WORKING PAPER

EVALUATION OF CONFIDENCE IN THE RESULTS OF NETWORK META-ANALYSIS

IMPRECISION, HETEROGENEITY AND INCOHERENCE

INTRODUCTION

Confidence In Network Meta-Analysis (CINeMA) is a framework to evaluate confidence in the treatment effects produced by network meta-analysis (NMA). A former description of the framework has been published (1) and sets the ground for its refinement, presented in this paper series. The previous paper in the series dealt with the domains of within-study bias, across-studies bias and indirectness, and described how the contribution of each study to each NMA treatment effect can be quantified and combined with study-level judgments (2).

Gain in precision is one of the advantages of NMA compared to pairwise meta-analysis (3). It results from the fact that indirect, on the top of direct, evidence is contributing to the estimation of NMA treatment effects and leads to narrower confidence intervals than their pairwise meta-analysis counterparts. In the GRADE system for pairwise meta-analysis, inconsistency refers to the variability across studies for a particular comparison (4); this variability reflects genuine differences among studies and is alternatively called heterogeneity. In NMA, dispersion in the relative treatment effects might appear either between studies within a comparison (heterogeneity) or between direct and indirect sources of evidence across comparisons (incoherence) (5–8). The two notions are different but closely related and incoherence can be seen as a special form of heterogeneity. In CINeMA we consider two separate domains for ***heterogeneity*** (variability between studies within each comparison) and ***incoherence*** (variability between direct and indirect evidence).

In this paper, we present ways for evaluating NMA treatment effects with respect to imprecision, heterogeneity and incoherence. The methods are exemplified using two examples; a network of trials comparing 18 antidepressants (9) introduced in (2) and a network of statins, presented in Box 1. All analyses have been undertaken in R software using the netmeta package and in the CINeMA web application *cinema.ispm.ch* (10,11).

IMPRECISION

A core aspect in the evaluation of imprecision is the definition of a range of relative treatment effects around the line of no effect that do not signify important clinical differences between the interventions. In the simplest case, this range would constitute

only by the point of no effect (0 in an additive scale, 1 in a ratio scale) if it is considered that even a small difference would be important enough to prefer one treatment over the other. If this is not the case, a larger 'range of equivalence' should be defined that will divide relative treatment effects into ranges 'in favor of A', 'not important differences between A and B', 'in favor of B'. Obviously, this division would not always be so clear and distinct but its comprehensive definition and easy application compensates more rigorous alternatives e.g. applying a spectrum of limits that would result to a continuum of conclusions. A range of equivalence can be symmetrical –a clinically important value is defined and its reciprocal constitutes the clinically important value for the opposite direction- or not –definition of clinically important values differs below and above the line of no effect-. For simplicity, we will assume throughout symmetrical ranges of equivalence although this will not be necessarily the case in NMA applications.

NMA treatment effects are estimated with an associated uncertainty, expressed in the 95% confidence interval, which gives an indication of where the true effect is likely to be placed. The greatest the –direct and indirect- information for a given comparison, the more narrow the confidence interval of the NMA treatment effect will be. The importance, however, of an imprecise treatment effect –expressed as a wide confidence interval- depends on its potential to lead to multiple decisions on the preferable treatment. Thus, the main driver for judging imprecision is considering whether the range of treatment effects included in the confidence interval would lead to a multitude of different clinical actions or treatment recommendations.

Consider for example the network of statins (figure 1) (12). Let us assume that odds ratios (ORs) greater than 1.05 (and accordingly less than $\frac{1}{1.05} = 0.95$) would lead to a recommendation over one of the two treatments. ORs between 0.95 and 1.05 would translate to an interpretation that no important differences in the safety profile of the two statins occur. The 95% confidence interval of pravastatin versus rosuvastatin is quite wide, including ORs from 1.09 to 1.82 (figure 2a). However, any treatment effect in this range would lead to the conclusion that pravastatin is safer than rosuvastatin. Thus, this apparent imprecision does not have important implications on the confidence to be placed on pravastatin versus rosuvastatin. The 95% confidence interval of pravastatin versus simvastatin is slightly wider (0.84, 1.42), but most importantly, its placement covers three

different areas, favoring pravastatin, favoring simvastatin and concluding that it is likely that they are equally safe. Less uncertain, but still not clear with respect to clinical decisions, is the rosuvastatin versus simvastatin comparison. Most treatment effects in its 95% confidence interval support simvastatin but the range of equivalence is also crossed. No, major and some concerns could characterize the evaluation of imprecision for these three NMA relative treatment effects with respect to imprecision.

In the network of antidepressants, Cipriani et al. defined as clinically important effect an OR of 0.8 and its reciprocal 1.25 (9). The range of equivalence (0.8, 1.25) is going to be adopted throughout our illustration of judging NMA treatment effects assuming that ORs lower than 0.8 and larger than 1.25 are considered to be clinically important. We will concentrate on three NMA ORs, clomipramine versus fluvoxamine, citalopram versus venlafaxine and amitriptyline versus paroxetine (table 1).

The 95% confidence interval of clomipramine versus fluvoxamine (0.75, 1.32) extends into clinically important effects in both directions implying uncertainty in clinical decisions. The lower confidence limit (0.75), the mean OR (0.99) and the upper confidence limit (1.32) are associated with three different decisions. Citalopram versus venlafaxine NMA OR is estimated at 1.12 with 95% confidence interval (0.90, 1.39) favoring venlafaxine. Values lower than 0.8 are not included in the confidence interval, which, however, includes values within the range of equivalence and extends into clinically important effects in favor of venlafaxine. The NMA OR of amitriptyline versus paroxetine is 0.96 in favor of amitriptyline. Despite the fact that it is close to 1, it is not associated with imprecision as its 95% confidence interval (0.82, 1.13) lies entirely within the range of equivalence.

HETEROGENEITY

There are several ways that heterogeneity can be measured. The estimation of τ^2, which represents the variance of the distribution of random effects, is a useful measure of the magnitude of heterogeneity. One can estimate heterogeneity variances from each pairwise meta-analysis and, under the usual assumption of a single heterogeneity variance across comparisons, a common heterogeneity variance for the whole network. The magnitude of τ^2 is usefully expressed using a prediction interval, that shows where the true effect of a new study is expected to lie (13).

Similarly to imprecision, the major driver for judging NMA treatment effects with respect to heterogeneity is whether it impacts on clinical decisions. Large variability in the included studies does not necessarily imply important or even any differences in clinical decisions while even small amounts of heterogeneity may in certain cases be important on how convincing the NMA treatment effect is. Compatibility between confidence and prediction intervals with respect to the range of equivalence can be used to capture the importance of heterogeneity and thus inform the evaluation of heterogeneity. For instance, a prediction interval may include values that would lead to different decisions than the decision suggested by the confidence interval; in such a case, heterogeneity may have important implications on the treatment effect and respectively on the associated confidence to place on it.

If a sizeable number of studies is not available, heterogeneity would not be adequately estimated and such an inadequate estimation of heterogeneity would also impact on the interpretation of prediction intervals. In such cases, informing heterogeneity through empirical distributions could be helpful. Turner et al. and Rhodes et al. analyzed 14886 and 6492 meta-analyses of binary and continuous outcomes respectively, categorized them according to the outcome and intervention comparison type and derived the –empirical- distributions of the heterogeneity values (14,15). The same empirical distributions can be used to aid the interpretation of the magnitude of heterogeneity, which can be considered complimentary to considerations based on prediction intervals.

In the example of statins (Figure 1), we have already assumed that the range of equivalence is (0.95, 1.05). The prediction interval of pravastatin versus simvastatin is wide (figure 2b). However, the confidence interval for the particular comparison already extended into clinically important effects in both directions; thus, the implications of heterogeneity is not particularly important as it does not change the decision to be made. The confidence interval of pravastatin versus rosuvastatin lies entirely above the equivalence range and was consequently previously considered as sufficiently precise. However, the prediction interval does cross both boundaries (0.95 and 1.05) extending to treatment effects that would imply different clinical decisions; in such a case we would claim placing major concerns on the impact of heterogeneity.

In the network of antidepressants, the prediction interval of clomipramine versus fluvoxamine does not add further uncertainty in clinical decisions than the uncertainty

already represented by the confidence interval. In conjunction with the small estimated value of heterogeneity for NMA ($\tau^2 = 0.03$) - only two studies compare directly clomipramine fluvoxamine and thus heterogeneity from pairwise meta-analysis cannot be adequately estimated- no concerns for heterogeneity apply for that comparison. The prediction interval of citalopram versus venlafaxine does cross the lower limit of the range of equivalence (0.74, 1.70) suggesting potential implications of heterogeneity. Again the availability of only two direct studies does not allow the estimation of heterogeneity from pairwise meta-analysis and its comparison to the respective empirical distribution. Heterogeneity could have some implications on our confidence on the comparison amitriptyline versus paroxetine as the prediction interval includes clinically important effects both in favor of amitriptyline and in favor of paroxetine (0.65, 1.42). The fact that the estimated heterogeneity from pairwise meta-analysis ($\tau^2 = 0.099$, estimated from 13 studies) is close to the median of the respective empirical predictive distribution (0.096, (14)) slightly mitigates the concerns regarding heterogeneity.

INCOHERENCE

Incoherence is in principle the statistical manifestation of intransitivity; when transitivity holds, it is expected that direct and indirect evidence will be in agreement (16,17). While transitivity is an untestable assumption, incoherence can be measured and tested. Consider, for example, the atorvastatin versus fluvastatin comparison in the example of statins. There are two studies directly comparing the two treatments resulting to a direct OR of 2 (1.05 to 3.79) favoring atorvastatin. The synthesis of the rest 99 studies that provide indirect evidence to the particular comparison gives an indirect OR of 0.89 (0.60, 1.33) favoring fluvastatin. The apparent disagreement between direct and indirect ORs can be quantified in an 'inconsistency factor' measured as the ratio of the two ORs (2.24) along with an approximate 95% confidence interval (1.05, 4.76). This approach can be applied in each comparison with mixed evidence and is called SIDE (Separate Indirect from Direct Evidence) splitting (originally termed node splitting) (5).

More generally, two sets of methods for testing incoherence exist; the first includes methods that examine the agreement between direct and indirect evidence in separate pieces of evidence while the second includes methods that examine incoherence in the

entire network. In the first set of methods where SIDE splitting belongs, called local methods, a number of inconsistency factors are estimated. The definitions of the inconsistency factors view incoherence as the disagreement between direct and indirect evidence either in each closed loop of evidence (loop-specific approach (18)) or in each pairwise comparison (SIDE splitting approach (5)). Global methods for incoherence include modeling simultaneously treatment effects and inconsistency factors and provide an omnibus test for inferring for incoherence globally in the network. The design by treatment interaction test is such a global method for incoherence that assesses the assumption that coherence holds for the entire network (7,8). In the network of statins, it results to a p-value of 0.76 that does not lead into rejecting the hypothesis of coherence. An overview of methods for testing incoherence can be found in (6,19).

We recommend the application of both local and global methods for assessing incoherence. However, incoherence tests are known for having low power and being interweaved with heterogeneity (20,21). Magnitude of inconsistency factors as well as their uncertainty and their potential implications are very important and need to be taken into account. For instance, although SIDE splitting test is not statistically significant, the direct OR of atorvastatin versus lovastatin is 80% larger than the indirect OR (ratio of ORs 1.80; 95% confidence interval (0.90, 3.57)). Despite the insignificant test, considering the fact that direct OR can be up to 3.57 times the indirect OR may lead placing some concerns on the NMA treatment effect with respect to incoherence.

The implication of incoherence to clinical decisions may be considered as an alternative or complimentary way for judging incoherence. Such a consideration may be aided by visual inspection of direct and indirect ORs with respect to the range of equivalence. Consider for instance the hypothetical examples in figure 3. The inconsistency factor between direct and indirect OR is exactly the same for the three examples, but their relative position affect the potential implications of incoherence. In the first case both direct and indirect ORs lie above the range of equivalence suggesting that A is favorable. It would be reasonable having no concerns with respect to incoherence in such a situation. In the second example, indirect OR straddles the range of equivalence with direct OR lying entirely above 1.05 suggesting potential implications of incoherence that could lead someone judge that some concerns are required. In the third example, the same relative disagreement between direct and indirect evidence may have important implications on the

interpretation of the NMA relative treatment effect, as direct and indirect OR point in different directions. This consideration would result into assigning major concerns with respect to incoherence. As a large inconsistency factor may be indicative of a biased direct or indirect estimate, judging its magnitude is always important on the top of evaluating potential implications of incoherence.

Note that in the three hypothetical examples above, both direct and indirect estimates exist. It could be, however, that only direct (e.g. venlafaxine versus vortioxetine in the network of antidepressants) or only indirect (e.g. agomelative versus vortioxetine in the network of antidepressants) exist. In such a case, we can neither estimate an inconsistency factor nor judge potential implications of incoherence with respect to their placement against the range of equivalence. Considerations on indirectness and intransitivity become even more important for this type of comparisons; statistically they can only be judged using the global design by treatment interaction test. In large networks, comparison of the extent of evidence of incoherence with the results from empirical studies (22,23) could also be useful.

In the network of antidepressants, direct OR is almost double the indirect OR of clomipramine versus fluvoxamine (ratio of ORs 1.94; 95% confidence interval (0.65, 5.73)). While values included in the 95% confidence interval of the ratio of ORs could be alarming, both direct and indirect ORs contain values that extend to clinically important values in both directions. Thus, incoherence may not have severe implications on the interpretation of the NMA treatment effect. Incoherence could imply major concerns for the confidence in citalopram versus venlafaxine NMA OR; the direct OR contains values within and above the range of equivalence while indirect OR includes values within and below the range of equivalence. The resulted estimated ratio of ORs is 2.08 with 95% confidence interval (1.03, 4.18) and the respective p-value of the test is 0.04. The ratio of direct to indirect amitriptyline versus paroxetine ORs is 1.05 (with 95% confidence interval (0.76, 1.46) and p-value 0.75) implying that the two sources of evidence appear to be in agreement. Direct and indirect ORs are very close in terms of mean OR, 95% confidence intervals and their placement with respect to the range of equivalence.

DISCUSSION

In this paper, we described ways to judge the domains imprecision, heterogeneity and incoherence when evaluating the NMA treatment effects. It is part of a two-papers series that introduces the CINeMA framework for placing confidence in NMA treatment effects.

The final step of the CINeMA framework is the integration of all the intermediate judgements to assign a confidence rating to each NMA treatment effect. As CINeMA considers only NMAs of randomized controlled trials (RCTs) (2), NMA treatment effects are initially assigned a high confidence rating. Then, intermediate considerations can lead to downgrading by one (moderate), two (low) or three (very low) levels. Intermediate judgments (Table 2 for the network of antidepressants) do not have a decisive role for downgrading the evidence but operate as a 'working table' for assigning the overall confidence rating. That is because some of the considered domains are interweaved; several aspects that could mitigate the confidence on an NMA treatment effect may appear in more than one domains.

Overlapping domains include indirectness – incoherence, heterogeneity – incoherence, heterogeneity – imprecision and heterogeneity – across-studies bias. Indirectness includes considerations on intransitivity that may manifest in the data as incoherence. Potential heterogeneity may affect the precision of the NMA treatment effects and across-studies bias can result to observe a large heterogeneity. The close connection of heterogeneity and incoherence is supported by the fact that incoherence can be seen a special form of heterogeneity; for example, within a loop that shows signs of incoherence, had each study included all arms would result to heterogeneity to some or all pairwise comparisons (8). Considering the above, separate judgements on the six CINeMA domains should be considered jointly rather than in isolation. The decision on the final confidence rating should take into account the entire NMA output with the help of the 'working table', placing particular attention not to downgrade twice for the same aspect.

CINeMA is primarily developed for evaluating confidence in the NMA treatment effects as a part of a systematic review with NMA. However, users of NMA can also evaluate confidence in the reported results using CINeMA but this would usually require outcome data. This limitation of the system is mitigated by the fact that 2 in 3 published NMAs (66%)

include their data in the manuscript (24). Subjectivity constitutes a further limitation of the system, which applies on several aspects of the evaluation such as setting the range of equivalence. We believe, however, that subjectivity is inevitable in such a process as it is associated with interpretation of the evidence. Indeed, any system that evaluates confidence in the results usually acknowledges the associated subjectivity and consequently the fact that different reviewers may not replicate initial judgments.

Although subjectivity cannot be eliminated, transparency is key in the framework described in this paper series. Judgments, although may differ across reviewers, are made using specified criteria and reasons for downgrading are provided. Among the strengths of the described strategy also lies its focus on the implications of potential limitations, in conjunction with the magnitude of the limitations. CINeMA web application *cinema.ispm.ch* can greatly facilitate the implementation of all steps described in this paper series constituting a further strength of the suggested framework (10).

The current paper series constitutes a refinement of a previously suggested framework (1) while a second approach has also been refined (25) since its initial introduction (26). The two methods (1,25,26) have similarities and differences. As an example of a difference, in (25,26) authors suggest a process of informing network estimates ratings through direct and potentially -depending on the sufficiency and certainty of direct evidence- indirect estimates. In contrast, in CINeMA, NMA relative treatment effects are evaluated only, without considering separately the sources from which the mixed evidence is produced. Considerations on defining the contribution of each piece of evidence also differs between the two methods. The impact of the differences between the two systems on the evaluation of NMA applications has not been formally investigated.

An alternative approach focuses on exploring how robust treatment recommendations are to potential degrees of bias in the evidence (27). While the starting point of this alternative approach is the summary estimates, as in our approach and the approach described in (25,26), the reliability of NMA is assessed using threshold analysis and do not make use of the five GRADE domains. The underlying theory of the approach considers a bias spectrum for each treatment effect, assumed known and without uncertainty, and determines the amount of bias that could be tolerated so that treatment recommendations are unchanged. While the process resembles setting ranges of

equivalence in the domains discussed in this paper, the question that threshold analysis aims to answer is different can be seen as complementary to our approach.

It is an exception rather than a rule for published NMAs to evaluate confidence in relative treatment effects (28). However, NMA's increasing use by national and international agencies (24,29) would render such an evaluation insightful as it would help decision makers, guideline developers and systematic reviewers, to interpret and critically appraise NMA evidence summaries. Evaluating confidence in each NMA treatment effect separately is important even when in total the systematic review has been appropriately conducted. That is because judgments on the confidence to be placed on NMA treatment effects may vary greatly within a network. For example, Cipriani et al. rated NMA treatment effects from moderate to very low, despite the fact that all resulted from the same systematic review and, thus, shared search strategy, eligibility criteria and, in general, research question (9).

In this paper series, we described the guiding principles of the CINeMA framework. We elaborated on ways that these guiding principles can be adopted in NMAs and provided examples of how they can be implemented. Although in (1) considerations had been made for critically appraising both relative treatment effects and treatment ranking, in this paper series we focused only on relative treatment effects. Suggested methodology for evaluating treatment hierarchy is incomplete and several issues, e.g. measuring imprecision in treatment ranking and quantifying treatment hierarchies' alterations due to within-study bias, remain unclear and constitute an area of further research. In conclusion, CINeMA is a transparent, rigorous and comprehensive framework for evaluating the confidence of NMA treatment effects.

Box 1 Description of the network used to exemplify the methods

Comparative tolerability and harms of individual statins

The aim of the systematic review by Naci et al. (12) was to determine the comparative tolerability and harms of 8 statins. The outcome we consider is the number of patients who discontinued due to adverse effects, measured as odds ratio. This outcome was evaluated in 101 studies. The network is presented in Figure 1 and the outcome data are given in Appendix Table 1.

Table 2 Results from direct, indirect and mixed evidence along with confidence and prediction intervals and incoherence ratio of odds ratios for the network of antidepressants. Odds ratios lower than 1 favor the first treatment. NMA: network meta-analysis, OR: odds ratio, PrI: prediction interval, CI: confidence interval.

Comparison	Direct OR (95% CI)	Indirect OR (95% CI)	Ratio of ORs (95% CI)	NMA OR (95% CI)	95% PrI of NMA OR
Clomipramine versus fluvoxamine	1.85 (0.65, 5.27)	0.96 (0.71, 1.29)	1.94 (0.65, 5.73)	0.99 (0.75, 1.32)	(0.63, 1.57)
Citalopram versus venlafaxine	1.72 (0.89, 3.32)	0.83 (0.66, 1.04)	2.08 (1.03, 4.18)	1.12 (0.90, 1.39)	(0.74, 1.70)
Amitriptyline versus paroxetine	1.07 (0.85, 1.36)	1.02 (0.82, 1.27)	1.05 (0.76, 1.46)	0.96 (0.82, 1.13)	(0.65, 1.42)

Table 3 Intermediate judgements for three network meta-analysis odds ratios from the network of antidepressants for the domains imprecision, heterogeneity and incoherence.

Comparison	Imprecision	Heterogeneity	Incoherence
Clomipramine versus fluvoxamine	Major concerns	No concerns	No concerns
Citalopram versus venlafaxine	Some concerns	Some concerns	Major concerns
Amitriptyline versus paroxetine	No concerns	Some concerns	No concerns

Figure 1 Network of randomised controlled trials comparing statins with respect to adverse effects. Nodes and edges are equally weighted across the network.

Figure 2 Network meta-analysis odds ratios from the network of statins, along with the range of equivalence, their 95% confidence intervals (black lines) and their 95% prediction intervals (red lines). NMA: network meta-analysis, OR: odds ratio, PrI: 95% prediction interval, CI: 95% confidence interval, vs: versus.

Figure 3 Hypothetical example illustrating situations where direct and indirect estimates may or may not imply different clinical decisions. OR: odds ratio.

Acknowledgements: The development of the software and part of the presented work was supported by the Campbell Collaboration. The sponsor had no role in the design and analysis or in the writing of the article.

REFERENCES

1. Salanti G, Del Giovane C, Chaimani A, Caldwell DM, Higgins JPT. Evaluating the quality of evidence from a network meta-analysis. PloS One. 2014;9(7):e99682.

2. Evaluation of confidence in the results of network meta-analysis: within-study bias, across-studies bias and indirectness. working paper

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

4. Guyatt GH, Oxman AD, Kunz R, Woodcock J, Brozek J, Helfand M, et al. GRADE guidelines: 7. Rating the quality of evidence--inconsistency. J Clin Epidemiol. 2011 Dec;64(12):1294–302.

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

6. Dias S, Welton NJ, Sutton AJ, Caldwell DM, Lu G, Ades AE. Evidence synthesis for decision making 4: inconsistency in networks of evidence based on randomized controlled trials. Med Decis Mak Int J Soc Med Decis Mak. 2013;33(5):641–56.

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

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

9. Cipriani A, Furukawa TA, Salanti G, Chaimani A, Atkinson LZ, Ogawa Y, et al. Comparative efficacy and acceptability of 21 antidepressant drugs for the acute treatment of adults with major depressive disorder: a systematic review and network meta-analysis. Lancet Lond Engl. 2018 Feb 20;

10. CINeMA: Confidence in Network Meta-Analysis. [Internet]. Institute of Social and Preventive Medicine, University of Bern.; 2017. Available from: cinema.ispm.ch

11. Rücker G, Schwarzer G, Krahn U, Konig J. netmeta: Network Meta-Analysis using Frequentist Methods. R package version 0.8-0. [Internet]. Available from: https://CRAN.R-project.org/package=netmeta

12. Naci H, Brugts J, Ades T. Comparative tolerability and harms of individual statins: a study-level network meta-analysis of 246 955 participants from 135 randomized, controlled trials. Circ Cardiovasc Qual Outcomes. 2013 Jul;6(4):390–9.

13. Riley RD, Higgins JPT, Deeks JJ. Interpretation of random effects meta-analyses. BMJ. 2011 Feb 10;342:d549.

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

15. Rhodes KM, Turner RM, Higgins JPT. Predictive distributions were developed for the extent of heterogeneity in meta-analyses of continuous outcome data. J Clin Epidemiol. 2015 Jan;68(1):52–60.

16. Lu G, Ades AE. Assessing Evidence Inconsistency in Mixed Treatment Comparisons. J Am Stat Assoc. 2006 Jun 1;101(474):447–59.

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

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

19. Dias S, Welton NJ, Sutton AJ, Caldwell DM, Lu G, Ades AE. Inconsistency in Networks of Evidence Based on Randomised Controlled Trials [Internet]. London: National Institute for Health and Care Excellence (NICE); 2014. (NICE Decision Support Unit Technical Support Documents). Available from: http://www.ncbi.nlm.nih.gov/books/NBK310372/

20. Veroniki AA, Mavridis D, Higgins JP, Salanti G. Characteristics of a loop of evidence that affect detection and estimation of inconsistency: a simulation study. BMC Med Res Methodol. 2014 Sep 19;14:106.

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

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

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

24. Petropoulou M, Nikolakopoulou A, Veroniki A-A, Rios P, Vafaei A, Zarin W, et al. Bibliographic study showed improving statistical methodology of network meta-analyses published between 1999 and 2015. J Clin Epidemiol. 2017 Feb;82:20–8.

25. Brignardello-Petersen R, Bonner A, Alexander PE, Siemieniuk RA, Furukawa TA, Rochwerg B, et al. Advances in the GRADE approach to rate the certainty in estimates from a network meta-analysis. J Clin Epidemiol. 2018 Jan;93:36–44.

26. Puhan MA, Schünemann HJ, Murad MH, Li T, Brignardello-Petersen R, Singh JA, et al. A GRADE Working Group approach for rating the quality of treatment effect estimates from network meta-analysis. BMJ. 2014 Sep 24;349:g5630.

27. Caldwell DM, Ades AE, Dias S, Watkins S, Li T, Taske N, et al. A threshold analysis assessed the credibility of conclusions from network meta-analysis. J Clin Epidemiol. 2016 Dec;80:68–76.

28. Zarin W, Veroniki AA, Nincic V, Vafaei A, Reynen E, Motiwala SS, et al. Characteristics and knowledge synthesis approach for 456 network meta-analyses: a scoping review. BMC Med. 2017 05;15(1):3.

29. Kanters S, Ford N, Druyts E, Thorlund K, Mills EJ, Bansback N. Use of network meta-analysis in clinical guidelines. Bull World Health Organ. 2016 Oct 1;94(10):782–4.

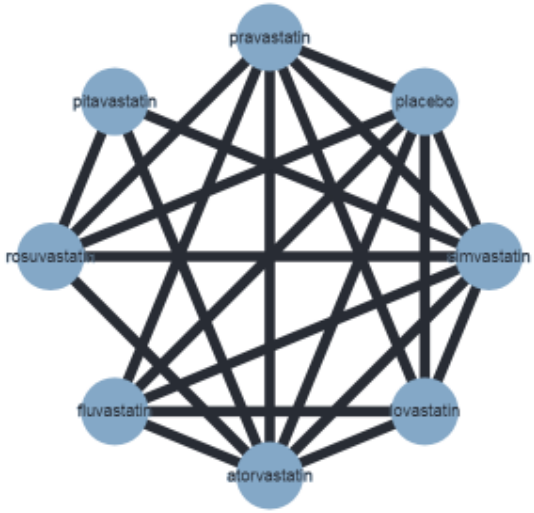
pravastatin
pitavastatin
placebo
rosuvastatin
simvastatin
fluvastatin
ovastatin
atorvastatin

Figure 1

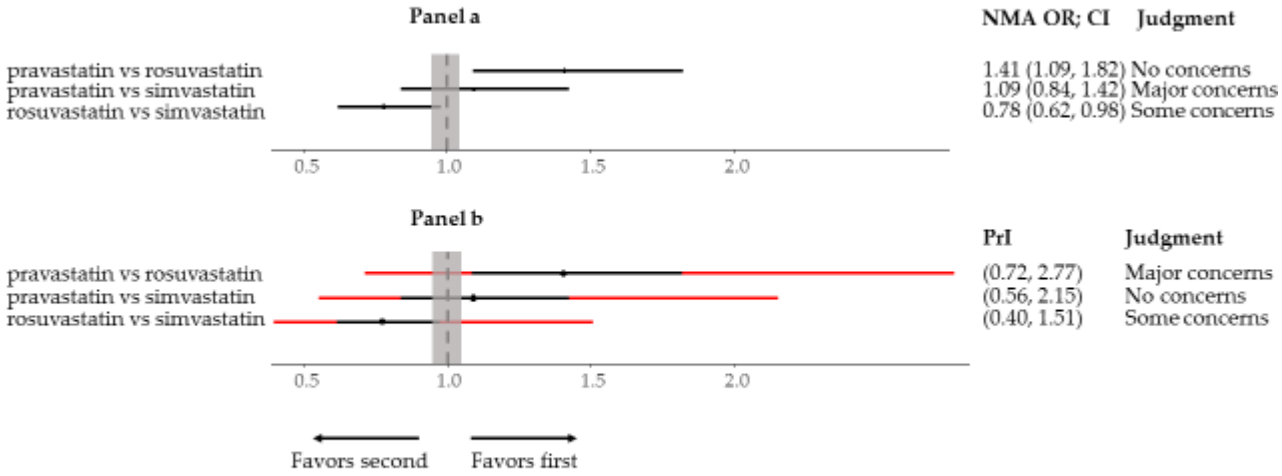
Panel a
NMA OR; CI
Judgment
pravastatin vs rosuvastatin
pravastatin vs simvastatin
rosuvastatin vs simvastatin
1.41 (1.09, 1.82) No concerns
1.09 (0.84, 1.42) Major concerns
0.78 (0.62, 0.98) Some concerns
0.5
1.0
1.5
2.0
Panel b
PrI
Judgment
pravastatin vs rosuvastatin
pravastatin vs simvastatin
rosuvastatin vs simvastatin
(0.72, 2.77) Major concerns
(0.56, 2.15) No concerns
(0.40, 1.51) Some concerns
0.5
1.0
1.5
2.0
Favors second
Favors first

Figure 2

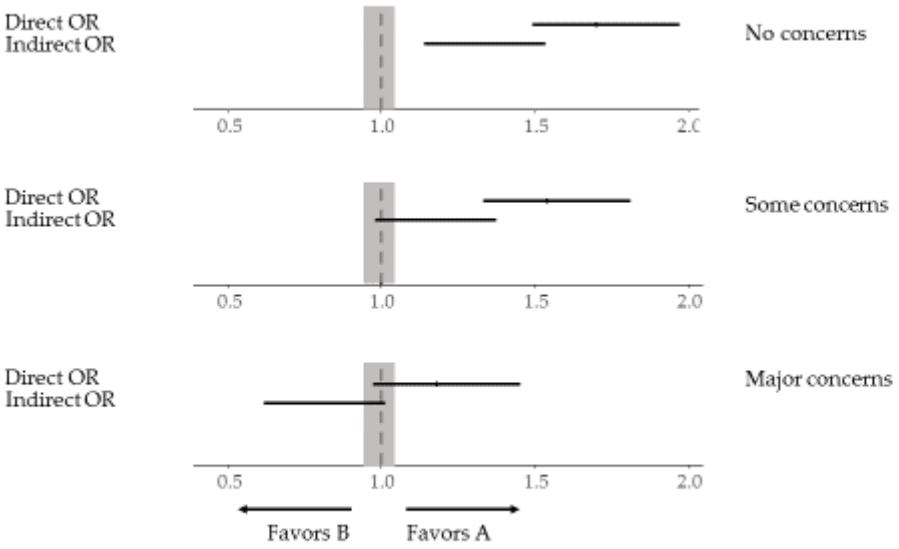
Direct OR
Indirect OR
No concerns
0.5
1.0
1.5
2.0
Direct OR
Indirect OR
Some concerns
0.5
1.0
1.5
2.0
Direct OR
Indirect OR
Major concerns
0.5
1.0
1.5
2.0
Favors B
Favors A

Figure 3