Sigma}^{-1}\boldsymbol{IF}$$

Statistical approaches to evaluate the assumption of consistency where $\boldsymbol{IF}$ is the vector of inconsistency factors estimated from the DBT model and $\boldsymbol{\Sigma}$ is their variance-covariance matrix. Under the consistency assumption the omnibus statistic for inconsistency has approximately a $\chi^2_{df_{DBT}}$ distribution as the W^{DBT} statistic. Although Lu et *al.* (71) have shown that the Q_{Inc} statistic can be calculated fitting either FE or RE model (assuming a common within-design heterogeneity) Rucker et al. recommended to use the FE model(71) and Krahn et *al.*(67) suggested that using a FE model within designs allows for better location of inconsistency.

Multivariate versions of the I^2 statistics that can be used to measure both heterogeneity and inconsistency have been suggested by White et *al.*(18), Rucker and Guido (72), and Jackson et *al.* (68). Jackson et al. (68) defined three different I^2 statistics that describe three different sources of variability:

$$I^2 = \frac{R^2 - 1}{R^2}$$

$$\text{with } R = \begin{cases} \left(\frac{|U_{RE,\text{DBT}}|}{|U_{RE,\text{CON}}|}\right)^{\frac{1}{2(S-1)}}, \text{ defines } I^2_{Inc} \\ \left(\frac{|U_{RE,\text{CON}}|}{|U_{FE,\text{CON}}|}\right)^{\frac{1}{2(S-1)}}, \text{ defines } I^2_{Het} \\ \left(\frac{|U_{RE,\text{DBT}}|}{|U_{FE,\text{CON}}|}\right)^{\frac{1}{2(S-1)}}, \text{ defines } I^2_{Het,Inc} \end{cases}$$

where $|U|$ is the determinant of the estimated variance-covariance matrix of $\hat{\mu}$, the subscript CON refers to the consistency model, DBT to the inconsistency DBT model and RE or FE to the random and fixed effects respectively. I^2_{Inc} describes the amount of variability attributed to inconsistency rather than sampling error or heterogeneity, I^2_{Het} describes the amount of variability attributed to heterogeneity rather than sampling error and $I^2_{Het,Inc}$ describes the amount of variability attributed to heterogeneity and inconsistency rather than sampling error.

3.5 Methods to evaluate local inconsistency

3.5.1 Loop-specific (LS) approach

The loop-specific (LS) method evaluates inconsistency in all closed loops of evidence formed by three or four treatments within each network, by contrasting direct with indirect estimates of a specific treatment effect. I call 'triangular loops' the closed paths involving three treatments and 'quadrilateral loops' the closed paths involving four treatments. Bucher et *al.* (2) described the method in an early paper and I refer to it as the 'loop-specific' approach. The LS method because of its simplicity and of not requiring specialised software to compute is so far the most commonly applied approach.

Inconsistency can be evaluated as the disagreement between different sources of evidence within a closed loop. In each network of treatments all triangular and quadrilateral loops are identified. In each loop we can estimate the treatment effects of all pairwise comparisons using conventional meta-analysis. As the LS method does not distinguish between different designs, I drop the respective subscript. Consider for example the triangular loop ABC formed by treatments $\mathrm{A, B, C}$ with available comparisons AB, AC and BC. In the RE model the observed treatment effect $y_{k,AB}$ is modeled as:

$$y_{k,\mathrm{AB}} = \mu_{\mathrm{AB}} + \delta_{k,\mathrm{AB}} + \varepsilon_{k,\mathrm{AB}}$$

where $\delta_{k,AB}$ is a random effect for study k of B relative to A and $\varepsilon_{k,AB}$ is the within-study sampling error. Similarly, for the other two comparisons in the loop:

$$y_{k,\mathrm{AC}} = \mu_{\mathrm{AC}} + \delta_{k,\mathrm{AC}} + \varepsilon_{k,\mathrm{AC}}$$

$$y_{k,\mathrm{BC}} = \mu_{\mathrm{BC}} + \delta_{k,\mathrm{BC}} + \varepsilon_{k,\mathrm{BC}}$$

To estimate all direct relative effects within the triangular loop ABC I perform a random-effects meta-analysis for each available comparison. I discuss the assumptions about the heterogeneity variances in section 2.3.3. Within each available loop, I evaluated whether the consistency assumption holds. Since in a single loop there may be only one inconsistency, the IF for the loop ABC is defined as (66,73)

$$\widehat{IF}_{\mathrm{ABC}}^{LS} = |\hat{\mu}_{\mathrm{BC}} - (\hat{\mu}_{\mathrm{AC}} - \hat{\mu}_{\mathrm{AB}})|$$

with $$Var(\widehat{IF}_{\mathrm{ABC}}^{LS}) = Var(\hat{\mu}_{\mathrm{BC}}) + Var(\hat{\mu}_{\mathrm{AC}}) + Var(\hat{\mu}_{\mathrm{AB}}),$$

I use the 'LS' superscript to denote the method in which IF is estimated. The direction of the estimated IF is irrelevant to the evaluation of inconsistency and only the magnitude of its absolute value is of interest. I assess the null hypothesis $H_0\colon IF_{\mathrm{ABC}} = 0$ using the z-test (74):

$$W_{\mathrm{ABC}}^{LS} = \frac{\widehat{IF}_{\mathrm{ABC}}^{LS}}{\sqrt{Var(\widehat{IF}_{\mathrm{ABC}}^{LS})}} \sim N(0,1)$$

A similar process is followed for all quadrilateral loops formed by four different head-to-head comparisons. However, if the quadrilateral loop is formed by two or more triangles, then only the triangles are evaluated. Since a multi-arm study is inherently consistent in an evidence loop, different parameterizations of its arms causes complications in the consistency assessment.

3.5.2 Separating one design from the rest (SODR)

The SODR method examines whether a specific design in the network can be responsible for inconsistency. In particular, the method evaluates whether the effect size for the same comparison differs when estimated using a particular study design b and other designs in the network. To estimate inconsistency I 'detach' the N_b studies of a specific design b from the network; then I estimate the $S_b - 1$ treatment effects from a) the N_b studies and b) the network without the N_b studies assuming consistency. The difference between the estimates is the magnitude of inconsistency for design b. The SODR method is a special case of the DBT model that includes only one design inconsistency term $\boldsymbol{IF}_b^{SODR}$ at the time; that corresponds to a vector with $S_b - 1$ elements for the differences between direct and indirect estimates for the comparisons it includes. If the detached design includes treatments A, B and C then $\boldsymbol{IF}_b^{SODR}$ is the vector $(IF_{b,\mathrm{AB}}^{SODR}, IF_{b,\mathrm{AC}}^{SODR})$.

Consider we want to estimate SODR inconsistency for a specific design b that includes the comparison AM. The observed treatment effect $y_{d,k,\mathrm{AM}}$ accounting for possible design inconsistency in b is modelled as

$y_{d,k,\mathrm{AM}} = \mu_{\mathrm{AM}}^{-b} + \delta_{d,k,\mathrm{AM}} + IF_{b,\mathrm{AM}}^{SODR} + \varepsilon_{d,k,\mathrm{AM}}$, for design $d = b$

$y_{d,k,\mathrm{AM}} = \mu_{\mathrm{AM}}^{-b} + \delta_{d,k,\mathrm{AM}} + \varepsilon_{d,k,\mathrm{AM}}$, for design $d \neq b$

where $IF_{b,\mathrm{AM}}^{SODR}$ measures the magnitude of inconsistency for design $d = b$ in comparison AM. The parameter μ_{AM}^{-b} is the mean treatment effect of M relative to A that comes from all available designs except design b whereas $\mu_{\mathrm{AM}}^{-b} + IF_{b,\mathrm{AM}}^{SODR}$ is the mean treatment effect for AM in design b. Note that this is different from the IF term included in the DBT approach that models inconsistency simultaneously for all designs. The number of inconsistency factors included in the model depends on the number of treatments in the detached design and is defined as $df_{SODR} = S_b - 1$, with df denoting the degrees of freedom. Note that df_{SODR} might be different for different designs.

In Table 2 I provide an illustrative example considering a network of three different sets of studies: AB, ABC, and ABCD studies. Setting A the reference the possible SODR inconsistencies are: design AB $df_{SODR} = 1$ ($IF^{SODR}_{AB,\mathrm{AB}}$), design ABC $df_{SODR} = 2$ ($IF^{SODR}_{ABC,\mathrm{AB}}$, $IF^{SODR}_{ABC,\mathrm{AC}}$), and design ABCD $df_{SODR} = 3$ ($IF^{SODR}_{ABCD,\mathrm{AB}}$ $IF^{SODR}_{ABCD,\mathrm{AC}}$, $IF^{SODR}_{ABCD,\mathrm{AD}}$). Note that $IF^{SODR}_{ABCD,\mathrm{AD}}$ is not estimable as no AD studies are available, so it should be omitted.

For each design $d = b$ under the null hypothesis is $H_0\colon \boldsymbol{IF}^{SODR}_b = \mathbf{0}$ an approximate test can be obtained using the χ^2-test with p degrees of freedom as:

$$W^{SODR}_b = \boldsymbol{IF}^{SODR\prime}_b \boldsymbol{\Sigma}^{-1} \boldsymbol{IF}^{SODR}_b$$

with $\boldsymbol{\Sigma}$ the $p \times p$ variance-covariance matrix of $\boldsymbol{IF}^{SODR}_b$. Note that SODR approach accounts for possible correlations in the likelihood in studies with multiple arms and is insensitive to their parameterisation.

Table 2. Consistency model and SODR method. Inconsistency is evaluated for designs AB, ABC and ABCD using the SODR method. In all cases A is the reference treatment.

Study	Type of study	Model study-specific treatment effects
Consistency Model		
1	AB	$y_{\mathrm{AB,1,AB}} = \mu_{\mathrm{AB}} + \delta_{\mathrm{AB,1,AB}} + \varepsilon_{\mathrm{AB,1,AB}}$
2	ABC	$\begin{pmatrix} y_{\mathrm{ABC,2,AB}} \\ y_{\mathrm{ABC,2,AC}} \end{pmatrix} = \begin{pmatrix} \mu_{\mathrm{AB}} \\ \mu_{\mathrm{AC}} \end{pmatrix} + \begin{pmatrix} \delta_{\mathrm{ABC,2,AB}} \\ \delta_{\mathrm{ABC,2,AC}} \end{pmatrix} + \begin{pmatrix} \varepsilon_{\mathrm{ABC,2,AB}} \\ \varepsilon_{\mathrm{ABC,2,AC}} \end{pmatrix}$
3	ABCD	$\begin{pmatrix} y_{\mathrm{ABCD,3,AB}} \\ y_{\mathrm{ABCD,3,AC}} \\ y_{\mathrm{ABCD,3,AD}} \end{pmatrix} = \begin{pmatrix} \mu_{\mathrm{AB}} \\ \mu_{\mathrm{AC}} \\ \mu_{\mathrm{AD}} \end{pmatrix} + \begin{pmatrix} \delta_{\mathrm{ABCD,3,AB}} \\ \delta_{\mathrm{ABCD,3,AC}} \\ \delta_{\mathrm{ABCD,3,AD}} \end{pmatrix} + \begin{pmatrix} \varepsilon_{\mathrm{ABCD,3,AB}} \\ \varepsilon_{\mathrm{ABCD,3,AC}} \\ \varepsilon_{\mathrm{ABCD,3,AD}} \end{pmatrix}$
Separating design AB		
1	AB	$y_{\mathrm{AB,1,AB}} = \mu^{-\mathrm{AB}}_{\mathrm{AB}} + \delta_{\mathrm{AB,1,AB}} + IF^{SODR}_{AB,\mathrm{AB}} + \varepsilon_{\mathrm{AB,1,AB}}$
2	ABC	$\begin{pmatrix} y_{\mathrm{ABC,2,AB}} \\ y_{\mathrm{ABC,2,AC}} \end{pmatrix} = \begin{pmatrix} \mu^{-\mathrm{AB}}_{\mathrm{AB}} \\ \mu^{-\mathrm{AB}}_{\mathrm{AC}} \end{pmatrix} + \begin{pmatrix} \delta_{\mathrm{ABC,2,AB}} \\ \delta_{\mathrm{ABC,2,AC}} \end{pmatrix} + \begin{pmatrix} \varepsilon_{\mathrm{ABC,2,AB}} \\ \varepsilon_{\mathrm{ABC,2,AC}} \end{pmatrix}$

3	ABCD	$\begin{pmatrix} y_{\text{ABCD,3,AB}} \\ y_{\text{ABCD,3,AC}} \\ y_{\text{ABCD,3,AD}} \end{pmatrix} = \begin{pmatrix} \mu_{\text{AB}}^{-\text{AB}} \\ \mu_{\text{AC}}^{-\text{AB}} \\ \mu_{\text{AD}}^{-\text{AB}} \end{pmatrix} + \begin{pmatrix} \delta_{\text{ABCD,3,AB}} \\ \delta_{\text{ABCD,3,AC}} \\ \delta_{\text{ABCD,3,AD}} \end{pmatrix} + \begin{pmatrix} \varepsilon_{\text{ABCD,3,AB}} \\ \varepsilon_{\text{ABCD,3,AC}} \\ \varepsilon_{\text{ABCD,3,AD}} \end{pmatrix}$
Separating design ABC		
1	AB	$y_{\text{AB,1,AB}} = \mu_{\text{AB}}^{-\text{ABC}} + \delta_{\text{AB,1,AB}} + \varepsilon_{\text{AB,1,AB}}$
2	ABC	$\begin{pmatrix} y_{\text{ABC,2,AB}} \\ y_{\text{ABC,2,AC}} \end{pmatrix} = \begin{pmatrix} \mu_{\text{AB}}^{-\text{ABC}} \\ \mu_{\text{AC}}^{-\text{ABC}} \end{pmatrix} + \begin{pmatrix} \delta_{\text{ABC,2,AB}} \\ \delta_{\text{ABC,2,AC}} \end{pmatrix} + \begin{pmatrix} IF_{ABC,\text{AB}}^{SODR} \\ IF_{ABC,\text{AC}}^{SODR} \end{pmatrix} + \begin{pmatrix} \varepsilon_{\text{ABC,2,AB}} \\ \varepsilon_{\text{ABC,2,AC}} \end{pmatrix}$
3	ABCD	$\begin{pmatrix} y_{\text{ABCD,3,AB}} \\ y_{\text{ABCD,3,AC}} \\ y_{\text{ABCD,3,AD}} \end{pmatrix} = \begin{pmatrix} \mu_{\text{AB}}^{-\text{ABC}} \\ \mu_{\text{AC}}^{-\text{ABC}} \\ \mu_{\text{AD}}^{-\text{ABC}} \end{pmatrix} + \begin{pmatrix} \delta_{\text{ABCD,3,AB}} \\ \delta_{\text{ABCD,3,AC}} \\ \delta_{\text{ABCD,3,AD}} \end{pmatrix} + \begin{pmatrix} \varepsilon_{\text{ABCD,3,AB}} \\ \varepsilon_{\text{ABCD,3,AC}} \\ \varepsilon_{\text{ABCD,3,AD}} \end{pmatrix}$
Separating design ABCD		
1	AB	$y_{\text{AB,1,AB}} = \mu_{\text{AB}}^{-\text{ABCD}} + \delta_{\text{AB,1,AB}} + \varepsilon_{\text{AB,1,AB}}$
2	ABC	$\begin{pmatrix} y_{\text{ABC,2,AB}} \\ y_{\text{ABC,2,AC}} \end{pmatrix} = \begin{pmatrix} \mu_{\text{AB}}^{-\text{ABCD}} \\ \mu_{\text{AC}}^{-\text{ABCD}} \end{pmatrix} + \begin{pmatrix} \delta_{\text{ABC,2,AB}} \\ \delta_{\text{ABC,2,AC}} \end{pmatrix} + \begin{pmatrix} \varepsilon_{\text{ABC,2,AB}} \\ \varepsilon_{\text{ABC,2,AC}} \end{pmatrix}$
3	ABCD	$\begin{pmatrix} y_{\text{ABCD,3,AB}} \\ y_{\text{ABCD,3,AC}} \\ y_{\text{ABCD,3,AD}} \end{pmatrix} = \begin{pmatrix} \mu_{\text{AB}}^{-\text{ABCD}} \\ \mu_{\text{AC}}^{-\text{ABCD}} \\ \mu_{\text{AD}}^{-\text{ABCD}} \end{pmatrix} + \begin{pmatrix} \delta_{\text{ABCD,3,AB}} \\ \delta_{\text{ABCD,3,AC}} \\ \delta_{\text{ABCD,3,AD}} \end{pmatrix} + \begin{pmatrix} IF_{ABCD,\text{AB}}^{SODR} \\ IF_{ABCD,\text{AC}}^{SODR} \\ IF_{ABCD,\text{AD}}^{SODR} \end{pmatrix} + \begin{pmatrix} \varepsilon_{\text{ABCD,3,AB}} \\ \varepsilon_{\text{ABCD,3,AC}} \\ \varepsilon_{\text{ABCD,3,AD}} \end{pmatrix}$

3.5.3 Separating indirect and direct evidence (SIDE)

The SIDE method has been presented and implemented in a Bayesian framework by Dias et *al*. (15) with the name ‘node-splitting’. The method examines whether a particular comparison might be associated with inconsistency, by separating the information of each network estimate into two distinct parts: the direct and the indirect evidence. The same process is repeated for all comparisons included in the network. Note that only comparisons that belong to closed loops are susceptible for inconsistency in the SIDE method. Suppose we want to separate comparison AX that belongs to at least one closed loop. The observed treatment effect $y_{k,\text{AM}}$ accounting for possible inconsistency between direct and indirect evidence for the AX comparison is modelled as

$y_{k,\text{AM}} = \mu_{\text{AM}}^{-\text{AX}} + \delta_{k,\text{AM}} + IF_{\text{AX}}^{SIDE} + \varepsilon_{k,\text{AM}}$, if study *k* includes A and M

$y_{k,\text{AM}} = \mu_{\text{AM}}^{-\text{AX}} + \delta_{k,\text{AM}} + \varepsilon_{k,\text{AM}}$, if study *k* does not include M

The parameter of interest IF_{AX}^{SIDE} measures the magnitude of inconsistency between the direct and indirect evidence for the comparison AX. I will call a comparison 'inconsistent' when the direct evidence disagrees with that of the remaining network beyond chance. The model above can also be seen as a special case of the LA model (and hence special case of DBT) where only one IF term is estimated at a time. For each comparison AX under the null hypothesis $H_0\colon IF_{\text{AX}}^{SIDE} = 0$ the approximate test can be obtained using the z-test:

$$W_{\text{AX}}^{SIDE} = \frac{\widehat{IF}_{\text{AX}}^{SIDE}}{\sqrt{\widehat{Var}(IF_{\text{AX}}^{SIDE})}} \sim N(0,1)$$

Note that if AX is part of a multi-arm study IF_{AX}^{SIDE} would be deferent depending on the parameterization of the multi-arm studies. This will be illustrated by a simple example. Consider one AB, one ABC and one AC study, as shown in Table 3 and let AB comparison be separated. The choice of the reference treatment in this dataset determines which comparisons from the three-arm study ABC will feature in the data. Let A be the reference treatment then ABC study will contribute with $y_{2,\text{AB}}$ and $y_{2,\text{AC}}$. There is no direct evidence on the BC comparison, $\mu_{\text{BC}}^{-\text{AB}}$ is not directly estimable and therefore inconsistency IF_{AB}^{SIDE} is not identifiable in this case. If we choose B to be the reference treatment then ABC study will contribute with $y_{2,\text{BA}}$ and $y_{2,\text{BC}}$. Now we have all three comparisons informed directly and hence inconsistency IF_{BA}^{SIDE} is identifiable. If BC studies were present, inconsistency would be identifiable in both parameterisations but its estimates would be different because the multi-arm study would provide information to only $\mu_{\text{AC}}^{-\text{AB}}$ or only $\mu_{\text{BC}}^{-\text{BA}}$ depending on the reference treatment.

Table 3. Evaluation of inconsistency for comparison AB using the SIDE method.

Study	Type of study	Model study-specific treatment effects
Reference Treatment A (parameterisation 1)		
1	AB	$y_{1,\text{AB}} = \mu_{\text{AB}}^{-\text{AB}} + \delta_{1,\text{AB}} + IF_{\text{AB}}^{SIDE} + \varepsilon_{1,\text{AB}}$
2	ABC	$\begin{pmatrix} y_{2,\text{AB}} \\ y_{2,\text{AC}} \end{pmatrix} = \begin{pmatrix} \mu_{\text{AB}}^{-\text{AB}} \\ \mu_{\text{AC}}^{-\text{AB}} \end{pmatrix} + \begin{pmatrix} \delta_{2,\text{AB}} \\ \delta_{2,\text{AB}} \end{pmatrix} + \begin{pmatrix} IF_{\text{AB}}^{SIDE} \\ 0 \end{pmatrix} + \begin{pmatrix} \varepsilon_{2,\text{AB}} \\ \varepsilon_{2,\text{AB}} \end{pmatrix}$
3	AC	$\text{y}_{3,\text{AC}} = \mu_{\text{AC}}^{-\text{AB}} + \delta_{3,\text{AC}} + \varepsilon_{3,\text{AC}}$

		with $\boldsymbol{\mu}_{\mathbf{AB}}^{-\mathbf{AB}} + \boldsymbol{IF}_{\mathbf{AB}}^{\boldsymbol{SIDE}} = \boldsymbol{\mu}_{\mathbf{AC}}^{-\mathbf{AB}} - \boldsymbol{\mu}_{\mathbf{BC}}^{-\mathbf{AB}}$
Reference Treatment B (parameterisation 2)		
1	AB	$y_{1,\mathrm{BA}} = \mu_{\mathrm{BA}}^{-\mathrm{BA}} + \delta_{1,\mathrm{BA}} + IF_{\mathrm{BA}}^{SIDE} + \varepsilon_{1,\mathrm{BA}}$
2	ABC	$\begin{pmatrix} y_{2,\mathrm{BA}} \\ y_{2,\mathrm{BC}} \end{pmatrix} = \begin{pmatrix} \mu_{\mathrm{BA}}^{-\mathrm{BA}} \\ \mu_{\mathrm{BC}}^{-\mathrm{BA}} \end{pmatrix} + \begin{pmatrix} \delta_{2,\mathrm{BA}} \\ \delta_{2,\mathrm{BC}} \end{pmatrix} + \begin{pmatrix} IF_{\mathrm{BA}}^{SIDE} \\ 0 \end{pmatrix} + \begin{pmatrix} \varepsilon_{2,\mathrm{BA}} \\ \varepsilon_{2,\mathrm{BC}} \end{pmatrix}$
3	AC	$\mathrm{y}_{3,\mathrm{AC}} = \mu_{\mathrm{AC}}^{-\mathrm{AB}} + \delta_{3,\mathrm{AC}} + \varepsilon_{3,\mathrm{AC}}$
		with $\boldsymbol{\mu}_{\mathbf{BA}}^{-\mathbf{BA}} + \boldsymbol{IF}_{\mathbf{BA}}^{\boldsymbol{SIDE}} = \boldsymbol{\mu}_{\mathbf{BC}}^{-\mathbf{BA}} - \boldsymbol{\mu}_{\mathbf{AC}}^{-\mathbf{BA}}$

3.6 Comparison of approaches to evaluate inconsistency

One of the drawbacks of the LS method is that inferences in loops are not independent, because different loops of the network share the same studies. To overcome this, Caldwell et *al.*(75) introduced a chi-squared test for the special case that all loops in the network share a single comparison. However, this can be applied only to specific parts of the network, and again yields multiple tests if all pieces of the network need to be tested. Another drawback of the LS approach is that indirect evidence is restricted to the information provided from a single loop. It is preferable to compare the direct evidence with the indirect estimate from the entire network, as is the approach taken in the SIDE method proposed by Dias et *al.*(15). All three methods outlined above are sensitive to the parameterization of multi-arm studies, and do not offer obvious ways to infer about network consistency. The only method to identify local inconsistency that is insensitive to the studies with multiple arms is the SODR method by Veroniki et *al.* (76). Among all the methods, LS is to date the most popular approach to evaluate inconsistency (6).

When NMA is applied within a Bayesian framework, investigators often contrast models with and without the consistency constraints with respect to fit and parsimony (77). This provides a global test for the plausibility of consistency in the entire network, but inferences are again sensitive to the parameterization of multi-arm studies. The DBT model is the only model that provides an omnibus test, can be fit in a frequentist setting and provides results insensitive to the parameterisation of multi-arm studies (17,18). Models that do not account for design inconsistency (e.g. LA (66) and the one presented by Lumley (78)) are special cases of the DBT model.

4. Evaluation of inconsistency in networks of interventions

4.1 Introduction

The assessment of the consistency assumption is vital to ensure that the NMA results are valid and interpreted appropriately. The need to define the levels of inconsistency in real life data led empirical studies to examine the prevalence of inconsistency between direct and indirect comparisons. Song et *al*.(29) carried out an empirical study applying the Bucher method and assuming different heterogeneity parameters in every comparison within each loop. They evaluated inconsistency in 112 loops of evidence formed by studies comparing pairs of three treatments and concluded that inconsistency was prevalent in 14% of the networks (29). In a response to comments on their article, Song et *al*. (79) alternatively assumed that all comparisons within each triangular loop share the same amount of heterogeneity and they observed that inconsistency was reduced to 12%. A recent analysis of 94 three-treatment networks in Cochrane Reviews found statistically significant inconsistency in 17% of the networks (19). However, no empirical evidence exists so far regarding the prevalence of inconsistency in more complex networks, primarily because no omnibus test was available until recently to evaluate the assumption of consistency in a network as a whole. A general model to detect inconsistency has been proposed, and called design-by-treatment interaction model (18) (see section 3.4.1). Inconsistency can be viewed not only as the disagreement between direct and indirect estimates in a loop, but also as the disagreement between studies involving different sets of treatments.

In a network of trials the detection of inconsistency can be hampered by the presence of heterogeneity. Large heterogeneity impacts on the uncertainty of the mean effect sizes, and hence statistical inconsistency is less likely to be detected. The estimation of the heterogeneity variance can vary using different methods (e.g. DL, REML (21)), which subsequently affects the ability to detect inconsistency. Assumptions about the heterogeneity being the same in different parts of the network or the same in the entire network may similarly impact on the detection of inconsistency. However, as factors that cause heterogeneity can also cause inconsistency, complete separation of the two is not always possible. In summary, large heterogeneity increases the chances of inconsistency being present, but decreases the chances of detecting it.

Both the presence and the detection of inconsistency may be affected by the use of different effect measures. Empirical studies have shown that ratio measures (odds ratio (OR) and risk ratio (RR)) are less heterogeneous than absolute effect measures (such as risk difference) and that the risk ratio for adverse outcomes is less likely to be heterogeneous than that for beneficial outcomes (22,23). These differences depend on the extent of variation in baseline risk across studies. If baseline risks are substantially different in different parts of a loop, then the underlying inconsistency may be greater for some effect measures than others; if baseline risks vary substantially within each comparison, then more or less heterogeneity may be present, depending on the effect measure. Caldwell et *al*. have also considered the choice of different effect measures in network meta-analysis and concluded that the choice of measure should be based on physiological understanding of the outcome and, if possible, after considering the model fit (13,80).

The aim of this chapter is to evaluate empirically the prevalence of inconsistency in published networks of interventions that compare at least four treatments, and to examine the extent to which this is acknowledged by the authors of the NMAs. I further aim to investigate the statistical considerations that might influence the statistical detection of inconsistency in these complex networks of evidence. I also explore whether different effect measures for dichotomous outcome data are associated with differences in inconsistency, and whether different ways to estimate heterogeneity impact upon the magnitude and detection of inconsistency. I explore the role of factors that may impact on inferences about inconsistency in a simple loop of evidence for a dichotomous outcome. The factors that I explore are associated with the amount of data available in the loop (such as number, size and distribution of trials across comparisons, frequency of events), the heterogeneity variance in the pairwise comparisons (presence or absence and estimation method) and the method for inference about pairwise summary effects (Wt or KH). I conduct a simulation study considering realistic scenarios and I evaluate the performance of the test for the assumption of consistency. I select the simulation scenarios relying on empirical findings (13,30).

4.2 Empirical study

4.2.1 Searching for network meta-analyses and data extraction

I searched in PubMed for research articles including networks with at least four treatments and dichotomous primary outcomes. I searched for articles published between March 1997 and February 2011 in which any form of indirect comparison was applied, according to their titles or abstracts. I used the search code:

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

In case I identified two or more networks on the same topic, I included only one in the dataset and preferred to the larger one. I extracted data regarding the year of publication, the methods applied for the indirect comparison, the number of trials and the number of arms of each study, as well as the total number of interventions involved in each network. From each network I extracted data for the primary outcome (as stated in the text or, if this was unclear, defined as the first outcome presented). I preferred data presented in 2×2 tables rather than as effect sizes with their measure of uncertainty, when both formats were reported. The extracted data include the name of each trial, as well as the number of events, the sample size and the treatment in every arm of each trial included in the network.

Two review authors (Areti Angeliki Veroniki and Georgia Salanti) independently assessed each article for the evaluation of the assumption of consistency. Differences in assessment results were discussed. For each network we extracted the statistical methods used by the original authors to evaluate consistency. We considered inappropriate methods a) the comparison of network estimates with the direct estimates, b) the informal comparison of the results with previously conducted meta-analyses, and c) the informal comparison of indirect estimates with the direct estimates.

4.2.2 Effect measures

I considered four effect measures for dichotomous outcomes: the OR, the risk difference (RD), the risk ratio of beneficial outcomes (RRB) and the risk ratio for harmful outcomes (RRH). The LS and DBT approaches were applied with OR, RRH, RRB, and RD measures, whereas SIDE, SODR and LA approaches were applied using the OR scale.

4.2.3 Estimation of the heterogeneity

I made assumptions about the heterogeneity variances, and I address first the LS approach. I used both common within-loop heterogeneity, τ^2_{loop}, and common within-network heterogeneity, τ^2_{ntw}, as already described in 2.3.3 section. In the DBT model I assume that all comparisons in the network share the same heterogeneity variance τ^2_{ntw}. Suppose the total number of treatments included in a study k is S_k, the variance-covariance matrix for the random effects has $(S_k - 1) \times (S_k - 1)$ dimension and is given by

$$\boldsymbol{T}_k = \tau^2_{ntw} \begin{pmatrix} 1 & \dots & 1/2 \\ \vdots & \ddots & \vdots \\ 1/2 & \dots & 1 \end{pmatrix}$$

If the k^{th} study is a two-arm study then the between studies variance-covariance matrix reduces to $\boldsymbol{T}_k = \tau^2_{ntw}$.

In general, when the number of studies included in the meta-analysis is large, the heterogeneity parameter is more precisely estimated (70). Therefore, it is likely that $\hat{\tau}^2_{ntw}$ is more precise than $\hat{\tau}^2_{loop}$. Assuming a common heterogeneity variance impacts also on the precision of the summary effects, and consequently on power for detecting inconsistency. For example, it is possible that the heterogeneity in a specific loop ABC is smaller than the heterogeneity in the rest of the network. Assuming the same heterogeneity in the network will then decrease precision for the summary estimates of the ABC loop and may therefore decrease the power to detect inconsistency. Similarly, assuming common within-network heterogeneity introduces heterogeneity in loops involving comparisons informed by a single study, decreasing the chance of identifying the presence of inconsistency. Although the assumption of the common within-network heterogeneity can underestimate the prevalence of substantial inconsistency, it allows for a more accurate representation of how the effects are combined in a network meta-analysis. On the contrary, it is possible that the common within-network heterogeneity increases precision for the summary estimates if the heterogeneity in a specific loop is bigger than the heterogeneity in the rest of the network, and hence statistical inconsistency can be evident.

I estimated inconsistency in a frequentist setting, where the heterogeneity τ^2 can be estimated by a variety of methods. The performance of the different estimators can differ in terms of bias and MSE, and they can over- or under-estimate the true heterogeneity variance (see section 2.3.1). As heterogeneity may affect the estimation of inconsistency, I evaluated inconsistency using different estimators of τ^2. I applied the different estimation methods using the OR effect measure. In the LS approach I used the DL (31), REML (51) and SJ (35) methods. I included the DL method because it is the most popular estimator in random-effects meta-analysis and is the default estimator in many meta-analysis' software (e.g. RevMan). I used the frequently applied REML method and the less popular SJ estimator because they are associated with large differences in their properties. In the DBT, SODR, SIDE and LA approaches only the DL, ML and REML estimators of the heterogeneity are available. I applied the ML and REML methods in DBT model, since the DL method is not appropriate when the augmentation technique is applied (18). For the LA, SIDE and SODR I selected the REML method relying on simulation studies that suggest the REML estimation as it performs better in terms of bias than the DL and ML estimators (see section 2.3).

4.2.4 Comparison of the approaches for the evaluation of inconsistency and model fit

For each network I estimated global inconsistency using the LA and DBT models and local inconsistency using LS, SIDE and SODR methods using the RE model. Total inconsistency and heterogeneity were also measured using I^2.

The DBT model estimates inconsistency in the entire network, whereas the LS approach evaluates each loop separately. It is therefore impossible to infer about the level of agreement between the two methods. I arbitrarily considered a network to be inconsistent under the loop-specific approach if at least 5% of its loops are inconsistent in order to describe the comparative performance of the two methods.

Loop inconsistency refers to a difference between direct and indirect estimates for the same comparison. However, the presence of multi-arm trials in a network of evidence complicates the evaluation of loop inconsistency, since loops formed within multi-arm trials are necessarily consistent. Consider for example a network comprising some AB studies, some AC studies and some three-arm ABC studies. Note that only two of the three possible treatment effects are sufficient to fully specify the results of the three-arm studies. If the two effects include the BC comparison, then loop inconsistency might be observed by contrasting it with an indirect estimate constructed from the other two groups of studies.

On the other hand, if the two effects from the three-arm studies are AB and AC, then an evaluation of inconsistency would not take place. I therefore exclude the comparison that is most frequent within the loop and evaluate loop inconsistency in the LS method. This can impact on the summary treatment effects though and hence on the evaluation of inconsistency for a network with many multi-arm studies. The LA and SIDE approaches are sensitive to the parameterisations of the multi-arm studies too. For LA and SIDE approaches I examined all possible parameterisations to account for differences in the results. In the LA model I considered a network inconsistent when it was found inconsistent with at least one parameterisation. Similarly, in SIDE approach I considered that a comparison is associated with inconsistency when it disagreed with the remaining network in at least one parameterisation of the multi-arm studies. Among the different W^{LA} test values that appeared using different parameterisations of the multi-arm trials I used the maximum W^{LA}. When estimating the heterogeneity in LA model I obtained a range of values that resulted from all possible parameterisations for the multi-arm trials. In all approaches I used fixed IF terms because there are often too few inconsistency parameters to get a reliable estimate for the inconsistency variance σ^2. Note that the IF is the logarithm of the ratio of odds ratios (ROR) from the two different sources of evidence for the same comparison, i.e. $ROR = \exp(|IF|)$. When there is no evidence for inconsistency the ROR is close to the unity.

I implemented LS method in software R 2.13.2 using the self-programmed routine *ifplot.fun*, which is available online (in http://www.mtm.uoi.gr/ under ‘How to do an MTM’). The four approaches LA, DBT, SODR, SIDE and I^2 measure were carried out in software STATA using the *mvmeta* (69) command.

4.2.5 Description of database

The search initially identified 817 relevant articles and after the screening process I ended up with 40 different networks. The full process is shown in the flow chart in Figure 1.

The original authors evaluated the assumption of inconsistency using appropriate statistical methodology in 15 (38%) networks. Out of these 15 networks, inconsistency for at least one comparison in the analysis was reported in 10 (67%). The most prevalent method (18%) of evaluating inconsistency was the LS approach. A large proportion of investigators (23%) seemed to be aware of the consistency assumption but used inappropriate methods to evaluate it, such as comparisons of direct and network estimates (see Appendix Table 1).

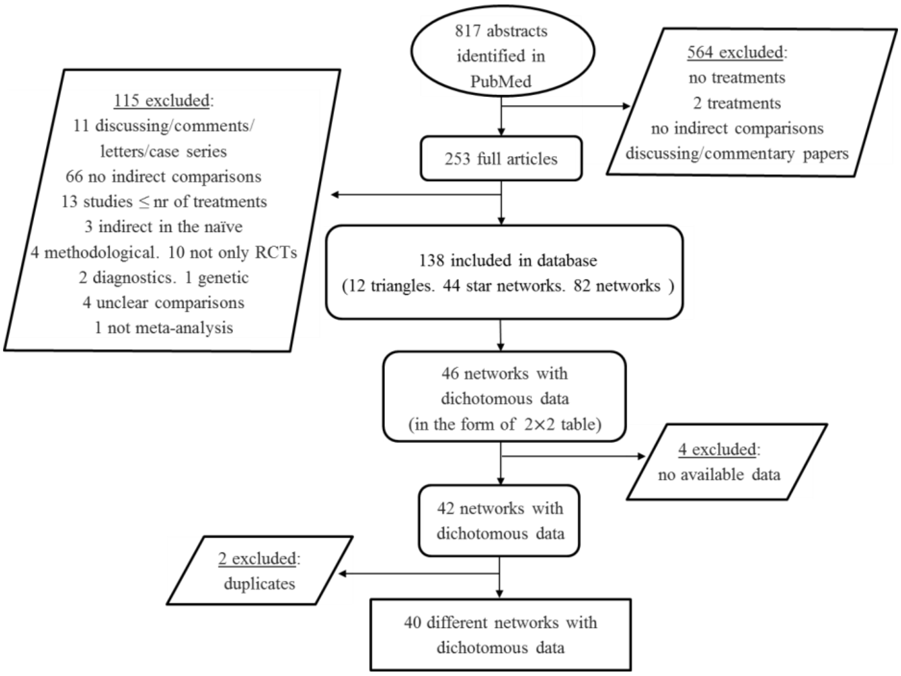
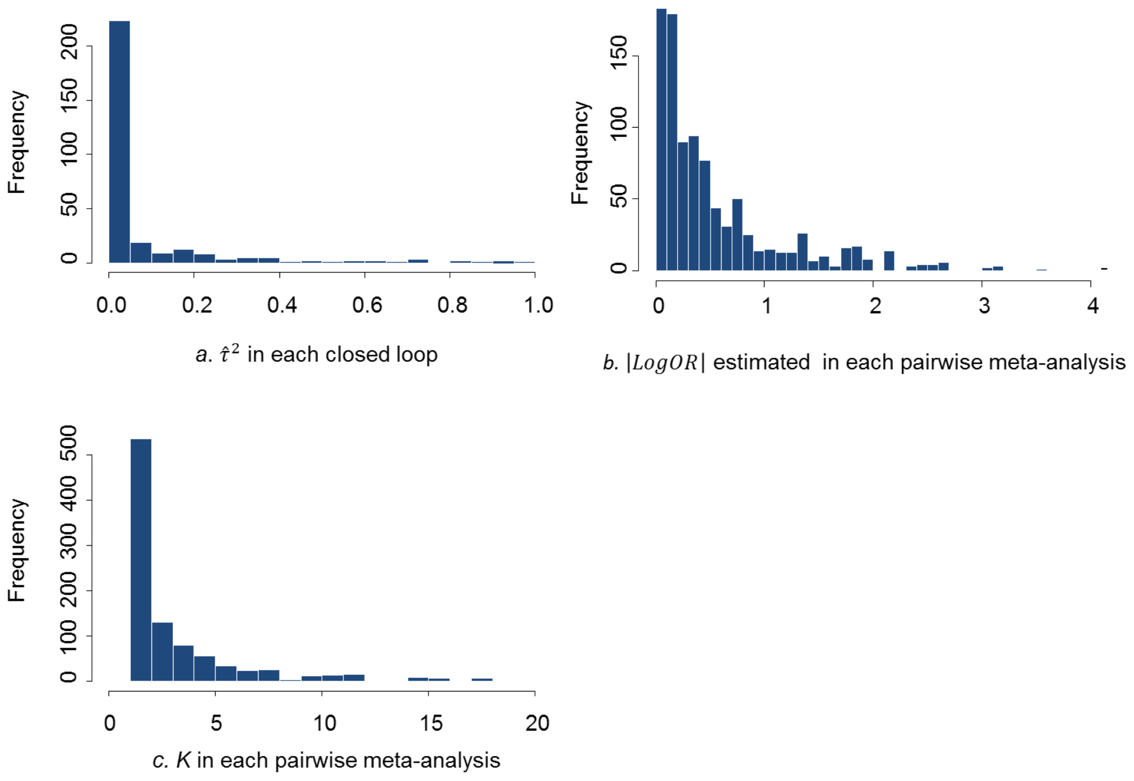

817 abstracts identified in PubMed
564 excluded:
no treatments
2 treatments
no indirect comparisons
discussing/commentary papers
253 full articles
115 excluded:
11 discussing/comments/ letters/case series
66 no indirect comparisons
13 studies ≤ nr of treatments
3 indirect in the naïve
4 methodological. 10 not only RCTs
2 diagnostics. 1 genetic
4 unclear comparisons
1 not meta-analysis
138 included in database
(12 triangles. 44 star networks. 82 networks)
46 networks with dichotomous data (in the form of 2×2 table)
4 excluded:
no available data
42 networks with dichotomous data
2 excluded:
duplicates
40 different networks with dichotomous data

Figure 1. Flow chart of the process of selecting network meta-analysis articles

Twenty-five (63%) networks used OR, 13 (33%) used RR, one (2%) used all of the three OR, RR and RD, and one (2%) used a hazard ratio. In only seven publications (18%) the authors explain their choice of effect measure. Most networks had a subjective primary outcome (43%), whereas 35% and 22% of the networks had semi-objective (e.g. cause-specific mortality, major morbidity event etc.) and all-cause mortality outcomes respectively.

The median number of studies per network is 23, ranging from 9 to 111. The median number of trials per loop is 8 and the median loop sample size is 2196; the respective median number of trials and sample size per comparison are 2 and 706. The number of treatments compared ranged from 4 to 17 with a median of 6. The majority of the networks (63%) compared pharmacological interventions *vs.* placebo. Multi-arm trials were included in most networks (34 networks, 85%). Thirty-three networks included three-arm trials and nine included four-arm trials. The number of included three-arm trials per network ranged from 0 to 12, whereas the number of included four-arm trials ranged from 0 to 6. The total number of loops obtained from the 40 networks is 303 and ranged from 1 to 70 per network. The characteristics of these networks are described in detail in Appendix Table 2. The 40 relevant NMAs included 348 different comparisons and 362 different designs. Out of the 362 designs 287 were designs including two-arm studies Each network included between one and 42 comparisons that could be separated to estimate SIDE inconsistency (median 10), and each comparison included between 1 and 47 studies (median 2). The median number of designs per network where inconsistency can be evaluated was 8 and ranged from 2 to 43, and each design was informed from 1 to 45 (median 1) studies. Most networks included at least one comparison (36 networks, 90%) or at least one design (37 networks, 93%) informed by a single study. In one network (66) the only present loop was informed by less than three independent comparisons and consequently the LA and SIDE approaches were not applicable.



a. $\hat{\tau}^2$ in each closed loop

b. $|LogOR|$ estimated in each pairwise meta-analysis

c. K in each pairwise meta-analysis

Figure 2. Histograms for the 40 published networks of evidence: a) the within-loop heterogeneity ($\hat{\boldsymbol{\tau}}^{\mathbf{2}}$), b) the mean treatment effect in the absolute log odds-ratio scale ($|\boldsymbol{LOR}|$), and c) the number of trials ($\boldsymbol{K}$) per meta-analysis Heterogeneity is estimated with the DerSimonian and Laird method.

In Figure 2 I summarise some of the attributes of the 303 loops of 40 published networks of interventions using the LOR scale. The majority of the pairwise meta-analyses (93%) included fewer than ten trials, and the median $|LOR|$ was 0.32 with interquartile range (IQR) (0.13, 0.75). In 91% of the loops the τ^2_{loop} using the DL estimator is estimated less than 0.5 and zero in 51% of the loops.

4.2.6 Models to evaluate global Inconsistency

Design-by-treatment interaction (DBT) model

In the DBT model the ML Wald tests for analyses of OR yielded 8 inconsistent networks out of the 40 networks (20%), whereas 11 (28%) of the networks were found to display inconsistency when analysed using each of the three effect measures RRH, RRB and RD (all pairwise comparisons between OR $vs.$ RRH, RRB or RD for inconsistent networks with the ML estimator using the McNemar test produced $P = 0.371$). The REML Wald test indicated 5 (13%), 6 (15%), 7 (17%) and 5 (13%) inconsistent networks using the OR, RRH, RRB and RD, respectively (all pairwise comparisons between OR $vs.$ RRH or RD for inconsistent networks with the REML estimator using the McNemar test produced $P = 1$, whereas OR vs. RRB produced $P = 0.617$) (see Table 4 and Table 5).

Table 4. Number of consistent networks that become inconsistent when changing from one effect size to another and vice versa, in the design-by-treatment interaction model and the restricted maximum likelihood (REML) and maximum likelihood (ML) estimators of the heterogeneity. RD: risk difference measure, RRH: risk ratio for harmful outcomes, RRB: risk ratio for beneficial outcomes, OR: odds ratio, C: consistent, I: inconsistent

IF^{DBT} estimated with ML								
		RRH		**RRB**		**RD**		**% of 40 networks**
OR		C	I	C	I	C	I	
	Consistent	28	4	28	4	28	4	80%
	Inconsistent	1	7	1	7	1	7	20%
	% of 40 networks	72%	28%	72%	28%	72%	28%	
IF^{DBT}estimated with REML								
		RRH		**RRB**		**RD**		**% of 40 networks**
OR		C	I	C	I	C	I	
	Consistent	33	2	32	3	32	3	87%
	Inconsistent	1	4	1	4	3	2	13%
	% of 40 networks	85%	15%	83%	17%	87%	13%	

Comparing the REML with the ML method, the former yielded fewer inconsistent networks (13% to 17% depending on effect measure) than the latter (20% to 28%

depending on effect measure), but there were no important differences (McNemar test under the comparison of ML estimator versus the REML estimator; OR: $P = 0.248$, RRH: $P = 0.074$, RRB: $P = 0.1336$, RD: $P = 0.041$) (see Table 5). This is probably because the ML method estimated slightly smaller values of the heterogeneity variance than the REML in almost all networks and all effect sizes.

Table 5. Number of consistent networks that become Inconsistent and vice versa, when heterogeneity is estimated with maximum likelihood (ML) or the restricted maximum likelihood (REML) method. Inconsistency is investigated with the design-by-treatment interaction (DBT) model for all four effect sizes. RD: risk difference measure, RRH: risk ratio for harmful outcomes, RRB: risk ratio for beneficial outcomes, OR: odds ratio

IF^{DBT} estimated with REML			IF^{DBT} estimated with ML		
			OR		
			Consistent	Inconsistent	% of 40 networks
	OR	Consistent	32	3	87%
		Inconsistent	0	5	13%
		% of 40 networks	80%	20%	
			RRH		
	RRH	Consistent	29	5	85%
		Inconsistent	0	6	15%
		% of 40 networks	72%	28%	
			RRB		
	RRB	Consistent	29	4	83%
		Inconsistent	0	7	17%
		% of 40 networks	72%	28%	
			RD		
	RD	Consistent	29	6	87%
		Inconsistent	0	5	13%
		% of 40 networks	72%	28%	

In Figure 3a and Figure 3b I present the heterogeneity estimates using ML and REML methods in consistency ($\hat{\tau}^2_{\mathrm{CON}}$) model against the DBT ($\hat{\tau}^2_{\mathrm{DBT}}$) model. Results are presented on the OR effect measure. On average the consistency models display higher heterogeneity than the DBT models, accounting probably for inconsistency in the data. Figure 3c shows the differences in $\sqrt{\hat{\tau}}$ values between DBT and consistency models estimated with REML method and the OR. The consistency model yielded higher heterogeneity values in 26 (65%) networks compared to the DBT model with a mean relative change $(mean((\hat{\tau}^2_{\mathrm{DBT}} - \hat{\tau}^2_{\mathrm{CON}})/mean(\hat{\tau}^2_{\mathrm{DBT}}, \hat{\tau}^2_{\mathrm{CON}})))$ -0.71. The fact that the consistency model often yielded higher heterogeneity estimates than the inconsistency model might indicate that the extra variability due to possible inconsistency in the former model is captured in the heterogeneity. Large relative drops in heterogeneity can be seen as an alternative approach to detect inconsistency. In Figure 3d I depict the relative change between consistency and DBT model against the squared root of the P values for the inconsistency parameters estimated in DBT model using the REML estimator and OR. The larger the heterogeneity in the consistency model compared to $\hat{\tau}_{\mathrm{DBT}}$ the more chances to find an inconsistent network.

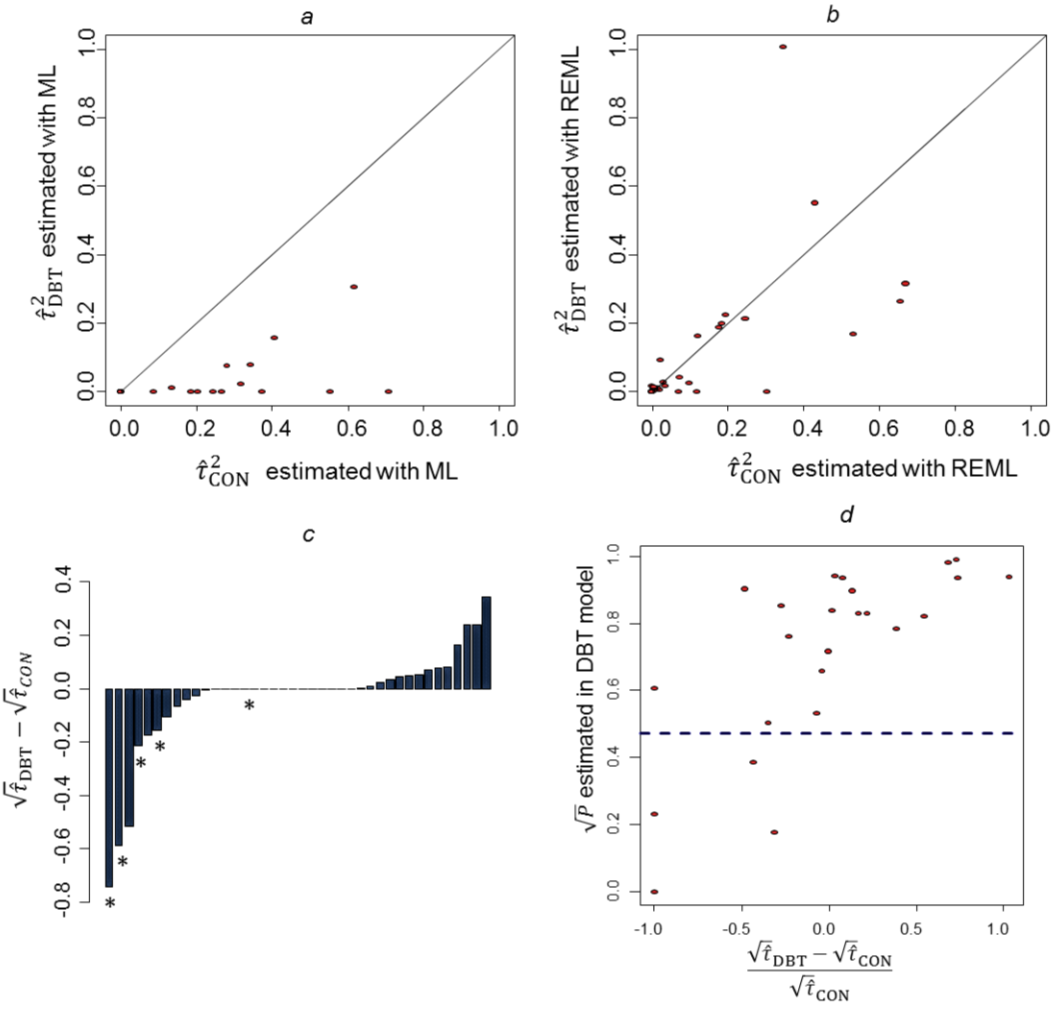


Figure 3. Plot of heterogeneity estimates ($\hat{\tau}^2$) with maximum likelihood (ML) (panel a) and restricted maximum likelihood (REML) (panel b) from the consistency (CON) model against heterogeneity estimates from the design-by-treatment interaction (DBT) model along with the equality line. c) Bar plot of the difference in the square root of the estimated heterogeneity standard deviation between consistency and DBT model. Negative values show greater heterogeneity in CON model, whereas positive values show greater heterogeneity in DBT model. Star points show the networks that were found inconsistent in DBT model using REML method. b) Plot of the squared root of the P values of the inconsistency estimated in DBT model against the relative change of the square root of the heterogeneity in the DBT model from the CON

model. The horizontal blue dashed line represents the cut-off value $P = 0.05$. Note that 15 networks with $\hat{\tau}_{\text{CON}} = 0$ could not be presented in the plot. All plots are presented on the odds ratio scale.

For fourteen networks (35%) I could not find any indication in the published articles that the authors evaluated the assumption of consistency. Four out of these networks were found to be inconsistent when I applied the DBT model using the REML method and the OR scale. A cause of concern is that one in three of the meta-analysis authors did not examine consistency since conclusions from NMA may not be valid when the consistency assumption does not hold.

Lu and Ades (LA) model

I applied the LA model in 39 networks in total. Inconsistency was prevalent in maximum 7 (18%) networks when I applied different parameterisations of the multi-arm studies (see also Appendix Table 8). A different parameterisation of the studies with multiple arms impacts on the inference about inconsistency and the impact is more pronounced when the network includes loops with comparisons informed by single studies (e.g. network of Imamura (81), Elliott (82)). Different parameterisation of the multi-arm studies impacts also on the estimation of heterogeneity in the LA model and I selected the maximum $\hat{\tau}^2_{\text{LA}}$ value (see the spread in the box plots presented in Figure 9). There is a large variation in the estimation of $\hat{\tau}_{\text{LA}}$ though when the multi-arm studies are differently parameterised. A large variability in the network might be expressed either as inconsistency or as heterogeneity when parameterising the multi-arm studies differently. For example, in one parameterisation the network by Salliot 2011 (83) is suggested consistent with $\hat{\tau}_{\text{LA}} = 0.14$, whereas in another parameterisation the network is suggested inconsistent with $\hat{\tau}_{\text{LA}} = 0.00$ (see Figure 9).

4.2.7 Methods to evaluate local Inconsistency

Loop-specific (LS) method

Inconsistency using different effect measures for dichotomous data

Out of the total 303 loops identified in the 40 networks, 23 were found to be inconsistent (8%) when analysed with OR, 26 (9%) with RRH, 29 (10%) with RRB and 29 (10%) as RD, for common within-loop heterogeneity ($\hat{\tau}^2_{loop}$) estimated using the DL method. Table 6 provides these results along with results under the assumption of common within-network heterogeneity ($\hat{\tau}^2_{ntw}$). Changing effect size when using $\hat{\tau}^2_{loop}$, some consistent loops became inconsistent and vice versa. These changes were more often observed between OR *vs.* RD and OR *vs.* RRB. Eleven (4%) consistent loops with OR changed to inconsistent with RD, whereas 5 (2%) loops that deviate from consistency using OR changed to consistent when RD was employed (see Table 6). The percentage of inconsistent loops was comparable across the four effect measures (McNemar test when $\hat{\tau}^2_{loop}$ was used; OR *vs.* RRH: $P = 0.505$, OR *vs.* RRB: $P = 0.239$, OR *vs.* RD: $P = 0.211$). In Appendix Table 3 I provide the inconsistency estimates in all four scales for all loops, along with their standard errors and W^{LS} values.

Table 6. Number of consistent (C) and inconsistent (I) loops using different effect measures and assumptions for the heterogeneity. I assume either common within-loop heterogeneity ($\hat{\tau}^2_{loop}$) estimated with DerSimonian and Laird method and network heterogeneity ($\hat{\tau}^2_{ntw}$) estimated with restricted maximum likelihood method.

		IF^{LS} estimated with $\hat{\tau}^2_{loop}$						
		RRH		RRB		RD		% of 303 loops
		C	I	C	I	C	I	
OR	Consistent	274	6	268	12	269	11	92%
	Inconsistent	3	20	6	17	5	18	8%
	% of 303 loops	91%	9%	91%	9%	91%	9%	
		IF^{LS} estimated with $\hat{\tau}^2_{ntw}$						
		RRH		RRB		RD		% of 303 loops
		C	I	C	I	C	I	
OR	Consistent	283	3	278	8	278	8	94%
	Inconsistent	2	15	7	10	9	8	6%
	% of 303 loops	94%	6%	94%	6%	95%	5%	

The 40 network dataset included 203 loops with at least one comparison being informed by a single study. Inconsistency was more likely to be found in these loops. For example, in the network of Elliot (82) I identified two inconsistent loops using the OR scale, which share the same comparison including only one study. It is possible that in such cases inconsistency is introduced by this particular study. Of the 203 loops 19 (9%) were found to be inconsistent with OR, whereas from the 100 remaining loops with comparisons including two or more studies only 4 (4%) were inconsistent ($P = 0.154$). The respective percentages of inconsistent loops for the other scales were 18 (9%) *vs.* 8 (8%) ($P = 0.972$)

with RRH, 21 (10%) vs.8 (8%) (P =0.657) with RRB and 20 (10%) vs. 9 (9%) (P =0.977) with RD.

A similar picture was observed assuming a common within-network heterogeneity ($\hat{\tau}^2_{ntw}$), although the total inconsistency rate dropped. Out of the 303 loops, I detected 16 (5%) inconsistent loops with OR, 19 (6%) with RRH, 18 (6%) with RRB and 16 (5%) with RD (see Table 6). In Appendix Table 4 I provide the inconsistency estimates using the four effect measures for all loops along with their standard errors and W^{LS} values. Again, there were no important differences in inconsistency between the four effect measures (McNemar test when $\hat{\tau}^2_{ntw}$ was used; OR vs. RRH: P = 0.371, OR vs. RRB: P = 0.789, OR vs. RD: P = 1).

Table 7. Number of consistent loops that become inconsistent in the loop-specific method (LS) when applying the common within-loop heterogeneity ($\hat{\tau}^2_{loop}$)) estimated with the DerSimonian and Laird method and network heterogeneity ($\hat{\tau}^2_{ntw}$) estimated with the restricted maximum likelihood method. RD: risk difference measure, RRH: risk ratio for harmful outcomes, RRB: risk ratio for beneficial outcomes, OR: odds ratio.

			IF^{LS} estimated with $\hat{\tau}^2_{loop}$		
IF^{LS} estimated with $\hat{\tau}^2_{ntw}$			OR		
			Consistent	Inconsistent	% of 303 loops
	OR	Consistent	280	7	95%
		Inconsistent	0	16	5%
		% of 303 loops	92%	8%	
			RRH		
	RRH	Consistent	275	10	94%
		Inconsistent	3	16	6%
		% of 303 loops	91%	9%	
			RRB		
	RRB	Consistent	273	13	94%
		Inconsistent	2	16	6%
		% of 303 loops	90%	10%	
			RD		
	RD	Consistent	273	15	95%
		Inconsistent	2	14	5%
		% of 303 loops	90%	10%	

Comparing the common within-loop and common within-network approaches I concluded that there are important differences in the number of inconsistent loops between the two methods, especially when OR, RRB or RD are applied (McNemar test when $\hat{\tau}^2_{loop}$

was used *vs.* when $\hat{\tau}^2_{ntw}$ was used; OR: $P \ = 0.023$, RRH: $P \ = 0.096$, RRB: $P \ = 0.010$, RD: $P \ = 0.004$). In Table 7 I provide the number of important IF^{LS} using the four effect measures assuming either τ^2_{loop} or τ^2_{ntw}.

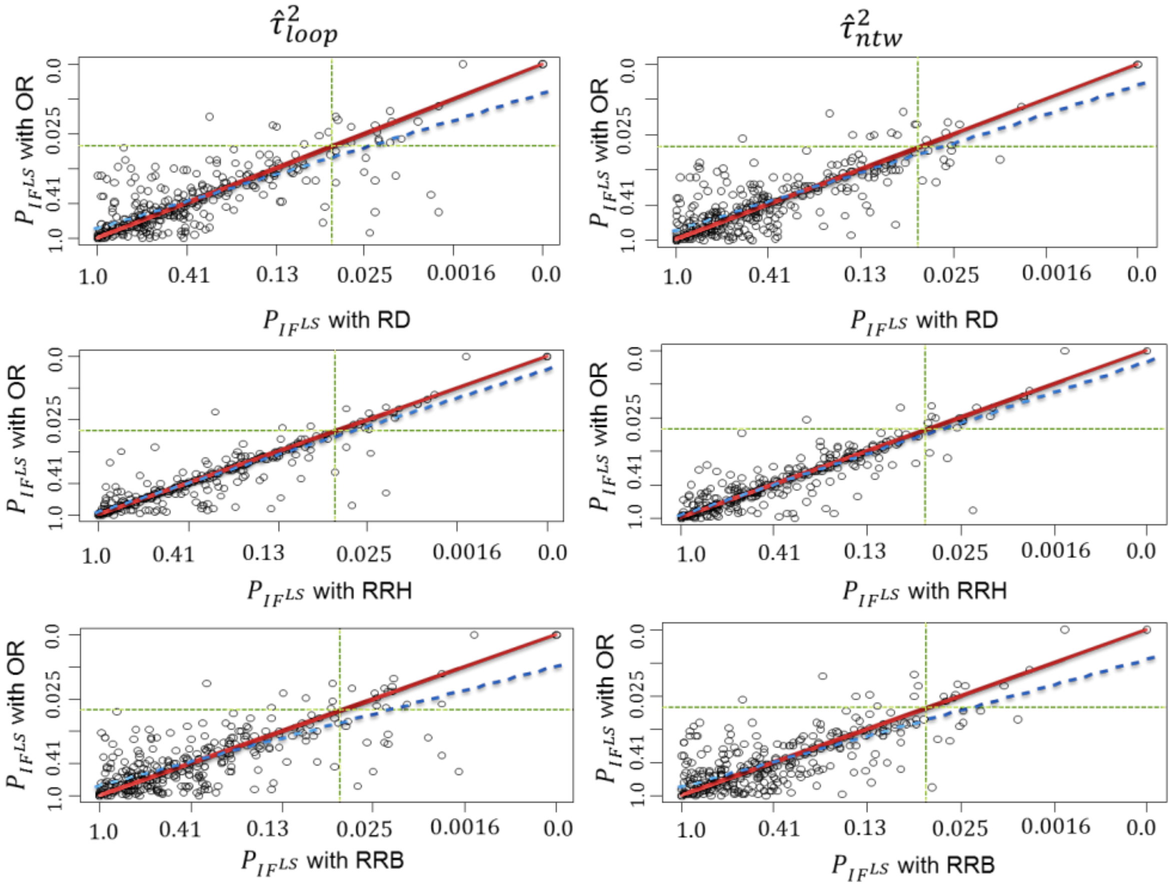


Figure 4. Two sided P values of IF^{LS} (fourth-root scale) for OR *vs.* RD, OR *vs.* RRH and OR *vs.* RRB effect measures with the DerSimonian and Laird method for τ^2_{loop} and the restricted maximum likelihood for τ^2_{ntw}. The red solid diagonal line indicates equality, the blue dashed diagonal line is the regression line and the two green dotted horizontal and vertical lines represent the P=0.05 threshold lines.

In Figure 2 the P values for the LS approach are presented for both τ^2_{loop} and $\tau^2_{ntw}$for the three pairs of effect measures; OR *vs.* RD, OR *vs.* RRH and OR *vs.* RRB. The two-sided P values are displayed on the fourth root scale (23). Among all six panels, agreement seems to be higher between OR and RRH as seen by less scatter around the equality line and a smaller number of discordant points. This is likely to be due to most outcomes being rare rather than common, so that OR is closer to RRH than to RRB. Heterogeneity estimates are in better agreement between OR and RRH (for $\hat{\tau}^2_{ntw}$: mean($|\tau^2_{RRH} - \tau^2_{OR}|/\tau^2_{OR}$) =52%, mean($|\tau^2_{RRB} - \tau^2_{OR}|/\tau^2_{OR}$) =63%, mean($|\tau^2_{RD} - \tau^2_{OR}|/\tau^2_{OR}$) =90%; for $\hat{\tau}^2_{loop}$: mean($|\tau^2_{RRH} - \tau^2_{OR}|/\tau^2_{OR}$) =51%, mean($|\tau^2_{RRB} - \tau^2_{OR}|/\tau^2_{OR}$) =79%, mean($|\tau^2_{RD} - \tau^2_{OR}|/$

τ^2_{OR}) =97%). In general, no substantial differences in inconsistency were observed between the effect measures.

Inconsistency using different estimators for the heterogeneity parameter

In Table 8 I present the number of inconsistent loops using three heterogeneity estimators for τ^2_{loop}: the DL, REML. SJ, and REML methods and the REML estimation for τ^2_{ntw}. Results are presented on the OR effect measure. I observed that both the DL and REML methods led to a greater number of inconsistent loops than the SJ method. This is due to the differences in their properties. As already discussed in section *Estimation methods for the* 2.3.2, in certain cases the DL and REML methods underestimate τ^2 whereas the SJ estimator overestimates τ^2. As noted earlier, I observed that inconsistency was more frequent in loops that include comparisons informed by a single study (Table 8). Using τ^2_{loop} 19 (9%) out of 203 loops with at least one comparison informed by one study were found to be inconsistent with the DL estimation, whereas only 4 (4%) were inconsistent of the remaining 100 loops (P =0.154). The respective percentages with the REML and SJ estimators are 18 (9%) versus 3 (3%) (P =0.099) and 12 (6%) *vs.* 2 (2%) (P =0.217). However, using τ^2_{ntw} the respective inconsistent loops were 4 (2%) *vs.* 12 (12%) (P =0.001) with the REML estimation. The evaluation of inconsistency assuming τ^2_{ntw} and REML in comparisons described by a single study decreases the inconsistency rate by 7% compared to τ^2_{loop}. This is because the amount of within-network heterogeneity in most inconsistent loops, and particularly those that include at least one comparison informed by a single study, is larger than τ^2_{loop}.

There was no evidence that inconsistency differs statistically among the three estimators when assuming a common within-loop heterogeneity (comparison of inconsistent loops with at least two studies per comparison: DL *vs*. REML: P =1, DL *vs*. SJ: P =0.679, SJ *vs*. REML: P =1; comparison of inconsistent loops with at least one comparison informed by a single study: DL *vs*. REML: P =1, DL *vs*. SJ: P =0.262, SJ *vs*. REML: P =0.343). However, inconsistency differs substantially between the common within-loop and the common within-network approach with the REML method (comparison of inconsistent loops with at least two studies per comparison: P =0.035; comparison of inconsistent loops with at least one comparison informed by a single study: P =0.003).

Table 8. Frequency of Inconsistent loops (IL) using the DerSimonian and Laird (DL), restricted maximum likelihood (REML) and Sidik-Jonkman (SJ) estimators for the heterogeneity. Inconsistency is estimated with the log odds-ratio scale using the loop-specific approach for both common within-loop heterogeneity ($\hat{\tau}^2_{loop}$) and network heterogeneity ($\hat{\tau}^2_{ntw}$). The number of IL is provided when $\hat{\tau}^2_{loop}$ or $\hat{\tau}^2_{ntw}$ is equal to zero, as well as when the closed loop involves one study in at least one comparison.

Estimator of τ^2	***IL***	***IL with $\hat{\tau}^2_{loop} = 0$***	***IL with1 study in at least one comparison***
		$\hat{\tau}^2_{loop}$	
DL	23 (8%)	14 (5%)	19 (9%)
REML	21 (7%)	18 (6%)	18 (9%)
SJ	14 (5%)	5 (2%)	12 (6%)
Total loops	303	303	203
		$\hat{\tau}^2_{ntw}$	
REML	17 (6%)	5 (2%)	5 (2%)
Total loops	303	303	203

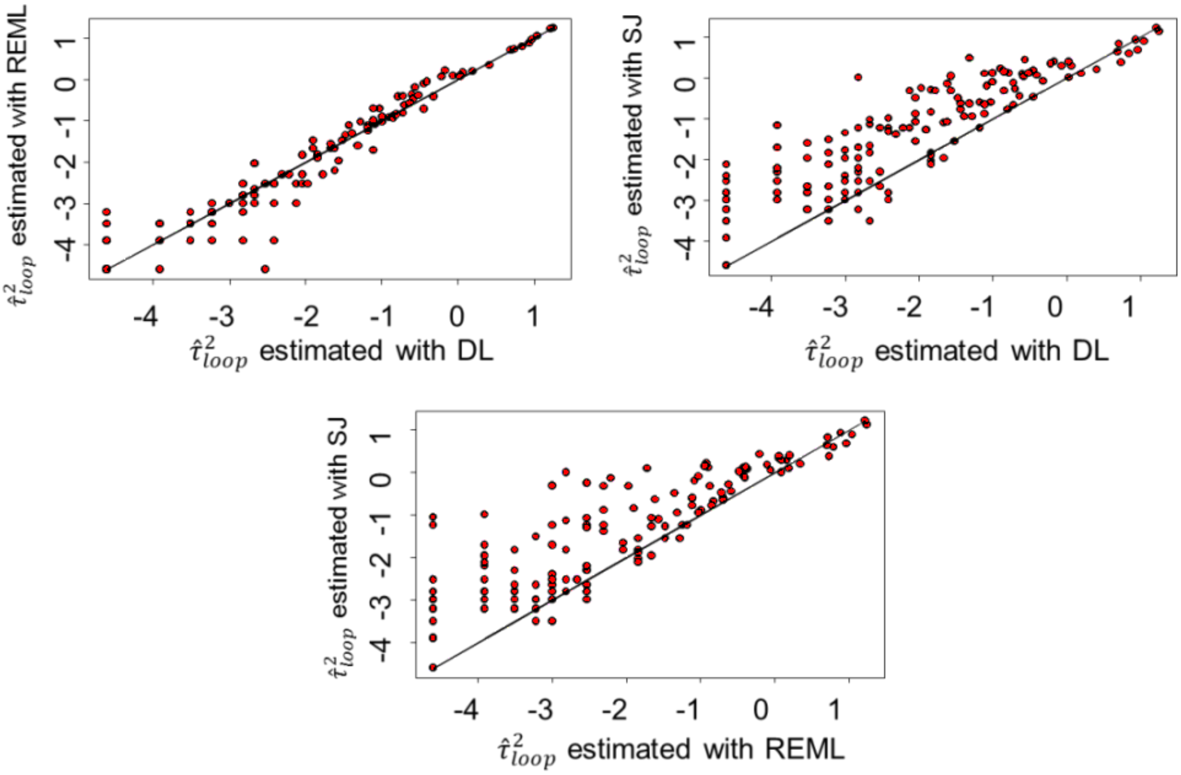


Figure 5. Comparison of the heterogeneity estimated with DerSimonian and Laird (DL), restricted maximum likelihood (REML) and Sidik-Jonkman (SJ) methods; the heterogeneity is presented on the log scale when applying the loop-specific approach (common within-loop heterogeneity variance, τ^2_{loop}) in the 303 loops.

In Figure 5 I compare the estimated heterogeneity on the log scale using the DL, REML and SJ methods. I show that the SJ method is associated with larger values of heterogeneity, leading to fewer inconsistent loops than the other two methods (see Table 8).

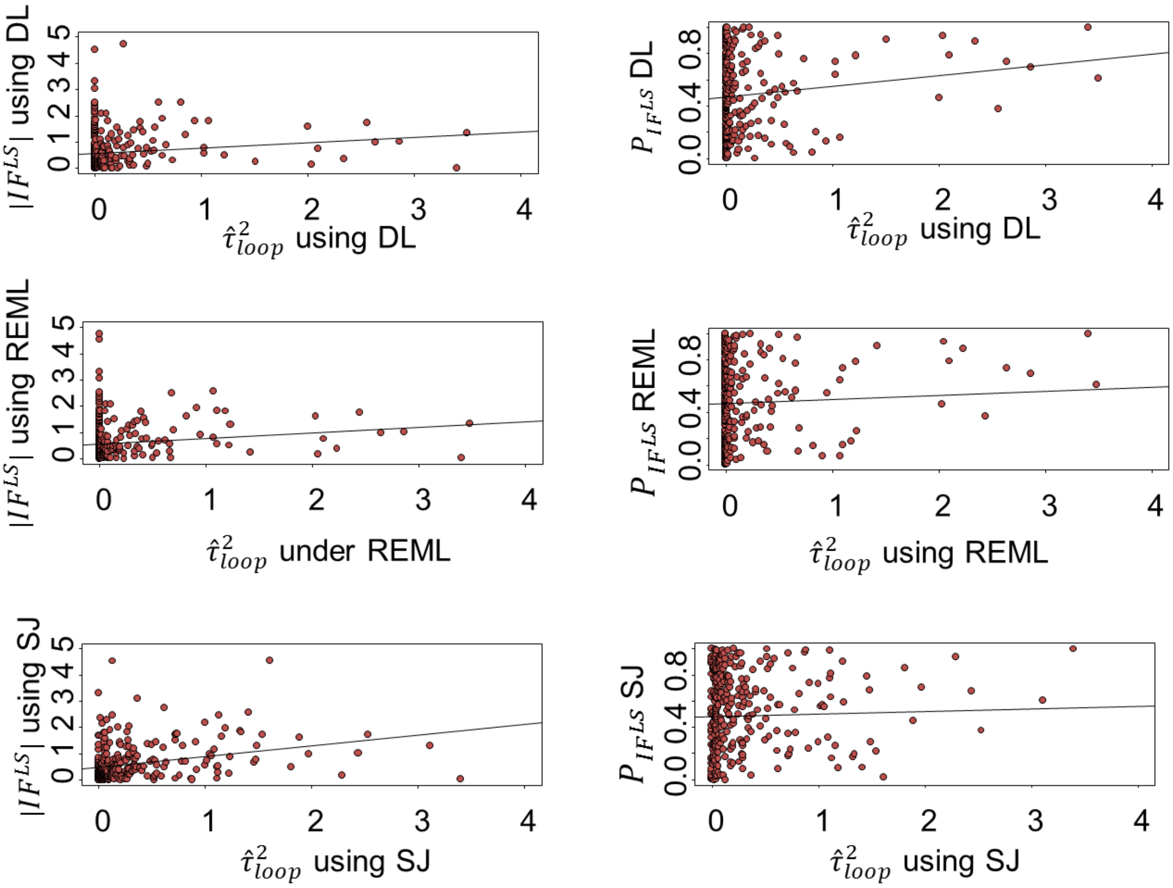


Figure 6. The left-hand side panels represent a plot of inconsistency estimate ($\widehat{IF}^{LS}$) against the heterogeneity ($\hat{\tau}^2$) and the right-hand side panels correspond to a plot of the P value of $\widehat{IF}^{LS}$ against $\hat{\tau}^2$. Inconsistency is estimated using the common within-loop heterogeneity variance and the DerSimonian and Laird (DL), restricted maximum likelihood (REML) and Sidik-Jonkman (SJ) methods.

For each loop, I compared the IF^{LS} and its P value with the estimated heterogeneity for each loop ($\hat{\tau}^2_{loop}$) using the three estimators (see Figure 6). Irrespective of the estimation method used, the magnitude of inconsistency increases slightly as the estimated heterogeneity increases. Conversely, lower values of the heterogeneity are associated with greater chances of identifying significant IF^{LS}, though the correlation coefficients between the P value or IF^{LS} and the heterogeneity variance are very small (correlation coefficients for $\widehat{IF}^{LS}$ versus $\hat{\tau}^2$: $r_{DL} = 0.14$ $r_{REML} = 0.15$, $r_{SJ} = 0.29$; correlation coefficients for P value of $\widehat{IF^{LS}}$ versus $\hat{\tau}^2$: $r_{DL} = 0.13$, $r_{REML} = 0.13$, $r_{SJ} = 0.04$).

Statistical approaches to evaluate the assumption of consistency

Separating indirect and direct evidence (SODR) method

The detachment of 31 (9%) designs (total 362) suggested important disagreement between those designs and the remaining network. In 16 (40%) networks I identified at least one inconsistent design (see Appendix Table 6). The median number of inconsistent designs per network was 2 and ranged from 0 to 4. In Table 9 I examine whether the magnitude of estimated heterogeneity and the number of studies informing each design are associated with more or less chances of identifying inconsistency. Although designs informed by a single study reject more often the assumption of consistency, the percentage of inconsistency did not significantly differ from designs informed by two or more studies (19/188 (10%) *vs*. 12/174 (7%); $P = 0.367$). The total heterogeneity was estimated equal to zero in the detachment of 117 (32%) designs after the detachment of the particular design in 24 (60%) networks. Twelve (10%) out of the 117 designs were found inconsistent (Table 9).

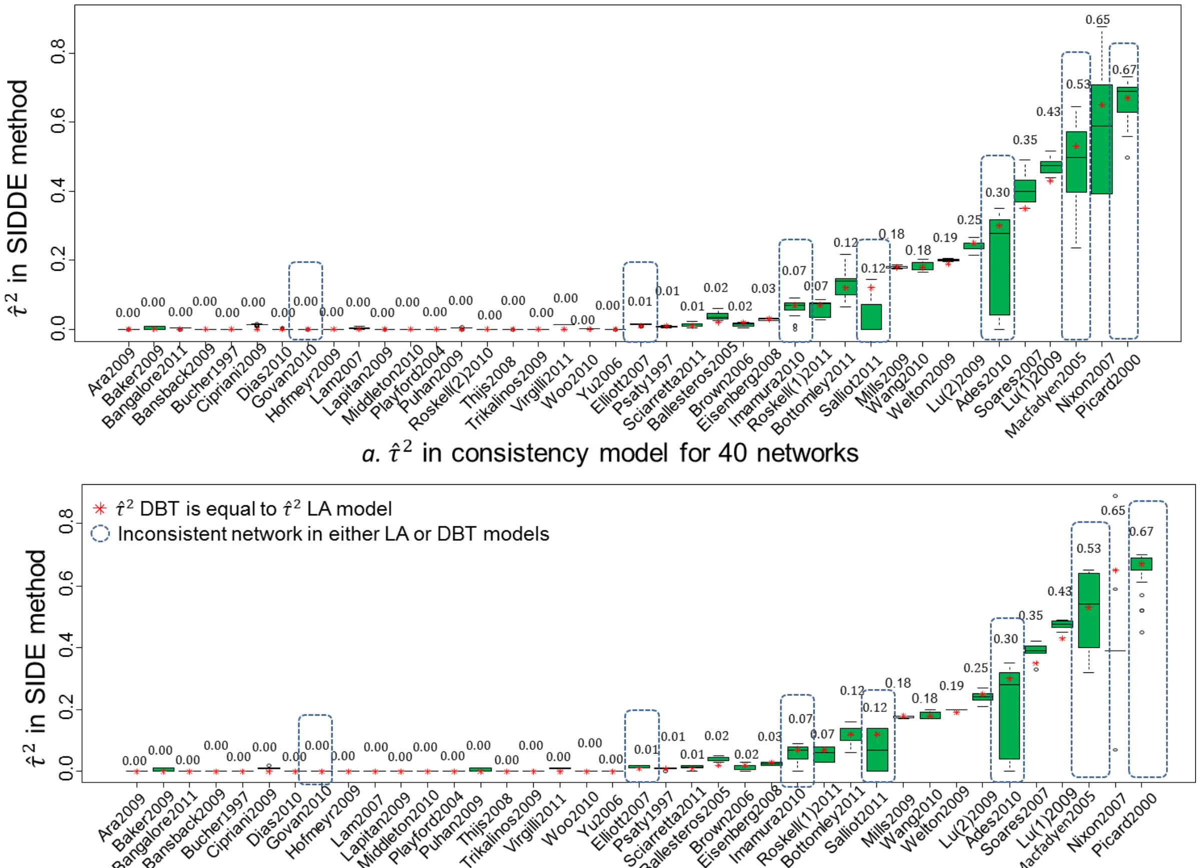
Table 9. Frequency of inconsistent designs and comparisons estimated using the SODR and SIDE methods. The number of designs/comparisons is provided when $\hat{\tau}^2$ is equal to zero, as well as when a design/comparison includes a single study. Comparisons that were found inconsistent with at least one parameterisation of the multi-arm studies were classified as inconsistent in SIDE method.

	SODR method				
Designs	$\hat{\tau}^2 = 0$	***Designs including a single study***	***Designs including a single study and*** $\hat{\tau}^2 = 0$		***Total***
Consistent	105 (90%)	167 (90%)	46 (89%)		331 (91%)
Inconsistent	12 (10%)	19 (10%)	6 (11%)		31 (9%)
Total	117 (32%)	188 (52%)	52 (14%)		362 (100%)
	SIDE method				
Comparisons	$\hat{\tau}^2 = 0$	***Comparisons including a single study***	***Comparisons including a single study and*** $\hat{\tau}^2 = 0$	***Comparisons included in at least one multi-arm trial***	***Total***
Consistent	124 (87%)	109 (89%)	37 (86%)	182 (89%)	309 (89%)
Inconsistent	18 (13%)	14 (11%)	6 (14%)	23 (11%)	39 (11%)
Total	142 (36%)	123 (35%)	43 (12%)	205 (59%)	348 (100%)

In Figure 7a I compare the distribution of the heterogeneity estimates in SODR method and I compare it to the heterogeneity estimated in the consistency model. In half of the networks with global inconsistency the heterogeneity of the consistency model is larger than the median $\hat{\tau}^2$ from the SODR model, suggesting again that in those networks the extra variability due to inconsistency might have inflated the heterogeneity parameter in the consistency model. As presented in Figure 7a heterogeneity can drop substantially when detaching some designs (see for example the network by Macfadyen 2005 (84)). Given that the power of inconsistency tests is suspected to be low, monitoring changes in heterogeneity might be used as an alternative; designs whose detachment leads to important drops in heterogeneity are suspected for inconsistency.

Separating indirect and direct evidence (SIDE) method

Inconsistency was prevalent in 39 (11%) comparisons (total 348) that belonged to 19 (49%) different networks out of the total 39 (see Appendix Table 7). The number of inconsistent comparisons per network ranged from 0 to 6 (median 2). In total 205 (59%) comparisons were informed by at least one multi-arm study in 76 (22%) of which the magnitude of heterogeneity changed when a multi-arm study was parameterised differently. Inference about inconsistency changed in 16 comparisons (5%) when parameterisation changed in multi-arm studies. Similarly to SODR method, inconsistency did not change when I restricted the analysis to comparisons for which at least two studies provide direct evidence (14/123 (11%) *vs*. 25/225(11%); $P = 1$) (Table 9). In Figure 7b I present the distribution of the heterogeneity estimated in SIDE method compared to the network heterogeneity estimated in the consistency model. Again, the detachment of some comparisons in networks can decrease heterogeneity a lot and this can be used as an alternative to the test for inconsistency.



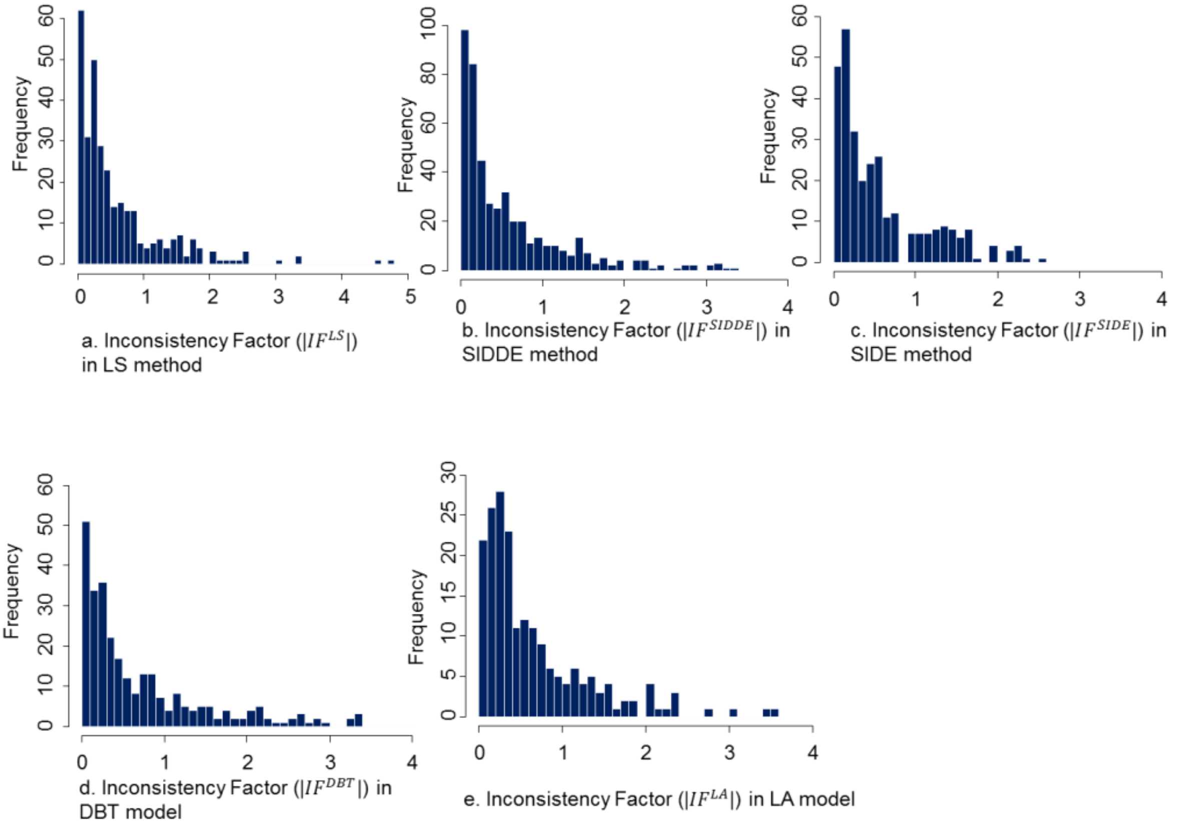
a. $\hat{\tau}^2$ in consistency model for 40 networks

b. $\hat{\tau}^2$ in consistency model for 39 networks

Figure 7. Comparison of the estimated heterogeneity ($\hat{\tau}^2$) in SODR method (panel a) and SIDE method (panel b) versus $\hat{\tau}^2$ in the consistency (NMA) model. Each boxplot represents the distribution of $\hat{\tau}^2$ when all designs (panel a) or comparisons (panel b) of a network were detached. Numbers on the top of the boxplots and red stars represent $\hat{\tau}^2$ in the consistency model. Note that in one network (66) inconsistency could not be evaluated using the SIDE method. Circled boxplots are inconsistent networks in either LA or DBT models. SODR: Separating one design from the rest, SIDE: separating indirect and direct evidence.

Comparison of approaches to evaluate inconsistency and distribution of I2 for inconsistency

In Figure 8 I present the histogram of the absolute IFs as estimated in each approach separately. To evaluate loop inconsistency using the LS method I used 303 closed loops of evidence. The median inconsistency using the LS method and $\hat{\tau}^2_{loop}$ with DL estimator was ROR=1.40 with IQR (1.16, 2.20). For SODR method I separated 362 different designs from their networks and I applied the SIDE method in 348 treatment comparisons. Figure 8b and Figure 8c show the histogram of the absolute IFs as estimated using the SODR and SIDE methods, respectively. Overall the evidence coming from different designs for the same comparison disagreed with a median ROR 1.37 with IQR (1.14, 2.18). Similar results were obtained with the SIDE method; direct and indirect evidence for the same comparison disagreed by a median ROR 1.44 with IQR (1.15, 2.16). Figure 8d and Figure 8e show the histogram of the absolute IFs for the DBT and LA models, respectively. For the LA model where different $\widehat{IF}$ values can occur with different parameterisation of multi-arm studies I display the maximum $\widehat{IF}$ per loop. Overall the evidence coming from different designs was found to disagree by a median ROR 1.49 with IQR (1.16, 2.84), whereas the LA model showed that direct evidence disagreed to indirect one in a closed loop by median ROR 1.48 with IQR (1.22, 2.46).



a. Inconsistency Factor ($|IF^{LS}|$) in LS method

b. Inconsistency Factor ($|IF^{SIDDE}|$) in SIDDE method

c. Inconsistency Factor ($|IF^{SIDE}|$) in SIDE method

d. Inconsistency Factor ($|IF^{DBT}|$) in DBT model

e. Inconsistency Factor ($|IF^{LA}|$) in LA model

Figure 8. Histogram of the absolute values of the inconsistency factors (IF) for the OR effect measure estimated using a) the LS method b) the SODR method, c) the SIDE method, d) the DBT model, and e) the LA model. In (b) and (e) histograms I display the maximum IF in case of multiple IFs for the same loop or comparison (due to different parameterisation of multi-arms studies).

In Table 10 I compare the number of inconsistent networks using the LA model and the DBT model. Two networks were found to be inconsistent in the LA model, but consistent with the DBT approach. This might be due to the differences in the estimation of the heterogeneity and the fact that I used the maximum test value from the different parameterisations of the multi-arm studies to infer about inconsistency in the LA model. In Figure 9 I compare the heterogeneity as estimated with the DBT and LA models. For networks with non-negligible heterogeneity (e.g. $\hat{\tau}^2$>0.05) the estimate from the DBT model is at the lower spectrum of values of τ^2 estimated with the LA method. This might suggest that in these networks there is design inconsistency which is not accounted for in the LA model and hence the variability associated with it is encompassed in the heterogeneity.

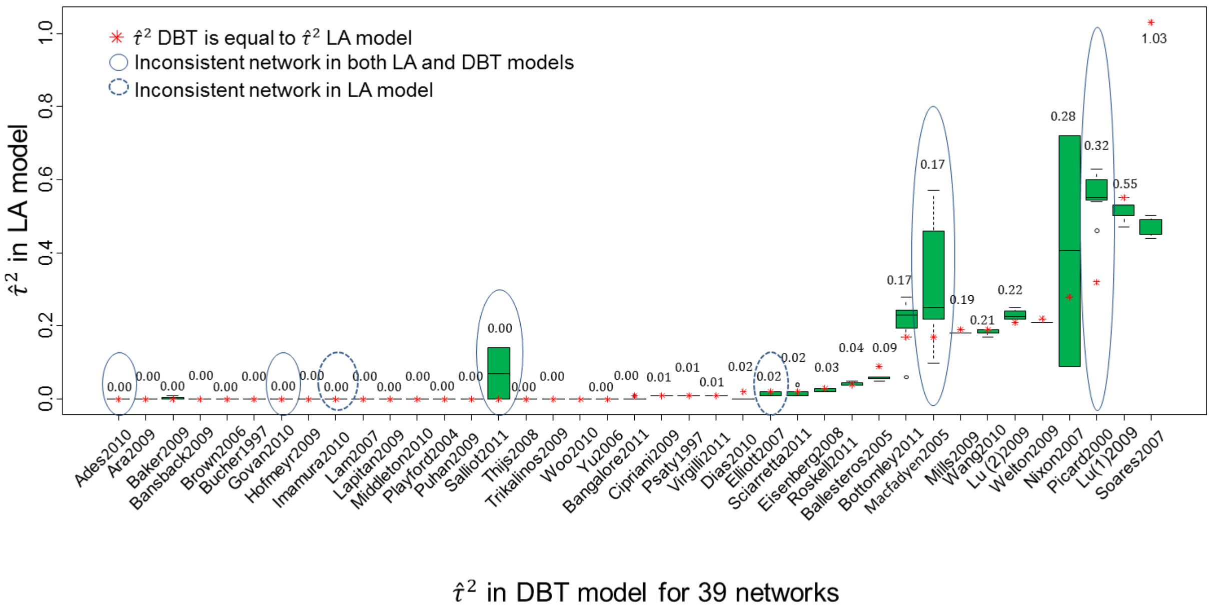


Figure 9. Comparison of the estimated heterogeneity ($\hat{\tau}^2$) in Lu and Ades (LA) model and design-by-treatment interaction (DBT) model. Each boxplot represents the estimates for $\hat{\tau}^2$ resulting from different parameterisations of the multi-arm studies included in each of the 39 networks. For one network39 inconsistency could not be evaluated using the LA model. Red points are the estimates of $\hat{\tau}^2$ in DBT model. Full circles denote inconsistent networks in both LA and DBT models and dashed circles denote inconsistent networks in LA model only. Networks found inconsistency with at least one parameterisation of the multi-arm studies were considered inconsistent.

Table 10. Number of consistent and inconsistent networks using the Lu and Ades (LA) and design-by-treatment interaction (DBT) models. Networks that were found inconsistent with at least one parameterisation of the multi-arm studies were classified as inconsistent in LA model. Note that in one network (66) inconsistency could not be evaluated using the LA approach.

		Lu and Ades model		
		Consistent	*Inconsistent*	*Total*
DBT model	*Consistent*	32	2	34 (87%)
	Inconsistent	0	5	5 (13%)
	Total	32 (82%)	7 (18%)	39 (100%)

In Figure 10 I present the distribution of I^2_{Inc} (median 50%, IQR(27%,71%)), I^2_{Het} (median 26%, IQR(0%,63%)), and $I^2_{Het,Inc}$ (median 72%, IQR(46%,86%)).

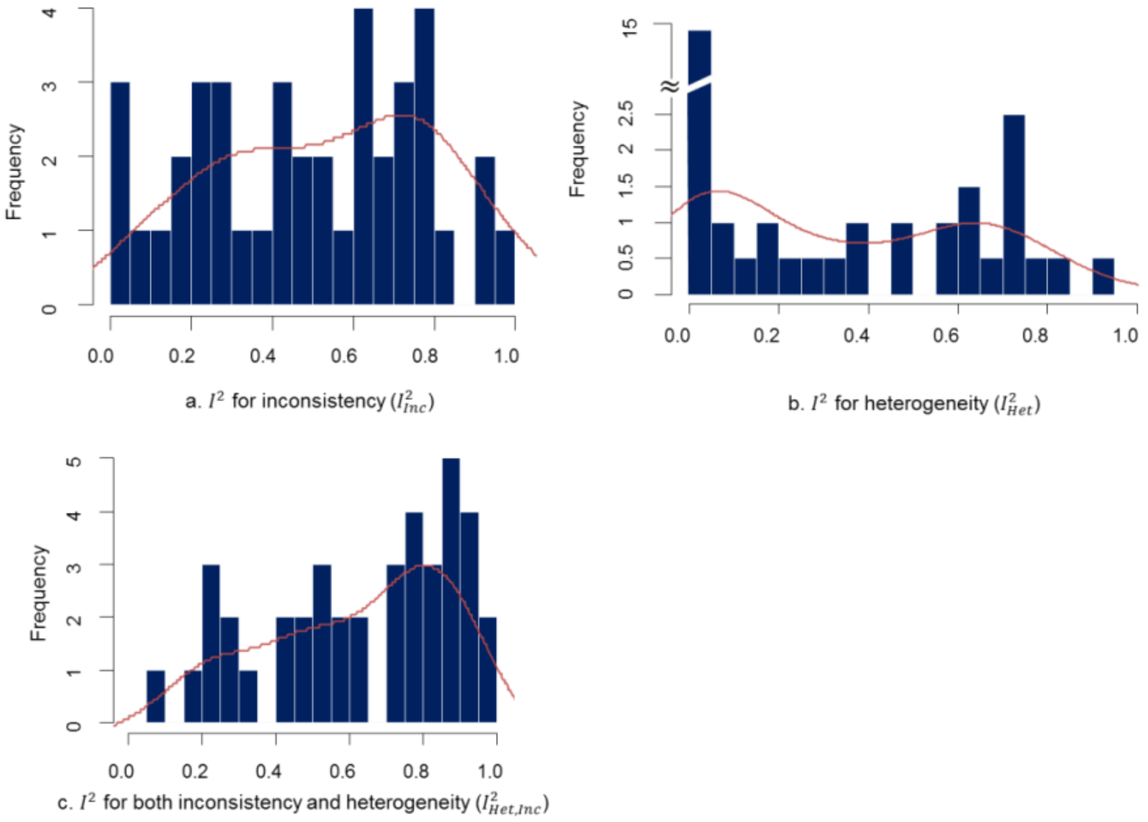


Figure 10. Histogram of the I^2 values when accounting for a) inconsistency, b) heterogeneity and c) both inconsistency and heterogeneity in the network.

4.3 Simulation study

4.3.1 Simulation study design

The simulation study was informed by our empirical study (for a description of the dataset see section 4.2.5). I restricted our analysis to dichotomous outcome data measured using a single scale as empirical evidence showed no differences in inconsistency using different effect sizes (13). I selected OR due to its good mathematical properties (22,23). Let us consider a three-treatment network ABC with AB, AC and BC trials. I assumed the summary treatment effect for the AB comparison $OR_{\mathrm{AB}} = 1/\exp(0.32) = 0.73$ and for the AC comparison OR_{AC} =1. I computed the OR for BC comparison as:

$$OR_{\mathrm{BC}} = \exp\{\log(OR_{\mathrm{AC}}) - \log(OR_{\mathrm{AB}}) + IF^{LS}_{ABC}\}.$$

I selected values $IF_{ABC}^{LS} = \{0, 0.3, 0.45, 0.6, 1\}$ to cover a range of plausible values for inconsistency as suggested by empirical data. I considered two different distributions for heterogeneity that pertain to a subjective outcome (the most frequently reported outcome in our data) and all-cause mortality for comparisons between pharmacological interventions and placebo; according to (30) these are $\tau^2 \sim LN(-2.13, 1.58^2)$ and $\tau^2 \sim LN(-4.06, 1.45^2)$ (median τ^2 = 0.02 with (IQR 0.01, 0.04)) (30).

Let k_1, k_2 and k_3 represent the number of studies included in the three comparisons AB, AC and BC respectively, with $k_1 = 1, \dots, K_{\text{AB}}$, $k_2 = 1, \dots, K_{\text{AC}}$ and $k_3 = 1, \dots, K_{\text{BC}}$. I examined both networks where all comparisons include the same number of trials $K_{\text{AB}} = K_{\text{AC}} = K_{\text{BC}} = K = 1, \dots, 7$ ('balanced' direct comparisons) and networks where each comparison is informed by a different number of trials with $K_{\text{AB}} = 1, K_{\text{AC}} = 4, K_{\text{BC}} = 7$ ('imbalanced' direct comparisons).

For each combination of OR, IF_{ABC}^{LS}, and τ^2 I simulated the trial-specific underlying relative treatment effects from a normal distribution as:

$$LOR_{\text{AB},k_1} \sim N(LOR_{\text{AB}}, \tau^2)$$

$$LOR_{\text{AC},k_2} \sim N(LOR_{\text{AC}}, \tau^2)$$

$$LOR_{\text{BC},k_3} \sim N(LOR_{\text{BC}}, \tau^2).$$

I generated arm-level data for each trial k_1, k_2 and k_3. For an AB trial I assumed equal trial sizes across arms: $n_{\text{A},k_1} = n_{\text{B},k_1} = n$ and I selected small, moderate and large trial sizes sampled from $n \sim U(20,50)$, $n \sim U(50,150)$ and $n \sim U(150,300)$, respectively. The number of events per arm, denoted with $r_{A,k_1}$and r_{B,k_1} are drawn from the binomial distributions:

$$r_{\text{A},k_1} \sim B\left(n_{\text{A},k_1}, p_{\text{A},k_1}\right)$$

$$r_{\text{B},k_1} \sim B\left(n_{\text{B},k_1}, p_{\text{B},k_1}\right)$$

where p_{A,k_1} and p_{B,k_1} are the probabilities of the outcome in each trial arm. To define these probabilities I make assumptions about the average risk (AR) of the outcome in the trial assuming both frequent and rare events. I simulated the AR from a uniform distribution as:

$AR_{\text{AB},k_1} \sim U(0.25, 0.75)$, for frequent event rates

$AR_{\mathrm{AB},k_1} \sim U(0.05, 0.15)$, for rare event rates.

I obtained the event probabilities in the arms as the solution to the equations:

$$AR_{\mathrm{AB},k_1} = \frac{p_{\mathrm{A},k_1} + p_{\mathrm{B},k_1}}{2}$$

$$LOR_{\mathrm{AB},k_1} = \log\left(\frac{p_{\mathrm{A},k_1}\left(1 - p_{\mathrm{B},k_1}\right)}{p_{\mathrm{B},k_1}\left(1 - p_{\mathrm{A},k_1}\right)}\right)$$

I then calculated the sample LOR and its variance as

$$LOR^{*}_{\mathrm{AB},k_1} = \log\left(\frac{r_{\mathrm{A},k_1}\left(n_{\mathrm{B},k_1} - r_{\mathrm{B},k_1}\right)}{r_{\mathrm{B},k_1}\left(n_{\mathrm{A},k_1} - r_{\mathrm{A},k_1}\right)}\right)$$

$$v^{*}_{\mathrm{AB},k_1} = \frac{1}{r_{\mathrm{A},k_1}} + \frac{1}{n_{\mathrm{A},k_1} - r_{\mathrm{A},k_1}} + \frac{1}{r_{\mathrm{B},k_1}} + \frac{1}{n_{\mathrm{B},k_1} - r_{\mathrm{B},k_1}}$$

If the simulated number of events in one of the study arms is zero, I add 0.5 to the cells of the 2×2 table. I repeated this process for all K_{AB} trials and then I perform a random-effects meta-analysis to obtain the summary effect size $\hat{\mu}_{\mathrm{AB}}$. I followed the same process for comparisons AC and BC and then I estimate the inconsistency factor.

I also considered an extra scenario representing the ‘typical’ loop; that is a loop with the characteristics most commonly encountered in our empirical study (13). Most three-treatment loops (65%) had at least one comparison informed by a single trial and a median sample size in the loop of 2310 participants. Based on the empirical distribution of trials in triangular loops I generate data for frequent events, $K_{\mathrm{AB}} = 1, K_{\mathrm{AC}} = 4, K_{\mathrm{BC}} = 3,$ and sample size per arm drawn from $n \sim U(120{,}160)$.

For each scenario I analysed 1000 simulated triangular networks. Assuming 0.05 the significance level, I estimated the power of the W^{LS}_{ABC} test when true inconsistency was present $(P(|\mathrm{z}| \geq 1.96 | IF^{LS}_{\mathrm{ABC}} \neq 0)$ and type I error when the null hypothesis was true $(P(|\mathrm{z}| \geq 1.96 | IF^{LS}_{\mathrm{ABC}} = 0)$. I computed the coverage probability for the CI of inconsistency, i.e. the probability that the estimated interval for IF^{LS} included its true value. I carried out the simulations in the freely available software R 2.15.2 using the self-programmed *sims.fun* function, which is available online (in http://www.mtm.uoi.gr/ under ‘Material from Publications’).

I estimated the uncertainty of the pairwise summary effects by employing four different strategies: the Wald type method using DL estimator (WtDL), REML (WtREML) and PM (WtPM) estimators for the heterogeneity and the Knapp-Hartung method with the DL estimator (KHDL). The main two differences between the Wt and KH methods are a) the Wt method estimates the variance of the overall treatment effect for a specific comparison, e.g. AB, using the inverse of the sum of the study weights, whereas the KH method derives the variance of the overall treatment effect as the ratio of a generalised Q statistic divided by the product of the degrees of freedom ($K_{AB} - 1$) and the sum of the random-effects study weights (for more details see sections 2.3 and 2.4). When a comparison was addressed by a single trial (so that the loop includes 3 trials in total) estimation of heterogeneity is impossible. In these cases I used the fixed-effect model (by setting $\tau^2 = 0$) and consequently all methods (WtDL, WtREML, WtPM and KHDL) would yield exactly the same results.

4.3.2 Estimation of the variance of inconsistency

In this section I explain how $Var(\widehat{IF}_{ABC}^{LS})$ depends on the magnitude of the heterogeneity as well as the number of trials included in the network. Without a loss of generalisation I use the Wt approach. The variances of the direct mean treatment effects are functions of the within-study variances $\widehat{v_i}$ and the heterogeneity τ^2. Let K_{AB}, K_{AC} and K_{BC} trials inform the AB, AC and BC comparisons respectively. Assuming the sampling variances are the same for all trials ($\hat{v}$), the variance of inconsistency is obtained by

$$Var(\widehat{IF}_{ABC}^{LS}) = \hat{v}\left(\frac{1}{K_{AB}} + \frac{1}{K_{AC}} + \frac{1}{K_{BC}}\right) + 3\hat{\tau}^2. \qquad (4.1)$$

Formula (4.1) shows that $Var(\widehat{IF}_{ABC}^{LS})$ increases with the heterogeneity and decreases with the number of the trials included in the network.

4.3.3 Type I error

The relatively small number of trials included in each pairwise meta-analysis (fewer than 7) and the magnitude of assumed heterogeneity for a subjective outcome (median $\tau^2 = 0.12$) make bias and MSE for τ^2 comparable between the three estimators WtDL, WtREML and WtPM. Type I error was therefore comparable between the WtDL, WtREML and WtPM methods (data not shown) and I present results only from WtDL and KHDL. Figure 11 and Figure 12 in display the estimated type I error for equal and different numbers of trials across comparisons. In general, type I error was close to the nominal level for the WtDL method, but larger than 5% for many scenarios analysed with KHDL. The KHDL method generally yielded smaller variances for IF^{LS} leading to larger type I errors. Type I error converged to the nominal level more rapidly when $\tau^2 = 0$ for both the WtDL and KHDL methods. The overall type I error approached the nominal level as the number of trials increases for the same trial size. In the current simulation scenarios, I did not find important differences between the three estimators of τ^2.

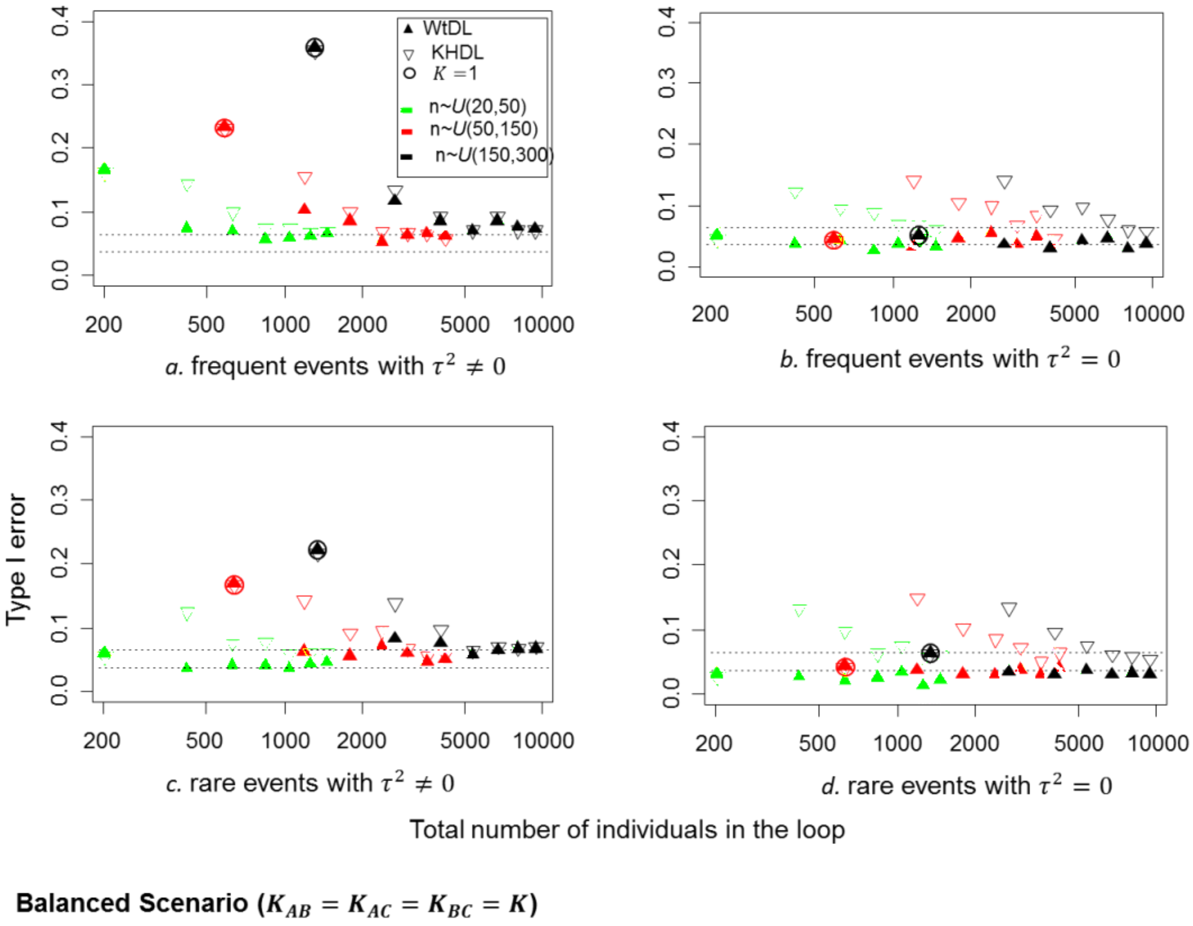


Figure 11. Type I error by sample size (n), frequency of events and loop sample size. Equal number of trials per comparison ($K_{AB} = K_{AC} = K_{BC} = K = 1, \dots, 7$) is assumed in the presence ($\tau^2 \neq 0$) and absence ($\tau^2 = 0$) of heterogeneity. Circled points correspond to loops with single study for which a fixed-effects model is employed. The region within the horizontal dotted lines defines the confidence interval for the 0.05 nominal level. n: sample size, WtDL: Wald type method using the DerSimonian and Laird estimator, KHDL: Knapp and Hartung method with DL estimator.

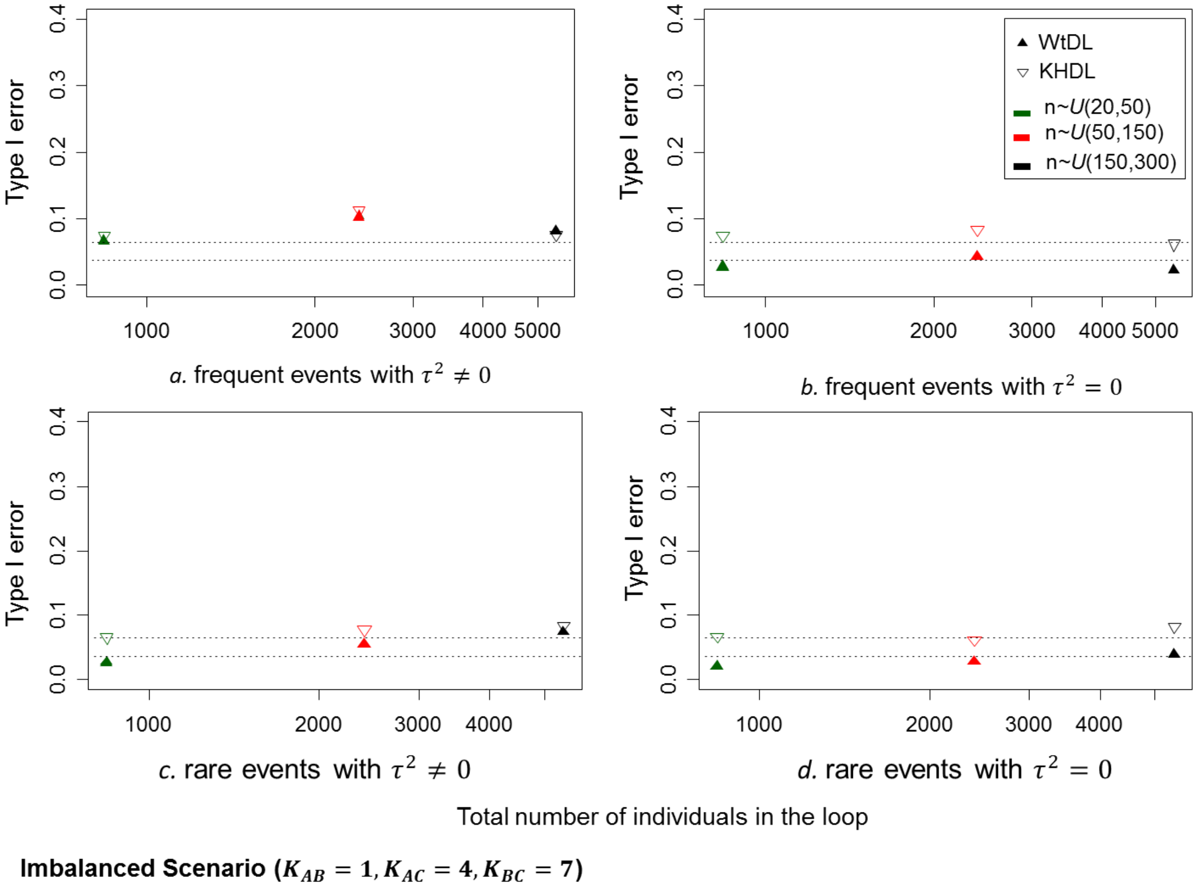


Figure 12. Type I error by sample size (n), frequency of events and loop sample size. Results are shown assuming different number of trials (K) per comparison ($K_{AB} = 1, K_{AC} = 4, K_{BC} = 7$). The region within the horizontal dotted lines defines the confidence interval for the 0.05 nominal level. WtDL: Wald type method with DerSimonian and Laird (DL) estimator, KHDL: Knapp and Hartung method with DL estimator.

In Table 11 I provide type I error values for various simulation scenarios. When the total number of individuals included in the network ranges from 2400 to 3000, type I error lied between 0.06–0.08. Type I error deviated from 0.05 considerably when an equal and small number of trials is considered across comparisons for all trial sizes.

For rare events, type I error departed from 0.05 in a greater extent than it does for frequent events. Type I error was lower than its nominal level in most cases for the WtDL method especially when $\tau^2 = 0$, probably due to overestimation of τ^2. The KHDL method resulted again in considerably larger type I errors, which is probably due to the small variances of the mean treatment effects. Type I error is closer to the nominal level for WtDL when $\tau^2 \neq 0$ for all sample sizes. All methods tend to improve their performance with increasing total number of trials included in the entire network (see Figure 11 and Figure 12).

Table 11. Type I error, power and coverage probability by sample size (n) and number of trials (K). Results are presented for frequent events and aggregated over different assumptions for heterogeneity and methods to estimate the variances of the mean treatment effects. In bold I present results from loops in which the total number of individuals is between 2400 and 3000.

	Balanced Scenario ($\boldsymbol{K_{AB} = K_{AC} = K_{BC} = K}$)							**Imbalanced Scenario**
	K=1	K=2	K=3	K=4	K=5	K=6	K=7	K_{AB}=1 K_{AC}=4 K_{Bc}=7
Type I error ($\boldsymbol{IF^{LS}}$= 0)								
n~U(20,50)	0.07	0.07	0.06	0.04	0.05	0.05	0.04	0.06
n~U(50,150)	0.10	0.07	0.06	0.06	**0.05**	**0.06**	0.04	**0.08**
n~U(150,300)	0.13	**0.07**	0.05	0.06	0.06	0.04	0.05	0.06
Power ($\boldsymbol{IF^{LS}}$ = 0.6)								
n~U(20,50)	0.13	0.15	0.18	0.23	0.27	0.33	0.37	0.16
n~U(50,150)	0.25	0.30	0.42	0.52	**0.62**	**0.70**	0.76	**0.32**
n~U(150,300)	0.42	**0.54**	0.70	0.79	0.84	0.88	0.89	0.49
Coverage Probability ($\boldsymbol{IF^{LS}}$ = 0.6)								
n~U(20,50)	0.96	0.96	0.97	0.98	0.97	0.97	0.97	0.97
n~U(50,150)	0.95	0.96	0.97	0.96	**0.96**	**0.96**	0.96	**0.95**
n~U(150,300)	0.93	**0.95**	0.94	0.94	0.96	0.95	0.95	0.95

4.3.4 Statistical Power

In Figure 13 and Figure 14 present the power for $IF^{LS} = \{0.3,0.45,0.6,1\}$ for both frequent and rare events when equal (Figure 13) and different (Figure 14) numbers of trials are included in comparisons. The overall power increases both with the number of trials included in each pairwise meta-analysis and with their sample size. Results were aggregated over all estimation methods for heterogeneity and the different methods to estimate the variance of the direct summary effects. In Table 11 I provide the power values for various simulation scenarios when IF^{LS} = 0.6 and frequent events are considered. When the total number of individuals included in the network ranges from 2400 to 3000, power ranged between 0.54 and 0.70 when an equal number of trials was assumed across comparisons but dropped to 0.32 when each comparison had a different number of trials. As can be seen in equation (4.1), the distribution of trials across comparisons affects the estimation of the variance of IF^{LS}. This has an impact on power and the test is more powerful when trials are distributed uniformly across comparisons. The comparison of frequent (Figure 13a) and rare (Figure 13b) events indicates that power is larger for frequent events. Rare events were associated with larger variability for the pairwise summary treatment effects and hence the chances of identifying potentially important inconsistency decrease. It should be noted that the first summary result of each power curve pertains to the case where there is only one trial per comparison and heterogeneity is set to be zero. This has an impact on monotonicity especially when IF^{LS} is low and sample size is large.

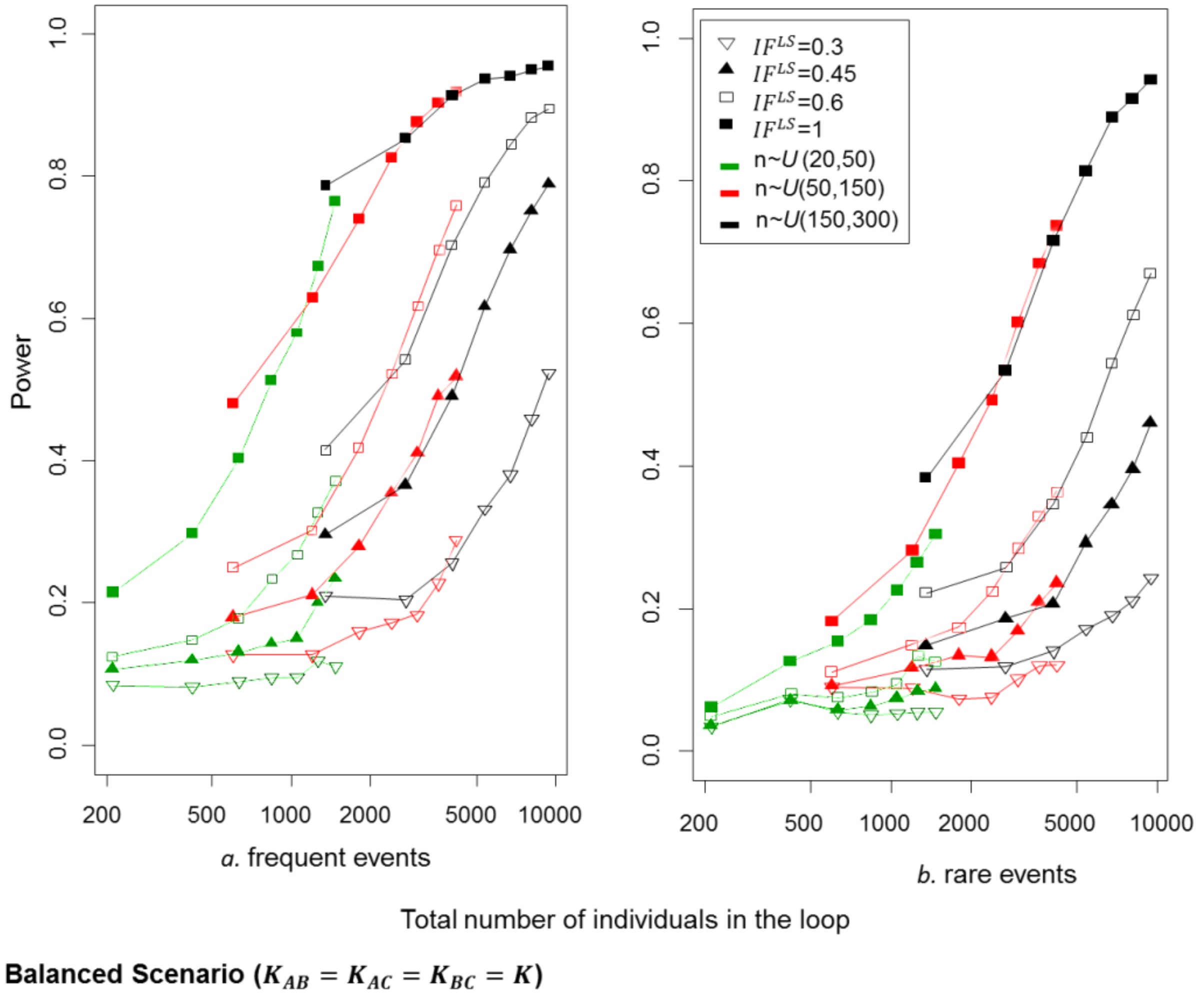


Figure 13. Power by magnitude of inconsistency factor (IF^{LS}), frequency of events and loop sample size. Power is presented for different sample sizes (n) assuming equal number of trials per comparison ($K_{AB} = K_{AC} = K_{BC} = K = 1, \dots, 7$). Results are aggregated over different assumptions for heterogeneity and methods to estimate the variance of the mean treatment effect. The first summary result in each power curve pertains to the case where there is a one trial per comparison and a fixed-effects model is employed.

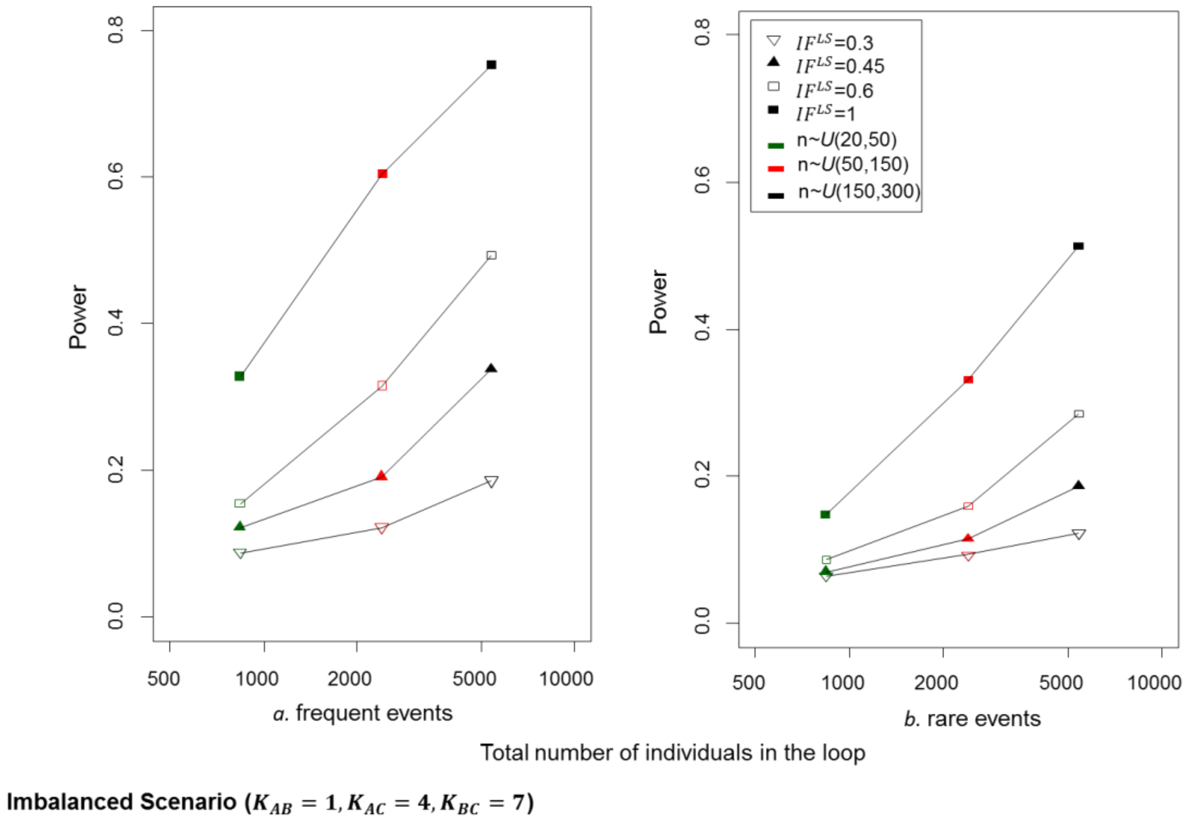


Figure 14.Power by magnitude of inconsistency factor (IF), frequency of events and loop sample size. I assume different number of trials (K) per comparison ($K_{AB} = 1, K_{AC} = 4, K_{BC} = 7$). Results are aggregated over different assumptions for the heterogeneity and methods to estimate the variances of the mean treatment effects.

In

Table **12** and Table 13 I present the power for the WtDL and KHDL methods. For frequent events the power to detect inconsistency does not vary significantly with the method used to estimate heterogeneity or to express uncertainty on the summary effects although the KH method is marginally more powerful, especially in the absence of heterogeneity. This is because, in many cases, the KH estimates smaller variances for inconsistency compared with the Wt method. The median inconsistency standard error is 0.33 (IQR 0.21, 0.50) for the KHDL method and 0.40 (IQR 0.27, 0.57) for the WtDL approach. These findings agree with a previous simulation study (25), which showed that when heterogeneity is zero the KH method yields a smaller variance for the mean treatment effects than the Wt method. As anticipated, when there is no heterogeneity, there is less uncertainty associated with each pairwise effect and the power to detect inconsistency increases for all IF^{LS} values.

Table 12. Power of the W^{LS} test aggregated over sample size. WtDL: Wald type method with DerSimonian and Laird (DL) estimator, KHDL: Knapp and Hartung method with DL estimator, IF: inconsistency factor

	Heterogeneity				No Heterogeneity			
$\boldsymbol{IF^{LS}}$	*0.3*	*0.45*	*0.6*	*1*	*0.3*	*0.45*	*0.3*	*0.45*
				Frequent Events				
WtDL	0.10	0.15	0.23	0.42	0.13	0.23	0.38	0.68
KHDL	0.11	0.17	0.24	0.42	0.19	0.31	0.44	0.73
				Rare Events				
WtDL	0.08	0.10	0.14	0.25	0.07	0.11	0.17	0.35
KHDL	0.11	0.12	0.16	0.28	0.12	0.17	0.25	0.44

The impact of heterogeneity is similar when the outcome is rare. Table 13 shows that the advantage of KHDL method when heterogeneity is zero becomes more pronounced for rare events. The WtREML and WtPM methods yielded similar power to WtDL.

Table 13. Power of the W^{LS} test aggregated over sample size and number of trials. Results are presented for equal number of trials across comparisons. IF^{LS}: inconsistency factor, WtDL: Wald type method with DerSimonian and Laird (DL) estimator, KHDL: the Knapp and Hartung method with DL estimator.

	Heterogeneity				No Heterogeneity			
$\boldsymbol{IF^{LS}}$	*0.3*	*0.45*	*0.6*	*1*	*0.3*	*0.45*	*0.6*	*1*
				Frequent Events				
WtDL	0.17	0.26	0.36	0.59	0.20	0.38	0.52	0.77
KHDL	0.19	0.27	0.37	0.60	0.27	0.44	0.58	0.80
				Rare Events				
WtDL	0.10	0.15	0.21	0.38	0.09	0.16	0.25	0.49
KHDL	0.13	0.18	0.24	0.41	0.16	0.23	0.33	0.55

4.3.5 Coverage Probability and Bias

I assess how often the 95% CI for inconsistency included the assumed IF^{LS} value used to generate the data. I plot the coverage probability for the 95% CI of IF^{LS} in Figure 15 and Figure 16. The coverage probability is close to the nominal level (95%) for most settings. Rare events were associated with larger uncertainty and therefore provide higher coverage than frequent events. In Table 11 I provide the coverage values for various simulation scenarios when IF^{LS}= 0.6. When the total number of individuals included in the network ranged from 2400 to 3000, coverage ranged from 95% to 96%. Coverage did not change considerably when an equal or different number of trials is assumed across comparisons.

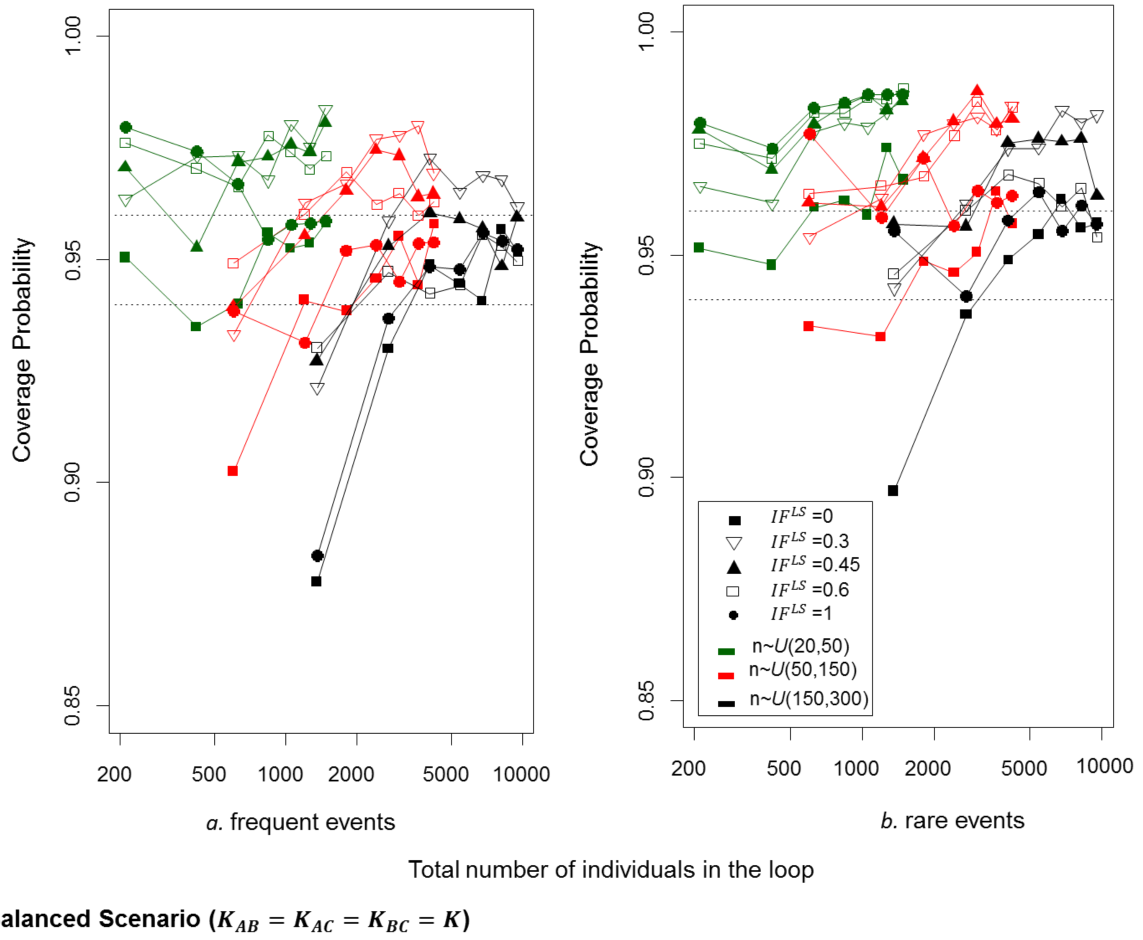


Figure 15. Coverage probabilities of the 95% confidence interval for the inconsistency factor (IF^{LS}) and loop sample size. Equal number of trials per comparison ($K_{AB} = K_{AC} = K_{BC} = K = 1, \dots, 7$) is assumed. Results are aggregated over different assumptions for the heterogeneity and methods to estimate the variances of the mean treatment effects. The region within the horizontal dotted lines defines the confidence interval for the 95% nominal level.

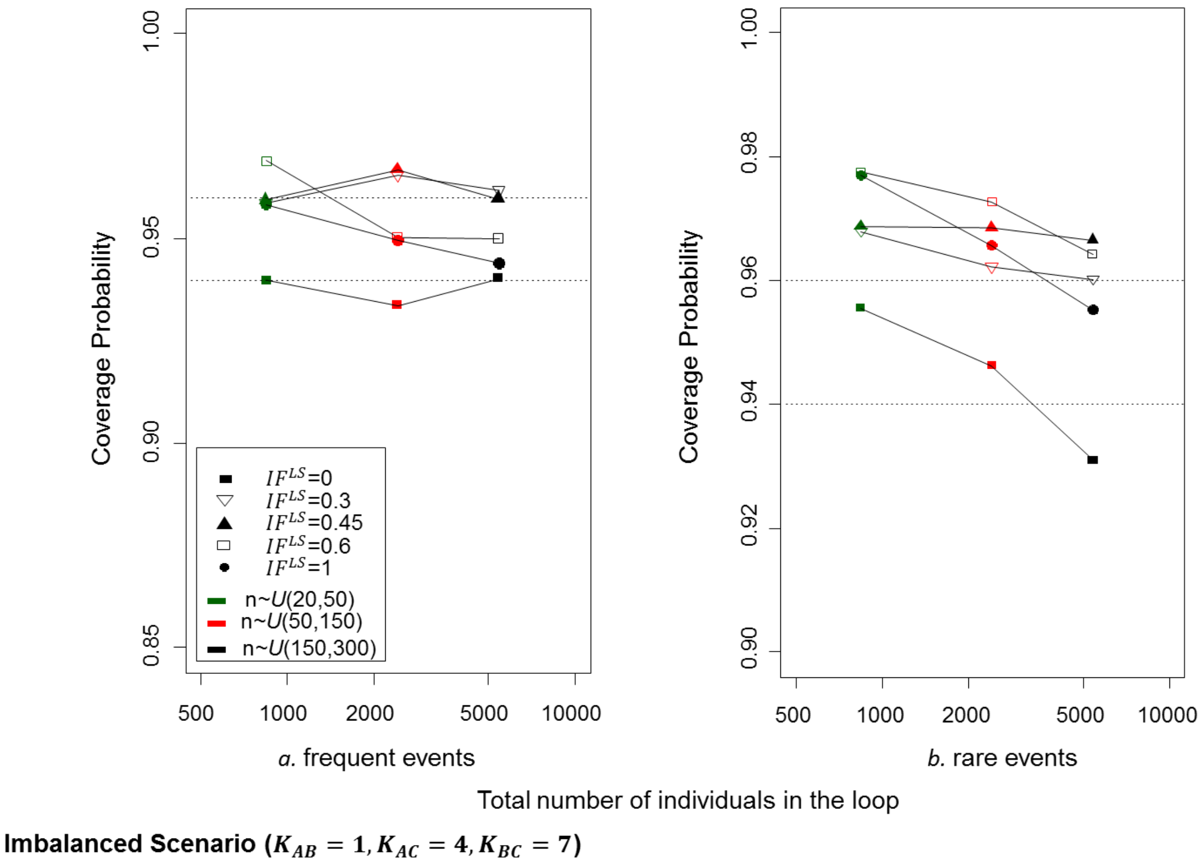


Figure 16. Coverage probabilities of the 95% confidence interval for the inconsistency factor (IF^{LS}). I assume different number of trials (K) per comparison ($K_{AB} = 1, K_{AC} = 4, K_{BC} = 7$). Results are aggregated over different assumptions for the heterogeneity and methods to estimate the variances of the mean treatment effects. The region within the horizontal dotted lines defines the confidence interval for the 95% nominal level. The first summary result in each coverage probability line pertains to the case where there is a single trial per comparison and a fixed-effects model is employed.

In Figure 17 and Figure 18 I present the average relative bias ($|\widehat{IF}^{LS} - IF^{LS}|/IF^{LS}$) when $IF^{LS} > 0$. Relative bias decreases with the total number of individuals included in the network, the total number of trials, and the assumed IF^{LS} value.

Table 14 and Table 15 present the coverage probability for the 95% CI of IF^{LS} using different methods to express uncertainty on the summary effects. The KHDL method reduces slightly the chances of including the true inconsistency in the 95% CI of IF^{LS}, especially when $\tau^2 = 0$, as the summary effects are more precise.

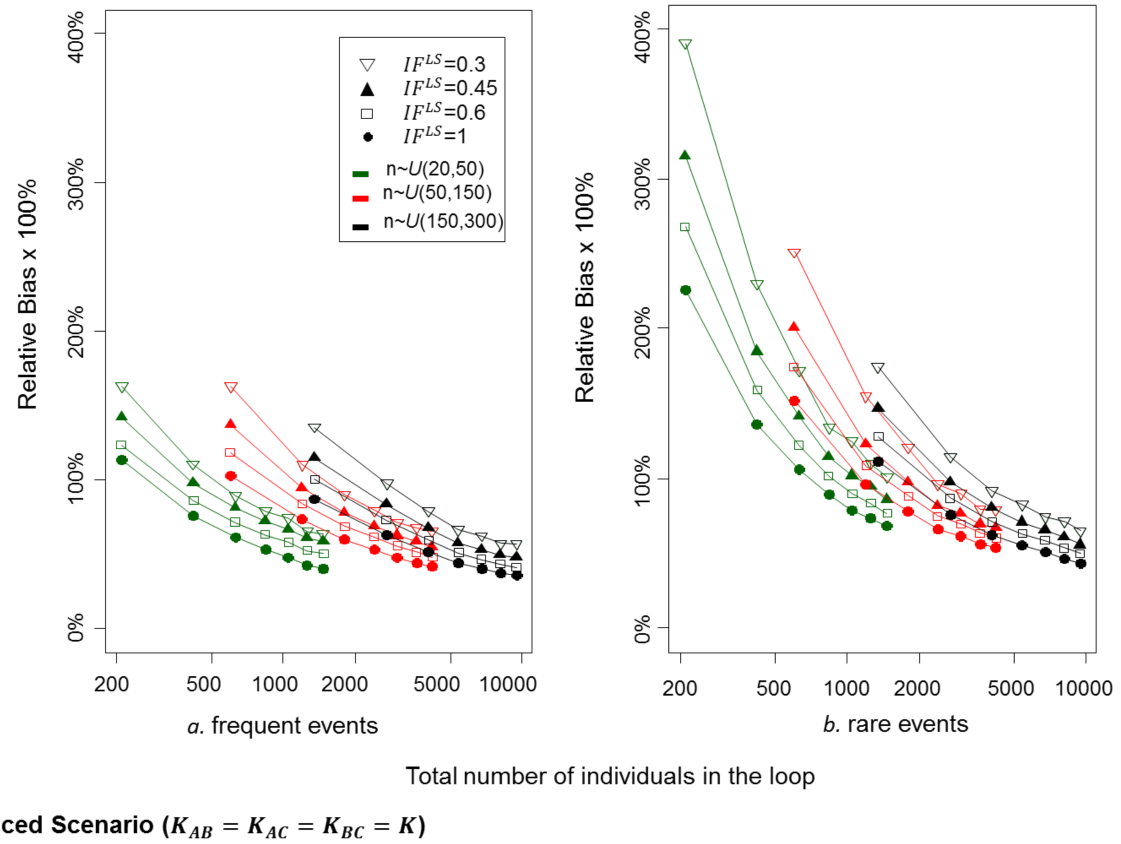


Figure 17. Averaged relative bias assuming various scenarios for the inconsistency factor (IF^{LS}) and the frequency of events. I assume equal number of trials per comparison ($K_{AB} = K_{AC} = K_{BC} = K = 1, \dots, 7$). Results are aggregated over different assumptions for the heterogeneity and methods to estimate the variances for the direct treatment effects.

Table 14. Coverage probability of the 95% confidence interval for the inconsistency factor (IF^{LS}). Results are aggregated over sample size and number of trials (assumed equal across comparisons). WtDL: Wald type method with DerSimonian and Laird (DL) estimator, KHDL: Knapp and Hartung method with DL estimator.

	Heterogeneity					**No Heterogeneity**				
IF^{LS}	*0*	*0.3*	*0.45*	*0.6*	*1*	*0*	*0.3*	*0.45*	*0.6*	*1*
					Frequent Events					
WtDL	0.90	0.94	0.94	0.94	0.93	0.96	0.98	0.97	0.97	0.97
KHDL	0.89	0.93	0.93	0.93	0.91	0.92	0.95	0.94	0.94	0.93
					Rare Events					
WtDL	0.93	0.96	0.96	0.97	0.96	0.97	0.98	0.99	0.98	0.96
KHDL	0.91	0.95	0.95	0.95	0.94	0.92	0.96	0.96	0.95	0.94

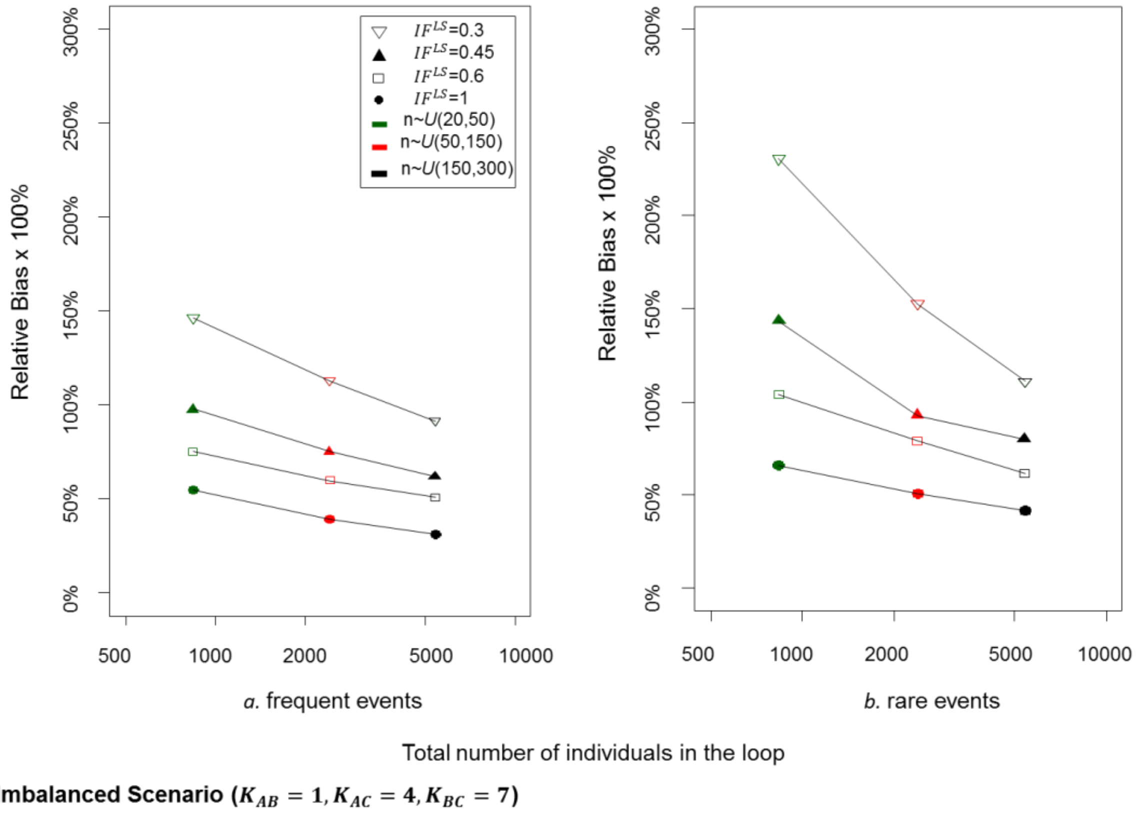


Figure 18. Averaged relative bias assuming various scenarios for the inconsistency factor (IF^{LS}) and the frequency of events. I assume different number of trials (K) per comparison ($K_{AB} = 1, K_{AC} = 4, K_{BC} = 7$). Results are aggregated over different assumptions for the heterogeneity and methods to estimate the variances of the mean treatment effects.

Table 15. Coverage probabilities of the 95% confidence interval for the inconsistency factor (IF^{LS}). WtDL: Wald type method with DerSimonian and Laird (DL) estimator, KHDL: Knapp and Hartung method with DL estimator.

	Heterogeneity					**No Heterogeneity**				
IF^{LS}	*0*	*0.3*	*0.45*	*0.6*	*1*	*0*	*0.3*	*0.45*	*0.6*	*1*
Frequent Events										
WtDL	0.92	0.96	0.96	0.96	0.95	0.97	0.98	0.98	0.97	0.97
KHDL	0.91	0.95	0.96	0.95	0.94	0.93	0.96	0.95	0.95	0.93
Rare Events										
WtDL	0.95	0.96	0.97	0.98	0.98	0.97	0.98	0.98	0.99	0.99
KHDL	0.93	0.95	0.96	0.96	0.96	0.93	0.96	0.96	0.96	0.95

4.3.6 Properties of loop-specific method in a ‘typical’ loop of evidence

The type I error in the ‘typical’ loop is 5% and 7% for subjective and all-cause mortality outcomes using the WtDL method and 11% and 12% using the KHDL estimation. The ‘typical’ loop of evidence with all-cause mortality outcome has considerably low power. The overall power ranged between 14% and 75% for the WtDL method and 21% and 78% for the KHDL approach depending on the magnitude of IF^{LS}. For a subjective outcome that pertains to larger heterogeneity power decreases to 14%-63% for WtDL and in 20% to 65% for KHDL. Coverage was close to the nominal level (see Table 16).

Table 16. Type I error, power and coverage probability for the W^{LS} test in a ‘typical’ loop of evidence. I assume a dichotomous frequent outcome, number of trials (K) per comparison $K_{AB} = 1, K_{AC} = 4, K_{BC} = 3$ and the sample size per arm is drown from $n \sim U(120{,}160)$. IF^{LS}: inconsistency factor, WtDL: Wald type method with DerSimonian and Laird (DL) estimator, KHDL: Knapp and Hartung method with DL estimator.

	Type I error	**Power**				**Coverage Probability**				
IF^{LS}	*0*	*0.3*	*0.45*	*0.6*	*1*	*0*	*0.3*	*0.45*	*0.6*	*1*
All-cause mortality outcome (median(τ^2) = 0.02)										
WtDL	0.05	0.14	0.23	0.38	0.75	0.95	0.97	0.99	0.98	0.95
KHDL	0.11	0.21	0.32	0.46	0.78	0.89	0.94	0.93	0.92	0.90
Subjective outcome (median(τ^2) = 0.11)										
WtDL	0.07	0.14	0.23	0.34	0.63	0.94	0.96	0.96	0.97	0.95
KHDL	0.12	0.20	0.29	0.41	0.65	0.88	0.93	0.93	0.92	0.91

5. Discussion

The increased use of NMA should be accompanied by caution when combining direct and indirect evidence. Evaluation of consistency is an important task in network meta-analysis (12). It has been shown though that it is not rare for reviewers to combine direct and indirect evidence in a network of interventions without evaluating the assumption of consistency (6,14). A recent survey showed that only 9% of the Cochrane review authors are aware of the prerequisite assumptions for a valid NMA (7). Empirical studies have shown that although NMA are increasingly conducted, the key assumptions are not always evaluated and reporting of the methodology applied is inadequate (6,27,28). Thus, there is a need to improve the quality of NMA regarding the assumptions and the methods that are reported. Protocols of NMA should present methods for the evaluation of inconsistency and define strategies to be followed when inconsistency is present. Several methodologies have been outlined in the literature to test inconsistency (15,17,66,75). In this research study I used a large-scale empirical dataset to evaluate the prevalence of inconsistency using five different approaches.

A key finding of our study is that heterogeneity plays an important role in the statistical detection of inconsistency and a lower heterogeneity is associated with higher rates of detected inconsistency. This suggests that heterogeneity might account for some disagreement between various sources of evidence. A general conclusion is that the changes in heterogeneity between consistency and inconsistency models can be used as an alternative to the tests for inconsistency. However, the choice of the heterogeneity assumption and estimation method can hamper the detection of inconsistency. On the contrary, although it has been suggested that a poor choice of the measurement scale, i.e. analysing data on a ‘preferred’ scale rather than on the ‘best’ scale, can increase the probability of finding inconsistency (80), this study showed that the three scales for dichotomous data are not associated with important inconsistency differences. It is advisable that the best approach is to choose the appropriate scale, relying on both type of outcome data and mathematical properties, and then transform the results to an alternative scale to aid interpretation.

Our empirical study suggests that inconsistency is prevalent in between 2% and 10% of the tested loops, depending on the effect measure and heterogeneity estimation method, and about one eighth of the networks is inconsistent. I also found that 9% of the tested designs and 11% of the total comparisons disagreed with the remaining network. I recommend using the DBT model to evaluate a network as a whole and then if inconsistency is detected the methods that evaluate which piece of evidence is responsible for a potential inconsistency in the network (e.g. the LS, SIDE or SODR methods) can be employed. I suggest the DBT model as it is the only method presented in the literature so far that can evaluate the entire network and is insensitive to the parameterization of studies with multiple arms, accounting also for potential design inconsistency. The LA is a special case of the DBT model and the main disadvantage of the approach is that different parameterisations of the multi-arm studies might conclude to different results. Hence, it is advisable that all possible scenarios should be used before making inference. Note that the LA approach does not account for design inconsistency and hence this variability is encompassed in the heterogeneity variance. This study suggested two networks inconsistent with the LA model but consistent with the DBT model. This might be due to the differences in the estimation of the heterogeneity and the fact that I used the maximum test value from the different parameterisations of the multi-arm studies to infer about inconsistency in the LA model. It should be also noted that the DBT model might lose power in detecting inconsistency as it has more degrees of freedom (greater number of parameters) in contrast to the LA approach. However, the W^{DBT} value is always greater than the W^{LA} one as LA is a special case of the DBT approach. A summary of the properties of the five different approaches is presented in Table 17.

If inconsistency is found, exploration of its possible causes is a key component of network meta-analysis and can raise research and editorial standards by shedding light on the strengths and weaknesses of the body of evidence. Results from statistical tests should however be interpreted with caution: the absence of statistical inconsistency does not provide reassurance that the NMA results are valid. The assumption of consistency should always be evaluated conceptually by identifying possible effect modifiers that differ across studies (10). In Table 18 I present a summary of recommendations on what should be applied when inconsistency is found.

Table 17. Overview of the properties of the loop-specific (LS), separating direct and indirect evidence (SIDE), separating one design from the rest (SODR), Lu and Ades (LA), and design-by-treatment interaction approaches.

	LS Method	**SIDE Method**	**SODR Method**	**LA Model**	**DBT Model**
Simple to compute	**Yes**	**No**	**No**	**No**	**No**
Insensitive to parameterization of multi-arm studies	**No**	**No**	**Yes**	**No**	**Yes**
Indirect estimate derived from the entire network	**No**	**Yes**	**Yes**	**Yes**	**Yes**
Does not suffer from multiple testing	**No**	**No**	**No**	**Yes**	**Yes**
Power	**No**	**No**	**Unclear**	**No**	**Unclear**

Table 18. Summary of recommendations when statistically significant inconsistency is found.

Action	**Inconsistency**
Check the data	Use LS, SIDE, or SODR inconsistency methods to identify studies with potential data extraction errors. Evidence loops that include comparisons informed by a single study are particularly suspicious for data errors.
Resign to it	Investigators may decide not to synthesize the network in the presence of excessive inconsistency
Explore it	Split the network into subgroups or use network meta-regression to account for differences across studies and comparisons.
Encompass it	Apply DBT or LA models that relax the consistency assumption.

In the simulation study I evaluated the properties of the LS method for detecting inconsistency comparing direct and indirect estimates in triangular networks. I informed the simulation scenarios by previous large-scale empirical studies, and I used the most commonly encountered meta-analytic tools for statistical inference regarding heterogeneity and the uncertainty of the mean treatment effects. The main advantage of this work is that it sheds light on factors that might affect the detection of inconsistency and have not been examined in the past, such as the use of KH approach for the CI of the direct summary effects. The simulation study shows that the inconsistency test has on average low power to detect inconsistency, in particular for rare outcomes. In the absence of heterogeneity and for a large number and size of trials the overall power for inconsistency might be adequate. A previous simulation study (3) also found that different ways to evaluate inconsistency (e.g. Lu and Ades model, separating indirect and direct method) have low power in particular under the RE models (see also Table 17). Our study suggests that power is improved if the KH method is used, although the type I error increases as well.

For the empirical assessment of consistency I evaluated articles included in PubMed and I restricted the analysis to dichotomous outcomes. Other network meta-analyses, such as those undertaken in technology appraisals for the National Institute for Health and Clinical Excellence (NICE) in the UK, are not included. I expect our findings regarding choice of effect measure and statistical techniques to be generalizable, although it is unclear whether our findings regarding prevalence of inconsistency are relevant to these settings. An empirical study for continuous outcomes will be needed to infer about possible differences in inconsistency between mean differences, standardized mean differences and ratios of means.

Network meta-analyses are typically restricted to evidence based on randomized controlled trials (RCTs). The randomized participant assignment to parallel treatment arms keeps study groups as similar as possible with known and unknown confounding factors balanced. Well-conducted RCTs are the gold standard of clinical information. However, including only RCTs in network meta-analysis a great deal of information from different study designs is ignored. It is widely accepted that non-randomised and quasi-randomised trials provide evidence from broader rages of settings and populations. If these studies have a certain level of quality, there is no technical reason not to include them in the network meta-analysis. Combining both randomized and observational evidence in network meta-analysis, while adjusting for potential biases due to study design, allows one to make an informed decision (85). However, the inclusion of observational trial in network meta-analysis might increase heterogeneity and inconsistency in the data. An additional empirical study including networks with nonrandomised trials to evaluate the extent of inconsistency would be valuable.

A limitation of the simulation study is that I did not account for the possible impact of multi-arm trials on inconsistency and I only reconsider triangular networks. Although I considered only the DL, REML and PM methods to estimate heterogeneity I do not anticipate that inclusion of different estimators would alter the conclusions. In fact, the three estimators considered here provided similar results because they gave comparable estimates for the heterogeneity in the scenarios considered. A further simulation study with more ‘extreme’ scenarios would potentially reveal differences between the choice of the estimator and its impact on the detection of inconsistency. Finally, a thorough investigation of the properties of all available methods for inconsistency and their sensitivity in the characteristics of the network would be needed for completeness.

The use of network meta-analysis is commonly performed on the basis of aggregated data. The benefits of using individual patient data rather than aggregated data have been previously examined and it has been suggested the use of individual patient data in network meta-analysis can reduce statistical heterogeneity across the network and hence can increase the precision of treatment effect estimates (86,87). This is because the parameter estimates of the individual patient data models are estimated using from both within-study and across-study evidence, whereas the results from the aggregated data models are only based on across-study associations. Jansen (87) showed that combining individual patient data with aggregated data minimizes the chances of confounding bias being evident in indirect comparison and network meta-analysis. One of the most important advantages of using individual patient data in network meta-analysis is that it is possible to identify interactions which cannot be detected when using aggregated data and hence evaluate the assumption of consistency. Donegan et al. (86) showed that using aggregated data in network meta-analysis to evaluate the consistency assumption did not reject the null hypothesis, whereas using individual patient data questioned the consistency and reliability of the results. This is because of the imbalance in patient-level effect modifiers across comparisons that could not be identified with aggregated data. Inconsistent evidence can also affect one of the most important properties of network meta-analysis, the ability to rank the treatments according to their efficacy (86). However, further research is needed to establish the benefits of individual patient data in various settings, as well as the properties of the individual patient data network meta-analysis in complex networks of interventions. This research study might be used to inform the development of strategies for the assessment of the assumption of consistency using individual patient data and examine whether this approach is more valuable than aggregated data network meta-analysis.

6. Conclusions

The findings of our research study can be used to inform the development of strategies to detect statistical inconsistency. Results from methods I examined appear to be sensitive to the estimation method and to assumptions made about heterogeneity, as well as the presence of the multi-arm studies. Consequently, investigators should interpret very carefully a statistically non-significant result and always consider the comparability of the studies in terms of potential effect modifiers. A sensitivity analysis using different methods for the heterogeneity and inconsistency is advisable, before concluding about the absence of statistical inconsistency.

Summary in English

Background: Network meta-analysis relies on the agreement between direct and indirect evidence defined as consistency. Empirical evidence about the prevalence of inconsistency is limited to simple loops of evidence about three interventions. No evidence exists so far regarding the extent of inconsistency and the factors that control its statistical detection in complex networks of interventions.

Aims: The objective is to evaluate empirically the prevalence of inconsistency in full networks using various approaches for the assessment of consistency and to explore factors that might control its statistical detection.

Methods: I evaluated inconsistency in 40 published networks with dichotomous data published in PubMed from March 1997 until February 2011 and involved at least four treatments and at least one closed loop. The networks included 303 loops of evidence, 362 trial designs - studies involving different sets of treatments- and 348 comparisons. I employed five approaches: 1) *loop-specific (LS):* I evaluated each loop in the network separately by contrasting direct and indirect estimates 2) *Separating one design from the rest* (*SODR*): I evaluated the agreement between studies of a particular design and the remaining network 3) *separating indirect and direct evidence (SIDE)*: I evaluated the agreement between a particular comparison and the remaining network 4) *Lu and Ades model (LA):* I jointly assessed all possible inconsistencies in the network to obtain an omnibus test 5) *Design-by-Treatment interaction model (DBT):* I evaluated the agreement between estimates from different designs in the network in an omnibus test. In LS and DBT approaches I used different effect measures, and various estimators and assumptions for the heterogeneity. I also carried out a simulation study to estimate the performance of the LS test.

Results: Inconsistency was prevalent in 1) between 2% and 10% of the tested loops depending on the effect measure, assumption and estimation method for heterogeneity, 2) 9% of the tested designs, 3) 11% of the total comparisons, 4) maximum seven (18%) networks depending on the parameterisation of the multi-arm studies, and 5) between 13% and 28% of the networks depending on the effect size and estimator for heterogeneity. Important heterogeneity was associated with a small decrease in statistical inconsistency, but different effect measures had no statistically significant impact on detecting inconsistency. The simulation study showed that the LS-test has generally low power that is positively associated with sample size and frequency of the outcome and negatively associated with the presence of heterogeneity. Type I error converges to the nominal level as the total number of individuals included in the loop increases. Coverage is close to the nominal level in most cases. Different estimation methods for heterogeneity do not greatly impact on test performance, but different methods to derive the variances of the direct estimates impact on the inconsistency inference.

Conclusions: This study suggests that changing effect measure might improve statistical consistency and that a sensitivity analysis in the assumptions and estimators of heterogeneity is needed before concluding the absence of statistical inconsistency, particularly in networks with few studies. Investigators should interpret every test results very carefully and always consider the comparability of the studies in terms of potential effect modifiers.

Περίληψη στα ελληνικά

Εισαγωγή: Στη μετα-ανάλυση πολλαπλών παρεμβάσεων κρίνεται απαραίτητη η αξιολόγηση της συνέπειας μεταξύ άμεσων και έμμεσων στοιχείων. Οι εμπειρικές μελέτες της βιβλιογραφίας που μελετούν την ασυνέπεια περιορίζονται σε δίκτυα τριών παρεμβάσεων, ενώ οι ιδιότητες των μεθόδων για τον έλεγχο της συνέπειας δεν έχουν μελετηθεί διεξοδικά.

Σκοπός: Η εμπειρική αξιολόγηση της ασυνέπειας σε περίπλοκα δίκτυα χρησιμοποιώντας διαφορετικές προσεγγίσεις και η διερεύνηση των παραγόντων που επηρεάζουν τη στατιστική ανίχνευσή της.

Μέθοδοι: Στην παρούσα ερευνητική εργασία η ύπαρξη της ασυνέπειας διερευνάται σε 40 δίκτυα πολλαπλών παρεμβάσεων δημοσιευμένα στην PubMed από το Μάρτιο του 1997 μέχρι και το Φεβρουάριο του 2011 που εμπεριέχουν τουλάχιστον 4 παρεμβάσεις και τουλάχιστον 1 κλειστό βρόχο. Τα δίκτυα αυτά περιλαμβάνουν 303 κλειστούς βρόχους που δημιουργούνται από τα στοιχεία των παρεμβάσεων, 348 συγκρίσεις, και 362 διαφορετικούς τύπους μελετών. Εφάρμοσα 5 διαφορετικές προσεγγίσεις: 1) Αξιολόγηση κάθε βρόχου ξεχωριστά, 2) Αξιολόγηση συμφωνίας μεταξύ στοιχείων μιας σύγκρισης και του υπόλοιπου δικτύου, 3) Αξιολόγηση συμφωνίας μεταξύ στοιχείων ενός τύπου μελετών με το υπόλοιπο δίκτυο, 4) Αξιολόγηση της ασυνέπειας χρησιμοποιώντας ένα γενικευμένο έλεγχο, 5) Αξιολόγηση της συμφωνίας μεταξύ των εκτιμήσεων διαφορετικών τύπων μελετών. Επιπλέον, διεξήχθη έρευνα προσομοίωσης για την αξιολόγηση της πρώτης προσέγγισης (μέθοδος (1)) σε δίκτυα τριών παρεμβάσεων ως προς το σφάλμα τύπου I, την ισχύ και την πιθανότητα επικάλυψης. Χρησιμοποιώντας τις μεθόδους (1) και (5) εξετάσαμε αν η χρήση διαφορετικών μέτρων σχέσης που περιγράφουν διχότομα δεδομένα και οι διαφορετικοί τρόποι εκτίμησης της ετερογένειας σχετίζονται με διαφορές στην εκτίμηση της ασυνέπειας.

Αποτελέσματα: Ανάλογα με τη μέθοδο εκτίμησης της ετερογένειας και το μέτρο σχέσης, ο αριθμός των εξεταζόμενων βρόχων που βρέθηκαν να είναι ασυνεπείς με τη μέθοδο (1) κυμαίνεται από 2% έως 10%. Οι μέθοδοι (2) και (3) έδειξαν πως το 11% των συγκρίσεων και το 9% των διαφορετικών τύπων μελετών δε συμφωνούν με το υπόλοιπο δίκτυο. Το μοντέλο (4) έδειξε ότι ο μέγιστος αριθμός δικτύων που μπορεί να είναι ασυνεπή είναι 18%

ανάλογα με τη μοντελοποίηση των διαφορετικών τύπων μελετών. Από το 13% έως το 28% των δικτύων δεν πληροί την υπόθεση της συνέπειας σύμφωνα με το μοντέλο (5), ανάλογα με τη μέθοδο εκτίμησης της ετερογένειας και το μέτρο σχέσης. Βρόχοι με συγκρίσεις που περιγράφονται από μία και μόνο μελέτη φαίνεται να απορρίπτουν συχνότερα την υπόθεση της συνέπειας. Παρόλο που ο λόγος αναλογιών είναι πιο συνεπές μέτρο σχέσης από το λόγο κινδύνων και τη διαφορά κινδύνων, δεν υπάρχει στατιστικά σημαντική διαφορά μεταξύ των μέτρων αυτών. Όμως, η υπερεκτίμηση της ετερογένειας μπορεί να οδηγήσει σε λανθασμένη αποδοχή της υπόθεσης ότι υπάρχει συνέπεια. Οι προσομοιώσεις έδειξαν ότι η ισχύς του τεστ αυξάνει με το μέγεθος δείγματος και τη συχνότητα των γεγονότων μιας έκβασης, ενώ μειώνεται με την ετερογένεια. Το σφάλμα τύπου I πλησιάζει το επίπεδο σημαντικότητας όσο αυξάνει ο συνολικός αριθμός των συμμετεχόντων, ενώ η πιθανότητα επικάλυψης παραμένει πάντα σε ικανοποιητικά επίπεδα.

Συμπεράσματα: Αποδεικνύεται πως μία εναλλαγή στα μέτρα σχέσης μπορεί να βελτιώσει τη στατιστική συνέπεια. Ίσως η χρήση της ανάλυσης ευαισθησίας στις διαφορετικές υποθέσεις-εκτιμητές της ετερογένειας θα βοηθούσε στο να αποφανθούμε αν υπάρχει ή όχι ασυνέπεια, ειδικότερα στην περίπτωση δικτύων με λίγες μελέτες. Η χρήση διαφορετικών μοντέλων μπορεί να δώσει μία διαφορετική εικόνα για την ύπαρξη ή όχι συνέπειας, καθώς αυτά συσχετίζονται με διαφορετικές ιδιότητες. Οι ερευνητές θα πρέπει να ερμηνεύουν τα αποτελέσματα με προσοχή αφού η ασυνέπεια μπορεί να υποεκτιμάται.

References

1. Caldwell DM, Ades AE, Higgins JPT. Simultaneous comparison of multiple treatments: combining direct and indirect evidence. BMJ. 2005 Oct 15;331(7521):897–900.

2. Bucher HC, Guyatt GH, Griffith LE, Walter SD. The results of direct and indirect treatment comparisons in meta-analysis of randomized controlled trials. J Clin Epidemiol. 1997 Jun;50(6):683–91.

3. Song F, Clark A, Bachmann MO, Maas J. Simulation evaluation of statistical properties of methods for indirect and mixed treatment comparisons. BMC Med Res Methodol. 2012 Sep 12;12(1):138.

4. Song F, Harvey I, Lilford R. Adjusted indirect comparison may be less biased than direct comparison for evaluating new pharmaceutical interventions. J Clin Epidemiol. 2008 May;61(5):455–63.

5. Mills EJ, Ghement I, O'Regan C, Thorlund K. Estimating the Power of Indirect Comparisons: A Simulation Study. PLoS ONE. 2011 Jan 21;6(1):e16237.

6. Nikolakopoulou A, Chaimani A, Veroniki AA, Vasiliadis HS, Schmid CH, Salanti G. Characteristics of Networks of Interventions: A Description of a Database of 186 Published Networks. PLoS ONE. 2014 Jan 22;9(1):e86754.

7. Abdelhamid AS, Loke YK, Parekh-Bhurke S, Chen Y-F, Sutton A, Eastwood A, et al. Use of indirect comparison methods in systematic reviews: a survey of Cochrane review authors. Res Synth Methods. 2012 Jun 1;3(2):71–9.

8. Li T, Puhan MA, Vedula SS, Singh S, Dickersin K, $author.lastName $author firstName. Network meta-analysis-highly attractive but more methodological research is needed. BMC Med. 2011 Jun 27;9(1):79.

9. Cipriani A, Higgins JPT, Geddes JR, Salanti G. Conceptual and Technical Challenges in Network Meta-analysis. Ann Intern Med. 2013 Jul 16;159(2):130–7.

10. Jansen JP, Naci H. Is network meta-analysis as valid as standard pairwise meta-analysis? It all depends on the distribution of effect modifiers. BMC Med. 2013 Jul 4;11(1):159.

11. Thompson SG, Higgins JPT. How should meta-regression analyses be undertaken and interpreted? Stat Med. 2002 Jun 15;21(11):1559–73.

12. Salanti G. Indirect and mixed-treatment comparison, network, or multiple-treatments meta-analysis: many names, many benefits, many concerns for the next generation evidence synthesis tool. Res Synth Methods. 2012 Jun 1;3(2):80–97.

13. Veroniki AA, Vasiliadis HS, Higgins JP, Salanti G. Evaluation of inconsistency in networks of interventions. Int J Epidemiol. 2013 Feb 1;42(1):332–45.

14. Donegan S, Williamson P, D’Alessandro U, Tudur Smith C. Assessing key assumptions of network meta-analysis: a review of methods. Res Synth Methods. 2013 Dec 1;4(4):291–323.

15. Dias S, Welton NJ, Caldwell DM, Ades AE. Checking consistency in mixed treatment comparison meta-analysis. Stat Med. 2010 Mar 30;29(7-8):932–44.

16. Dias S, Welton NJ, Sutton AJ, Caldwell DM, Lu G, Ades AE. Evidence Synthesis for Decision Making 4 Inconsistency in Networks of Evidence Based on Randomized Controlled Trials. Med Decis Making. 2013 Jul 1;33(5):641–56.

17. Higgins JPT, Jackson D, Barrett JK, Lu G, Ades AE, White IR. Consistency and inconsistency in network meta-analysis: concepts and models for multi-arm studies. Res Synth Methods. 2012 Jun 1;3(2):98–110.

18. White IR, Barrett JK, Jackson D, Higgins JPT. Consistency and inconsistency in network meta-analysis: model estimation using multivariate meta-regression. Res Synth Methods. 2012 Jun 1;3(2):111–25.

19. Xiong T, Parekh-Bhurke S, Loke YK, Abdelhamid A, Sutton AJ, Eastwood AJ, et al. Overall similarity and consistency assessment scores are not sufficiently accurate for predicting discrepancy between direct and indirect comparison estimates. J Clin Epidemiol. 2013 Feb;66(2):184–91.

20. Veroniki AA, Mavridis D, Higgins JPT, Salanti G. Statistical evaluation of inconsistency in a loop of evidence: a simulation study informed by empirical evidenc. Appear. 2014;

21. Sidik K, Jonkman JN. A comparison of heterogeneity variance estimators in combining results of studies. Stat Med. 2007 Apr 30;26(9):1964–81.

22. Deeks JJ. Issues in the selection of a summary statistic for meta-analysis of clinical trials with binary outcomes. Stat Med. 2002 Jun 15;21(11):1575–600.

23. Engels EA, Schmid CH, Terrin N, Olkin I, Lau J. Heterogeneity and statistical significance in meta-analysis: an empirical study of 125 meta-analyses. Stat Med. 2000 Jul 15;19(13):1707–28.

24. Sidik K, Jonkman JN. A simple confidence interval for meta-analysis. Stat Med. 2002 Nov 15;21(21):3153–9.

25. Sánchez-Meca J, Marín-Martínez F. Confidence intervals for the overall effect size in random-effects meta-analysis. Psychol Methods. 2008 Mar;13(1):31–48.

26. Song F, Altman DG, Glenny A-M, Deeks JJ. Validity of indirect comparison for estimating efficacy of competing interventions: empirical evidence from published meta-analyses. BMJ. 2003 Mar 1;326(7387):472.

27. Sobieraj DM, Cappelleri JC, Baker WL, Phung OJ, White CM, Coleman CI. Methods used to conduct and report Bayesian mixed treatment comparisons published in the medical literature: a systematic review. BMJ Open. 2013 Jul 1;3(7):e003111.

28. Bafeta A, Trinquart L, Seror R, Ravaud P. Analysis of the systematic reviews process in reports of network meta-analyses: methodological systematic review. BMJ. 2013 Jul 1;347(jul01 1):f3675–f3675.

29. Song F, Xiong T, Parekh-Bhurke S, Loke YK, Sutton AJ, Eastwood AJ, et al. Inconsistency between direct and indirect comparisons of competing interventions: meta-epidemiological study. BMJ. 2011 Aug 16;343(aug16 2):d4909–d4909.

30. Turner RM, Davey J, Clarke MJ, Thompson SG, Higgins JP. Predicting the extent of heterogeneity in meta-analysis, using empirical data from the Cochrane Database of Systematic Reviews. Int J Epidemiol. 2012 Jun;41(3):818–27.

31. DerSimonian R, Laird N. Meta-analysis in clinical trials. Control Clin Trials. 1986 Sep;7(3):177–88.

32. Paule RC, Mandel J. Consensus Values and Weighting Factors [Internet]. National Institute of Standards and Technology; 1982 [cited 2014 Mar 25]. Available from: http://archive.org/details/jresv87n5p377_A1b

33. Raudenbush SW. Analyzing Effect Sizes: Random Effects Models. In: Cooper H, Hedges LV, Valentine JC, editors. The Handbook of Research Synthesis and Meta-Analysis. Russell Sage Foundation; 2009.

34. Hardy RJ, Thompson SG. A likelihood approach to meta-analysis with random effects. Stat Med. 1996 Mar 30;15(6):619–29.

35. Sidik K, Jonkman JN. Simple heterogeneity variance estimation for meta-analysis. J R Stat Soc Ser C Appl Stat. 2005 Apr 1;54(2):367–84.

36. Smith TC, Spiegelhalter DJ, Thomas A. Bayesian approaches to random-effects meta-analysis: a comparative study. Stat Med. 1995 Dec 30;14(24):2685–99.

37. Veroniki AA, Bender R, Bowden J, Higgins J, Jackson D, Kuss O, et al. Tutorial in Biostatistics: Methods to estimate heterogeneity variance, its uncertainty and to draw inference on the meta-analysis summary effect. Appear. 2014;

38. Knapp G, Hartung J. Improved tests for a random effects meta-regression with a single covariate. Stat Med. 2003 Sep 15;22(17):2693–710.

39. Brockwell SE, Gordon IR. A simple method for inference on an overall effect in meta-analysis. Stat Med. 2007 Nov 10;26(25):4531–43.

40. Henmi M, Copas JB. Confidence intervals for random effects meta-analysis and robustness to publication bias. Stat Med. 2010 Dec 20;29(29):2969–83.

41. Thompson SG, Sharp SJ. Explaining heterogeneity in meta-analysis: a comparison of methods. Stat Med. 1999 Oct 30;18(20):2693–708.

42. Villar J, Mackey ME, Carroli G, Donner A. Meta-analyses in systematic reviews of randomized controlled trials in perinatal medicine: comparison of fixed and random effects models. Stat Med. 2001 Dec 15;20(23):3635–47.

43. Cramér H. Mathematical Methods of Statistics. Princeton University Press; 1999. 596 p.

44. Lehmann EL, Casella G. Theory of Point Estimation. Springer; 1998. 611 p.

45. Rao CR. Linear Statistical Inference and its Applications. John Wiley & Sons; 2009. 657 p.

46. Viechtbauer W. Bias and Efficiency of Meta-Analytic Variance Estimators in the Random-Effects Model. J Educ Behav Stat. 2005 Sep 21;30(3):261–93.

47. Rukhin AL. Estimating heterogeneity variance in meta-analysis. J R Stat Soc Ser B Stat Methodol. 2013 Jun 1;75(3):451–69.

48. Bowden J, Tierney JF, Copas AJ, Burdett S. Quantifying, displaying and accounting for heterogeneity in the meta-analysis of RCTs using standard and generalised Q statistics. BMC Med Res Methodol. 2011 Apr 7;11(1):41.

49. Makambi KH. The effect of the heterogeneity variance estimator on some tests of treatment efficacy. J Biopharm Stat. 2004 May;14(2):439–49.

50. Jackson D, Bowden J, Baker R. How does the DerSimonian and Laird procedure for random effects meta-analysis compare with its more efficient but harder to compute counterparts? J Stat Plan Inference. 2010 Apr;140(4):961–70.

51. Morris CN. Parametric Empirical Bayes Inference: Theory and Applications. J Am Stat Assoc. 1983 Mar;78(381):47.

52. DerSimonian R, Kacker R. Random-effects model for meta-analysis of clinical trials: an update. Contemp Clin Trials. 2007 Feb;28(2):105–14.

53. Rukhin AL, Biggerstaff BJ, Vangel MG. Restricted maximum likelihood estimation of a common mean and the Mandel–Paule algorithm. J Stat Plan Inference. 2000 Feb 1;83(2):319–30.

54. Panityakul T, Bumrungsup C, Knapp G. On Estimating Residual Heterogeneity in Random-Effects Meta-Regression: A Comparative Study. J Stat Theory Appl. 2013;12(3):253.

55. Berkey CS, Hoaglin DC, Mosteller F, Colditz GA. A random-effects regression model for meta-analysis. Stat Med. 1995 Feb 28;14(4):395–411.

56. Thorlund K, Wetterslev J, Awad T, Thabane L, Gluud C. Comparison of statistical inferences from the DerSimonian–Laird and alternative random-effects model meta-analyses – an empirical assessment of 920 Cochrane primary outcome meta-analyses. Res Synth Methods. 2011 Dec 1;2(4):238–53.

57. Lu G, Ades A. Modeling between-trial variance structure in mixed treatment comparisons. Biostat Oxf Engl. 2009 Oct;10(4):792–805.

58. Normand SL. Meta-analysis: formulating, evaluating, combining, and reporting. Stat Med. 1999 Feb 15;18(3):321–59.

59. Hartung J. An Alternative Method for Meta-Analysis. Biom J. 1999 Dec 1;41(8):901–16.

60. Sidik K, Jonkman JN. On Constructing Confidence Intervals for a Standardized Mean Difference in Meta-analysis. Commun Stat - Simul Comput. 2003;32(4):1191–203.

61. Pullenayegum EM. An informed reference prior for between-study heterogeneity in meta-analyses of binary outcomes. Stat Med. 2011 Nov 20;30(26):3082–94.

62. Kontopantelis E, Springate DA, Reeves D. A Re-Analysis of the Cochrane Library Data: The Dangers of Unobserved Heterogeneity in Meta-Analyses. PLoS ONE. 2013 Jul 26;8(7):e69930.

63. Novianti PW, Roes KCB, van der Tweel I. Estimation of between-trial variance in sequential meta-analyses: A simulation study. Contemp Clin Trials. 2014 Jan;37(1):129–38.

64. Friedrich JO, Adhikari NKJ, Beyene J. Ratio of means for analyzing continuous outcomes in meta-analysis performed as well as mean difference methods. J Clin Epidemiol. 2011 May;64(5):556–64.

65. Franchini AJ, Dias S, Ades AE, Jansen JP, Welton NJ. Accounting for correlation in network meta-analysis with multi-arm trials. Res Synth Methods. 2012 Jun 1;3(2):142–60.

66. Lu G, Ades AE. Assessing Evidence Inconsistency in Mixed Treatment Comparisons. J Am Stat Assoc. 2006;101(474):447–59.

67. Krahn U, Binder H, König J. A graphical tool for locating inconsistency in network meta-analyses. BMC Med Res Methodol. 2013 Mar 9;13(1):35.

68. Jackson D, Barrett J, Rice S, White I, Higgins J. A design-by-treatment interaction model for network metaanalysis with random inconsistency effects. Appear. 2014;

69. White IR. Multivariate random-effects meta-regression: Updates to mvmeta. Stata J. 2011;11(2):255–70.

70. Introduction to Meta Analysis. New York, NY: [John Wiley and Sons Ltd]; 2009.

71. Lu G, Welton NJ, Higgins JPT, White IR, Ades AE. Linear inference for mixed treatment comparison meta-analysis: A two-stage approach. Res Synth Methods. 2011 Mar 1;2(1):43–60.

72. Rucker G, Guido S. A graph-theoretical approach to multi-armed studies in frequentist network meta-analysis. Appear. 2014;

73. Salanti G, Higgins JPT, Ades AE, Ioannidis JPA. Evaluation of networks of randomized trials. Stat Methods Med Res. 2008 Jun;17(3):279–301.

74. Salanti G, Marinho V, Higgins JPT. A case study of multiple-treatments meta-analysis demonstrates that covariates should be considered. J Clin Epidemiol. 2009 Aug;62(8):857–64.

75. Caldwell DM, Welton NJ, Ades AE. Mixed treatment comparison analysis provides internally coherent treatment effect estimates based on overviews of reviews and can reveal inconsistency. J Clin Epidemiol. 2010 Aug;63(8):875–82.

76. Veroniki AA, Higgins J, White I, Salanti G. Evaluation of novel methods to detect inconsistency in a network of interventions. Appear. 2014;

77. Spiegelhalter DJ, Best NG, Carlin BP, Van Der Linde A. Bayesian measures of model complexity and fit. J R Stat Soc Ser B Stat Methodol. 2002 Oct 1;64(4):583–639.

78. Lumley T. Network meta-analysis for indirect treatment comparisons. Stat Med. 2002 Aug 30;21(16):2313–24.

79. Song F, Xiong T, Parekh-Bhurke S, Sutton A, Eastwood A, Holland R, et al. Inconsistency between direct and indirect estimates remains more prevalent than previous observed. Httpwwwbmjcomrapid-Response20111103inconsistency--Direct--Indirect-Estim-Remains-More-Prevalent. 2011;

80. Caldwell DM, Welton NJ, Dias S, Ades A. Selecting the best scale for measuring treatment effect in a network meta-analysis: a case study in childhood nocturnal enuresis. Res Synth Methods. 2012 Jun 1;3(2):126–41.

81. Imamura M, Abrams P, Bain C, Buckley B, Cardozo L, Cody J, et al. Systematic review and economic modelling of the effectiveness and cost-effectiveness of non-surgical treatments for women with stress urinary incontinence. Health Technol Assess Winch Engl. 2010 Aug;14(40):1–188, iii–iv.

82. Elliott WJ, Meyer PM. Incident diabetes in clinical trials of antihypertensive drugs: a network meta-analysis. Lancet. 2007 Jan 20;369(9557):201–7.

83. Salliot C, Finckh A, Katchamart W, Lu Y, Sun Y, Bombardier C, et al. Indirect comparisons of the efficacy of biological antirheumatic agents in rheumatoid arthritis in patients with an inadequate response to conventional disease-modifying antirheumatic drugs or to an anti-tumour necrosis factor agent: a meta-analysis. Ann Rheum Dis. 2011 Feb;70(2):266–71.

84. Macfadyen CA, Acuin JM, Gamble CL. Topical antibiotics without steroids for chronically discharging ears with underlying eardrum perforations. In: The Cochrane Collaboration, editor. Cochrane Database of Systematic Reviews [Internet]. Chichester, UK: John Wiley & Sons, Ltd; 2005 [cited 2014 Mar 26]. Available from: http://summaries.cochrane.org/CD004618/a-cochrane-systematic-review-assessing-topical-antibiotics-without-steroids-for-treating-chronically-discharging-ears-with-underlying-eardrum-perforations-in-participants-of-any-age

85. Schmitz S, Adams R, Walsh C. Incorporating data from various trial designs into a mixed treatment comparison model. Stat Med. 2013 Jul 30;32(17):2935–49.

86. Jansen JP. Network meta-analysis of individual and aggregate level data. Res Synth Methods. 2012 Jun 1;3(2):177–90.

87. Donegan S, Williamson P, D'Alessandro U, Tudur Smith C. Assessing the consistency assumption by exploring treatment by covariate interactions in mixed treatment comparison meta-analysis: individual patient-level covariates versus aggregate trial-level covariates. Stat Med. 2012 Dec 20;31(29):3840–57.

Appendix

Appendix Table 1. Characteristics of included networks regarding the assessment of inconsistency in the original reviews

id	*Network*	*Assumption of consistency was evaluated*	*Method to detect inconsistency*	*Inconsistency reported as present*
1	Ades[1]	Unclear	Model comparison in fit and parsimony - unclear whether this was specific to the assumption of consistency	Unclear
2	Ara[2]	No	Not reported	Not reported
3	Baker[3]	Inappropriate method*****	Comparison of network estimates to direct estimates	No
4	Ballesteros[4]	Yes	Loop-based approach	No
5	Bangalore[5]	Inappropriate method*****	Comparison of network estimates to direct estimates	No
6	Bansback[6]	No	Not reported	Not reported
7	Bottomley[7]	No	Not reported	Not reported
8	Brown[8]	Yes	Loop-based approach	No
9	Bucher[9]	Yes	Loop-based approach	No
10	Cipriani[10]	Yes	Loop-based approach	Yes
11	Dias[11]	Yes	Node-splitting & back-calculation	Yes
12	Eisenberg[12]	No	Not reported	Not reported
13	Elliott[13]	Yes	Lumley's method	Yes
14	Govan[14]	No	Not reported	Not reported
15	Hofmeyr[15]	Inappropriate method*****	Informal comparison of the results to previously conducted meta-analyses	No
16	Imamura[16]	No	Not reported	Not reported
17	Lam[17]	Inappropriate method*****	Comparison of network estimates to direct estimates	No
18	Lapitan[18]	Inappropriate method*****	Informal comparison of the results to previously conducted meta-analyses	No
19	Lu (1)[19]	Yes	Lu and Ades model	No
20	Lu (2)[19]	Yes	Model comparison in fit and parsimony	No
21	Macfayden [22]	No	Not reported	Not reported
22	Middleton[23]	No	Not reported	Not reported

23	Mills[24]	Yes	Loop-based approach	No
24	Nixon[25]	No	Not reported	Not reported
25	Picard[26]	No	Not reported	Not reported
26	Playford[27]	Yes	Loop-based approach	No
27	Psaty[28]	Yes	Lumley’s method	Yes
28	Puhan[29]	Inappropriate method*****	Informal comparison of the results to previously conducted meta-analyses	No
29	Roskell (1)[31]	Inappropriate method*****	Comparison of network estimates to direct estimates	No
30	Roskell (2)[30]	Inappropriate method*****	Comparison of network estimates to direct estimates	Yes
31	Salliot[32]	No	Not reported	Not reported
32	Sciarretta[33]	Yes	Lu and Ades model	Yes
33	Soares-Weiser[34]	No	Not reported	Not reported
34	Thijs[35]	Yes	Lumley’s method	No
35	Trikalinos[36]	Yes	Lumley’s method	Yes
36	Virgili[37]	Yes	Loop-based approach	No
37	Wang[38]	Inappropriate method*****	Informal comparison of the results to previously conducted meta-analyses	No
38	Welton[39]	Unclear	Model comparison in fit and parsimony - unclear whether this was specific to the assumption of consistency	Unclear
39	Woo[40]	No	Not reported	Not reported
40	Yu[41]	No	Not reported	Not reported

* Some systematic reviews compared estimates from meta-analysis to the estimates obtained from network meta-analysis. I consider this to be an inappropriate method to evaluate consistency.
**Inconsistency has been previously assessed[21]
***Inconsistency has been previously assessed[20]

Appendix Table 2. Characteristics of networks with at least one closed loop included in the database. I define K the total number of studies and S the total number of treatments included in each network. (NMA = network meta-analysis; GLM = generalized linear model, HR = hazard ratio, RR = risk ratio, OR = odds ratio, RD = risk difference).

id	*Network*	*loops*	*K*	*S*	*Disease/ Condition*	*Outcome*	*Type of Treatments*	*2arm trials*	*3arm trials*	*4arm trials*	*Indirect Method*	*Effect Measure used by reviewers*
1	Ades[1]	3	15	9	Schizophrenia	Relapse	Antipsychotic treatments	15	0	0	Bayesian NMA	HR
2	Ara[2]	5	12	5	Hypercholesterolaemi a	Effectiveness in reducing LDL-c.	Statins	10	0	1	Bayesian NMA	RR
3	Baker[3]	12	39	8	Chronic obstructive pulmonary disease (COPD>=1)	Exacebration episodes	Pharmacological treatments	29	3	6	Bayesian NMA	OR
4	Ballesteros[4]	2	9	4	Dysthymia	Efficacy (50% reduction in depressive symptoms since baseline, or similar criteria)	Antidepressants	6	3	0	GLM	OR , RR , RD
5	Bangalore[5]	18	49	8	High blood pressure	Cancer and cancer-related deaths	Antihypertensive drugs	45	4	0	Bayesian NMA	OR
6	Bansback[6]	2	22	8	Moderate to severe plaque psoriasis	Psoriasis area and severity index (PASI)	Treatments for psoriasis	21	1	0	Bayesian NMA	RR
7	Bottomley[7]	4	10	7	Moderately severe scalp psoriasis	Investigator's global assessment	Topical therapies	8	1	1	Meta-regression	RR
8	Brown[8]	6	40	6	Non-steroidal anti-inflammatory drug-induced gastrointestinal toxicity	Serious GI complications	Pharmacological interventions	36	2	0	Bucher	RR

9	Bucher[9]	2	18	4	Pseudocystis carinii in HIV infected patients	Number of pseudocystis carinii pneumonia (prophylaxis against pneumocystis carinii in HIV infected patients)	Pharmacological prophylaxis for pseudocystis carinii	18	0	0	Bucher	OR
10	Cipriani[10]	70	111	12	Unipolar major depression in adults	The proportion of patients who responded to or dropped out of the allocated treatment	Antidepressants	109	2	0	Bayesian NMA	OR
11	Dias[11]	11	50	9	Acute myocardial infraction	Death	Thrombolytic drugs and angioplasty	48	2	0	NMA for trial-level and summary-level data	OR
12	Eisenberg[12]	1	61	5	Smoking	Smoking abstinence	Pharmacotherapies for smoking cessation	59	3	0	Bayesian NMA	OR
13	Elliott[13]	16	22	6	Hypertension, high-risk patients	Proportion of patients who developed diabetes.	Antihypertensive drugs	18	4	0	GLM	OR
14	Govan[14]	2	31	5	Stroke	Death	Types of stroke unit care	25	3	0	Bayesian NMA	OR
15	Hofmeyr[15]	1	24	4	Postpartum haemorrhage	Maternal death	Misoprostol or other uterotonic medication	18	1	0	Bucher	RR
16	Imamura[16]	26	38	13	Stress urinary incontinence	Cure	Non surgical treatments	31	5	2	Bayesian NMA	OR

17	Lam[17]	3	12	5	Left ventricular dysfunction	Mortality	Combined resynchronisation and implantable defibrillator therapy	9	2	0	Bayesian NMA	OR
18	Lapitan[18]	5	22	9	Urinary incontinence in women	Number not cured within first year	Treatments for urinary incontinence in women	19	2	1	Not reported	RR
19	Lu (1)[19]	4	24	4	Smoking	Cessation	Smoking cessation interventions	22	2	0	Bayesian NMA	OR
20	Lu (2)[19]	4	40	6	Gastroesophageal reflux disease	Effectiveness	Gastroesophageal reflux disease therapies	38	2	0	Bayesian NMA	OR
21	Macfayden[22]	2	13	4	Chronically discharging ears with underlying eardrum perforations	Resolution of discharge	Topical antibiotics without steroids	10	3	0	Not reported	RR
22	Middleton[23]	1	20	4	Heavy menstrual bleeding	Dissatisfaction at 12 months	Second line treatment	20	0	0	Logistic regression	OR
23	Mills[24]	2	89	4	Smoking	Abstinence from smoking at at least 4 weeks post-target quit date	Pharmacotherapies	86	3	0	Bucher	OR
24	Nixon[25]	2	11	9	Rheumatoid arthritis	American college of rheumatology (ACR) response criteria at 6 months or beyond	Cytokine antagonists	10	1	0	NMA & meta-regression	OR
25	Picard[26]	33	43	8	Pain on injection with propofol	No pain	Drugs, physical measurements, and combinations	28	12	3	Not reported	RR
26	Playford[27]	1	10	5	Fungal infections in	Mortality	Antifungal agents	10	0	0	Not reported	RR

					solid organ transplant recipients							
27	Psaty[28]	10	28	7	Coronary heart disease (CHD)	Fatal and nonfatal events	Antihypertensive therapy	24	4	0	GLM	RR
28	Puhan[29]	7	34	5	Stable chronic obstructive pulmonary disease	Exacerbation	Inhaled drug regimes	27	1	6	Logistic regression	OR
29	Roskell (1)[31]	6	17	11	Atrial fibrillation	Stroke prevention	Anticoagulants	15	1	1	Mixed log-binomial model	RR
30	Roskell (2)[30]	3	12	10	Fibromyalgia	30% improvement in pain response	Pharmacological interventions	6	6	0	Mixed log-binomial model	RR
31	Salliot[32]	1	15	5	Rheumatoid arthritis (with inadequate response to conventional disease-modifying AR drugs or to anti-tumour necrosis factor agent)	ACR50 response rate	Biological antirheumatic agents	14	1	0	Bucher	OR
32	Sciarretta[33]	13	26	8	Heart fealure	Prevention of heart failure	Antihypertensive treatments	24	2	0	Bayesian NMA	OR
33	Soares-Weiser[34]	4	14	8	Bipolar disorder	All relapses	Pharmacological interventions for the prevention of relapse in people with bipolar disorder	10	4	0	Logistic regression & Bayesian NMA	OR
34	Thijs[35]	3	24	5	Transient ischaemic attack or stroke	Prevention of serious vascular events	Antiplatelets	20	3	0	GLM	OR
35	Trikalinos[36]	1	63	4	Non-acute coronary artery disease	Death	Percutaneous coronary interventions	62	0	0	GLM	RR
36	Virgili[37]	1	10	5	Neovascular age-related macular degeneration	Visual acuity loss	Pharmacological Treatments	10	0	0	Logistic regression & Bayesian NMA	OR

37	Wang[38]	4	43	9	Catheter-related infections	Catheter colonisation	Different central venous catheters	41	2	0	Bayesian NMA	OR
38	Welton[39]	4	36	17	Coronary heart disease	All-cause mortality	Psychological Interventions	31	4	0	Logistic regression & Bayesian NMA	OR
39	Woo[40]	3	19	10	Chronic hepatidis B	HBV DNA levels	Nucleostides	16	3	0	Bayesian NMA	OR
40	Yu[41]	5	14	6	Cardiac surgery	Cardiac ischemic complications and mortality	Inhaled anesthetics	11	2	1	Not reported	OR

Appendix Table 3. Inconsistency estimates (IF^{LS}) along with their standard error ($SE(IF^{LS})$) and W^{LS} values estimated in the loop specific approach for the four effect sizes. Within each loop, inconsistency is estimated assuming the network heterogeneity ($\hat{\tau}^2_{ntw}$). The amount of heterogeneity is estimated with the restricted maximum likelihood estimator in the design-by-treatment interaction model. RD: risk difference measure, $LRRH$: log risk ratio for harmful outcomes, $LRRB$: log risk ratio for beneficial outcomes and LOR: log odds ratio.

		LOR				***LRRH***				***LRRB***				***RD***			
Network	*total loops*	*Inconsistent loops*	*heterogeneity*	IF^{LS} $(SE(IF^{LS}))$	W^{LS} *(P value)*	*Inconsistent loops*	*heterogeneity*	IF^{LS} $(SE(IF^{LS}))$	W^{LS} *(P value)*	*Inconsistent loops*	*heterogeneity*	IF^{LS} $(SE(IF^{LS}))$	W^{LS} *(P value)*	*Inconsistent loops*	*heterogeneity*	$IF^{LS}($ $SE(IF^{LS}))$	W^{LS} *(P value)*
Ades[1]	3	0	0.30			0	0.22			1	0.01	0.38 (0.16)	-2.42 (0.020)	1	0.01	0.29 (0.14)	2.03 (0.040)
Ara[2]	5	0	0.00				0.00				0.00				0.00		
Baker[3]	12	0	0.00				0.00				0.00				0.00		
Ballesteros[4]	2	0	0.02				0.00				0.04				0.00		
Bangalore[5]	18	0	0.00			0	0.00			2	0.00	0.02 (0.01) 0.02 (0.01)	-2.74 (0.010) 2.27 (0.020)	2	0.00	0.02 (0.01) 0.02 (0.01)	2.67 (0.010) -2.20 (0.030)
Bansback[6]	2	0	0.00				0.35				0.05				0.00		
Bottomley[7]	4	0	0.12				0.02				0.02				0.01		
Brown[8]	6	0	0.02				0.02				0.00				0.00		
Bucher[9]	2	0	0.00				0.00				0.00				0.00		
Cipriani[10]	70	3	0.00	0.69 (0.28) 1.15 (0.51) 0.61 (0.24)	-2.49 (0.013) -2.27 (0.023) -2.51 (0.012)	2	0.00	0.57 (0.28) 0.31 (0.15) 0.23 (0.11)	2.00 (0.045) 2.00 (0.045) -2.19 (0.028)	3	0.00	0.38 (0.15) 0.58 (0.27)	-2.63 (0.009) -2.19 (0.029)	3	0.00	0.18 (0.08) 0.29 (0.13) 0.14 (0.06)	-2.28 (0.022) -2.17 (0.030) -2.18 (0.029)
Dias[11]	11	1	0.00	1.2 (0.41)	-2.92 (0.003)	1	0.00	1.15 (0.40)	-2.90 (0.004)	1	0.00	0.05 (0.02)	2.86 (0.004)	1	0.00	0.05 (0.02)	-2.91 (0.004)
Eisenberg[12]	1	0	0.03			0	0.00			0	0.02			0	0.00		

Elliott[13]	16	2	0.01	0.83 (0.3)	2.78 (0.005)	2	0.01	0.80 (0.28)	2.82 (0.005)	0	0.00			0	0.00	
				0.71 (0.33)	2.18 (0.030)			0.70 (0.31)	2.27 (0.024)							
Govan[14]	2	1	0.00	0.90 (0.39)	2.29 (0.022)	1	0.00	0.82 (0.33)	2.49 (0.013)	0	0.00			0	0.00	
Hofmeyr[15]	1	0	0.00			0	0.00			0	0.00			0	0.00	
Imamura[16]	26	5	0.07	4.74 (1.19)	-3.99 (<0.001)	6	0.01	3.35 (0.97)	3.45 (0.001)	5	0.05	3.34 (1.00)	3.33 (0.001)	2	0.02	0.79 (0.20)	3.88 (<0.001)
				2.56 (1.13)	-2.26 (0.024)			1.72 (0.78)	2.22 (0.026)			1.74 (0.83)	2.09 (0.037)			0.74 (0.19)	3.86 (<0.001)
				4.52 (0.99)	-4.56 (<0.001)			1.68 (0.46)	3.70 (<0.001)			1.81 (0.52)	3.51 (<0.001)				
				3.06 (1.24)	2.48 (0.013)			1.36 (0.59)	2.33 (0.020)			1.28 (0.64)	2.01 (0.045)				
				1.9 (0.85)	2.24 (0.025)			2.37 (1.00)	-2.37 (0.018)			2.37 (1.04)	-2.28 (0.023)				
								1.14 (0.56)	-2.03 (0.042)								
Lam[17]	3	0	0.00			0	0.00			0	0.00			0	0.00		
Lapitan[18]	6	0	0.00			0	0.00			0	0.00			1	0.00	0.30 (0.14)	2.16 (0.030)
Lu (1)[19]	4	0	0.43			0	0.02			0	0.26			0	0.01		
Lu (2)[19]	4	0	0.25			0	0.03			0	0.07			0	0.01		
Macfayden[22]	2	0	0.53			0	0.05			0	0.15			0	0.04		
Middleton[23]	1	0	0.00			0	0.00			0	0.00			0	0.00		
Mills[24]	2	0	0.18			0	0.02			0	0.09			0	0.01		
Nixon[25]	2	0	0.65			0	0.06			0	0.30			0	0.03		
Picard[26]	33	2	0.67	1.9 (0.94)	2.01 (0.045)	4	0.15	0.91 (0.41)	-2.20 (0.028)	1	0.13	1.08 (0.51)	-2.11 (0.035)	2	0.03	0.43 (0.19)	2.22 (0.027)
				2.5 (1.17)	-2.13 (0.033)			1.13 (0.57)	-1.99 (0.047)							0.50 (0.25)	-2.02 (0.044)

								1.20 (0.61) 1.38 (0.65)	1.97 (0.049) -2.12 (0.034)								
Playford[27]	1	0	0.00			0	0.00			0	0.00			0	0.00		
Psaty[28]	10	1	0.01	0.77 (0.31)	-2.47 (0.013)	1	0.01	0.71 (0.28)	-2.50 (0.012)	2	0.00	0.03 (0.01) 0.03 (0.01)	2.04 (0.041) 2.14 (0.032)	2	0.00	0.02 (0.01) 0.03 (0.01)	-1.98 (0.047) -2.09 (0.037)
Puhan[29]	7	0	0.00			0	0.00			1	0.00	0.15 (0.07)	2.23 (0.026)	1	0.00	0.08 (0.04)	-2.17 (0.030)
Roskell (1)[31]	6	0	0.07			0	0.07			0	0.00			0	0.00		
Roskell (2)[30]	3	0	0.00			0	0.00			0	0.00			0	0.00		
Salliot[32]	1	1	0.12	0.87 (0.4)	2.18 (0.029)	0	0.00			1	0.09	0.70 (0.32)	2.17 (0.03)	0	0.00		
Sciarretta[33]	13	0	0.01			1	0.01	0.61 (0.30)	2.05 (0.040)	0	0.00			0	0.00		
Soares-Weiser[34]	4	0	0.35			0	0.03			0	0.13			0	0.02		
Thijs[35]	3	0	0.00			0	0.00			0	0.00			0	0.00		
Trikalinos[36]	1	0	0.00			0	0.00			0	0.00			0	0.00		
Virgili[37]	1	0	0.00			0	0.01			0	0.00			0	0.00		
Wang[38]	4	0	0.18			0	0.10			1	0.00	1.00 (0.44)	2.26 (0.02)	1	0.01	0.45 (0.20)	-2.23 (0.030)
Welton[39]	4	0	0.19			0	0.16			0	0.00			0	0.00		
Woo[40]	3	0	0.00			0	0.07			0	0.08			0	0.01		
Yu[41]	5	0	0.00			0	0.00			0	0.00			0	0.00		

Appendix Table 4 Inconsistency estimates (IF^{LS}) along with their standard error ($SE(IF^{LS})$) and W^{LS} values estimated in the loop specific approach for the four effect sizes. Within each loop, inconsistency is estimated assuming a common heterogeneity for each comparison ($\hat{\tau}^2_{loop}$). The amount of heterogeneity is estimated with the DerSimonian and Laird estimator in the random-effects model. RD: risk difference measure, $LRRH$: log risk ratio for harmful outcomes, $LRRB$: log risk ratio for beneficial outcomes and LOR: log odds ratio.

		LOR				$LRRH$				$LRRB$				RD			
Network	*total loops*	*Inconsistent loops*	*heterogeneity*	IF^{LS} $(SE(IF^{LS}))$	W^{LS} *(P value)*	*Inconsistent loops*	*heterogeneity*	IF^{LS} $(SE(IF^{LS}))$	W^{LS} *(P value)*	*Inconsistent loops*	*heterogeneity*	IF^{LS} $(SE(IF^{LS}))$	W^{LS} *(P value)*	*Inconsistent loops*	*heterogeneity*	$IF^{LS}($ $SE(IF^{LS}))$	W^{LS} *(P value)*
Ades[1]	3	2	0.00	1.59 (0.41)	3.91 (0.000)	1	0.00	1.21 (0.32)	3.76 (0.000)	1	0.00	0.37 (0.09)	-4.14 (0.000)	1	0.00	0.28 (0.07)	4.26 (0.000)
			0.00	2.07 (1.00)	2.06 (0.039)												
Ara[2]	5	0				0				0				0			
Baker[3]	12	0				2	0.001	0.12 (0.06)	1.97 (0.049)	0				0			
							0.00	0.12 (0.06)	2.25 (0.024)								
Ballesteros[4]	2	0				0				0				0			
Bangalore[5]	18	2	0.00	0.21 (0.10)	2.12 (0.034)	2	0.00	0.21 (0.10)	2.12 (0.034)	2	0.00	0.02 (0.01)	-2.72 (0.006)	2	0.00	0.02 (0.01)	2.5 (0.012)
			0.00	0.19 (0.09)	2.18 (0.029)		0.00	0.19 (0.09)	2.18 (0.029)		0.00	0.02 (0.01)	2.54 (0.011)		0.00	0.02 (0.01)	2.57 (0.010)
Bansback[6]	2	0				0				1	0.00	0.91 (0.38)	2.37 (0.018)	0			
Bottomley[7]	4	0				0				0				0			
Brown[8]	6	0				0				0				0			
Bucher[9]	2	0				0				0				0			
Cipriani[10]	70	3	0.02	0.71 (0.33)	-2.14 (0.032)	3	0.02	0.71 (0.33)	-2.17 (0.030)	4	0.00	0.38 (0.13)	-2.86 (0.004)	3	0.00	0.18 (0.08)	-2.37 (0.018)
			0.00	1.15 (0.51)	-2.27 (0.023)		0.00	1.15 (0.51)	-2.27 (0.023)		0.00	0.58 (0.26)	-2.27 (0.024)		0.00	0.29 (0.12)	-2.35 (0.019)

			0.00	0.61 (0.24)	-2.51 (0.012)		0.00	0.61 (0.24)	-2.51 (0.012)		0.00	0.23 (0.1)	-2.33 (0.02)		0.00	0.14 (0.06)	-2.48 (0.013)
											0.00	0.28 (0.14)	-2.01 (0.045)				
Dias[11]	11	1	0.00	1.20 (0.41)	-2.93 (0.003)	1	0.00	1.15 (0.40)	-2.90 (0.004)	1	0.00	0.05 (0.02)	2.89 (0.004)	1	0.00	0.05 (0.02)	-2.96 (0.003)
Eisenberg[12]	1	0				0				0				0			
Elliott[13]	16	2	0.01	0.83 (0.30)	2.79 (0.005)	3	0.00	0.58 (0.29)	1.99 (0.046)	3	0.00	0.02 (0.01)	2.90 (0.004)	3	0.00	0.02 (0.01)	2.86 (0.004)
			0.00	0.71 (0.27)	2.64 (0.008)		0.01	0.80 (0.29)	2.79 (0.005)		0.00	0.02 (0.01)	-2.23 (0.026)		0.00	0.01 (0.01)	2.33 (0.020)
							0.00	0.70 (0.26)	2.68 (0.007)		0.00	0.03 (0.01)	-2.41 (0.016)		0.00	0.03 (0.01)	2.45 (0.014)
Govan[14]	2	1	0.00	0.90 (0.39)	2.29 (0.022)	1	0.00	0.82 (0.33)	2.49 (0.013)	0				0			
Hofmeyr[15]	1	0				0				0				0			
Imamura[16]	26	5	0.27	4.71 (1.30)	-3.61 (0.000)	6	0.02	3.35 (0.98)	3.41 (0.001)	6	0.02	3.35 (0.98)	3.41 (0.001)	7	0.03	0.8 (0.24)	3.32 (0.001)
			0.00	2.52 (1.06)	-2.38 (0.017)		0.00	1.72 (0.77)	2.24 (0.025)		0.00	1.72 (0.77)	2.24 (0.025)		0.00	0.45 (0.21)	2.12 (0.034)
			0.00	4.52 (0.95)	-4.76 (0.000)		0.01	1.68 (0.45)	3.71 (0.000)		0.01	1.68 (0.45)	3.71 (0.000)		0.00	0.69 (0.14)	4.79 (0.000)
			0.00	3.05 (1.18)	2.59 (0.010)		0.03	1.31 (0.62)	2.1 (0.036)		0.03	1.31 (0.62)	2.1 (0.036)		0.00	0.17 (0.08)	-1.99 (0.046)
			0.00	1.90 (0.75)	2.53 (0.011)		0.00	2.37 (1.00)	-2.38 (0.017)		0.00	2.37 (1.00)	-2.38 (0.017)		0.01	0.45 (0.23)	-2.01 (0.044)
							0.00	1.12 (0.55)	-2.05 (0.040)		0.00	1.12 (0.55)	-2.05 (0.040)		0.00	0.37 (0.13)	-2.73 (0.006)
															0.00	0.37 (0.16)	2.28 (0.023)
Lam[17]	3	0				0				0				0			
Lapitan[18]	6	0				0				1	0.00	0.33 (0.16)	-2.02 (0.043)	1	0.00	0.30 (0.13)	2.24 (0.025)
Lu (1)[19]	4	0				0				0				0			

Lu (2)[19]	4	0				0				0				0			
Macfayden[22]	2	0				0				0				0			
Middleton[23]	1	0				0				0				0			
Mills[24]	2	0				0				0				0			
Nixon[25]	2	1	0.00	2.36 (0.52)	4.59 (0.000)	1	0.00	0.65 (0.16)	-4.08 (0.000)	1	0.00	1.72 (0.39)	4.36 (0.000)	1	0.00	0.45 (0.09)	5.21 (0.000)
Picard[26]	33	2	0.64	1.89 (0.93)	2.03 (0.042)	3	0.14	0.89 (0.40)	-2.22 (0.027)	1	0.00	1.58 (0.73)	-2.18 (0.029)	1	0.04	0.43 (0.20)	2.11 (0.035)
			0.81	2.52 (1.25)	-2.02 (0.043)		0.09	1.21 (0.54)	2.25 (0.025)								
							0.17	1.39 (0.68)	-2.06 (0.040)								
Playford[27]	1	0				0				0				0			
Psaty[28]	10	1	0.00	0.76 (0.29)	2.66 (0.008)	1	0.00	0.70 (0.26)	2.72 (0.007)	1	0.00	0.03 (0.01)	2.33 (0.020)	2	0.00	0.05 (0.03)	2.00 (0.046)
															0.00	0.03 (0.01)	2.33 (0.02)
Puhan[29]	7	0				0				1	0.00	0.15 (0.06)	2.36 (0.018)	1	0.00	0.08 (0.04)	2.22 (0.026)
Roskell (1)[31]	6	1	0.00	0.77 (0.32)	2.43 (0.015)	1	0.00	0.75 (0.3)	2.45 (0.014)	1	0.00	0.03 (0.01)	-2.39 (0.017)	1	0.00	0.03 (0.01)	2.33 (0.020)
Roskell (2)[30]	3	0				0				0				0			
Salliot[32]	1	1	0.02	0.86 (0.35)	2.44 (0.015)	0				1	0.03	0.70 (0.3)	2.36 (0.018)	0			
Sciarretta[33]	13	0				0				2	0.00	0.02 (0.01)	-2.14 (0.032)	2	0.00	0.01 (0.10)	2.08 (0.037)
											0.00	0.01 (0.01)	-2.13 (0.033)		0.00	0.01 (0.00)	2.06 (0.040)
Soares-Weiser[34]	4	0				1	0.01	0.38 (0.16)	2.39 (0.017)	0				0			
Thijs[35]	3	0				0				0				0			
Trikalinos[36]	1	0				0				0				0			

Virgili[37]	1	0				0				0				0			
Wang[38]	4	1	0.11	2.08 (1.00)	2.07 (0.038)	0				1	0.01	0.99 (0.44)	2.26 (0.024)	1	0.00	0.45 (0.19)	2.36 (0.018)
Welton[39]	4	0				0				0				0			
Woo[40]	3	0				0				0				0			
Yu[41]	5	0				0				0				0			

Appendix Table 5. W^{DBT} (P value) results according to design-by-treatment interaction model (DBT) using the restricted maximum likelihood (REML) and maximum likelihood (ML) estimators when applying all four effect measures. RD: risk difference measure, RRH: risk ratio for harmful outcomes, RRB: risk ratio for beneficial outcomes, OR: odds ratio.

Network	***OR***		***RRH***		***RRB***		***RD***	
	REML	*ML*	*REML*	*ML*	*REML*	*ML*	*REML*	*ML*
Ades[1]	19.60 (<0.001)	19.52 (<0.001)	13.20 (0.004)	18.32 (<0.001)	22.63 (<0.001)	22.63 (<0.001)	22.03 (<0.001)	22.03 (<0.001)
Ara[2]	1.72 (0.944)	1.76 (0.941)	1.75 (0.941)	1.75 (0.941)	1.11 (0.981)	1.83 (0.935)	2.41 (0.878)	2.41 (0.878)
Baker[3]	16.07 (0.1883)	17.61 (0.128)	25.02 (0.015)	26.24 (0.01)	15.13 (0.235)	15.13 (0.235)	11.70 (0.470)	13.58 (0.328)
Ballesteros[4]	1.78 (0.776)	3.20 (0.526)	3.07 (0.547)	4.36 (0.359)	2.86 (0.582)	6.06 (0.194)	1.96 (0.744)	3.57 (0.467)
Bangalore[5]	7.7 (0.935)	14.36 (0.499)	14.17 (0.513)	20.49 (0.154)	16.82 (0.330)	16.83 (0.329)	18.86 (0.220)	18.86 (0.220)
Bansback[6]	2.16 (0.339)	2.16 (0.340)	2.22 (0.330)	2.35 (0.310)	7.15 (0.028)	7.15 (0.028)	1.30 (0.523)	1.47 (0.480)
Bottomley[7]	5.57 (0.473)	22.59 (0.001)	6.92 (0.328)	31.18 (<0.001)	5.52 (0.479)	16.89 (0.01)	5.26 (0.511)	24.90 (<0.001)
Brown[8]	5.75 (0.675)	5.85 (0.664)	5.50 (0.703)	5.57 (0.695)	5.45 (0.709)	5.45 (0.709)	5.91 (0.657)	5.91 (0.657)
Bucher[9]	0.74 (0.692)	0.73 (0.695)	0.70 (0.706)	0.70 (0.706)	1.04 (0.594)	1.35 (0.508)	1.13 (0.567)	1.49 (0.474)
Cipriani[10]	30.79 (0.577)	32.25 (0.504)	28.4 (0.696)	37.04 (0.288)	32.7 (0.482)	38.85 (0.223)	30.37 (0.599)	39.72 (0.196)
Dias[11]	9.90 (0.449)	12.78 (0.236)	9.90 (0.449)	12.60 (0.247)	8.41 (0.589)	11.49 (0.321)	8.73 (0.558)	12.18 (0.273)
Eisenberg[12]	2.65 (0.265)	3.27 (0.195)	3.19 (0.203)	3.76 (0.153)	3.23 (0.199)	4.24 (0.120)	3.09 (0.214)	3.66 (0.161)
Elliott[13]	19.61 (0.106)	31.70 (0.003)	20.09 (0.093)	31.27 (0.003)	9.53 (0.732)	31.78 (0.003)	9.00 (0.773)	32.33 (0.002)
Govan[14]	12.12 (0.017)	12.1 (0.017)	12.67 (0.013)	12.67 (0.013)	7.69 (0.104)	8.23 (0.083)	9.07 (0.059)	9.50 (0.050)
Hofmeyr[15]	3.44 (0.179)	3.44 (0.179)	3.47 (0.177)	3.47 (0.177)	2.72 (0.257)	2.92 (0.232)	2.72 (0.256)	2.94 (0.230)
Imamura[16]	32.47 (0.070)	26.84 (0.140)	11.16 (0.934)	33.17 (0.032)	21.71 (0.357)	23.56 (0.262)	15.85 (0.726)	45.81 (0.001)
Lam[17]	2.92 (0.404)	2.92 (0.404)	2.78 (0.427)	2.78 (0.427)	0.21 (0.977)	0.57 (0.904)	0.16 (0.983)	0.35 (0.949)
Lapitan[18]	6.09 (0.193)	6.49 (0.166)	5.85 (0.211)	5.85 (0.211)	8.97 (0.062)	8.97 (0.062)	9.49 (0.050)	9.49 (0.050)
Lu (1)[19]	5.11 (0.646)	6.76 (0.455)	4.57 (0.713)	5.87 (0.555)	5.19 (0.637)	6.97 (0.432)	5.64 (0.582)	7.48 (0.381)
Lu (2)[19]	11.24 (0.081)	6.06 (0.195)	11.86 (0.065)	14.53 (0.024)	10.32 (0.112)	13.92 (0.031)	12.05 (0.061)	16.76 (0.010)
Macfayden [22]	13.14 (0.022)	20.74 (0.001)	15.23 (0.009)	15.23 (0.009)	0.00 (<0.001)	27.22 (<0.001)	3.69 (0.595)	14.38 (0.013)

Middleton[23]	2.18 (0.140)	2.17 (0.141)	1.90 (0.168)	1.90 (0.168)	2.76 (0.097)	2.76 (0.097)	2.87 (0.091)	2.87 (0.091)
Mills[24]	1.75 (0.782)	2.02 (0.732)	3.14 (0.535)	3.53 (0.473)	1.14 (0.889)	1.29 (0.863)	1.94 (0.746)	2.19 (0.700)
Nixon[25]	7.25 (0.065)	29.51 (<0.001)	14.92 (0.002)	21.76 (<0.001)	5.09 (0.165)	28.05 (<0.001)	12.37 (0.006)	39.33 (<0.001)
Picard[26]	60.27 (0.001)	101.29 (<0.001)	60.67 (0.001)	127.27 (<0.001)	50.24 (0.016)	50.24 (0.016)	62.85 (0.001)	123.81 (<0.001)
Playford[27]	1.53 (0.217)	1.52 (0.218)	1.49 (0.222)	1.49 (0.222)	0.94 (0.333)	0.94 (0.333)	0.81 (0.369)	1.11 (0.291)
Psaty[28]	10.71 (0.38)	13.62 (0.191)	5.99 (0.816)	10.32 (0.413)	10.21 (0.423)	18.10 (0.053)	9.64 (0.473)	16.76 (0.080)
Puhan[29]	9.4 (0.226)	7.15 (0.413)	8.52 (0.289)	8.52 (0.289)	6.37 (0.498)	9.51 (0.218)	6.49 (0.418)	8.19 (0.316)
Roskell (1)[31]	4.57 (0.335)	8.03 (0.090)	4.54 (0.337)	8.23 (0.084)	3.56 (0.469)	5.66 (0.226)	3.45 (0.486)	5.86 (0.210)
Roskell (2)[30]	0.2 (0.906)	0.20 (0.906)	1.31 (0.520)	1.31 (0.520)	0.51 (0.776)	0.51 (0.776)	0.82 (0.663)	0.82 (0.663)
Salliot[32]	11.81 (0.003)	11.81 (0.003)	2.74 (0.254)	2.76 (0.252)	10.44 (0.005)	13.34 (0.001)	5.11 (0.078)	5.11 (0.078)
Sciarretta[33]	12.89 (0.456)	22.25 (0.052)	14.33 (0.351)	14.33(0.351)	42.75 (<0.001)	42.75 (<0.001)	50.80 (<0.001)	50.80 (<0.001)
Soares-Weiser[34]	1.98 (0.961)	7.97 (0.336)	1.33 (0.988)	21.62 (0.003)	2.86 (0.898)	7.17 (0.411)	1.91 (0.965)	12.62 (0.082)
Thijs[35]	1.66 (0.893)	1.66 (0.893)	1.87 (0.867)	1.87 (0.867)	1.61 (0.9)	1.91 (0.861)	1.64 (0.896)	1.86 (0.868)
Trikalinos[36]	0.66 (0.415)	0.73 (0.393)	0.68 (0.411)	0.68 (0.411)	0.01 (0.905)	0.01 (0.906)	0.04 (0.850)	0.04 (0.850)
Virgili[37]	0.09 (0.766)	0.13 (0.714)	0.01 (0.910)	0.01 (0.910)	2.39 (0.122)	2.39 (0.122)	1.50 (0.221)	1.59 (0.207)
Wang[38]	5.68 (0.577)	8.46 (0.294)	5.64 (0.582)	8.76 (0.270)	6.21 (0.515)	8.01 (0.331)	6.05 (0.534)	8.28 (0.309)
Welton[39]	4.14 (0.845)	4.48 (0.812)	4.01 (0.857)	4.30 (0.829)	6.33 (0.611)	8.13 (0.420)	6.57 (0.584)	8.25 (0.410)
Woo[40]	5.52 (0.238)	5.59 (0.232)	2.13 (0.711)	3.51 (0.477)	10.69 (0.030)	24.39 (<0.001)	4.89 (0.299)	8.10 (0.088)
Yu[41]	3.28 (0.858)	3.28 (0.858)	3.27 (0.859)	3.27 (0.859)	2.71 (0.910)	2.71 (0.910)	2.82 (0.901)	2.82 (0.901)

Appendix Table 6. Inconsistency factors (IF^{SIDDE}) along with their standard error (SE(IF^{SIDDE})), W^{SIDDE} and heterogeneity as estimated when detaching each design. Heterogeneity has been estimated using the restricted maximum likelihood estimator.

Network	*Number of designs in the network*	IF^{SIDDE} *(SE(IF^{SIDDE}))*	W^{SIDDE}*(P-value)*	*heterogeneity*
Ades[1]	8	1.63 (0.39)	17.79 (<0.001)	0.00
		1.54 (0.51)	9.17 (0.002)	0.04
		1.54 (0.51)	9.17 (0.002)	0.04
Ara[2]	8	*Consistent*		
Baker[3]	12	*Consistent*		
Ballesteros[4]	5	*Consistent*		
Bangalore[5]	18	0.20 (0.08)	6.53 (0.011)	0.00
Bansback[6]	2	*Consistent*		
Bottomley[7]	7	1.34 (0.55)	6.04 (0.014)	0.06
Brown[8]	11	*Consistent*		
Bucher[9]	5	*Consistent*		
Cipriani[10]	43	0.43 (0.17)	6.24 (0.013)	0.01
Dias[11]	15	1.19 (0.41)	8.45 (0.004)	0.00
Eisenberg[12]	3	*Consistent*		
Elliott[13]	15	0.77 (0.27) 0.65 (0.27)	9.48 (0.009)	0.01
		0.29 (0.14)	4.19 (0.041)	0.01
Govan[14]	6	0.73 (0.35)	4.31 (0.038)	0.00
		0.91 (0.39)	5.28 (0.022)	0.00
		0.91 (0.39)	5.28 (0.022)	0.00
Hofmeyr[15]	3	*Consistent*		
Imamura[16]	24	3.11 (1.54)	4.07 (0.044)	0.06
		1.91 (0.80)	5.75 (0.016)	0.01
		0.93 (0.46)	4.09 (0.043)	0.01
		1.44 (0.69)	4.35 (0.037)	0.00
Lam[17]	3	*Consistent*		
Lapitan[18]	5	*Consistent*		
Lu (1)[19]	8	*Consistent*		
Lu (2)[19]	10	1.59 (0.72)	4.90 (0.027)	0.21
		0.72 (0.49) 0.67 (0.51)	6.01 (0.049)	0.22
Macfayden[22]	6	1.87 (1.25) 3.13 (1.30)	7.05 (0.029)	0.23
Middleton[23]	3	*Consistent*		
Mills[24]	5	*Consistent*		
Nixon[25]	7	*Consistent*		
Picard[26]	22	3.03 (1.01) 2.79 (0.95) 1.25 (1.02)	12.59 (0.006)	0.50
		2.29 (1.12)	4.17 (0.041)	0.60
		1.28 (1.12) 3.20 (1.24) 2.75 (1.16)	8.94 (0.030)	0.56
Playford[27]	3	*Consistent*		
Psaty[28]	11	0.52 (0.25)	4.22 (0.040)	0.01

		0.69 (0.30)	5.24 (0.022)	0.00
Puhan[29]	8	0.31 (0.14)	4.89 (0.027)	0.00
Roskell (1)[30]	8	*Consistent*		
Roskell (2)[31]	2	*Consistent*		
Salliot[32]	3	0.26 (0.34) 1.04 (0.32)	11.81(0.003)	0.00
		0.91 (0.27)	11.20 (0.001)	0.00
Sciarretta[33]	17	0.49 (0.22)	4.80 (0.028)	0.01
		0.55 (0.28)	3.96 (0.047)	0.00
Soares-Weiser[34]	9	*Consistent*		
Thijs[35]	8	*Consistent*		
Trikalinos[36]	3	*Consistent*		
Virgili[37]	3	*Consistent*		
Wang[38]	10	2.10 (1.05) 2.10 (1.05)	3.97 (0.046)	0.17
		2.10 (1.05) 2.10 (1.05)	3.97 (0.046)	0.17
Welton[39]	11	*Consistent*		
Woo[40]	4	*Consistent*		
Yu[41]	8	*Consistent*		

Appendix Table 7. Inconsistency factors (IF^{SIDE}) and their standard error ($SE(IF^{SIDE})$), W^{SIDE} values and common-within network heterogeneity estimated in SIDE method. Heterogeneity has been estimated using the restricted maximum likelihood estimator. Note that in one network (21) SIDE inconsistency could not be evaluated. Whenever a comparison was part of at least one multi-arm study, I estimated inconsistency after re-parameterisation of the multi-arm studies and I present the maximum W^{SIDE} value. N: number of comparisons that can be separated in the network.

Network	*Is the comparison included in a multi-arm trial?*	*N*	*IF (SE(IF))*	*W^{SIDE} (P-value)*	*heterogeneity*
Ades[1]	NO	8	1.63 (0.39)	4.22 (<0.001)	0.00
	NO		1.54 (0.51)	3.03 (0.002)	0.04
	NO		1.54 (0.51)	3.03 (0.002)	0.04
Ara[2]		8	*Consistent*		
Baker[3]		14	*Consistent*		
Ballesteros[4]		5	*Consistent*		
Bangalore[5]	NO	18	0.20 (0.08)	2.56 (0.011)	0.00
Bansback[6]		4	*Consistent*		
Bottomley[7]	YES	6	1.34 (0.55)	2.46 (0.014)	0.06
Brown[8]	YES	10	1.27 (0.63)	2.02 (0.044)	0.00
Bucher[9]		5	*Consistent*		
Cipriani[10]	NO	42	0.43 (0.17)	2.50 (0.013)	0.01
Dias[11]	NO	15	1.19 (0.41)	2.91 (0.004)	0.00
Eisenberg[12]		1	*Consistent*		
Elliott[13]	YES	14	0.71 (0.23)	3.05 (0.002)	0.01
	YES		0.27 (0.11)	2.54 (0.011)	0.01

Govan[14]	NO	5	0.91 (0.39)	2.30 (0.022)	0.00
	YES		0.73 (0.35)	2.08 (0.038)	0.00
	NO		0.91 (0.39)	2.30 (0.022)	0.00
Hofmeyr[15]		1		*Consistent*	
Imamura[16]	YES	21	1.39 (0.69)	2.01 (0.044)	0.03
	YES		1.42 (0.6)	2.36 (0.018)	0.02
	NO		1.91 (0.80)	2.40 (0.016)	0.01
	YES		1.51 (0.69)	2.17 (0.030)	0.00
	YES		1.61 (0.51)	3.19 (0.001)	0.00
	NO		1.44 (0.69)	2.09 (0.037)	0.00
Lam[17]		7		*Consistent*	
Lapitan[18]		1		*Consistent*	
Lu (1)[19]		6		*Consistent*	
Lu (2)[19]	NO	9	1.60 (0.72)	2.22 (0.026)	0.21
Macfayden[22]	YES	5	1.49 (0.75)	2.00 (0.045)	0.32
	YES		2.98 (1.49)	2.01 (0.045)	0.32
Middleton[23]		3		*Consistent*	
Mills[24]		5		*Consistent*	
Nixon[25]	YES	6	1.94 (0.50)	3.88 (<0.001)	0.07
Picard[26]	YES	23	1.51 (0.47)	3.17 (0.002)	0.52
	YES		1.53 (0.53)	2.91 (0.004)	0.52
	YES		1.63 (0.79)	2.07 (0.039)	0.61
	YES		2.57 (0.70)	3.7 (<0.001)	0.45
	YES		2.11 (0.85)	2.48 (0.013)	0.57
Playford[27]		3		*Consistent*	
Psaty[28]	YES	10	0.46 (0.22)	2.09 (0.036)	0.00
	NO		0.52 (0.25)	2.05 (0.04)	0.01
Puhan[29]	YES	8	0.29 (0.13)	2.26 (0.024)	0.00
Roskell (1)[31]		8		*Consistent*	
Roskell (2)[30]				*Inconsistency could not be estimated*	
Salliot[32]	YES	3	0.91 (0.27)	3.35 (0.001)	0.00
	YES		1.83 (0.55)	3.35 (0.001)	0.00
Sciarretta[33]	NO	15	0.55 (0.28)	1.99 (0.047)	0.00
	NO		0.49 (0.22)	2.19 (0.028)	0.01
Soares-Weiser[34]		8		*Consistent*	
Thijs[35]		7		*Consistent*	
Trikalinos[36]		3		*Consistent*	
Virgili[37]		3		*Consistent*	
Wang[38]	NO	9	2.10 (1.05)	1.99 (0.046)	0.17
	NO		2.10 (1.05)	1.99 (0.046)	0.17
Welton[39]		9		*Consistent*	
Woo[40]	YES	9	0.63 (0.30)	2.10 (0.036)	0.00

	YES	1.26 (0.60)	2.10 (0.035)	0.00
Yu[41]	9	*Consistent*		

Appendix Table 8. Results according to Lu and Ades (LA) model and I^2 measures. Heterogeneity has been estimated ($\hat{\tau}^2$) using the restricted maximum likelihood estimator. Note that in one network (21) inconsistency could not be evaluated using the LA model. In bold I present the networks for which the test for inconsistency was statistically significant. For the LA model I applied all different parameterisations of the multi-arm studies and I present the maximum W^{LA} value. df: degrees of freedom

	Lu and Ades model				***I^2 measures***		
Network	df	W^{LA}	p-value	$\hat{\tau}^2$	I^2_{Inc}	I^2_{Het}	$I^2_{Het,Inc}$
Ades[1]	**3**	**19.60**	**<0.001**	**0.00**	0.00	0.71	0.71
Ara[2]	4	0.97	0.914	0.00	0.73	0.00	0.73
Baker[3]	6	12.14*	0.059*	0.00	0.70	0.23	0.77
Ballesteros[4]	2	1.51*	0.471*	0.06	0.84	0.32	0.89
Bangalore[5]	10	9.34*	0.501*	0.00	0.91	0.38	0.95
Bansback[6]	1	0.02	0.883	0.00	0.29	0.00	0.29
Bottomley[7]	2	2.90*	0.235*	0.11	0.64	0.74	0.91
Brown[8]	5	4.96*	0.421*	0.00	0.43	0.14	0.51
Bucher[9]	2	0.74	0.692	0.00	0.21	0.00	0.21
Cipriani[10]	31	29.96	0.519	0.01	0.75	0.18	0.80
Dias[11]	5	7.34*	0.197	0.00	0.94	0.08	0.95
Eisenberg[12]	1	2.47*	0.116*	0.02	0.30	0.35	0.55
Elliott[13]	**9**	**18.10**	**0.034**	**0.01**	0.68	0.62	0.88
Govan[14]	**2**	**10.93**	**0.004**	**0.00**	0.58	0.00	0.58
Hofmeyr[15]	1	3.25*	0.071*	0.00	0.06	0.00	0.06
Imamura[16]	**14**	**25.25**	**0.032**	**0.00**	0.52	0.18	0.61
Lam[17]	1	0.08	0.773	0.00	0.45	0.00	0.45
Lapitan[18]	1	0.16	0.691	0.00	0.48	0.00	0.48
Lu (1)[19]	3	2.85*	0.416*	0.47	0.76	0.84	0.96
Lu (2)[19]	4	8.44*	0.077*	0.22	0.62	0.64	0.86
Macfayden [22]	**2**	**11.08**	**0.004**	**0.11**	0.39	0.75	0.85
Middleton[23]	1	2.18	0.140	0.00	0.46	0.00	0.46
Mills[24]	2	1.52*	0.468	0.18	0.17	0.78	0.81
Nixon[25]	2	3.03	0.220	0.39	0.00	0.91	0.88
Picard[26]	**11**	**23.73**	**0.008**	**0.46**	0.63	0.72	0.89
Playford[27]	1	1.53	0.217	0.00	0.23	0.00	0.23
Psaty[28]	6	10.63*	0.100*	0.05	0.68	0.46	0.83
Puhan[29]	4	8.66*	0.070*	0.00	0.63	0.28	0.73
Roskell (1)[31]	4	4.57*	0.335*	0.04	0.12	0.60	0.65
Roskell (2)[30]	-	-	-	-	0.41	0.00	0.41
Salliot[32]	**1**	**11.20**	**<0.001**	**0.00**	0.00	0.60	0.19
Sciarretta[33]	9	11.16*	0.264*	0.02	0.96	0.60	0.98
Soares-Weiser[34]	2	1.15*	0.562*	0.44	0.78	0.72	0.94

Thijs[35]	2	0.68*	0.877*	0.00	0.76	0.00	0.76
Trikalinos[36]	1	0.66	0.415	0.00	0.25	0.00	0.25
Virgili[37]	1	0.09	0.766	0.01	0.28	0.07	0.33
Wang[38]	3	5.59*	0.134*	0.17	0.28	0.66	0.75
Welton[39]	4	1.86*	0.762*	0.21	0.20	0.46	0.57
Woo[40]	2	5.02*	0.081*	0.00	0.51	0.00	0.51
Yu[41]	4	2.16*	0.706*	0.00	0.72	0.00	0.72

* a different parameterisation of the multi-arm studies results in different chi-square tests, but the significance of the test does not change.

References to the included networks

1. Ades AE, Mavranezouli I, Dias S, et al. Network meta-analysis with competing risk outcomes. *Value Health* 2010;**13**:976-83.

2. Ara R, Pandor A, Stevens J, et al. Early high-dose lipid-lowering therapy to avoid cardiac events: a systematic review and economic evaluation. *Health Technol Assess* 2009;**13**:1-118.

3. Baker WL, Baker EL, Coleman CI. Pharmacologic treatments for chronic obstructive pulmonary disease: a mixed-treatment comparison meta-analysis. *Pharmacotherapy* 2009;**29**:891-905.

4. Ballesteros J. Orphan comparisons and indirect meta-analysis: a case study on antidepressant efficacy in dysthymia comparing tricyclic antidepressants, selective serotonin reuptake inhibitors, and monoamine oxidase inhibitors by using general linear models. *J Clin Psychopharmacol* 2005;**25**:127-31.

5. Bangalore S, Kumar S, Kjeldsen SE, et al. Antihypertensive drugs and risk of cancer: network meta-analyses and trial sequential analyses of 324,168 participants from randomised trials. *Lancet Oncol* 2011;12:65-82.

6. Bansback N, Sizto S, Sun H, et al. Efficacy of systemic treatments for moderate to severe plaque psoriasis: systematic review and meta-analysis. *Dermatology* 2009;**219**:209-18.

7. Bottomley JM, Taylor RS, Ryttov J. The effectiveness of two-compound formulation calcipotriol and betamethasone dipropionate gel in the treatment of moderately severe scalp psoriasis: a systematic review of direct and indirect evidence. *Curr Med Res Opin* 2011;27:251-68.

8. Brown TJ, Hooper L, Elliott RA, et al. A comparison of the cost-effectiveness of five strategies for the prevention of non-steroidal anti-inflammatory drug-induced gastrointestinal toxicity: a systematic review with economic modelling. *Health Technol Assess* 2006;**10**:iii-xiii, 1.

9. 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**:683-91.

10. Cipriani A, Furukawa TA, Salanti G, et al. Comparative efficacy and acceptability of 12 new-generation antidepressants: a multiple-treatments meta-analysis. *Lancet* 2009;**373**:746-58.

11. Dias S, Welton NJ, Caldwell DM, et al. Checking consistency in mixed treatment comparison meta-analysis. *Stat Med* 2010;**29**:932-44.

12. Eisenberg MJ, Filion KB, Yavin D, et al. Pharmacotherapies for smoking cessation: a meta-analysis of randomized controlled trials. *CMAJ* 2008;179:135-44.

13. Elliott WJ, Meyer PM. Incident diabetes in clinical trials of antihypertensive drugs: a network meta-analysis. *Lancet* 2007;**369**:201-7.

14. Govan L, Ades AE, Weir CJ, et al. Controlling ecological bias in evidence synthesis of trials reporting on collapsed and overlapping covariate categories. *Stat Med* 2010;**29**:1340-56.

15. Hofmeyr G J, Gulmezoglu A M, Novikova N, et al. Misoprostol to prevent and treat postpartum haemorrhage: a systematic review and meta-analysis of maternal deaths and dose-related effects. *Bull World Health Organ* 2009;645-732.

16. Imamura M, Abrams P, Bain C, et al. Systematic review and economic modelling of the effectiveness and cost-effectiveness of non-surgical treatments for women with stress urinary incontinence. *Health Technol Assess* 2010;14:1-iv.

17. Lam SK, Owen A. Combined resynchronisation and implantable defibrillator therapy in left ventricular dysfunction: Bayesian network meta-analysis of randomised controlled trials. *BMJ* 2007;**335**:925.

18. Lapitan MC, Cody JD, Grant A. Open retropubic colposuspension for urinary incontinence in women. *Cochrane Database Syst Rev* 2009;CD002912.

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

20. Lu G, Ades AE, Sutton AJ, et al. Meta-analysis of mixed treatment comparisons at multiple follow-up times. *Stat Med* 2007;**26**:3681-99.

21. Lu GB, Ades AE. Assessing evidence inconsistency in mixed treatment comparisons. *Journal of the American Statistical Association* 2006;**101**:447-59.

22. Macfadyen CA, Acuin JM, Gamble C. Topical antibiotics without steroids for chronically discharging ears with underlying eardrum perforations. *Cochrane Database Syst* Rev 2005;CD004618.

23. Middleton LJ, Champaneria R, Daniels JP, et al. Hysterectomy, endometrial destruction, and levonorgestrel releasing intrauterine system (Mirena) for heavy

menstrual bleeding: systematic review and meta-analysis of data from individual patients. *BMJ* 2010;**341**:c3929.

24. Mills EJ, Wu P, Spurden D, et al. Efficacy of pharmacotherapies for short-term smoking abstinance: a systematic review and meta-analysis. *Harm Reduct* J 2009;6:25.

25. Nixon R, Bansback N, Brennan A. The efficacy of inhibiting tumour necrosis factor alpha and interleukin 1 in patients with rheumatoid arthritis: a meta-analysis and adjusted indirect comparisons. *Rheumatology* (Oxford) 2007;**46**:1140-7.

26. Picard P, Tramer MR. Prevention of pain on injection with propofol: a quantitative systematic review. *Anesth Analg* 2000;**90**:963-9.

27. Playford EG, Webster AC, Sorell TC, et al. Antifungal agents for preventing fungal infections in solid organ transplant recipients. *Cochrane Database Syst Rev* 2004;CD004291.

28. Psaty BM, Smith NL, Siscovick DS, et al. Health outcomes associated with antihypertensive therapies used as first-line agents. A systematic review and meta-analysis. *JAMA* 1997;**277**:739-45.

29. Puhan MA, Bachmann LM, Kleijnen J, et al. Inhaled drugs to reduce exacerbations in patients with chronic obstructive pulmonary disease: a network meta-analysis. *BMC* Med 2009;**7**:2.

30. Roskell NS, Beard SM, Zhao Y, et al. A meta-analysis of pain response in the treatment of fibromyalgia. *Pain Pract* 2011;**11**:516-27.

31. Roskell NS, Lip GY, Noack H, et al. Treatments for stroke prevention in atrial fibrillation: a network meta-analysis and indirect comparisons versus dabigatran etexilate. *Thromb Haemost* 2010;**104**:1106-15.

32. Salliot C, Finckh A, Katchamart W, et al. Indirect comparisons of the efficacy of biological antirheumatic agents in rheumatoid arthritis in patients with an inadequate response to conventional disease-modifying antirheumatic drugs or to an anti-tumour necrosis factor agent: a meta-analysis. *Ann Rheum Dis* 2011;70:266-71.

33. Sciarretta S, Palano F, Tocci G, et al. Antihypertensive treatment and development of heart failure in hypertension: a Bayesian network meta-analysis of studies in patients with hypertension and high cardiovascular risk. *Arch Intern Med* 2011;**171**:384-94.

34. Soares-Weiser K, Bravo VY, Beynon S, et al. A systematic review and economic model of the clinical effectiveness and cost-effectiveness of interventions for preventing relapse in people with bipolar disorder. *Health Technol Assess* 2007;**11**:iii-206.

35. Thijs V, Lemmens R, Fieuws S. Network meta-analysis: simultaneous meta-analysis of common antiplatelet regimens after transient ischaemic attack or stroke. *Eur Heart J* 2008;**29**:1086-92.

36. Trikalinos TA, Alsheikh-Ali AA, Tatsioni A, et al. Percutaneous coronary interventions for non-acute coronary artery disease: a quantitative 20-year synopsis and a network meta-analysis. *Lancet* 2009;**373**:911-8.

37. Virgili G, Novielli N, Menchini F, et al. Pharmacological treatments for neovascular age-related macular degeneration: can mixed treatment comparison meta-analysis be useful? *Curr Drug Targets* 2011;**12**:212-20.

38. Wang H, Huang T, Jing J, et al. Effectiveness of different central venous catheters for catheter-related infections: a network meta-analysis. *J Hosp Infect* 2010;**76**:1-11.

39. Welton NJ, Caldwell DM, Adamopoulos E, et al. Mixed treatment comparison meta-analysis of complex interventions: psychological interventions in coronary heart disease. *Am J Epidemiol* 2009;**169**:1158-65.

40. Woo G, Tomlinson G, Nishikawa Y, et al. Tenofovir and entecavir are the most effective antiviral agents for chronic hepatitis B: a systematic review and Bayesian meta-analyses. *Gastroenterology* 2010;**139**:1218-29.

41. Yu CH, Beattie WS. The effects of volatile anesthetics on cardiac ischemic complications and mortality in CABG: a meta-analysis. *Can J Anaesth* 2006;**53**:906-18.