Int J Epidemiol. Author manuscript; available in PMC 2017 May 01.

Published in final edited form as:
Int J Epidemiol. 2013 February ; 42(1): 332–345. doi:10.1093/ije/dys222.

Evaluation of inconsistency in networks of interventions

Areti Angeliki Veroniki,
Department of Hygiene and Epidemiology, University of Ioannina School of Medicine, Ioannina, Greece

Haris S. Vasiliadis,
Department of Orthopaedics, School of Medicine, University of Ioannina, Greece; Molecular Cell Biology and Regenerative Medicine, Sahlgrenska Academy, University of Gothenburg, Sweden

Julian P.T. Higgins, and
MRC Biostatistics Unit, Cambridge, UK; Centre for Reviews and Dissemination, University of York, York, UK

Georgia Salanti
Department of Hygiene and Epidemiology, University of Ioannina School of Medicine, Ioannina, Greece

Abstract

Background—The assumption of consistency, defined as agreement between direct and indirect sources of evidence, underlies the increasingly popular method of network meta-analysis. No evidence exists so far regarding the extent of inconsistency and the factors that control its statistical detection in full networks of interventions.

Methods—In this paper the prevalence of inconsistency is evaluated using 40 published networks of interventions involving 303 loops of evidence. Inconsistency is evaluated in each loop by contrasting direct and indirect estimates and by employing an omnibus test of consistency for the entire network. We explore whether different effect measures for dichotomous outcomes are associated with differences in inconsistency and evaluate whether different ways to estimate heterogeneity impact on the magnitude and detection of inconsistency.

Results—Inconsistency was detected in between 2% and 9% of the tested loops, depending on the effect measure and heterogeneity estimation method. Loops that included comparisons informed by a single study were more likely to show inconsistency. About one eighth of the networks were found to be inconsistent. The proportions of inconsistent loops do not materially change when different effect measures are employed. Important heterogeneity or overestimation of the heterogeneity was associated with a small decrease in the prevalence of statistical inconsistency.

Conclusions—The study suggests that changing effect measure might improve statistical consistency and that a sensitivity analysis to the assumptions and estimator of heterogeneity might

Correspondence to: Georgia Salanti Department of Hygiene and Epidemiology University of Ioannina School of Medicine University Campus Ioannina 45110 - Greece gsalanti@cc.uoi.gr.

Europe PMC Funders Author Manuscripts

be needed before concluding about the absence of statistical inconsistency, particularly in networks with few studies.

Keywords

mixed-treatment comparison; multiple treatments meta-analysis; loops; heterogeneity; odds ratio; coherence

Europe PMC Funders Author Manuscripts

1. Introduction

To inform health-care decision making the comparison of many relevant interventions is required. A commonly encountered problem in evaluating the efficacy of multiple interventions is the lack of trials (or very few available) that directly compare the treatments of interest. In such cases indirect evidence can be used via a common comparator. Bucher *et al.*[1] were early proponents of the use of indirect evidence in meta-analysis when head-to-head evidence is not available. The application of indirect comparison rests on the assumption of transitivity, requiring that the pairwise comparisons are similar in factors which could affect the relative treatment effects.

An extension of conventional meta-analysis is network meta-analysis. Network meta-analysis is used to combine the results of clinical trials that undertake different comparisons of treatments[2-5]. The method involves the simultaneous analysis of both direct comparisons within trials and indirect comparisons across trials. When combining the results of direct and indirect comparisons, the extent to which they are consistent (in agreement) with each other should be examined. Network meta-analysis is most justifiable under an assumption of consistency between different sources of evidence. The evaluation of evidence inconsistency is therefore an important aspect in network meta-analysis. In a network of treatments, different pairwise comparisons can form ‘evidence cycles’, also called ‘loops’, within which inconsistency can be evaluated[6].

Empirical studies have examined the prevalence of inconsistency between direct and indirect comparisons. Song *et al.*[7,8] 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 detected in 14% of the networks[8]. In a response to comments on their article, Song *et al.*[9] alternatively assumed that all comparisons within each triangular loop share the same amount of heterogeneity and they observed that inconsistency was reduced to 12%. However, no empirical evidence exists 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[10]. 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. A large heterogeneity variance in the treatment effects leads to greater

uncertainty in estimates of the mean effect sizes, and statistical inconsistency is less likely to be detected. The estimation of the heterogeneity variance can vary under different methods (e.g. DerSimonian and Laird, restricted maximum likelihood[11]), 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 ratios and risk ratios) are less heterogeneous than absolute effect measures (such the risk difference) and that the risk ratio for adverse outcomes is less likely to be heterogeneous than that for beneficial outcomes[12,13]. 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, with the same consequences as discussed in the previous paragraph. 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[14].

The aim of this paper 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 meta-analyses. We further aim to investigate the statistical considerations that might influence the statistical detection of inconsistency in these complex networks of evidence. We 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.

2. Methods

To assess inconsistency in a network we use two methods. The first 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. Bucher *et al.* described the method in an early paper[1] and we will refer to it, and its extensions employed in this paper, as the ‘loop-specific approach’. The second 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’)[10]. 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

Europe PMC Funders Author Manuscripts

design inconsistency reflects the possibility that they might give different estimates for the same comparisons they make (*AB, AC* and *BC*).

We chose the loop-based approach as it is simple and can be easily applied without specialised software in a frequentist setting, and is so far the most commonly applied approach. Moreover, the results obtained from this method can be compared directly with findings from other empirical studies[8]. We chose the design-by-treatment interaction approach as it is the only approach of which we are aware that does not require arbitrary assumptions on inclusion of trials with more than two treatment arms. It provides a generalization to the method earlier proposed by Lu and Ades[6]. Both the loop-specific and the design-by-treatment interaction approaches are employed under various effect measures for dichotomous outcome data and various estimators for the heterogeneity variance.

2.1 Loop-specific approach

Inconsistency can be evaluated as the disagreement between different sources of evidence within a closed loop. In each network of treatments we identified all triangular loops (closed paths involving three different treatments) as well as all quadrilateral loops (closed paths involving four different treatments).

We first estimate treatment effects of all pairwise comparisons in each loop using standard meta-analysis. Consider for example the triangular loop *ABC* formed by treatments *A, B, C* with available comparisons *AB, AC* and *BC*. Let $y_{i,AB}$ be the observed effect size (e.g. log-odds ratio) of treatment *B* relative to treatment *A* in study *i*, with an estimated variance $v_{i,AB}$. Under the random-effects model the observed treatment effect $y_{i,AB}$ is modeled as

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

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

$$\begin{aligned} y_{i,AC} &= \mu_{AC}+\delta_{i,AC}+\varepsilon_{i,AC} \\ y_{i,BC} &= \mu_{BC}+\delta_{i,BC}+\varepsilon_{i,BC} \end{aligned}$$

To estimate all direct relative effects within the triangular loop *ABC* we performed a random-effects meta-analysis for each available comparison. Under the random-effects model it is assumed that

$$\begin{aligned} &\delta_{i,AB}\sim N\left(0,\tau^2_{AB}\right), \quad \delta_{i,AC}\sim N\left(0,\tau^2_{AC}\right), \quad \delta_{i,BC}\sim N\left(0,\tau^2_{BC}\right) \\ &\varepsilon_{i,AB}\sim N\left(0,\nu_{i,AB}\right), \quad \varepsilon_{i,AC}\sim N\left(0,\nu_{i,AC}\right), \quad \varepsilon_{i,BC}\sim N\left(0,\nu_{i,BC}\right) \end{aligned}$$

where τ^2_{AB}, τ^2_{AC} and τ^2_{BC} are the heterogeneity variances in the *B vs.A* , *C vs.A* and *C vs.B* comparisons, respectively. The variances $v_{i,AB}$, $v_{i,AC}$ and $v_{i,BC}$ are assumed known and

uncorrelated with the effect sizes. We discuss assumptions about the heterogeneity variances in section 2.4.

Within each available loop, we evaluated whether the consistency assumption[6]

$$\mu_{BC}=\mu_{AC}-\mu_{AB}$$

holds. Since in a single loop there may be only one inconsistency, the inconsistency estimate (*IF*) for the loop *ABC* is defined as[6,15]

$$\text{with} \qquad \begin{array}{l} \widehat{IF}_{ABC}=\hat{\mu}_{BC}-\left(\hat{\mu}_{AC}-\hat{\mu}_{AB}\right) \\ \hat{\sigma}^2_{IF_{ABC}}= \quad Var\left(\hat{\mu}_{BC}\right)+Var\left(\hat{\mu}_{AC}\right)+Var\left(\hat{\mu}_{AB}\right), \end{array}$$

Under the null hypothesis that there is no inconsistency (H_0: IF_{ABC} = 0) the approximate test can be obtained as

$$Z=\frac{\widehat{IF}_{ABC}}{\hat{\sigma}_{IF_{ABC}}}\sim N\left(0,1\right)$$

We define a loop as statistically inconsistent when $|z| > 1.96$[16].

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, it causes complications and we therefore exclude the comparison that is most frequent within the loop. This can impact on the summary treatment effects and subsequently on the evaluation of inconsistency for a network with many multi-arm studies.

The loop-specific approach was carried out in software R 2.13.2 [17] using the *ifplot.fun* function, which is available online (in http://www.mtm.uoi.gr/ under ‘How to do an MTM’).

2.2 Design-by-treatment interaction approach

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. To overcome these problems, a different type of inconsistency has been proposed, known as design inconsistency. This refers to the differences in the estimated effect sizes for the same comparison from studies that involve

different sets of treatments. The design-by-treatment interaction model is an extension of the previous approach assessing not only ‘loop inconsistency’ but also ‘design inconsistency’.

Consider a network consisting of treatments in the set $T = \{A, B, C, D, \ldots\}$ including different studies that compare subsets of T named ‘designs’ and denoted by $des = 1, \ldots, Des$. Let T_{des}, with $T_{des} \in T$, define the set of treatments in design des. The dataset includes in total N studies, where each design des is present in n_{des} studies indexed $i = 1, \ldots, n_{des}$.

The network meta-analysis model is defined as a multivariate random-effects meta-analysis. Assume A is an arbitrarily chosen reference treatment and T is some treatment in the set $T_{des} = \{B, C, D \ldots\}$. The observed effect size $y_{des,i,AT}$ of treatment T relative to treatment A of study i with design des is modelled under the consistency assumption as

$$y_{des,i,AT} = \mu_{AT} + \delta_{des,i,AT} + \varepsilon_{des,i,AT} \quad (1)$$

The inconsistency model is an extension of model (1) and is defined as a multivariate random-effects meta-regression with additional covariates for the different designs:

$$y_{des,i,AT} = \mu_{AT} + \delta_{des,i,AT} + IF_{des,AT} + \varepsilon_{des,i,AT} \quad (2)$$

where $IF_{des,AT}$ represents inconsistency in comparison AT for design des, which may correspond with either design or loop inconsistency. As described in detail elsewhere[18,19] not all possible $IF_{des,AT}$ covariates are required, since otherwise the model is overparameterised. For designs that do not include the reference treatment, a data augmentation technique is applied[10]. This is basically imputing data for arm A that contains a very small amount of information, such as 0.01 successes out of 0.1 individuals. The study random errors are normally distributed $\boldsymbol{\varepsilon}_{des,i} \sim N(0, \boldsymbol{S}_i)$, where $\boldsymbol{S}_i$ is the within study variance-covariance matrix.

$$\delta_{des,i} \sim N(0, \boldsymbol{\Sigma})$$

where $\boldsymbol{\Sigma}$ is the between studies variance-covariance matrix involving the heterogeneity variance for each treatment comparison. We discuss the structure of $\boldsymbol{\Sigma}$ in section 2.4.

If a design-by-treatment interaction model has l independent inconsistency parameters, then under the null hypothesis $H_0: \widehat{IF}_1 = \cdots = \widehat{IF}_l = 0$, the joint statistical significance of the l inconsistency parameters is tested by the χ^2-test

$$W = \sum_{j=1}^{l} \frac{\widehat{IF}_j^2}{\hat{\sigma}_j^2} \sim \chi_l^2$$

We estimated inconsistency by fitting model (2) in STATA using the *mvmeta* command[10].

The design-by-treatment interaction approach estimates inconsistency in the entire network, whereas the loop-specific approach evaluates each loop separately. It is therefore impossible to infer about the level of agreement between the two methods. We 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 how the two methods perform.

2.3 Effect measures

We restrict our investigation of inconsistency to dichotomous outcomes. We consider four effect measures; the odds ratio (*OR*), the risk difference (*RD*), the risk ratio of beneficial outcomes (*RRB*) and the risk ratio for harmful outcomes (*RRH*). It has been shown that the choice of the effect measure can impact on the heterogeneity variance[12,13], which subsequently might impact on the estimation of inconsistency.

2.4 Estimation of the heterogeneity

Let us define as τ^2_{XY} the heterogeneity in the *Yvs.X* comparison. We made assumptions about these heterogeneity variances, and we address first the loop-specific approach. Consider the network defined by two triangular loops, *ABC* and *BCD*, informed by *AB*, *AC*, *BC*, *BD* and *CD* comparisons. Heterogeneity might be present in each comparison, and the amount of heterogeneity is estimated either by considering the loop to which the comparison belongs (common within-loop heterogeneity) or by considering the entire network (common within-network heterogeneity). Under the common within-loop heterogeneity (τ^2_{loop}) approach all comparisons in a particular loop have the same amount of heterogeneity; *ABC* loop: $\tau^2_{AB}=\tau^2_{AC}=\tau^2_{BC}=\tau^2_{loop,1}$, *BCD* loop: $\tau^2_{BC}=\tau^2_{BD}=\tau^2_{CD}=\tau^2_{loop,2}$. 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, we set τ^2_{loop} equal to zero. Note that in our analyses, τ^2_{loop} may be different for the same comparison when it is involved in different loops.

In the design-by-treatment interaction model, we assume that all comparisons in the network share the same heterogeneity variance (common within-network heterogeneity), i.e. $\tau^2_{AB}=\tau^2_{AC}=\tau^2_{BC}=\tau^2_{BD}=\tau^2_{CD}=\cdots=\tau^2_{ntw}$. Suppose the total number of treatments included in a network is *p*, the variance-covariance matrix for the random effects is therefore given by

$$\boldsymbol{\Sigma}_{(p-1)\times(p-1)}=\tau^2_{ntw}\begin{pmatrix} 1 & \dots & 1/2 \\ \vdots & \ddots & \vdots \\ 1/2 & \dots & 1 \end{pmatrix}$$

In general, when the number of studies included in a meta-analysis is large, the heterogeneity parameter is more precisely estimated[20]. 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 a more accurate representation of how the effects are being combined in a network meta-analysis.

The heterogeneity variance (τ^2) can be estimated by a variety of methods[21]. The performance of the different estimators can differ in terms of bias and mean squared error (MSE), and they can over- or under-estimate the true heterogeneity variance. As heterogeneity may affect the estimation of inconsistency, we evaluate inconsistency under different estimators of τ^2. We apply the different estimation methods under the *OR* measure. In the loop-based approach we used the DerSimonian and Laird (DL)[21,22], restricted maximum likelihood (REML)[21,23] and Sidik-Jonkman (SJ)[24] methods. We include the DL method because it is frequently used in random-effects meta-analysis and is the default estimator in STATA *metan* command[25] and RevMan[26]. The DL estimator performs well for small values of τ^2, but underestimates the true heterogeneity variance when τ^2 is large or the number of studies is relatively small producing a large negative bias[24,27,28]. The popular REML method is less biased than the DL method (except for small values of τ^2 that the methods are comparable)[11,29] , but underestimates τ^2 when data are sparse[29,30]. The less popular SJ estimator has been shown to overestimate τ^2 when the true heterogeneity variance is relatively small[31]. The SJ method is one of the best methods when the true heterogeneity variance is large producing small bias and substantially smaller than the DL estimator[11,24]. Between the three estimators the DL method is less variable in terms of the MSE in meta-analysis with small to moderate heterogeneity[11].

In the design-by-treatment interaction model only DL, maximum likelihood (ML)[21,32] and REML[21,23] estimators of Σ are available. We apply the ML and REML methods, since the DL method is not appropriate when the augmentation technique is applied[19]. The ML method underestimates τ^2 when the number of studies is small to moderate producing a relatively large amount of negative bias[11,23]. It has been shown that the REML method is less biased with larger MSE than the ML method[11,29].

2.5 Other methods to evaluate inconsistency

Several other methodologies to evaluate consistency have been outlined in the literature (for a review see NICE DSU Technical Support Document 4[33]). The methods can be broadly categorised into methods that contrast direct and indirect evidence for a particular comparison within a network (as the loop-specific approach outlined above) and methods that evaluate inconsistency in a network as a whole (such as the design-by-treatment model). Methods in the former category are useful to locate sources of inconsistency whereas methods in the latter category provide global tests.

One of the drawbacks of the loop-based 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.*[34] 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 loop-based 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 node-splitting method proposed by Dias *et al.*[35]. The node-splitting approach is computationally intensive and to our knowledge has not yet been automated, making it impractical for large networks. 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. Among all the methods, the loop-based approach is, despite its shortcomings, to date the most popular approach to evaluate inconsistency.

When network meta-analyses are fit within a Bayesian framework, investigators often contrast models with and without the consistency constraints with respect to fit and parsimony[36]. 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 design-by-treatment interaction model is the only method that provides an omnibus test, can be fit in a frequentist setting and provides results insensitive to the parameterisation of multi-arm studies[18,19]. Models that do not account for design inconsistency (such as those presented in Lu and Ades[37] and Lumley[38]) are special cases of the design-by-treatment interaction model.

2.6 Searching for network meta-analyses and data extraction

We searched in PubMed for research articles including networks with at least four treatments and dichotomous primary outcomes. We 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. The search code we used was ‘(network OR mixed treatment* OR multiple treatment* OR mixed comparison* OR indirect comparison* OR umbrella OR simultaneous comparison*) AND (meta-analysis)’.

We extracted data regarding the year of publication, the methods applied for the indirect comparison, the number of studies and the number of arms the studies included, as well as the total number of interventions involved in each network. From each network we extracted the trial data for the primary outcome (as stated in the text or, if this was unclear, defined as the first outcome presented). We preferred data presented in 2 × 2 tables rather than as effect sizes and precisions, when both formats were reported. The extracted trial 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.

3 Results

3.1 Database

Eight hundred and seventeen relevant articles were initially identified and after the screening process we ended up with 40 networks. The full process is shown in the flow chart of Figure 1. The 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 loop-based 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 (Appendix Table 1).

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%) did the authors explain why they chose the employed effect measure. The median number of studies per network is 23, ranging from 9 to 111. The number of treatments compared ranged from 4 to 17 with a median of 6. 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 Appendix Table 2.

3.2 Loop-specific approach

3.2.1 Inconsistency under the four effect measures for binary data—Out of the total of 303 loops, 23 were found to be inconsistent (8%) when analysed as *OR*, 26 (9%) as *RRH*, 29 (10%) as *RRB* and 29 (10%) as *RD*, for common within-loop heterogeneity ($\hat{\tau}^2_{loop}$) estimated using the DL method. Table 1 provides these results along with results under the assumption of common within-network heterogeneity ($\hat{\tau}^2_{ntw}$) which we discuss later. When we changed from one effect size to another under $\hat{\tau}^2_{loop}$, some consistent loops became inconsistent and vice versa. Such changes were mostly observed between *OR vs. RD* and *OR vs. RRB*. Eleven (4%) consistent loops under *OR* changed to inconsistent under *RD*, whereas 5 (2%) loops that deviate from consistency under *OR* changed to consistent when *RD* is employed (see Table 1). The percentage of inconsistent loops was comparable across the four effect measures (McNemar test under the within-loop heterogeneity; *OR vs. RRH*: $P = 0.505$, *OR vs. RRB*: $P = 0.239$, *OR vs. RD*: $P = 0.211$). In Appendix Table 3 we provide the inconsistency estimates under the four scales for all loops, along with their standard errors and z-scores.

Our database includes 203 loops with at least one comparison being informed by a single study. Inconsistency was more likely to be found in such loops. For example, in the network of Elliot[39] we identified two inconsistent loops under 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 under *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 effect measures were 18 (9%) versus 8 (8%) (*P* = 0.972) under *RRH*, 21 (10%) versus 8 (8%) (*P*=0.657) under *RRB* and 20 (10%) versus 9 (9%) (*P* =0.977) under *RD*.

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

Comparing the τ^2_{loop} and τ^2_{ntw} approaches we 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 under the common within-loop heterogeneity versus the common within-network heterogeneity; *OR*: *P* = 0.023, *RRH*: *P* = 0.096, *RRB*: *P* = 0.010, *RD*: *P* = 0.004). In Appendix Table 5 we provide the number of *IF* with a 95%CI incompatible with zero under the four effect measures when we assume either τ^2_{loop} or τ^2_{ntw}.

In Figure 2 the *P* values for the loop-specific approach are presented under the common within-loop and the common within-network heterogeneity 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[40,41]. 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* (under the within-network heterogeneity:
$\text{mean}\left(|\tau^2_{RRH}-\tau^2_{OR}|/\tau^2_{OR}\right)=52\%$, $\text{mean}\left(|\tau^2_{RRB}-\tau^2_{OR}|/\tau^2_{OR}\right)=63\%$,
$\text{mean}\left(|\tau^2_{RD}-\tau^2_{OR}|/\tau^2_{OR}\right)=90\%$; under the within-loop heterogeneity:
$\text{mean}\left(|\tau^2_{RRH}-\tau^2_{OR}|/\tau^2_{OR}\right)=51\%$, $\text{mean}\left(|\tau^2_{RRB}-\tau^2_{OR}|/\tau^2_{OR}\right)=79\%$,
$\text{mean}\left(|\tau^2_{RD}-\tau^2_{OR}|/\tau^2_{OR}\right)=97\%$). In general, no substantial differences in inconsistency were observed between the effect measures.

3.2.2 Inconsistency under different estimators for the heterogeneity parameter

—In Table 2 we present the number of inconsistent loops under the three heterogeneity estimators for τ^2_{loop}, as well as under the REML method for τ^2_{ntw}, using the *OR* effect measure. We observed that both DL and REML methods led to a greater number of inconsistent loops than the SJ method. This is because under certain circumstances the first two methods underestimate τ^2 whereas SJ overestimates the true heterogeneity variance. As noted earlier, we observed that inconsistency was more frequent in loops that include comparisons informed by only one study (Table 2). Under the assumption of a common

within-loop heterogeneity, 19 (9%) out of the 203 loops with at least one comparison informed by a single study were found to be inconsistent under DL, whereas only 4 (4%) were inconsistent of the remaining 100 loops (*P*=0.154). The respective percentages under the REML and SJ estimators are 18 (9%) versus 3 (3%) (*P*=0.099) and 12 (6%) versus 2 (2%) (*P*=0.217). However, assuming a common within-network heterogeneity the respective inconsistent loops were 4 (2%) versus 12 (12%) (*P*=0.001) under REML. 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 under 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).

In Figure 3 we compare the estimated heterogeneity variance on the log scale under the DL, REML and SJ methods, showing that the SJ method is associated with larger values of heterogeneity variance, leading to fewer inconsistent loops than the other two methods.

Among the three estimation methods, SJ is less likely to estimate τ^2_{loop} equal to zero (comparison of inconsistent loops when the within-loop heterogeneity is estimated equal to zero; DL *vs.* REML: *P*=0.586, DL *vs.* SJ: *P*=0.062, REML *vs.* SJ: *P*=0.011) (see Table 2).

For each loop, we compared the *IF* and its *P* value with the estimated heterogeneity variance for each loop ($\hat{\tau}^2_{loop}$) under the three estimators (see Appendix Figure 1). We observe that, irrespective of the estimation method used, the magnitude of inconsistency increases slightly as the estimated heterogeneity variance increases. Conversely, lower values of the heterogeneity variance are associated with a greater chance of identifying *IF* with a 95%CI incompatible with zero, though the correlation coefficients between the *P* value or *IF* and the heterogeneity variance are very small (correlation coefficients for $\widehat{IF}$ versus $\hat{\tau}^2$: r_{DL} = 0.14 r_{REML} = 0.15, r_{SJ} = 0.29; correlation coefficients for *P* value of $\widehat{IF}$ versus $\hat{\tau}^2$: r_{DL} = 0.13, r_{REML} = 0.13, r_{SJ} = 0.04).

The median *IF* under the common within-loop heterogeneity ($\hat{\tau}^2_{loop}$) and the DL estimator was 0.34 with an interquartile range (0.15, 0.79). A histogram of the estimated *IF* is given in Figure 4.

3.3 Design-by-treatment interaction approach

On applying the design-by-treatment interaction approach, 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 under 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 under *OR*, *RRH*, *RRB* and *RD*, respectively (all pairwise comparisons between *OR vs.* *RRH* or *RD* for inconsistent networks under the REML estimator using the McNemar test produced $P = 1$, whereas *OR vs.* *RRB* produced $P = 0.617$) (see Appendix Table 6 and Appendix Table 7). Comparing the REML with the ML method, the former yielded fewer inconsistent networks (12% 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 Appendix Table 8). This is probably because the ML method estimated slightly smaller values of the heterogeneity variance than the REML in almost all networks and under all effect sizes.

For fourteen networks (35%) we 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 we applied the design-by-treatment interaction model using the REML method and the *OR* scale. That one in three of the meta-analysis authors did not examine consistency is a cause of concern, since conclusions from combining direct and indirect evidence may not be valid when consistency does not hold.

In Figure 5 we present a plot of the heterogeneity variance estimated under the consistency and inconsistency models considering both ML and REML methods under the *OR* effect measure. On average the consistency models display higher heterogeneity than the inconsistency models, accounting probably for inconsistency in the data.

3.4 Comparing loop-specific and design-by-treatment interaction model

In Table 3 we compare the number of inconsistent networks under the loop-specific approach with τ^2_{ntw} and the design-by-treatment interaction approach when the *OR* is considered, assuming that if at least 5% of the loops are inconsistent then the network is inconsistent. The design-by-treatment interaction approach suggested fewer inconsistent networks (13%) than our ad hoc approach based on loop-specific assessments (20%). One network was inconsistent under the design-by-treatment interaction model while it was consistent with the loop-specific approach. That network was associated with design inconsistency, which was not accounted for in the loop-based method.

4 Discussion

Evaluation of consistency is an important task in network meta-analysis[42]. Protocols of network meta-analysis should ideally describe the methods for such an evaluation and outline the strategy that is to be followed if important inconsistency is detected. In this study

we undertook a large-scale empirical evaluation of the prevalence of inconsistency, focusing both on closed loops of evidence within a network and on entire networks of interventions.

Our study confirms previous assumptions that heterogeneity plays an important role in the statistical detection of inconsistency. We found that lower heterogeneity was associated with higher rates of detected inconsistency, but the estimated magnitude of inconsistency is lesser. This suggests that heterogeneity might account for some disagreement between various sources of evidence. The use of τ^2_{ntw} in the loop-specific approach provides a fair reflection of heterogeneity[43] and decreases the prevalence of inconsistency compared with τ^2_{loop}. We further found that in some cases inconsistency might be reduced when changing the effect measure, but in general the three scales for dichotomous data present the same inconsistency rates. It has been shown that a poor choice of the measurement scale, i.e. analysing data on a 'preferred' scale rather than on the 'best' scale (a scale where the treatment effects can be assumed to be linear), can increase the probability of finding inconsistency[14]. It is advisable 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.

Inconsistency was detected in 2% to 9% of the tested loops, depending on the effect measure and heterogeneity estimation method, and about one eighth of the networks were found to be inconsistent. We regard the two methods used in the paper as complementary methods rather than competing ones. The identification of inconsistency in a network of evidence as a whole using the design-by-treatment interaction approach provides an omnibus test and should lead to a careful examination of all parts of the network. It is advisable to employ methods that can indicate which piece of evidence is responsible for this disagreement (e.g. the 'loop-based method' used here, the 'node-splitting' method[35] or the chi-squared test if possible[34]) alongside the evaluation of the network as a whole[33]. 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[42].

When few studies are included in a loop, the choice of the heterogeneity estimator might impact on inferences about inconsistency. The presence of a comparison informed by a single study was associated with higher prevalence of inconsistency when τ^2_{loop} was employed. This is in line with findings from a recent simulation study[44] and previous empirical evidence[8]. Such cases should prompt further investigation of the comparability of studies in the loop, although the finding might be indicative of data extraction errors. The use of several techniques (e.g. predictive cross-validation) is probably required to decide whether the study is a statistical outlier[45]. Results from statistical tests should however be interpreted with caution: the absence of statistical inconsistency does not provide reassurance that the network meta-analysis results are valid. The assumption of consistency should always be evaluated conceptually by identifying possible effect modifiers that differ across studies[42,46].

In the present study we evaluated articles included in PubMed and we restricted the analysis to dichotomous outcomes. Other network meta-analyses, such as those undertaken in

Europe PMC Funders Author Manuscripts

technology appraisals for the National Institute for Health and Clinical Excellence (NICE) in the UK, are not included. We 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.

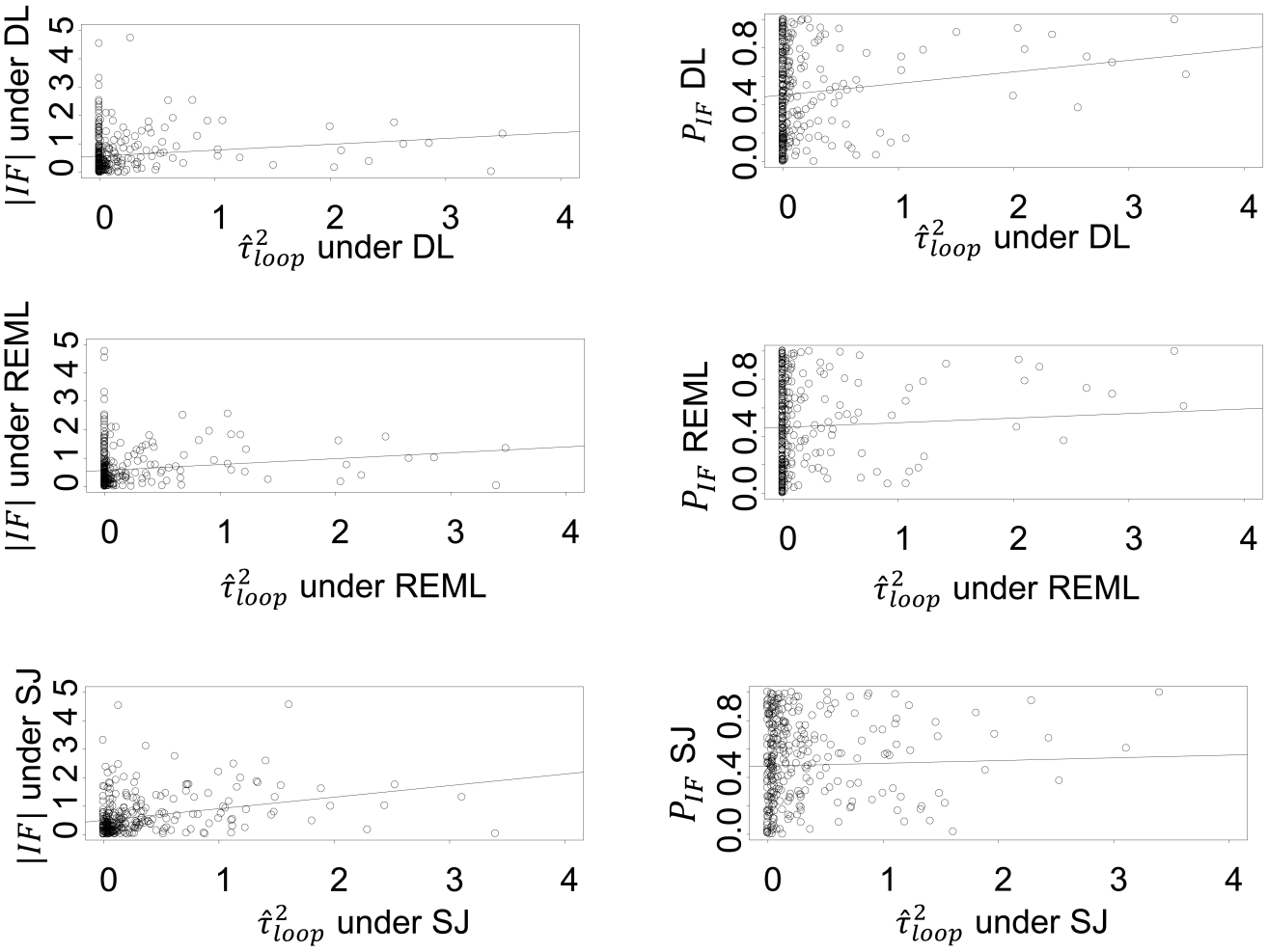
The findings of our study can be used to inform the development of strategies to detect and address statistical inconsistency. Results from methods we examined appear to be sensitive to the estimation method and to assumptions made about heterogeneity. Consequently, caution is needed when over-conservative or over-liberal estimation approaches are employed for the heterogeneity parameter, and often a sensitivity analysis might be necessary. Further empirical evidence is needed to evaluate the performance of other methods to detect inconsistency not included in this article. More importantly, understanding of the power of both approaches under different assumptions regarding the heterogeneity parameter would benefit from an extensive simulation study.

Acknowledgments

This work was supported by the European Research Council (IMMA, Grant nr 260559 to AAV and GS). JPTH was supported by the UK Medical Research Council (unit program number U105285807).

Europe PMC Funders Author Manuscripts

Appendix



Appendix Figure 1.

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

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

id	Network	Assumption of consistency was evaluated	Method to detect inconsistency	Inconsistency reported as present
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

id	Network	Assumption of consistency was evaluated	Method to detect inconsistency	Inconsistency reported as present
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. We consider this to be an inappropriate method to evaluate consistency.

** Inconsistency has been previously assessed[21]

*** Inconsistency has been previously assessed[20]

Europe PMC Funders Author Manuscripts

Europe PMC Funders Author Manuscripts

Appendix Table 2

Characteristics of networks with at least one closed loop included in the database. We define *N* the total number of studies and *T* 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	*N*	*T*	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

Europe PMC Funders Author Manuscripts

id	Network	loops	*N*	*T*	Disease/ Condition	Outcome	Type of Treatments	2arm trials	3arm trials	4arm trials	Indirect Method	Effect Measure used by reviewers
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 solid organ transplant recipients	Mortality	Antifungal agents	10	0	0	Not reported	RR
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

Europe PMC Funders Author Manuscripts

id	Network	loops	*N*	*T*	Disease/ Condition	Outcome	Type of Treatments	2arm trials	3arm trials	4arm trials	Indirect Method	Effect Measure used by reviewers
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

Europe PMC Funders Author Manuscripts

Appendix Table 3

Inconsistency estimates (*IF*) along with their standard error (*SE*(*IF*)) and z-scores under 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 (REML) estimator under the design-by-treatment interaction model. *RD* is the risk difference measure, *RRH* is the risk ratio for harmful outcomes, *RRB* is the risk ratio for beneficial outcomes and *OR* is the odds ratio.

		logOR				*logRRH*				*logRRB*				*RD*			
Network	no. loops	Inconsistent loops	heterogeneity	*IF* (*SE*(*IF*))	z-score (*P* value)	Inconsistent loops	heterogeneity	*IF* (*SE*(*IF*))	z-score (*P* value)	Inconsistent loops	heterogeneity	*IF* (*SE*(*IF*))	z-score (*P* value)	Inconsistent loops	heterogeneity	*IF* (*SE*(*IF*))	z-score (*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)	−2.74 (0.010)	2	0.00	0.02 (0.01)	2.67 (0.010)
												0.02 (0.01)	2.27 (0.020)			0.02 (0.01)	−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)	−2. 49 (0.013)	2	0.00	0.57 (0.28)	2.00 (0.045)	3	0.00	0.38 (0.15)	−2.63 (0.009)	3	0.00	0.18 (0.08)	−2.28 (0.022)
				1.15 (0.51)	−2.27 (0.023)			0.31 (0.15)	2.00 (0.045)			0.58 (0.27)	−2.19 (0.029)			0.29 (0.13)	−2.17 (0.030)
				0.61 (0.24)	−2.51 (0.012)			0.23 (0.11)	−2.19 (0.028)							0.14 (0.06)	−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)				

Europe PMC Funders Author Manuscripts

		logOR				*logRRH*				*logRRB*				*RD*			
Network	**no. loops**	**Inconsistent loops**	**heterogeneity**	***IF (SE(IF))***	**z-score (*P* value)**	**Inconsistent loops**	**heterogeneity**	***IF (SE(IF))***	**z-score (*P* value)**	**Inconsistent loops**	**heterogeneity**	***IF (SE(IF))***	**z-score (*P* value)**	**Inconsistent loops**	**heterogeneity**	***IF (SE(IF))***	**z-score (*P* value)**
				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)	1.97 (0.049)								
								1.38 (0.65)	−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)	2.04 (0.041)	2	0.00	0.02 (0.01)	−1.98 (0.047)
												0.03 (0.01)	2.14 (0.032)			0.03 (0.01)	−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		

		logOR				*logRRH*				*logRRB*				*RD*			
Network	**no. loops**	**Inconsistent loops**	**heterogeneity**	***IF (SE(IF))***	***z*-score (*p* value)**	**Inconsistent loops**	**heterogeneity**	***IF (SE(IF))***	***z*-score (*p* value)**	**Inconsistent loops**	**heterogeneity**	***IF (SE(IF))***	***z*-score (*p* value)**	**Inconsistent loops**	**heterogeneity**	***IF (SE(IF))***	***z*-score (*p* value)**
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		

Europe PMC Funders Author Manuscripts

Appendix Table 4

Inconsistency estimates (*IF*) along with their standard error (*SE*(*IF*)) and z-scores under 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 (DL) estimator under the random-effects model. *RD* is the risk difference measure, *RRH* is the risk ratio for harmful outcomes, *RRB* is the risk ratio for beneficial outcomes and *OR* is the odds ratio.

		logOR				*logRRH*				*logRRB*				*RD*			
Network	**no. loops**	**Inconsistent loops**	**heterogeneity**	***IF* (*SE*(*IF*))**	**z-score (*P* value)**	**Inconsistent loops**	**heterogeneity**	***IF* (*SE*(*IF*))**	**z-score (*P* value)**	**Inconsistent loops**	**heterogeneity**	***IF* (*SE*(*IF*))**	**z-score (*P* value)**	**Inconsistent loops**	**heterogeneity**	***IF* (*SE*(*IF*))**	**z-score (*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)

Europe PMC Funders Author Manuscripts

		logOR				*logRRH*				*logRRB*				*RD*			
Network	**no. loops**	**Inconsistent loops**	**heterogeneity**	***IF (SE(IF))***	**z-score (*P* value)**	**Inconsistent loops**	**heterogeneity**	***IF (SE(IF))***	**z-score (*P* value)**	**Inconsistent loops**	**heterogeneity**	***IF (SE(IF))***	**z-score (*P* value)**	**Inconsistent loops**	**heterogeneity**	***IF (SE(IF))***	**z-score (*P* value)**
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			

Europe PMC Funders Author Manuscripts

		logOR				***logRRH***				***logRRB***				***RD***			
Network	**no. loops**	**Inconsistent loops**	**heterogeneity**	***IF (SE(IF))***	**z-score (*P* value)**	**Inconsistent loops**	**heterogeneity**	***IF (SE(IF))***	**z-score (*P* value)**	**Inconsistent loops**	**heterogeneity**	***IF (SE(IF))***	**z-score (*P* value)**	**Inconsistent loops**	**heterogeneity**	***IF (SE(IF))***	**z-score (*P* value)**
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

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

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

Europe PMC Funders Author Manuscripts

Int J Epidemiol. Author manuscript; available in PMC 2017 May 01.

Appendix Table 6

Results according to Wald test of consistency under the restricted maximum likelihood (REML) and maximum likelihood (ML) estimators when applying all four effect measures. *RD* is the risk difference measure, *RRH* is the risk ratio for harmful outcomes, *RRB* is the risk ratio for beneficial outcomes and *OR* is the odds ratio.

	Design-by-treatment interaction approach							
	OR		*RRH*		*RRB*		*RD*	
Network	**REML Wald test (*P* value)**	**ML Wald test (*P* value)**	**REML Wald test (*P* value)**	**ML Wald test (*P* value)**	**REML Wald test (*P* value)**	**ML Wald test (*P* value)**	**REML Wald test (*P* value)**	**ML Wald test (*P* value)**
Ades[1]	19.52 (<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]	1.76 (0.941)	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]	16.02 (0.191)	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]	8.91 (0.882)	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 (0.340)	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]	5.65 (0.464)	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]	5.77 (0.673)	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]	0.73 (0.695)	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]	32.25 (0.504)	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]	19.62 (0.105)	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]	12.1 (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]	26.84 (0.140)	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]	6.06 (0.195)	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 (0.647)	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]	11.19 (0.083)	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]	12.20 (0.032)	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]	2.17 (0.141)	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)

Europe PMC Funders Author Manuscripts

	Design-by-treatment interaction approach							
	OR		***RRH***		***RRB***		***RD***	
Network	**REML Wald test (*P* value)**	**ML Wald test (*P* value)**	**REML Wald test (*P* value)**	**ML Wald test (*P* value)**	**REML Wald test (*P* value)**	**ML Wald test (*P* value)**	**REML Wald test (*P* value)**	**ML Wald test (*P* value)**
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]	7.45 (0.059)	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]	60.43 (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]	1.52 (0.218)	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]	6.13 (0.525)	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]	4.54 (0.337)	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]	0.20 (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]	12.99 (0.449)	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]	1.94 (0.963)	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]	0.73 (0.393)	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 (0.759)	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]	5.71 (0.574)	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]	5.59 (0.232)	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 7

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

IF under ML								**Percentage out of the total 40 networks**
		RRH		***RRB***		***RD***		
OR		Consistent	Inconsistent	Consistent	Inconsistent	Consistent	Inconsistent	
	Consistent	28	4	28	4	28	4	80%
	Inconsistent	1	7	1	7	1	7	20%
	Percentage out of the total 40 networks	72%	28%	72%	28%	72%	28%	
IF under REML								
		RRH		*RRB*		*RD*		
OR		Consistent	Inconsistent	Consistent	Inconsistent	Consistent	Inconsistent	
	Consistent	33	2	32	3	32	3	87%
	Inconsistent	1	4	1	4	3	2	13%
	Percentage out of the total 40 networks	85%	15%	83%	17%	87%	13%	

Appendix Table 8

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

	IF under ML				Percentage out of the total 40 networks
IF under REML			***OR***		
			Consistent	Inconsistent	
	OR	Consistent	32	3	87%
		Inconsistent	0	5	13%
		Percentage out of the total 40 networks	80%	20%	
			RRH		
	RRH	Consistent	29	5	85%
		Inconsistent	0	6	15%
		Percentage out of	72%	28%	

		the total 40 networks			
			RRB		
	RRB	Consistent	29	4	83%
		Inconsistent	0	7	17%
		Percentage out of the total 40 networks	72%	28%	
			RD		
	RD	Consistent	29	6	87%
		Inconsistent	0	5	13%
		Percentage out of the total 40 networks	72%	28%	

5 Reference List

(1). 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. Jun; 1997 50(6): 683–91. [PubMed: 9250266]

(2). Caldwell DM, Ades AE, Higgins JP. Simultaneous comparison of multiple treatments: combining direct and indirect evidence. BMJ. 2005; 331(7521):897–900. [PubMed: 16223826]

(3). Higgins JP, Whitehead A. Borrowing strength from external trials in a meta-analysis. Stat Med. 1996; 15(24):2733–49. [PubMed: 8981683]

(4). Lu G, Ades AE. Combination of direct and indirect evidence in mixed treatment comparisons. Stat Med. 2004; 23(20):3105–24. [PubMed: 15449338]

(5). Coleman, CI., Phung, OJ., Cappelleri, JC., et al. Use of Mixed Treatment Comparisons in Systematic Reviews [Internet]. Appendix A, Verbatim Quotes From Guidance Documents. 2012. Available from: http://www.ncbi.nlm.nih.gov/books/NBK107337/

(6). Lu GB, Ades AE. Assessing evidence inconsistency in mixed treatment comparisons. J. Amer. Statist. Assoc. 2006; 101(474):447–59.

(7). Song F, Altman DG, Glenny AM, Deeks JJ. Validity of indirect comparison for estimating efficacy of competing interventions: empirical evidence from published meta-analyses. BMJ. 2003; 326(7387):472. [PubMed: 12609941]

(8). 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; 343:d4909. [PubMed: 21846695]

(9). Song, F., Chen, YF., Loke, Y., Eastwood, A., Altman, D. Inconsistency between direct and indirect estimates remains more prevalent than previous observed. 2011. http://www.bmj.com/rapid-response/2011/11/03/inconsistency-between-direct-and-indirect-estimates-remains-more-prevalent

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

(11). Sidik K, Jonkman JN. A comparison of heterogeneity variance estimators in combining results of studies. Stat Med. 2007; 26(9):1964–81. [PubMed: 16955539]

(12). Deeks JJ. Issues in the selection of a summary statistic for meta-analysis of clinical trials with binary outcomes. Stat Med. 2002; 21(11):1575–600. [PubMed: 12111921]

(13). 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; 19(13):1707–28. [PubMed: 10861773]

(14). Caldwell DM, Welton NJ, Dias S, Ades AE. Selecting the best scale for measuring treatment effect in a network meta-analysis: a case study in childhood nocturnal enuresis. Research Synthesis Methods. 2012; 3:126–41. [PubMed: 26062086]

Europe PMC Funders Author Manuscripts

(15). Salanti G, Higgins JP, Ades AE, Ioannidis JP. Evaluation of networks of randomized trials. Stat Methods Med Res. 2008; 17(3):279–301. [PubMed: 17925316]

(16). Salanti G, Marinho V, Higgins JP. A case study of multiple-treatments meta-analysis demonstrates that covariates should be considered. J Clin Epidemiol. 2009; 62(8):857–64. [PubMed: 19157778]

(17). R: A language and environment for statistical computing. R Foundation for Statistical Computing; Vienna, Austria: 2011. ISBN 3-900051-07-0, URL http://www.R-project.org

(18). Higgins JPT, Jackson D, Barret JK, Lu G, Ades AE, White IR. Consistency and inconstency in network meta-analysis: concepts and models for multi-arm studies. Research Synthesis Methods. 2012; 3(2):98–110. [PubMed: 26062084]

(19). White IR, Barret JK, Jackson D, Higgins JPT. Consistency and inconsistency in multiple treatments meta-analysis: model estimation using multivariate meta-regression. Research Synthesis Methods. 2012; 3(2):111–25. [PubMed: 26062085]

(20). Borenstein, M., Hedges, LV., Higgins, JPT., Rothstein, HR. Introduction to Meta-analysis. 1st edition. John Wiley&Sons; Chichester, UK: 2009.

(21). Raudenbush, SW. Analyzing Effect Sizes: Random Effects Models. In: Cooper, H.Hedges, LV., Valentine, JC., editors. The Handbook of Research Synthesis and Meta-Analysis. 2nd edition. Russell Sage Foundation; New York: 2009. p. 295-315.

(22). DerSimonian R, Laird N. Meta-analysis in clinical trials. Control Clin Trials. 1986; 7(3):177–88. [PubMed: 3802833]

(23). Viechtbauer W. Bias and Efficiency of Meta-Analytic Variance Estimators in the Random-Effects Model. Journal of Educational and Behavioral Statistics. 2005; 30(3):261–93.

(24). Sidik K, Jonkman JN. Simple heterogeneity variance estimation for meta-analysis. Journal of the Royal Statistical Society Series C (Applied Statistics). 2005; 54:367–84.

(25). Harris R, Bradburn M, Deeks J, Harbord R, Altman D, Sterne J. metan: fixed- and random-effects meta-analysis. Stata Journal. 2008; 8(1):3–28.

(26). The Cochrane Collaboration.Review Manager (RevMan). Version 5.1. Copenhagen: 2011. Available at http://ims.cochrane.org

(27). Sanchez-Meca J, Marin-Martinez F. Confidence intervals for the overall effect size in random-effects meta-analysis. Psychol Methods. 2008; 13(1):31–48. [PubMed: 18331152]

(28). Sidik K, Jonkman JN. Robust variance estimation for random effects meta-analysis. Computational Statistics & Data Analysis. 2006; 50(12):3681–701.

(29). Thompson SG, Sharp SJ. Explaining heterogeneity in meta-analysis: a comparison of methods. Stat Med. 1999; 18(20):2693–708. [PubMed: 10521860]

(30). Berkey CS, Hoaglin DC, Mosteller F, Colditz GA. A random-effects regression model for meta-analysis. Stat Med. 1995; 14(4):395–411. [PubMed: 7746979]

(31). Sidik K, Jonkman JN. A note on variance estimation in random effects meta-regression. J Biopharm Stat. 2005; 15(5):823–38. [PubMed: 16078388]

(32). Hardy RJ, Thompson SG. A likelihood approach to meta-analysis with random effects. Stat Med. 1996; 15(6):619–29. [PubMed: 8731004]

(33). Dias, S., Welton, NJ., Sutton, AJ., Ades, AE. NICE DSU Technical Support Document 4: inconsistency in networks of evidence based on randomised controlled trials. NICE Decision Support Unit; 2011. Technical Support Document series No. 4available from http://www.nicedsu.org.ukTechnical Support Document

(34). Caldwell DM, Welton NJ, Ades AE. Mixed treatment comparison analysis provides internally coherent treatment effect estimates based on overviews of reviews and can reveal inconsistency. J Clin Epidemiol. 2010; 63(8):875–82. [PubMed: 20080027]

(35). Dias S, Welton NJ, Caldwell DM, Ades AE. Checking consistency in mixed treatment comparison meta-analysis. Stat Med. 2010; 29(7-8):932–44. [PubMed: 20213715]

(36). Spiegelhalter DJ, Best NG, Carlin BP, Van der Linde A. Bayesian measures of model complexity and fit. Journal of the Royal Statistical Society. 2002; 64(4):583–639.

(37). Lu G, Ades AE. Assessing Evidence Inconsistency in Mixed Treatment Comparisons. Journal of American Statistical Association. 2006; 101(474):447–59.

(38). Lumley T. Network meta-analysis for indirect treatment comparisons. Stat Med. 2002; 21(16): 2313–24. [PubMed: 12210616]

(39). Elliott WJ, Meyer PM. Incident diabetes in clinical trials of antihypertensive drugs: a network meta-analysis. Lancet. 2007; 369(9557):201–7. [PubMed: 17240286]

(40). 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; 19(13):1707–28. [PubMed: 10861773]

(41). Friedrich JO, Adhikari NK, Beyene J. Ratio of means for analyzing continuous outcomes in meta-analysis performed as well as mean difference methods. J Clin Epidemiol. 64(5):556–64. [PubMed: 21447428]

(42). 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. Research Synthesis Methods. 2012; 3(2):80–97. [PubMed: 26062083]

(43). Lu G, Ades A. Modeling between-trial variance structure in mixed treatment comparisons. Biostatistics. 2009; 10(4):792–805. [PubMed: 19687150]

(44). Mills EJ, Ghement I, O'Regan C, Thorlund K. Estimating the power of indirect comparisons: a simulation study. PLoS One. 2011; 6(1):e16237. [PubMed: 21283698]

(45). Madan J, Stevenson MD, Cooper KL, Ades AE, Whyte S, Akehurst R. Consistency between direct and indirect trial evidence: is direct evidence always more reliable? Value Health. 14(6): 953–60.

(46). Song F, Loke YK, Walsh T, Glenny AM, Eastwood AJ, Altman DG. Methodological problems in the use of indirect comparisons for evaluating healthcare interventions: survey of published systematic reviews. BMJ. 2009; 338:b1147. [PubMed: 19346285]

1 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. [PubMed: 20825617]
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. [PubMed: 19637942]
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. [PubMed: 15738743]
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. [PubMed: 21123111]
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. [PubMed: 19657180]
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. [PubMed: 21142838]
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. [PubMed: 9250266]

Europe PMC Funders Author Manuscripts

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. [PubMed: 19185342]
11. Dias S, Welton NJ, Caldwell DM, et al. Checking consistency in mixed treatment comparison meta-analysis. Stat Med. 2010; 29:932–44. [PubMed: 20213715]
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. [PubMed: 18625984]
13. Elliott WJ, Meyer PM. Incident diabetes in clinical trials of antihypertensive drugs: a network meta-analysis. Lancet. 2007; 369:201–7. [PubMed: 17240286]
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. [PubMed: 20191599]
15. Hofmeyr GJ, Gulmezoglu AM, 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. [PubMed: 17932160]
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. [PubMed: 19687150]
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. [PubMed: 17285571]
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. [PubMed: 16235370]
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. [PubMed: 20713583]
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. [PubMed: 19761618]
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. [PubMed: 17478472]
26. Picard P, Tramer MR. Prevention of pain on injection with propofol: a quantitative systematic review. Anesth Analg. 2000; 90:963–9. [PubMed: 10735808]
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. [PubMed: 15266524]
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. [PubMed: 9042847]
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. [PubMed: 19144173]

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. [PubMed: 21199320]
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. [PubMed: 20967400]
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. [PubMed: 21097801]
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. [PubMed: 21059964]
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. [PubMed: 18349026]
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. [PubMed: 19286090]
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. [PubMed: 20887240]
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. [PubMed: 20638155]
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. [PubMed: 19258485]
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. [PubMed: 20600036]
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. [PubMed: 16960269]

Europe PMC Funders Author Manuscripts

Key messages

- A challenge in network meta-analysis is that there may be inconsistency between direct and indirect evidence for a particular treatment comparison.
- Based on empirical examination of a large sample of published network meta-analyses, inconsistency occurs in 2%-9% of triangular and quadrilateral loops of evidence about three and four treatments and in one in eight networks of multiple treatments.
- The choice of the heterogeneity estimation method will impact to a small extent on the detection and estimation of inconsistency.
- Lower statistical heterogeneity is associated with more chances to detect inconsistency but the estimated magnitude of inconsistency is lower.
- Evidence loops that include comparisons informed by a single study are more likely to show inconsistency.

Europe PMC Funders Author Manuscripts

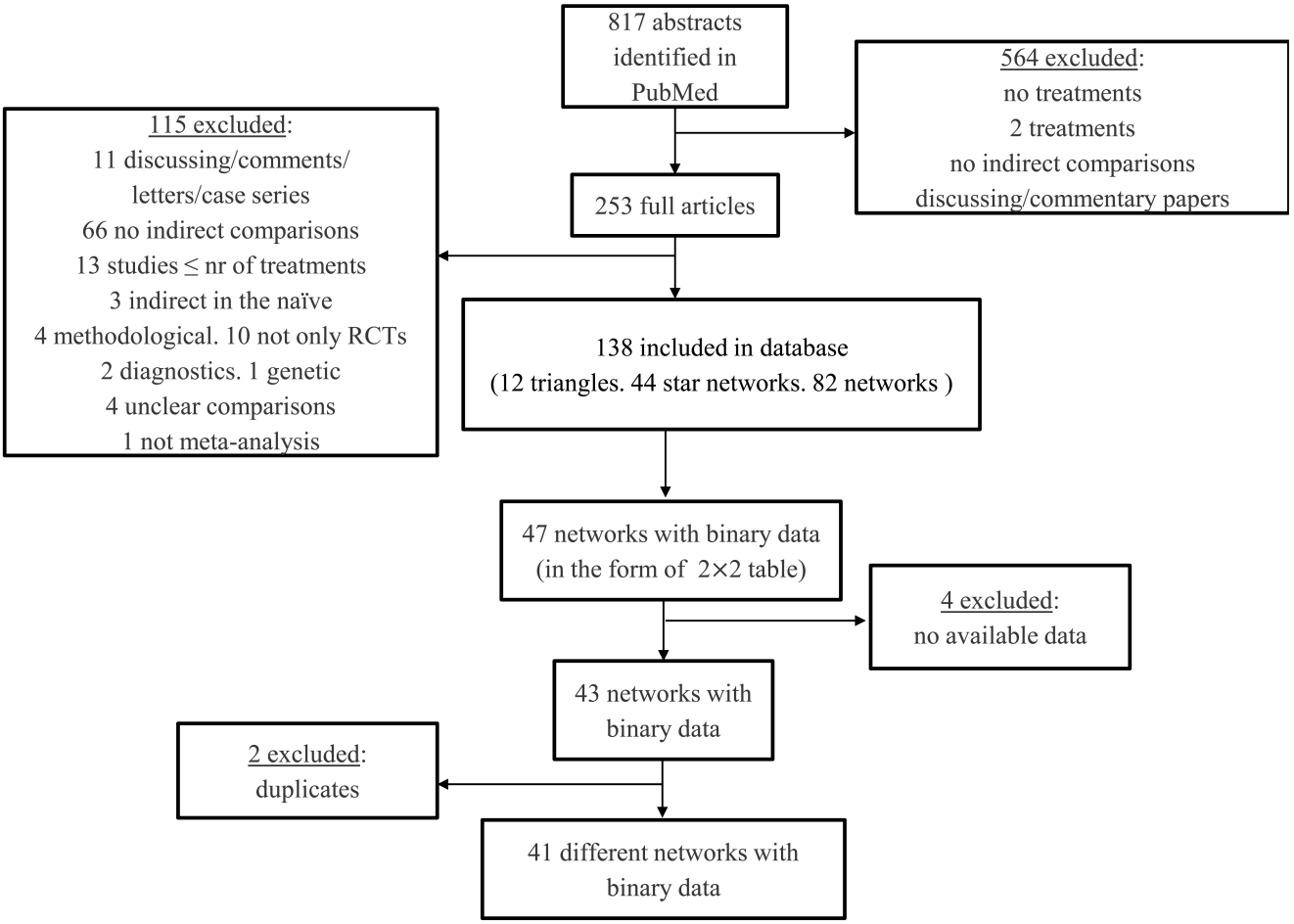


Figure 1.
Flow chart of the process of selecting articles describing network analyses.

Europe PMC Funders Author Manuscripts

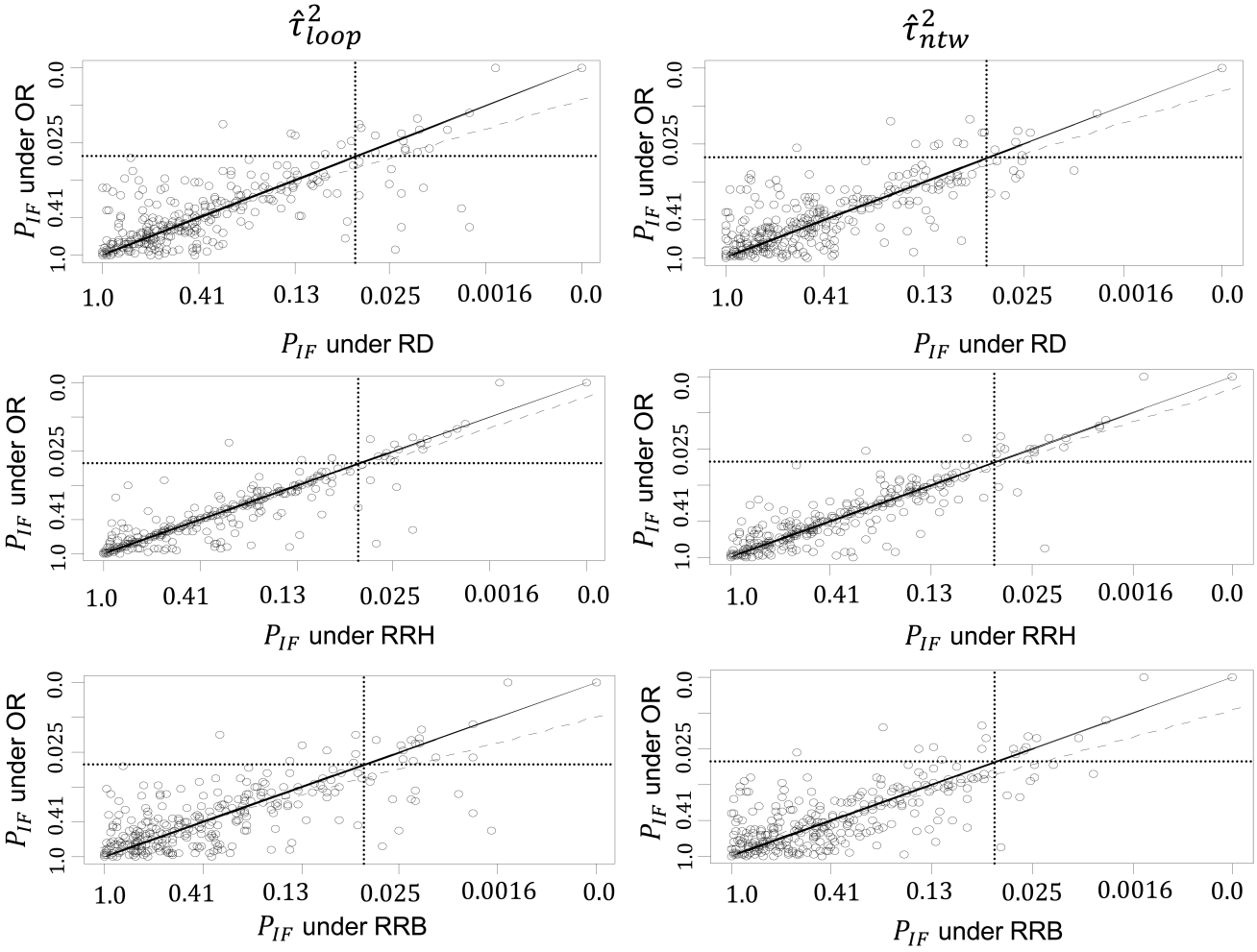


Figure 2.

Plot of the two sided *P* values of *IF* (fourth-root scale) for *OR* vs. *RD*, *OR* vs. *RRH* and *OR* vs. *RRB* effect measures under the DerSimonian and Laird method for τ^2_{loop} and the restricted maximum likelihood for τ^2_{ntw}. The solid diagonal line indicates equality, the dashed diagonal line is the regression line and the two dotted horizontal and vertical lines represent the *P*=0.05 threshold lines. *RD* is the risk difference measure, *RRH* is the risk ratio for harmful outcomes, *RRB* is the risk ratio for beneficial outcomes and *OR* is the odds ratio.

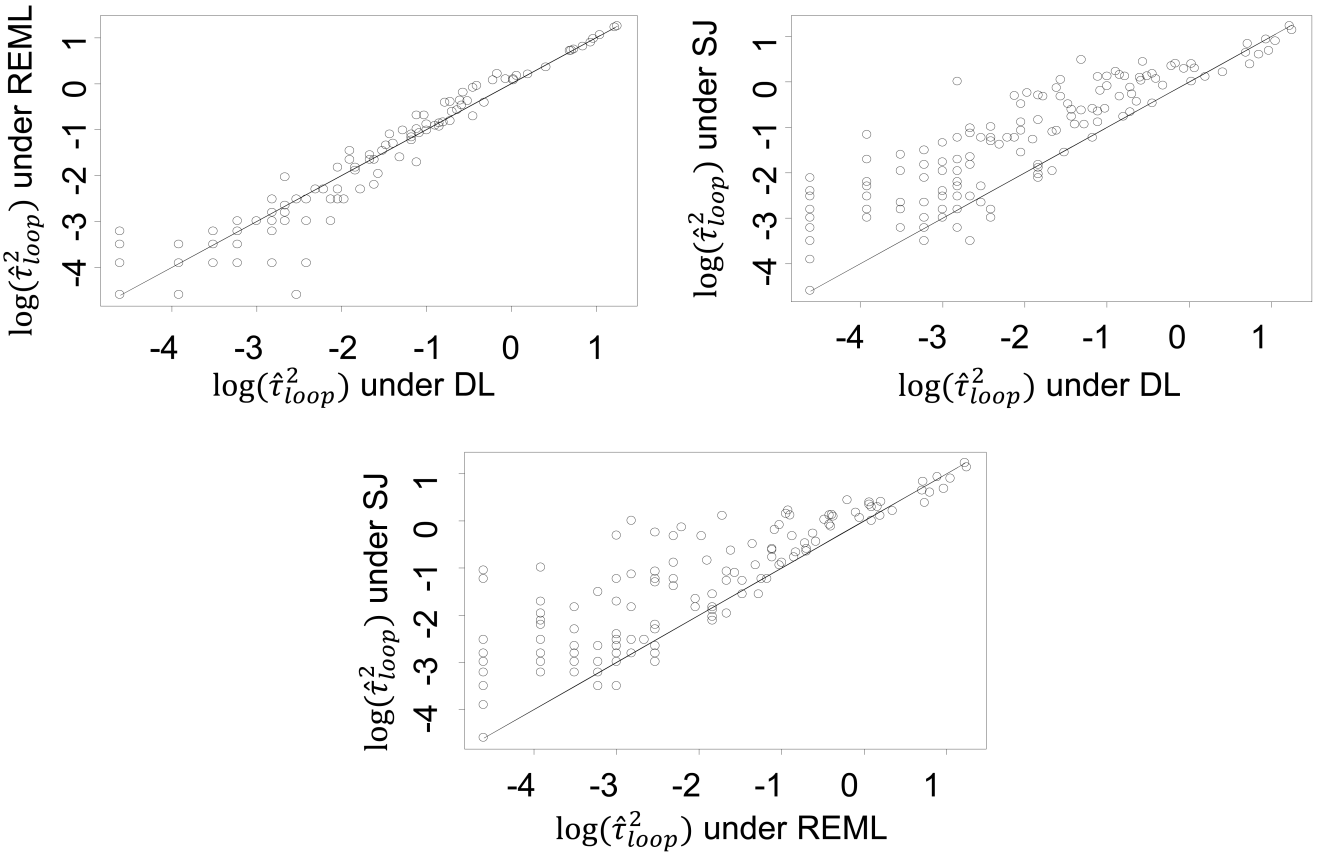


Figure 3.

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

Europe PMC Funders Author Manuscripts

Europe PMC Funders Author Manuscripts

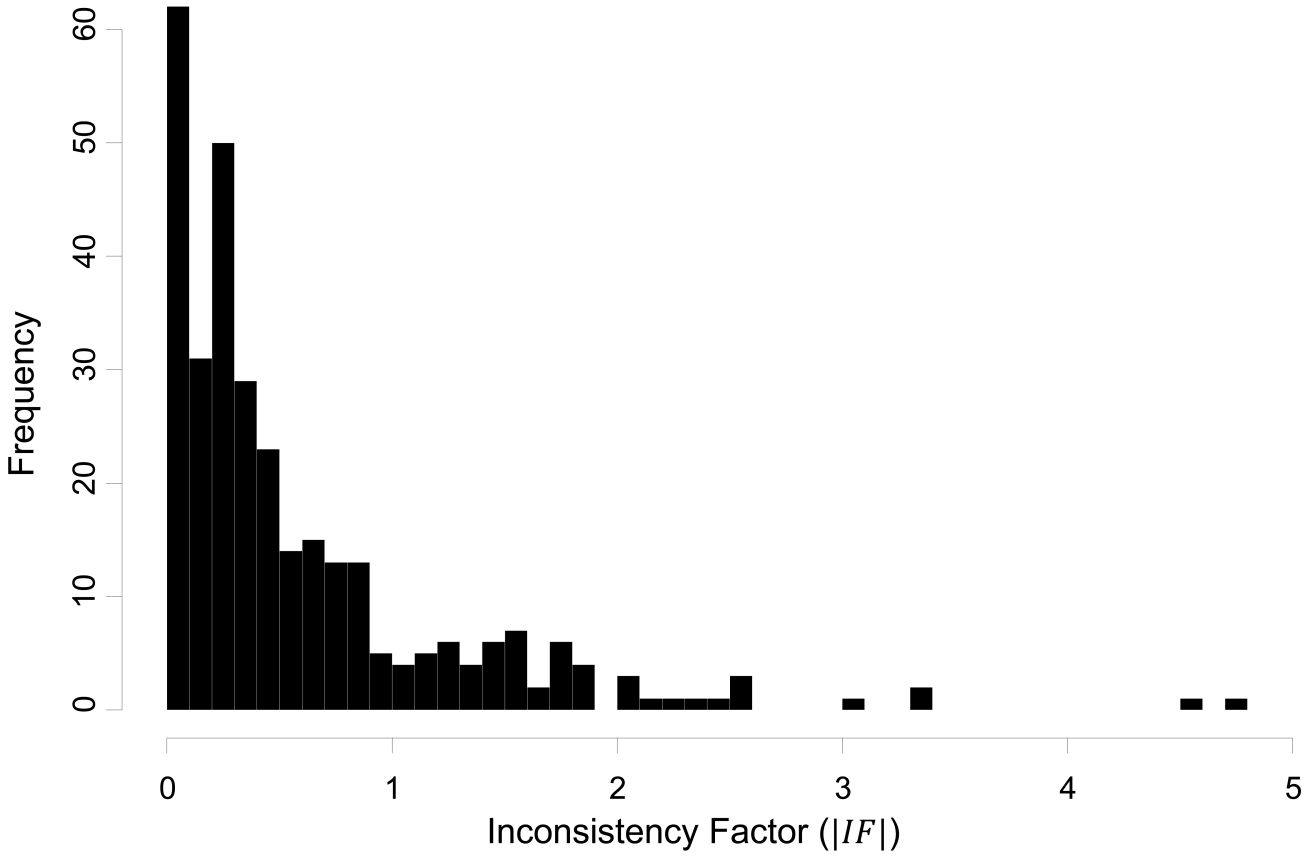


Figure 4.

Histogram of the absolute values of the inconsistency factors (*IF*) for the *OR* effect measure estimated under the common within-loop heterogeneity variance, τ^2_{loop}, estimated with the DerSimonian and Laird method.

Europe PMC Funders Author Manuscripts

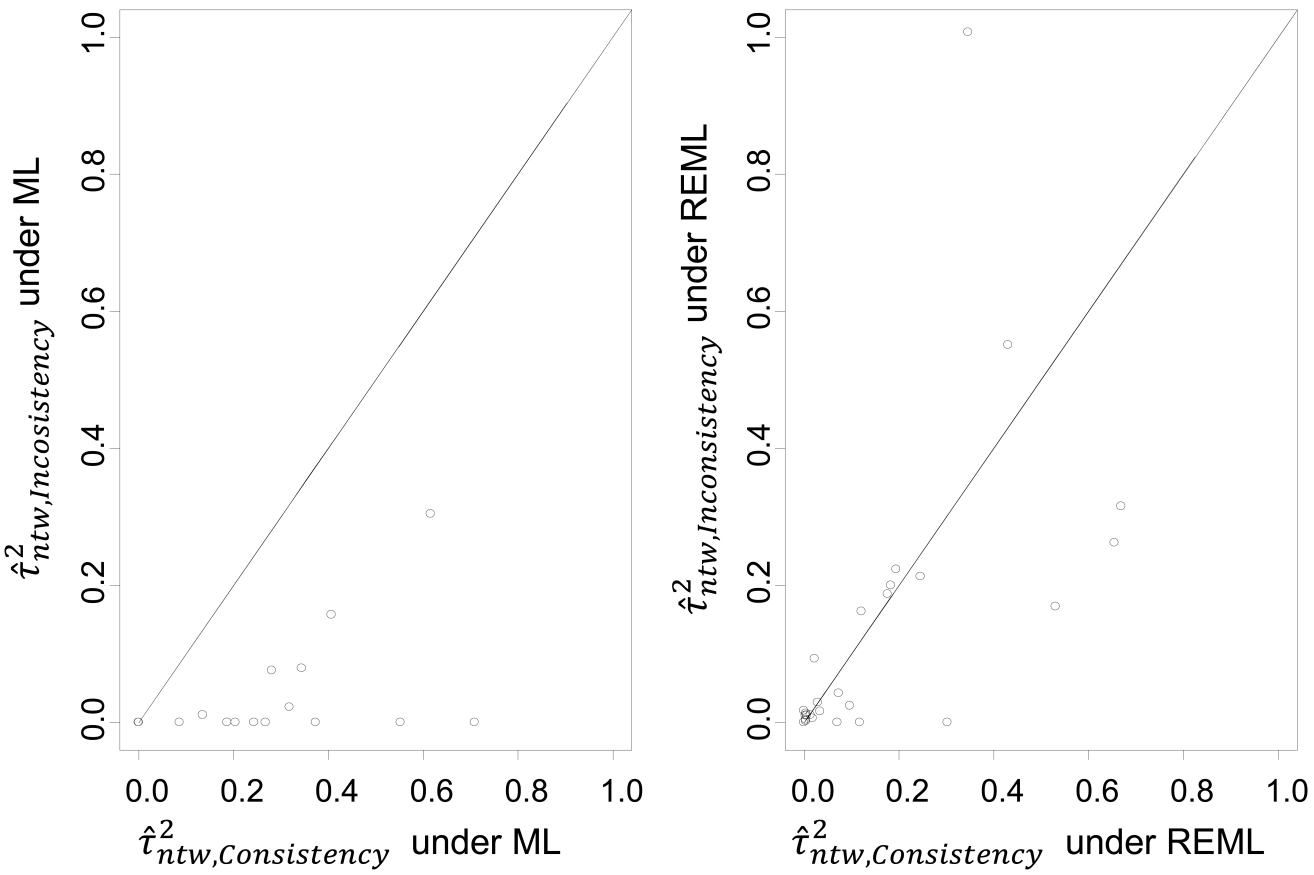


Figure 5.

Plot of heterogeneity estimates from the consistency model against heterogeneity estimates from the inconsistency model under the design-by-treatment interaction approach, along with the equality line. Heterogeneity is estimated under maximum likelihood (1st panel) and restricted maximum likelihood (2nd panel) methods when the effect measure is the odds ratio (*OR*).

Europe PMC Funders Author Manuscripts

Table 1

Number of consistent loops (C) that become inconsistent (I) when changing from one effect size to another and vice versa, assuming both common within-loop heterogeneity $\left(\hat{\tau}^2_{loop}\right)$ estimated under the DerSimonian and Laird method and network heterogeneity $\left(\hat{\tau}^2_{ntw}\right)$ estimated under the restricted maximum likelihood method. *RD* is the risk difference, *RRH* the risk ratio for harmful outcomes, *RRB* the risk ratio for beneficial outcomes and *OR* the odds ratio.

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

Table 2

Frequency of Inconsistent loops under the DerSimonian and Laird (DL), restricted maximum likelihood (REML) and Sidik-Jonkman (SJ) estimators for the heterogeneity variance. Inconsistency is estimated under 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 inconsistent loops 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	Inconsistent loops	Inconsistent loops with $\hat{\tau}^2_{loop} = 0$	Inconsistent loops including 1 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

Table 3

Number of consistent networks that become inconsistent under the loop-specific and design-by-treatment interaction approach when the effect measure is the odds ratio. The common within-network heterogeneity ($\hat{\tau}^2_{ntw}$) is estimated with the restricted maximum likelihood method. Under the loop-specific approach the networks that involve at least 5% inconsistent loops out of their total loops are considered as inconsistent. We define as 'C' the consistent networks and as 'I' the inconsistent networks.

	Loop-specific approach - $\hat{\tau}^2_{ntw}$			**Percentage out of the total 40 networks**
Design-by-treatment interaction approach- $\hat{\tau}^2_{ntw}$		C	I	
	C	30	4	85%
	I	2	4	15%
	Percentage out of the total 40 networks	80%	20%	