WORKING PAPER

EVALUATION OF CONFIDENCE IN THE RESULTS OF NETWORK META-ANALYSIS

WITHIN-STUDY BIAS, ACROSS-STUDIES BIAS AND INDIRECTNESS

INTRODUCTION

While network meta-analysis (NMA) has become an increasingly popular tool in making reimbursement recommendations and forming treatment guidelines, less than 1 per cent of published NMAs attempt to make an evaluation of the credibility of their conclusions [1]. Two systems have been presented so far that can be used for this purpose [2,3]. However, their complexity and lack of suitable software to speed-up and simplify the process have considerably limited their uptake.

Puhan et al. presented a system closely based on the GRADE framework (Grading of Recommendations Assessment, Development and Evaluation)[2,4]. It evaluates the confidence in NMA results considering the possibility that the evidence base is compromised by study limitations, indirectness, imprecision, inconsistency and publication bias. The starting point of an alternative system by Salanti et al. (termed CINeMA, Confidence in Network Meta-Analysis) is also GRADE but has important conceptual and semantical differences [3]. CINeMA comprises six domains; within-study bias (referring to the impact of risk of bias in the included studies), across-studies bias (referring to publication and reporting bias), indirectness, imprecision, heterogeneity and incoherence. Each domain is assigned an intermediate or "temporary" judgement (no concerns, some concerns, major concerns). Judgements within each domain are then summarized across to obtain four possible levels of confidence for each NMA relative treatment effect: very low, low, moderate or high.

The present article and its companion clarify the methodology underpinning CINeMA and present advances that have been recently implemented in a freely available web application (cinema.ispm.ch [4]) in two articles. Both articles make some assumptions. While Salanti et al described how to evaluate confidence in treatment ranking, we will address only the case of evaluating the confidence in the relative treatment effects. The vast majority of NMAs include only randomized controlled trials (RCTs), so we will consider only this study design. We further assume that evaluation of the credibility of results takes place when all primary analyses and sensitivity analyses have been undertaken. In NMA context this involves the integration of direct and indirect evidence in an entire network of relevant trials. While some researchers might argue that it is preferable to use only direct or only indirect evidence for a comparison based on the quality of the respective piece of evidence, we do not recommend this approach. We assume that reviewers have a-priory defined the study inclusion criteria (that potentially include risk of bias considerations) and have obtained the best possible relative treatment effects. Then the question is how to make judgements about the credibility of these relative treatment effects considering that trials of variable risk of bias, precision, relevance and heterogeneity contribute information to the final result. Finally, we assume that the assumption of transitivity has been deemed reasonable and that statistical synthesis of study results using NMA is appropriate.

The aim of this first paper is to propose ways of forming judgements about the first three CINeMA components; within-study bias, across-studies bias and indirectness. The methods are exemplified using two examples; a network of trials that compare various diagnostic strategies for patients with low risk of acute coronary syndrome [7] and a network of trials comparing 18 antidepressants [8]. The two examples are presented in Box 1. All analyses have been undertaken in R software using the *netmeta* package and the CINeMA web application [5,6].

WITHIN-STUDY BIAS

Within-study bias refers to the shortcomings in the design or conduct of a study that can result into an estimated relative treatment effect that is different from the true. The Cochrane Collaboration has developed and established a tool that the majority of the published systematic reviews use to evaluate the risk of bias in the included RCTs [9]. Studies can be classified as having low, moderate/unclear or high risk of bias after summarizing judgements across the various bias components (such as allocation concealment, attrition, blinding etc.).

While it is easy to make assumptions about how within-study biases might impact on the summary relative treatment effect in a pairwise meta-analysis [10], in NMA studies contribute data to the estimation of each summary effect in a complicated manner, and the influence of a study depends both on its precision and location in the network. A treatment comparison directly evaluated in studies with low risk of bias, can be also estimated indirectly (via a common comparator) using studies at high risk of bias and vice versa. While studies at low risk of bias are expected to provide more credible results, is not always possible to restrict the database. The treatment comparison of interest might have not been tested directly in any trial, or in a few small trials with high risk of bias.

Consider for example the network in Figure 1 ; no studies have compared Exercise ECG and CMR and judgement has to be made by considering that three studies (at low, moderate and high risk of bias) collectively provide indirect evidence to estimate the odds ratio (OR) 0.73 via Standard care). Even when direct evidence is present, judgments about the NMA relative treatment effect cannot ignore the risk of bias in the studies providing indirect evidence. Direct evidence from the single study comparing Exercise ECG with Standard care is at low risk of bias (study 16); one might argue there are no study limitations when interpreting the direct OR 0.42 (Table 1). However, the NMA OR 0.52 is estimated also by using indirect information via seven studies that compare Standard care and CCTA and one study comparing Exercise ECG and CCTA. The risk of bias in these eight studies providing indirect evidence is variable; a total of 2162 study participants are randomized to high risk of bias studies, 2355 to low risk of bias and 60 to moderate risk of bias. In these two examples, risk of bias in the indirect evidence via an intermediate comparator (termed here "one-step loop") is important to consider along the direct evidence when evaluating study limitations.

In complex networks with many interventions and loops of evidence, it is neither practical nor desirable to derive judgements considering the risk of bias in studies in a single 'one-step loop' [2,4]. Indirect evidence can be obtained via several 'routes' going beyond the 'one-step loop'. For example, in Figure 1, indirect evidence about Exercise ECG versus SPECT-MPI can be obtained from two "one-step loops" (via CCTA or via Standard Care) and three "two-step loops" (via CCTA-Standard Care, Stress Echo-Standard Care, Standard Care-CCTA). In each loop of evidence, a different subgroup of studies contributes indirect information and their size and risk of bias vary considerably. The only studies that do not contribute information to the NMA OR between Exercise ECG and SPECT-MPI are the two studies comparing Standard Care and CMR. As a general rule, most studies in a network contribute, to some extent, indirect information to every NMA relative treatment effect.

Not all studies have however the same impact on the estimation of an NMA relative treatment effect. Studies contribute more when their results are precise (e.g. large studies), when they provide direct evidence or when they are located close to the targeted comparison (the treatment comparison we want to evaluate). To compare Stress Echo and

CCTA, indirect evidence comes from eight studies via Standard Care. To estimate the NMA OR 4.31 in, Study 8 (at low risk of bias) with sample size 1392 will be more influential than study 10 (at moderate risk of bias) with sample size 60. For the NMA OR between Exercise ECG and SPECT-MPI study 5 will be more influential than study 8 because study 5 is closer to the comparison of interest and facilitates "one-step" indirect evidence (via Standard care). In summary, estimation of the contribution of every study requires complicated calculations that involve the study precision and location. CINeMA uses the contribution matrix to approximate the contribution of each study.

The percentage contribution matrix is a matrix that quantifies how much a study contributes in an NMA relative treatment effect on a 0 to 100 percentage scale. It is an approximated transformation of the H matrix in a two-step NMA model [11,12]. The technicalities of the matrix estimation are detailed in a technical article [13].

Table 2 shows the matrix for the network and data of Figure 1. The columns represent the studies, grouped by comparison. The rows represent all NMA relative treatment effects. The matrix entries show how much each study contributes to the estimation of each NMA relative treatment effect. Combination of the studies contributions with risk of bias judgements is the main instrument we propose to evaluate study limitation for each NMA relative treatment effect. This can be presented in the form of a bar chart as in

Table 2 The percentage contribution matrix for the network presented in Figure 1. The columns refer to the studies (grouped by comparison) and the rows refer to the NMA relative treatment effects (grouped into mixed and indirect evidence). The entries show how much each study contributes (as percentage) to the estimation of each NMA relative treatment effects. . It shows that the indirect odds ratio between ECG versus SPECT-MPI is estimated by synthesizing data from studies at high, moderate and low risk of bias with contributions 31%, 55% and 14%. Then, we will need to make a total judgment about the actual study limitations in this comparison.

Different options are possible when deriving study limitation judgements from the risk of bias bar chart. Depending on the source of bias suspected one can be more or less "tolerant" to the contribution from studies at high and moderate risk of bias. In some applications it might be handy to use thresholds; concerns can be no serious, serious or very serious for study limitation according to the contribution from studies at high/moderate risk of bias exceeding values x_1% and x_2%.

Using thresholds to derive judgements is practical when many comparisons need to be evaluated (see for example [8]) but should be used with caution. The thresholds should be pre-specified to avoid spurious conclusions and should be informed by a sensitivity analysis. It is part of the good practice to compare the results from NMA with all studies to those obtained when only low risk of bias studies are included [14,15]. If the design shortcomings are shown (in empirical studies and in the sensitivity analysis) to produce results different to those found in studies with low risk of bias, then judgements would be stricter, and thresholds should be defined accordingly. Reviewers can choose to characterize the NMA relative treatment effect as having very serious concerns for study limitations even when the contribution of the high risk of bias studies is small if the results from a sensitivity analysis, after excluding studies at high risk of bias, are different from those obtain from the entire dataset.

Forming judgements about within-study bias in the network of antidepressants would have been impossible without the contribution matrix as direct and evidence via many different intermediate comparators are present. Note that, with 18 active treatments, there are 153 NMA odds ratios to evaluate and for most of them it is impossible to choose the most influential "one-step loop" as suggested in [4]. For fluoxetine versus paroxetine there are up to 13 "one-step loops" providing indirect evidence while half of the treatment comparisons have indirect evidence from at least four one-step loops to choose from (Appendix Table 2). More importantly, the amount of information coming from loops involving more than one step is substantial. Even if all "one-step loops" were accounted for, only 56% of the information in the network would have been considered (Appendix Table 3).

We will focus on evaluating the results for three comparisons; amitriptyline vs milnacipran (one direct study at low and one at moderate risk of bias), mirtazapine versus paroxetine (three direct studies at low risk of bias and two at moderate) and amitriptyline vs clomipramine (no direct studies). The response ORs are presented in Table 4. Study contributions with the risk of bias judgements for each study produce the graph plot in Figure 4.

For the first two treatment comparisons in Table 4, the sensitivity analysis excluding studies at moderate risk of bias provides results comparable to those obtained from all studies (Table 4). Consequently, we can employ a "generous" threshold for summarizing each bar in Figure 4 whereby concerns about study limitations are expressed only if the contribution of moderate risk of bias studies exceeds, for example, 90%. However, for amitriptyline versus clomipramine we might want to employ stricter threshold (e.g. of 60%) because the OR from the sensitivity analysis is quite different to the one obtained from all studies. Considering these two thresholds, one can derive the interim judgements; no concerns for amitriptyline vs milnacipran and mirtazapine versus paroxetine and some concerns for amitriptyline versus clomipramine. Cipriani et al employed a single threshold of 70% for the contribution of the moderate risk of bias studies, based on their observation that in the vast majority of the comparisons the sensitivity analysis gives similar results to those from all studies.

ACROSS-STUDIES BIAS

Across-studies bias occurs when the studies included in the systematic review are not a representative sample of the studies undertaken. This phenomenon can be the result of the suppression of statistically significant (or “negative”) findings (publication bias), their delayed publication (time-lag bias) or omission of unfavorable study results (outcome reposting bias). The presence and the impact of such biases has been well documented (cite). However, across-studies bias is a missing data problem, and hence it is impossible to conclude with certainty for or against its presence in a given dataset. Consequently, and in agreement with the GRADE system, CINeMA assumes two possible descriptions for across-studies bias: undetected and suspected.

Deciding about the risk of across-studies bias is not easy and follows considerations applicable to the pairwise meta-analysis described in [16]. Without aiming to be exhaustive we list below some conditions that can be associated with the presence of across-studies bias:

- Failure to include unpublished data and data from grey literature in the review
- The treatment comparison includes an agent newly introduced in the market, as early evidence is likely to overestimate its efficacy and safety (cite).
- The treatment comparison is studied in few small trials with positive early findings
- The treatment comparison is studied exclusively or primarily in industry-funded trials

- There is previous empirical evidence in the field documenting the presence of reporting bias (such as, for example the study by Turner et al. showing publication bias in the placebo-controlled antidepressant trials [17]).

Note that while the presence of few small studies might be associated with high risk of reporting bias, the absence of trials for a given comparison (so that the treatment effect is estimated only indirectly) is not necessarily subject to bias.

Across-studies bias can be undetected when

- Data from unpublished studies have been identified and their findings agree with those in published studies
- There is a tradition of prospective trial registration in the field and protocols or clinical trial registries do not indicate important discrepancies with published reports
- Empirical examination of patterns of results between small and large studies (e.g. using the comparison-adjusted funnel plot [18,19]), regression models [20] or selection models [21] do not indicate that results from small studies are likely to differ from those in published studies.

The literature search in the antidepressants review retrieved supplementary and unpublished information from clinical trial registries, regulatory agencies' repositories and drug companies' websites (particularly for the newest and most recently marketed antidepressants). Results between published and unpublished studies did not differ materially. A comparison-adjusted funnel plots for all drugs against fluoxetine was drawn, as (fluoxetine was the most often studied drug and has been standard pharmacological treatment for most of countries over time); no asymmetry was observed. Network meta-regression on study variance did not indicate an important association between study precision and study OR. However, the authors decided that they cannot completely rule out the possibility that some studies are still missing because the field of antidepressant trials has been shown to be prone to publication bias. Consequently, the review team decided to characterise the across-studies bias as suspected for all drug comparisons.

Indirectness

In the GRADE framework for pairwise meta-analysis, indirectness refers to the relevance of the included studies to the research question [22]. Study populations, interventions, outcomes and study settings as reported in studies should match the inclusion criteria of the systematic review but might not be fully representative of the settings, populations or outcomes that reviewers want to make inferences about. A systematic review aiming to provide evidence about treating adults, might identify studies in elderly patients; these studies will have an indirect relevance.

We suggest that each study included in the network is evaluated according to its relevance to the research question and classified into low, high or moderate indirectness. Note that only patient, outcome and intervention characteristics that can act as effect modifiers shall be considered; these are variables that modify the relative effect of an intervention against another. Then, the study-level judgements can be combined with the contribution matrix to produce a bar plot similar to this presented in

Table 2 The percentage contribution matrix for the network presented in Figure 1. The columns refer to the studies (grouped by comparison) and the rows refer to the NMA relative treatment effects (grouped into mixed and indirect evidence). The entries show how much each study contributes (as percentage) to the estimation of each NMA relative treatment effects. . Finally, the evaluation of indirectness for each NMA relative treatment effect is obtained by judging whether the contribution from studies of high or moderate indirectness is important.

We suggest that this approach also addresses the assumption of transitivity. The assumption of transitivity is fulfilled when the distribution of effect modifiers is similar across the treatment comparisons linked in a network [23]. Consider the fictional case of indirect comparison between B and C presented in Table 3. In the first scenario, the distribution of the effect modifier “age” is comparable in the two sets of studies, so that there are no concerns about intransitivity. In the second scenario, there is intransitivity. In both scenarios 1 and 2, the bar chart will indicate that half of the information comes from studies with some form of indirectness. Note that it is impossible to have a situation where the bar plot will not indicate the presence of indirectness while the distribution of effect modifiers is unequal. Consequently, if indirectness is evaluated using the bar plot, it will also reflect concerns about intransitivity. Another limitation is reporting; values of important effect modifiers might not always be available in trial reports.

Evaluation of distribution of effect modifiers is possible when enough studies are available per comparison and inference is challenging or even impossible for interventions poorly connected to the network. This approach shall be used with care in sparse networks (when the studies are few compared to the total number of treatments). In scenario 3 of

Table 3 we cannot conclude anything about transitivity because there is only one study to represent each distribution. For this reason, we recommend that the network structure and the amount of available data are consider and that judgements are on the stricter side.

Cipriani et al concluded that there is no indirectness in any of the studies included and that the distribution of modifiers was similar across studies and comparisons. However, they decided to downgrade evidence about drugs that are poorly connected to the network. Vortioxetine features in a single study (against venlafaxine) and consequently it is difficult to infer about the comparability of effect modifiers in the comparisons against vortioxetine. Consequently, Cipriani et al. some concerns for indirectness for all ORs against vortioxetine.

Discussion

We presented a framework that enables forming judgements about the within-study biases and indirectness in NMA relative treatment effects that avoids selective use of the indirect evidence, while considering the characteristics of all studies included in the network. The approach can be operationalized and is easy-to-implement even for very large networks. The use of thresholds about the percentage contribution from studies at high and moderate risk of bias and indirectness can further speed-up the process. As deviations in studies from the targeted population and setting are related to the notion of transitivity, evaluation of the indirectness domain also addresses the core assumption of network meta-analysis.

Every method to evaluate confidence in evidence synthesis results involves some subjective judgements. Our approach is no exception to this. The use of thresholds to summarize bar charts speeds-up the process, but their justification is difficult. Further limitations of the framework are associated with the fact that published data are used to make judgements and they do not necessarily reflect the way studies were undertaken. For instance, judging indirectness requires study data to be collected on pre-specified effect modifiers; reporting limitations will inevitably impact on the reliability of the judgements. Finally, further research is needed to operationalize the way that judgements about across-studies bias is evaluated.

Our approach offers some important advantages. Difficult-to-make choices, such as choosing the most influential “one-step loop” to represent indirect evidence and selective use of the data, are not needed. The use of percentage contribution matrix is the only viable option available to date to estimate the relative impact of each study included in a network. The impact of study shortcomings (either in the form of bias or indirectness) can be easily visualized in a bar chart. The framework naturally includes the results from sensitivity analyses in the interpretation of the bar charts. A further advantage of the contribution matrix is that is can be generalized to present risk of bias in the entire network of interventions; the entries in the contribution matrix can be re-scaled to estimate the percentage contribution of each study to the entire network.

With the use of open-source free software, our approach can be routinely applied to any NMA [5]. Despite some inevitable subjectivity involved in the judgements, our approach is step forward into transparency and reproducibility. The suggested framework operationalizes, simplifies and speeds-up the process of evaluation of results from large and complex networks without compromising in statistical and methodological rigor.

Diagnostic strategies for patients with low risk of acute coronary syndrome

The aim of the systematic review by Siontis et al was to evaluate the differences between the non-invasive diagnostic modalities used to detect coronary artery disease in patients presenting with symptoms suggestive of low risk acute coronary syndrome. As outcome we will consider the number of referrals for invasive coronary angiography out of the total number of randomized patients. For this outcome, 18 studies were included. The network is presented in Figure 1 and the data in Appendix Table 1.

Antidepressants for moderate and major depression

The aim of a systematic review published by Cipriani et al. was to compare 18 commonly prescribed antidepressants studied in 179 head-to-head randomized trials involving patients diagnosed with major/moderate depression [8]. The primary efficacy outcome was response measured as 50% reduction in the symptoms scale between baseline and 8 weeks of follow-up. According to the inclusion criteria specified in the protocol only studies at low or moderate risk of bias were included [24]. The methodological and statistical details presented in the published article and its appendix. Here, we will focus on how judgements about credibility of the results were derived. The network is presented in Figure 2 and the data is available in Mendeley Data (DOI:10.17632/83rthbp8ys.2).

Box 1 Description of networks used to exemplify the methods

Table 1 Results from pairwise (upper triangle) and network meta-analysis (lower triangle) from the network of non-invasive diagnostic strategies for the detection of coronary artery disease in Figure 1. Odds ratios and their 95% confidence intervals are presented for referrals for invasive coronary angiography. Odds ratios in the lower triangle less than one favor the strategy in the column; odds ratios in the upper triangle less than one favor the strategy in the row.

CCTA	.	2.25 [1.04; 4.90]	1.04 [0.70; 1.55]	1.23 [1.00; 1.50]	.
3.07 [1.46; 6.45]	**CMR**	.	.	0.38 [0.18; 0.78]	.
2.24 [1.22; 4.11]	0.73 [0.28; 1.88]	**Exercise ECG**	.	0.42 [0.14; 1.30]	1.93 [1.39; 2.67]
1.27 [1.01; 1.60]	0.42 [0.20; 0.87]	0.57 [0.30; 1.07]	**SPECT-MPI**	0.87 [0.71; 1.06]	.
1.17 [0.97; 1.40]	0.38 [0.18; 0.78]	0.52 [0.28; 0.96]	0.92 [0.76; 1.10]	**Standard care**	2.95 [0.97; 8.98]
4.31 [2.23; 8.32]	1.40 [0.53; 3.74]	1.93 [1.39; 2.66]	3.38 [1.71; 6.68]	3.69 [1.90; 7.17]	**Stress Echo**

Table 2 The percentage contribution matrix for the network presented in Figure 1. The columns refer to the studies (grouped by comparison) and the rows refer to the NMA relative treatment effects (grouped into mixed and indirect evidence). The entries show how much each study contributes (as percentage) to the estimation of each NMA relative treatment effects.

Direct comparisons (number of studies)	**CCTA vs Exercise ECG (1)**	**CCTA vs SPECT-MPI (2)**		**CCTA vs Standard care (7)**							**CMR vs Standard care (2)**		**Exercise ECG vs Standard care (1)**	**Exercise ECG vs Stress Echo (4)**				**SPECT-MPI vs Standard care (2)**		**Standard care vs Stress Echo (1)**
NMA Estimates/study IDs	**3**	**2**	**9**	**1**	**10**	**13**	**14**	**4**	**7**	**8**	**11**	**6**	**12**	**12**	**15**	**16**	**17**	**18**	**5**	**12**
Mixed estimates																				
CCTA:Exercise ECG	52	1	1	3	0	3	1	3	4	4	0	0	14	0	3	0	2	1	1	6
CCTA:SPECT-MPI	1	18	16	5	1	5	1	6	7	7	0	0	0	0	0	0	0	22	10	0
CCTA:Standard care	1	4	4	13	2	13	3	15	18	17	0	0	1	0	0	0	0	6	3	0
CMR:Standard care	0	0	0	0	0	0	0	0	0	0	60	40	0	0	0	0	0	0	0	0
Exercise ECG:Standard care	23	1	1	3	0	3	1	4	5	4	0	0	30	1	6	1	3	2	1	11
Exercise ECG:Stress Echo	1	0	0	0	0	0	0	0	0	0	0	0	1	5	52	8	29	0	0	2
SPECT-MPI:Standard care	0	5	4	1	0	1	0	2	2	2	0	0	0	0	0	0	0	57	26	0
Standard care:Stress Echo	14	1	1	2	0	2	1	2	3	3	0	0	14	2	16	2	9	1	1	27
Indirect estimates	--	--	--	--	--	--	--	--	--	--	--	--	0	0	--	--	--	--	--	0
CCTA:CMR	1	3	2	6	1	7	2	8	9	8	28	19	1	0	0	0	0	4	2	0
CCTA:Stress Echo	24	1	1	3	0	3	1	3	4	4	0	0	8	2	18	3	10	1	1	13
CMR:Exercise ECG	16	1	1	2	0	2	1	3	3	3	22	15	15	0	4	1	2	1	1	7
CMR:SPECT-MPI	0	3	3	1	0	1	0	1	1	1	28	19	0	0	0	0	0	28	13	0
CMR:Stress Echo	11	1	1	1	0	2	0	2	2	2	20	14	9	1	11	2	6	1	0	13
Exercise ECG:SPECT-MPI	21	7	6	1	0	1	0	1	2	2	0	0	15	0	4	1	2	20	9	7
SPECT-MPI:Stress Echo	14	5	4	1	0	1	0	1	1	1	0	0	9	1	13	2	7	19	9	13

Table 3 Fictional case of indirect comparison between B and C with three scenarios where intransitivity, indirectness or both can occur. It is assumed that all studies are of the same precision.

	A vs B studies	A vs C studies	Distribution of effect modifiers	Bar plot will indicate indirectness in NMA relative effect for B vs C
Scenario 1	4 studies in elderly 4 studies in adults	4 studies in elderly 4 studies in adults	similar	yes
Scenario 2	5 studies in adults	5 studies in elderly	dissimilar	yes
Scenario 3	1 study in adults	1 study in adults	unknown?	no

Table 4 NMA summary odds ratios comparing six antidepressants and sensitivity analysis excluding studies at moderate risk of bias. Differences between the estimates suggests that a stricter threshold should be employed for the contribution of studies are high risk of bias.

Comparison	**Response odds ratio [95% confidence interval]**	
	All studies (179 studies)	*Only studies at low risk of bias (83 studies)*
amitriptyline versus milnacipran	1.11 [0.85; 1.43]	1.10 [0.77; 1.59]
mirtazapine versus paroxetine	1.07 [0.88; 1.30]	1.08 [0.83; 1.39]
amitriptyline versus clomipramine	1.24 [0.97; 1.59]	0.96 [0.59; 1.57]

Figure 1 Network of randomised controlled trials comparing non-invasive diagnostic strategies for the detection of coronary artery disease in patients with low risk acute coronary syndrome. The edges are propositional to the number of patients randomised in each comparison. ECG: electrocardiogram; echo: echocardiography; SPECT-MPI: single photon emission computed tomography-myocardial perfusion imaging; CCTA: coronary computed tomographic angiography; CMR: cardiovascular magnetic resonance. In square brackets are the study IDs as presented in Appendix Table 1.

Figure 2 Network of randomised controlled trials comparing active antidepressants in patients with moderate/major depression. The edges are propositional to the number of patients randomised in each comparison.

Figure 3 Risk of bias bar chart. Each bar represents an NMA relative treatment effect estimated using the data in the network in Figure 1. Each bar shows the percentage contribution from studies judged to be at low (green), moderate (yellow) and high(red) risk of bias.

Figure 4 Risk of bias bar chart for the comparison of five antidepressants using data from a network of 18 antidepressants. Different thresholds for the contribution of studies at moderate risk of bias are used for each comparison (red dashed lines) after considering the results from the sensitivity analysis presented in Table 4. For reference we present the threshold used by Cipriani et al. (same threshold for all comparisons).

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. 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;15:3.

2. 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;349:g5630.

3. 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:e99682.

4. 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;93:36–44.

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

6. 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

7. Siontis GC, Mavridis D, Greenwood JP, Coles B, Nikolakopoulou A, Jüni P, et al. Outcomes of non-invasive diagnostic modalities for the detection of coronary artery disease: network meta-analysis of diagnostic randomised controlled trials. BMJ. 2018;360:k504.

8. 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;

9. Higgins JPT, Altman DG, Gøtzsche PC, Jüni P, Moher D, Oxman AD, et al. The Cochrane Collaboration's tool for assessing risk of bias in randomised trials. BMJ. 2011;343:d5928.

10. Guyatt GH, Oxman AD, Vist G, Kunz R, Brozek J, Alonso-Coello P, et al. GRADE guidelines: 4. Rating the quality of evidence--study limitations (risk of bias). J Clin Epidemiol. 2011;64:407–15.

11. Krahn U, Binder H, König J. Visualizing inconsistency in network meta-analysis by independent path decomposition. BMC Med Res Methodol [Internet]. 2014 [cited 2015 Apr 7];14. Available from: http://www.biomedcentral.com/1471-2288/14/131/abstract

12. Lu G, Welton NJ, Higgins JPT, White IR, Ades AE. Linear inference for mixed treatment comparison meta-analysis: A two-stage approach. ResSynthMeth. 2011;2:43–60.

13. Papakonstantinou T, Nikolakopoulou A, Rücker G, Chaimani A, Schwarzer G, Egger M, et al. Estimating the contribution of studies in network meta-analysis: paths, flows and streams. F1000Research. 2018;7:610.

14. Caldwell DM. An overview of conducting systematic reviews with network meta-analysis. Syst Rev. 2014;3:109.

15. Jansen JP, Fleurence R, Devine B, Itzler R, Barrett A, Hawkins N, et al. Interpreting indirect treatment comparisons and network meta-analysis for health-care decision making: report of the ISPOR Task Force on Indirect Treatment Comparisons Good Research Practices: part 1. ValueHealth. 2011;14:417–28.

16. Guyatt GH, Oxman AD, Montori V, Vist G, Kunz R, Brozek J, et al. GRADE guidelines: 5. Rating the quality of evidence--publication bias. J Clin Epidemiol. 2011;64:1277–82.

17. Turner EH, Matthews AM, Linardatos E, Tell RA, Rosenthal R. Selective publication of antidepressant trials and its influence on apparent efficacy. N Engl J Med. 2008;358:252–60.

18. Chaimani A, Salanti G. Using network meta-analysis to evaluate the existence of small-study effects in a network of interventions. Res Synth Methods. 2012;3:161–176.

19. Chaimani A, Salanti G. Visualizing assumptions and results in network meta-analysis: The network graphs package. 2015;15:905–50.

20. Mavridis D, Efthimiou O, Leucht S, Salanti G. Publication bias and small-study effects magnified effectiveness of antipsychotics but their relative ranking remained invariant. J Clin Epidemiol. 2015;

21. Mavridis D, Sutton A, Cipriani A, Salanti G. A fully Bayesian application of the Copas selection model for publication bias extended to network meta-analysis. Stat Med. 2013;32:51–66.

22. Guyatt GH, Oxman AD, Kunz R, Woodcock J, Brozek J, Helfand M, et al. GRADE guidelines: 8. Rating the quality of evidence--indirectness. J Clin Epidemiol. 2011;64:1303–10.

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

24. Furukawa TA, Salanti G, Atkinson LZ, Leucht S, Ruhe HG, Turner EH, et al. Comparative efficacy and acceptability of first-generation and second-generation antidepressants in the acute treatment of major depression: protocol for a network meta-analysis. BMJ Open. 2016;6:e010919.

Figure 1

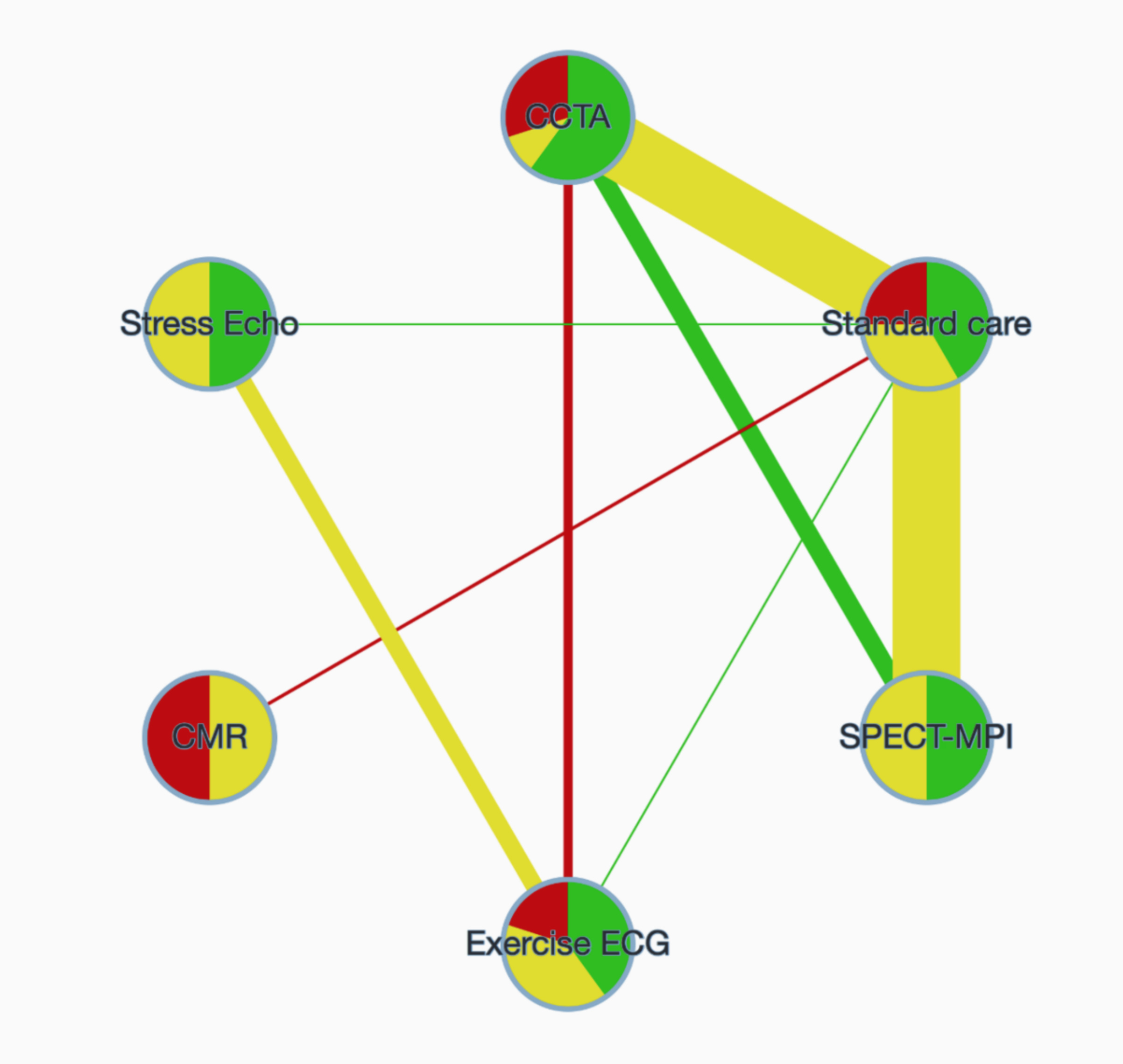


Figure 2

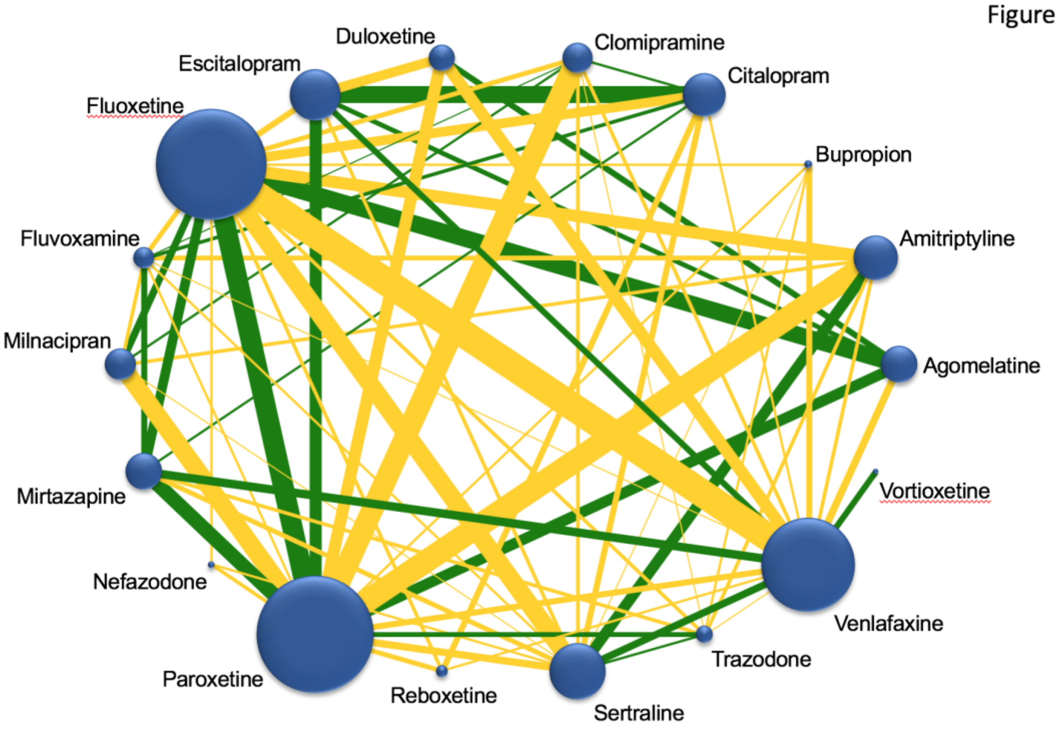


Figure 3

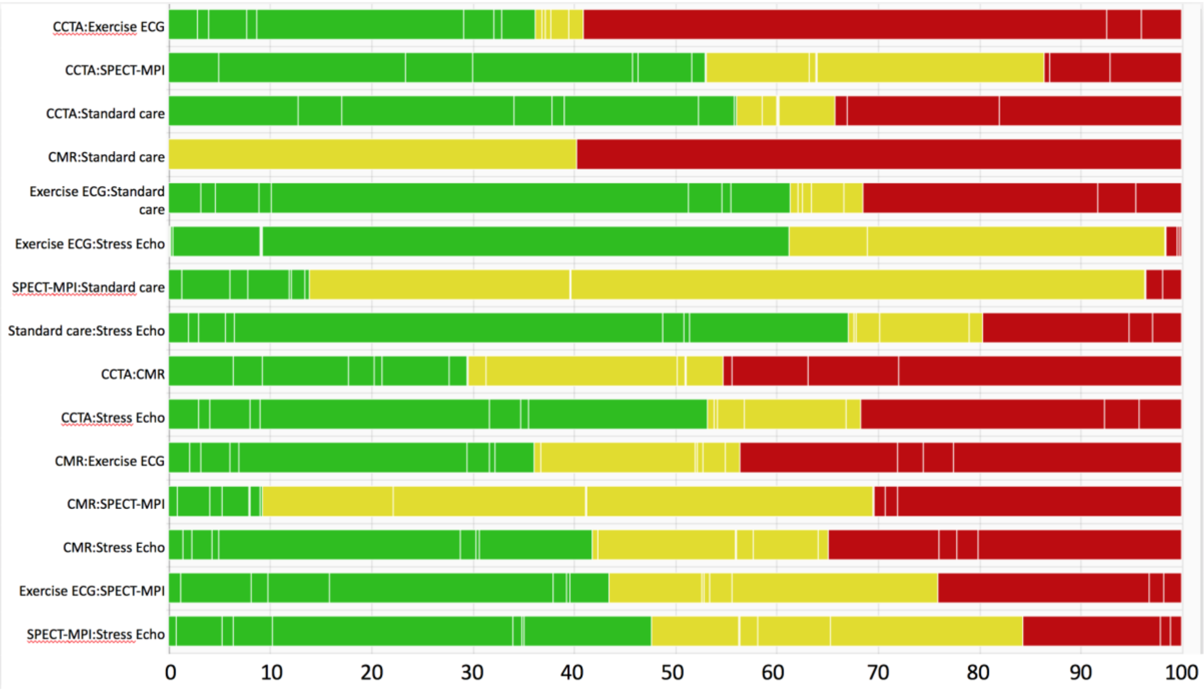
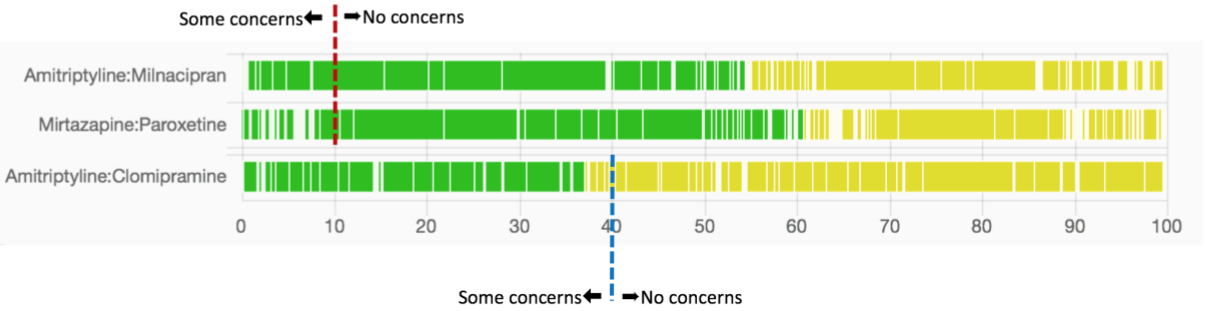


Figure 4


Some concerns
No concerns
Amitriptyline:Milnacipran
Mirtazapine:Paroxetine
Amitriptyline:Clomipramine
0
10
20
30
40
50
60
70
80
90
100
Some concerns
No concerns