UNIVERSITÄT
BERN

Graduate School for Health Sciences

University of Bern

Methods for ranking competing treatments in network meta-analysis

PhD Thesis submitted by

Virginia Chiocchia

for the degree of

PhD in Health Sciences (Epidemiology and Biostatistics)

Thesis advisor
Prof. Georgia Salanti
Institute of Social and Preventive Medicine
Faculty of Medicine, University of Bern

Co-referee
Prof. Tianjing Li
Department of Ophthalmology,
School of Medicine, University of Colorado Denver

Accepted by the Faculty of Medicine and the Faculty of Human Sciences of the University of Bern

Bern, Dean of the Faculty of Medicine

Bern, Dean of the Faculty of Human Sciences

Table of contents

Abstract

Background

One of the main objectives of comparative effectiveness research is to identify the most preferable treatments for a specific condition. Network meta-analysis (NMA) has been increasingly used for this purpose as it enables synthesis of data about competing interventions compared directly and indirectly in many studies, that form a network of evidence. NMA is used to estimate the relative treatment effect of each intervention versus all the others, and can produce statistical *ranking metrics* that lead to a *treatment hierarchy* from the least preferable to the most preferable option. Treatment hierarchies have been increasingly reported in published NMAs, and their use and reporting are recommended by international guidance and reporting guidelines. However, several methodological issues have been debated. For instance, the agreement between hierarchies from different ranking metrics has not been explored empirically. Methods to rank treatments for multiple outcomes, accounting for both efficacy and safety as well as individual preferences simultaneously, are underdeveloped. Also, it is unclear how to critically appraise a treatment hierarchy, since a rigorous framework to assess reporting bias in NMA is lacking.

Aim

The aim of this thesis is to fill some of these methodological gaps on the topic of ranking metrics and reporting bias in NMA. The first objective is to study the agreement between different rankings from an empirical perspective and to aid the interpretation and use of existing ranking metrics. The second objective is to extend the existing ranking methodology to account for multiple efficacy and safety outcomes, as well as specific preferences and trade-offs between benefits and harms. The third objective is to develop a methodological framework to evaluate the risk of reporting bias in network meta-analysis.

Methods

An empirical evaluation of the level of agreement between hierarchies obtained from existing ranking metrics is carried out by re-analysing over 200 previously published networks of four or more interventions. We explore how agreement is affected by the amount of information present in a network in terms of average variance, differences in the variance estimates, and the total sample size over the number of interventions of a network.

To expand on the existing ranking methodology, we combine a recently developed ranking metric, accounting for both multiple outcomes and individual preferences, with a trade-off value defining the compromise between positive and negative outcomes.

For evaluating the risk of reporting bias in NMA we combine the risk of bias due to missing evidence in pairwise comparisons with that of the network estimates. For the latter, we consider the contribution matrix, the unobserved comparisons, and the presence of small-study effects as evaluated by network meta-regression. We also develop an online web-application to facilitate this evaluation.

Results

The level of agreement between treatment hierarchies obtained by different ranking metrics can be affected by the amount of information present in a network. Differences in level of agreement become more evident when there are large imbalances in the precision of the estimates, though we find that such imbalances are rare in practice. We also developed rankings based on relative treatment effects against a fictional treatment of average performance, which are useful in networks of interventions where a natural reference treatment does not exist. We provide recommendations for reporting the treatment hierarchies obtained from different ranking metrics, avoiding misinterpretation, and properly addressing "treatment hierarchy questions" in the decision-making context.

We extended the existing ranking methodology by combining the standardised area within spie charts with different trade-offs between benefits and harms. The obtained quantity is useful to show variation in the ranking for a whole range of trade-off values and a specific set of individual preferences.

We developed the first risk of bias tool to evaluate the risk of bias due to missing evidence in NMA and we facilitate its use with a user-friendly web application that automates some of the required steps.

Conclusions

In this thesis we made significant contributions to the evidence synthesis field, providing knowledge and tools that assist clinicians, policy makers and patients in choosing the most preferable treatment for a specific condition. Our results are a step forward in the direction of actively translating knowledge into practical use, although more implementation research in clinical practice is still needed to guide decision-making processes.

Introduction

Ranking treatments in network meta-analysis

Clinicians must regularly make decisions about their patients' care, particularly with regards to the choice of treatments for a specific condition [1,2]. To make such decisions they usually refer to current clinical guidelines which are informed by the best available evidence. Guideline recommendations are normally produced using the efficacy and safety results of quantitative evidence synthesis techniques, in particular pairwise and network meta-analysis (NMA) [3–5]. The latter is an extension of pairwise meta-analysis for more than two interventions that enables synthesis of direct evidence (i.e. treatments compared directly within a study) and indirect evidence from studies comparing the treatments of interest with one or more intermediate comparators [6,7]. NMA produces all relative treatment effects of each intervention versus another and, in turn, these can be used to create a hierarchy from the most to the least preferable treatment for a specific outcome of interest.

The statistical quantity measuring the performance of an intervention on a specific outcome and producing a specific ranking is referred to as a *ranking metric*. Ranking metrics can be probabilistic or non-probabilistic according to whether the entire distribution of each estimated treatment effect is taken into account or not.

Let us imagine we have a set $\mathbb{T}$ of T competing treatments and the relative effects of treatment i over j, μ_{ij}, estimated in NMA, for a given outcome of interest, specifically a negative outcome, e.g. mortality. A hierarchy of treatments ranked based on their estimated mean relative effects against a common comparator (e.g. placebo), $\hat{\mu}_{iP}$, is indeed a non-probabilistic hierarchy as the uncertainty with which these effects are estimated is not considered in the ranking process. The distribution of μ_{ij} is estimated either as the posterior distribution in a Bayesian setting or using resampling in a frequentist setting. Several probabilistic ranking metrics have been developed and employed to date and, among the most popular, we find the probability that treatment i produces the best value ($p_{i,BV}$) for a given outcome. The best value is the smallest value (for a negative outcome) compared to all other competing treatments, so $p_{i,BV}$ is defined as

$$p_{i,BV} \coloneqq p_{i,1} = P\big(\mu_{ij} < 0 \ \forall\, j \in \mathbb{T}\big).$$

Similarly, one can calculate the probability that treatment i produces the *R*th-best value, defined as the probability that treatment i will outperform exactly $T - R$ treatments

$$p_{i,R} = \sum_{\mathbb{R}} P\left(\left(\mu_{ij} < 0 \; \forall \, j \; \in \; \mathbb{R}\right) \cap \left(\mu_{ij} \geq 0 \; \forall \, j \; \notin \; \mathbb{R}\right)\right)$$

where the sum is over all possible $\mathbb{R}$ subsets of $\mathbb{T}$, $\mathbb{R} \subset \mathbb{T}$, of size $T - R$.

Table 1 shows the ranking probabilities for a network meta-analysis investigating the effects of five antihypertensive drugs and placebo on the incidence of diabetes [8].

Table 1: Ranking probabilities, SUCRA values, mean and median ranks for the network of antihypertensive drugs and placebo effects on the incidence of diabetes. ARB = angiotensin-receptor blockers; ACE = angiotensin-converting-enzyme; CCB = calcium-channel blockers; Bblocker = Beta blocker; SUCRA = surface under the cumulative ranking curve.

	Ranks						SUCRA	Mean rank	Median rank
	1	2	3	4	5	6			
ARB	0.26	0.67	0.07	0.01	0.00	0.00	0.84	1.82	1
ACE	0.73	0.25	0.02	0.00	0.00	0.00	0.94	1.30	1
Placebo	0.00	0.00	0.00	0.01	0.81	0.19	0.16	4.07	4
CCB	0.00	0.02	0.27	0.71	0.01	0.00	0.46	3.70	3
Bblocker	0.00	0.00	0.00	0.00	0.18	0.82	0.04	0.92	5
Diuretic	0.01	0.07	0.64	0.28	0.00	0.00	0.56	3.19	2

The ranking probabilities can also be presented graphically as so-called rankograms, treatment-specific plots showing the distribution of ranking probabilities for each intervention [9]. The rankograms for the network of antihypertensive treatments is shown in Figure 1.

One downside of these ranking probabilities – particularly the $p_{i,BV}$ – is that a treatment with a high probability of producing the best value may also have a high probability of producing worse values i.e. being ranked last. To overcome this issue, one option is to calculate the cumulative probabilities that the treatment i will be ranked in the top *R* positions,

$$c_{i,R} = \sum_{r=1}^{R} p_{i,r} \; .$$

The cumulative probabilities are often presented in the cumulative probability plots and are also used to produce the surface under the cumulative ranking curve,

$$SUCRA_i = \frac{\sum_{r=1}^{T-1} c_{i,r}}{T-1} \; .$$

$SUCRA_i$ represents the area below the step function of the cumulative probability plots – the larger the area, the higher the probability that treatment i is the best-performing [9].

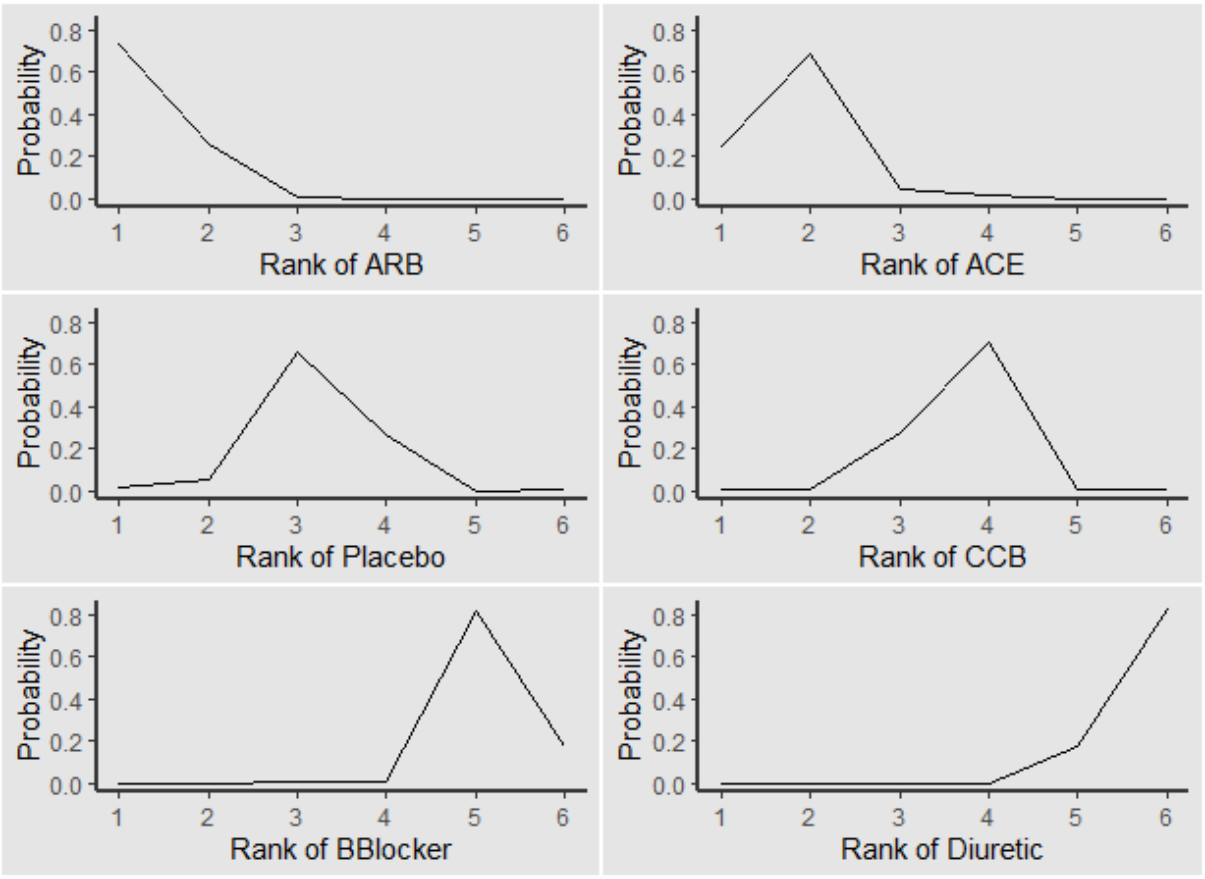


Figure 1: Rankograms for the network of antihypertensive drugs and placebo effects on the incidence of diabetes. ARB = angiotensin-receptor blockers; ACE = angiotensin-converting-enzyme; CCB = calcium-channel blockers; Bblocker = Beta blocker.

In this way, $SUCRA_i$ summarises the full information of the treatment effectiveness into a single value that can be interpreted as the average proportion of treatments worse than treatment i [10]. Cumulative probability plots for the network of antihypertensive drugs' effects on the incidence of diabetes are shown in Figure 2.

The posterior distribution of each treatment's rank can also be summarised by its mean or median, producing this way two additional ranking metrics: the mean rank $MeanR_i = \sum_{r=1}^{T-1} p_{i,r} \times r$, and the median rank, $medianR_i$, defined as the largest value satisfying $\sum_{R=1}^{medianR_i} p_{i,R} \leq \frac{1}{2}$. The SUCRA is essentially an inversely scaled transformation of the mean rank and can indeed also be defined as $SUCRA_i = \frac{T - MeanR_i}{T-1}$ [10]. SUCRA values, mean and median ranks for the network of antihypertensive drugs' effects on the incidence of diabetes are reported in Table 1.

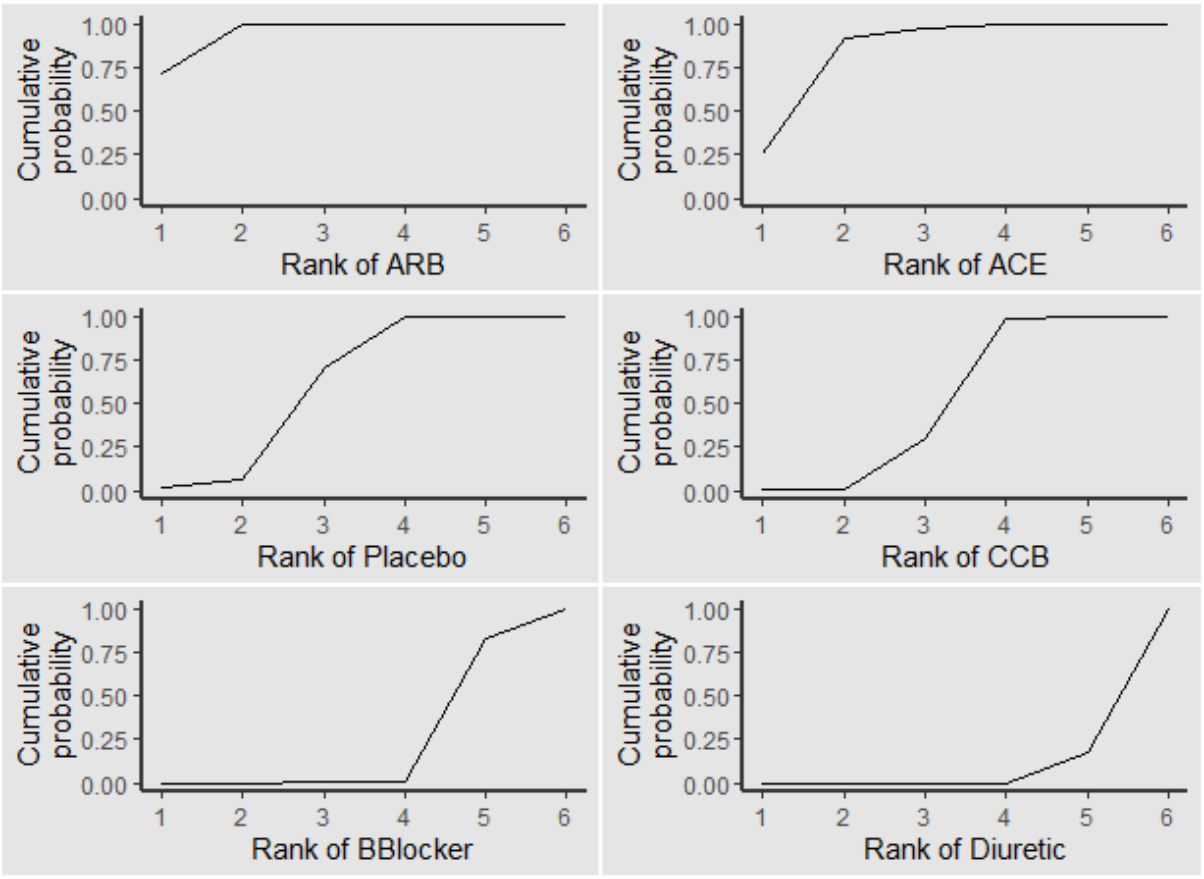


Figure 2: Cumulative probability plots for the network of antihypertensive drugs and placebo effects on the incidence of diabetes. ARB = angiotensin-receptor blockers; ACE = angiotensin-converting-enzyme; CCB = calcium-channel blockers; Bblocker = Beta blocker.

A frequentist version of the SUCRA, the *P-score*, has been developed as a function of the one-sided p-values based on the estimated mean relative treatment effects and standard error. It has been proven that the two ranking metrics are equivalent in terms of results produced and interpretation, at least when the assumption of normality for the distribution of the effects holds [10].

Treatment hierarchies have been increasingly reported in published NMAs [11] but they have not been exempt from criticism. This mainly states concerns about the instability and limited interpretability of the treatment ranks, particularly due to the fact they do not encompass uncertainty nor account for bias in the evidence [12–17]. Part of this criticism, specifically regarding uncertainty of the rank ordering, has already been addressed: while some claim that uncertainty in rankings should be reported either as confidence/credible intervals for the ranking metric or as a complete presentation of rank probabilities (e.g. using rankograms or cumulative ranking curve [14,18]), others question whether these suggestions are actually feasible and useful [15]. Reporting all rank probabilities or rankograms can prove difficult for

NMAs involving many treatments and outcomes, and uncertainty intervals may not be informative or interpretable. One reason for this argument is that ranking statistics such as SUCRA and P-score are not considered population parameters and therefore do not have a distribution. Besides, the calculation of SUCRA values already incorporates uncertainty of the treatment effects.

Salanti *et al.* also claim that some of the criticism and confusion about treatment hierarchies was inappropriate as it implicitly assessed ranking metrics as if one is better that the others, while they address different problems or, as the authors present them, different treatment hierarchy questions [19]. A *treatment hierarchy question* is based on a clear definition of "best treatment" which must be specified by the researcher in a given setting, so that the relevant ranking does not give misleading results. A bibliographic study found that the majority of network meta-analyses that presented some form of treatment hierarchy calculated the probability of producing the best value, $p_{i,BV}$, followed by SUCRA, but the choice of the ranking metric was not justified [11]. The level of agreement between rankings obtained using different metrics in practice is also unknown, as no study investigating the empirical agreement between treatment hierarchies has been carried out. While it is clear that decisions cannot be made on rankings alone, as they are no substitute for relative treatment effects and their confidence or credible intervals, international guidance and guidelines support the use and reporting of treatment hierarchies [20,21].

A major issue of treatment hierarchies in NMA is that they usually refer to a single outcome, making it difficult to summarise the performance of an intervention from both an effectiveness and a safety perspective. Furthermore, when multiple outcomes are considered, individual preferences reflecting the relative importance of different benefit and harm outcomes should also be accounted for. In healthcare decision-making, several methods for benefit-risk assessment are available that can incorporate multiple outcomes as well as patients preferences [22,23]. Probably the most popular decision analytic approach is multicriteria decision analysis (MCDA), whose objective function is a linear combination of absolute treatment effects of various outcomes that can be weighted in different ways to reflect preferences. Within the last decade, there have been some examples of combining network meta-analysis results within MCDA to rank treatment alternatives, but the use of this method in clinical practice still faces challenges and it is unclear which preference elicitation technique would be more suitable [24–26].

An ideal ranking metric should consider all these important aspects and potentially other characteristics of interest. Some extensions to the existing ranking metrics and new methods have been developed in recent years to account for multiple outcomes, individual preferences and to distinguish between clinically important and unimportant treatment effects. Mavridis *et al.* extended the P-score method to multiple outcomes and modified it so that the obtained ranking can also reflect clinical important differences [27]. Intuitive visualisation tools have also been proposed to present results for multiple outcomes such as: clustered ranking plots [28]; the Kilim plot [29], that also allow to specify clinical important values for the outcomes; and, more recently, the Vitruvian plot [30]. However, these visualisation methods do not provide a quantitative measure to rank the treatments considering jointly all outcomes of interest.

Daly *et al.* introduced the spie charts framework, which measures the effectiveness or safety of each treatment on multiple outcomes [31]. The outcome measures are plotted on a treatment-specific spie chart as sectors, whose angles represent the importance of each outcome. The area inside a spie chart represents the quantity by which to rank the treatments but the authors recommend not to plot benefit and harms outcomes on the same spie chart as that could mask important safety information.

Another measure, the *Probability Of Selecting a Treatment to Recommend* (POST-R) was proposed to account for other important factors which are often of interest for treatment guidelines but can be incorporated only in a qualitative manner [32]. One of these characteristics is the confidence in the evidence, or credibility of the network meta-analysis results, which can be evaluated using the Confidence in Network Meta-Analysis (CINeMA) framework [33] or the Grading of Recommendations Assessment, Development and Evaluation (GRADE) approach [34]. These evaluations consider several domains, including the risk of bias, to assign to each treatment comparison four possible ratings of confidence, high, moderate, low or very low, which are usually summarised in a table, as recommended by reporting guidelines [20,35]. POST-R, which uses a Markov chain model, can incorporate this information by translating it into initial probabilities as prior probabilities, and subsequently obtain probabilities of recommending each treatment through its stationary probability distribution [32]. However, this method can combine one outcome (i.e. efficacy) with only one of the other characteristics.

Bias due to missing evidence

The above-mentioned CINeMA framework has been increasingly used to evaluate confidence in network meta-analysis results and it consists of six domains: within-study bias, reporting bias, indirectness, imprecision, heterogeneity, and incoherence. However, in its first version, the CINeMA domain for reporting bias was underdeveloped compared to the other domains, mainly due to the lack of a rigorous methodology to evaluate it. Indeed, in agreement with the GRADE approach [34], two possible judgements – suspected and undetected – are assigned to each network meta-analysis estimate. However, the current documentation for CINeMA suggests basing the reporting bias assessment only on qualitative conditions which include: previous evidence indicating the presence of reporting bias; the inclusion of data from grey literature and unpublished sources; and the characteristics of the treatment comparison, e.g. whether it involves new versus old drugs or the studies investigating are primarily industry-funded. Therefore, no specific guidance is presented to evaluate reporting bias in the network meta-analysis context.

Reporting bias arises if the non-reporting of results is related to the nature of the results. It can occur when a study is not reported at all, commonly referred to as *publication bias*, or when some results are not reported, usually known as *outcome reporting bias* or *selective non-reporting of results* [36]. In both cases, the data included in a meta-analysis differ systematically from the missing results, which threatens the validity of the meta-analysis conclusions [37,38]. In pairwise meta-analysis, reporting bias has been extensively studied, and several methods have been developed to investigate the corresponding risk of bias [39]. Approaches include comparisons of study protocols with published reports to identify outcomes measured but not reported and comparing results obtained from published versus unpublished sources [39]. Other approaches include graphical methods (e.g. funnel plots [40–42]), tests for systematic differences between effects in smaller versus larger studies ("small-study effects", e.g. Egger's test and its counterparts [40,43–45]), regression-based adjustment methods, and selection models [46–49].

Selection models account for the mechanism (i.e. the selection process) with which studies are selected for publication. In particular, in the Copas selection models, this process is defined by a latent variable describing the "propensity of publication" which is correlated with the study effect size and is assumed as a function of the study variance through a regression model [46]. However, since the number of unpublished studies is unknown, the model for the

propensity for publication is difficult to identify. Therefore, a sensitivity analysis is usually required to compute the pooled intervention effect and the probability of a study to be published under various possible assumptions about the severity of selection bias. Regression models describing treatment effects as a function of the standard error [50,51] and limit meta-analysis [52] do not require such assumptions and may therefore performed better than bias-adjustment methods [53].

Several of the numerical approaches to evaluate reporting bias developed for pairwise meta-analysis have been adapted to the network meta-analysis setting [54,55], such as the comparison-adjusted funnel plot [28] and the extension of the Copas selection model [56,57]. However, these methods have limitations and are only meaningful when specific assumptions can be made, such as the characteristics associated with small-study effects in the comparison-adjusted funnel plots and the direction of publication bias for the Copas model in a network of interventions. Also, the model has not been used frequently in practice, probably due its complexity compared to other models and techniques. Most importantly, the numerical evaluation plays only a small role when judging the presence of bias. Comprehensive searches and qualitative approaches are key to preventing and identifying potential biases due to missing results [39].

The Risk of Bias due to Missing Evidence (ROB-ME) tool, in its preliminary version, has been recently developed to integrate all these considerations into a structured framework for the overall assessment of risk of bias due to missing evidence in meta-analysis [58]. However, ROB-ME only applies to pairwise meta-analysis as it was not designed to encompass the added complexities of NMA such as the role of indirect evidence and the contribution of the direct comparisons to each network estimate.

Hypothesis and aim of the thesis

This PhD thesis is a cumulative work of publications set to address the unanswered questions on the topic of ranking metrics and reporting bias in network meta-analysis. Specifically:

- the existing literature lacks empirical studies on the level of agreement between treatment hierarchies using different ranking metrics;
- there is a need to produce a hierarchy of treatments in NMA that account for multiple efficacy and safety outcomes, as well as other important aspects such as specific preferences and trade-offs between benefits and harms;
- the methodology for assessing reporting bias in a network of interventions is incomplete.

Therefore, the overall aim of the thesis is to provide answers to these questions with three projects set to achieve the specific objectives:

- the first project aims to study the agreement between different rankings from an empirical perspective (Articles 1 and 2) to aid the interpretation and use of existing ranking metrics (Article 3);
- the objective of the second project is to extend the existing ranking metrics (Article 4) to address complex treatment hierarchy problems;
- the third project aims to provide a methodological framework to assess the risk of bias due to missing evidence in the context of a network of interventions (Article 5).

Article 1: Agreement between ranking metrics in network meta-analysis: an empirical study

This article presents the results from an empirical evaluation of the level of agreement between hierarchies obtained from existing ranking metrics by reanalysing over 200 published networks of four or more interventions. We hypothesised that some network features could influence this agreement and show how it is affected by the amount of information present in a network in terms of average variance, differences in the variance estimates, and the total sample size over the number of interventions of a network.

Article 2: Network meta-analysis results against a fictional treatment of average performance: treatment effects and ranking metric

In this article, we introduce a new ranking metric, defined as the probability that a treatment is better than a fictional treatment of average performance by using an alternative parameterisation of the NMA model. Using the methodology employed in Article 1, we also compare the hierarchies obtained with the new ranking metric with those obtained from the existing ranking metrics.

Article 3: The complexity underlying treatment rankings: how to use them and what to look at

This invited article aims to guide in interpreting and presenting the treatment hierarchies obtained from different ranking metrics by avoiding common mistakes and considering the whole output of a network meta-analysis, including the quality of evidence. The article also refers to the recently introduced concept of treatment hierarchy question, which is based on a specific definition of "best treatment", and how to address it in the decision-making context.

Article 4:

This article presents a numerical quantity to explore the changes in rankings with different trade-offs between benefits and harms of treatments and a specific set of preferences. We extend the concept of the standardised area within a spie chart by combining it with a trade-off value defining the compromise between positive and negative outcomes. We illustrate the method as a sensitivity analysis by demonstrating how the rankings for benefits and harms of three real network examples varies for the whole range of the trade-off values.

Article 5: ROB-MEN: a tool to assess risk of bias due to missing evidence in network meta-analysis

We aim to develop the first tool for the evaluation of risk of bias due to missing evidence in network meta-analysis (ROB-MEN). We illustrate the methodology of the framework underlying the tool, that combines the risk of bias due to missing evidence in pairwise comparisons with that of the network estimates by considering the contribution matrix, the unobserved comparisons, and the presence of small-study effects as evaluated by network meta-regression. We also introduce the online web-application that facilitates and semi-

automates some of the tool assessments and produces the output table. We present the final tool with an application in two published network meta-analysis.

References Introduction

1 Del Fiol G, Workman TE, Gorman PN. Clinical questions raised by clinicians at the point of care: a systematic review. *JAMA Intern Med* 2014;**174**:710–8. doi:10.1001/jamainternmed.2014.368

2 Ely JW. A taxonomy of generic clinical questions: classification study. *BMJ* 2000;**321**:429–32. doi:10.1136/bmj.321.7258.429

3 Dias S, Ades AE, Welton NJ, *et al. Network meta-analysis for decision making*. Hoboken, NJ: : Wiley 2018.

4 Wells GA, Sultan SA, Chen L, *et al.* Indirect Evidence: Indirect Treatment Comparisons in Meta-Analysis. Ottawa: : Canadian Agency for Drugs and Technologies in Health 2009. https://www.cadth.ca/indirect-evidence-indirect-treatment-comparisons-meta-analysis

5 National Institute for Health and Care Excellence. Guide to the methods of technology appraisal 2013. London: : National Institute for Health and Clinical Excellence 2013.

6 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**:80–97. doi:10.1002/jrsm.1037

7 Salanti G, Higgins JP, Ades A, *et al.* Evaluation of networks of randomized trials. *Stat Methods Med Res* 2008;**17**:279–301. doi:10.1177/0962280207080643

8 Elliott WJ, Meyer PM. Incident diabetes in clinical trials of antihypertensive drugs: a network meta-analysis. *The Lancet* 2007;**369**:201–7. doi:10.1016/S0140-6736(07)60108-1

9 Salanti G, Ades AE, Ioannidis JPA. Graphical methods and numerical summaries for presenting results from multiple-treatment meta-analysis: an overview and tutorial. *Journal of Clinical Epidemiology* 2011;**64**:163–71. doi:10.1016/j.jclinepi.2010.03.016

10 Rücker G, Schwarzer G. Ranking treatments in frequentist network meta-analysis works without resampling methods. *BMC Medical Research Methodology* 2015;**15**:58. doi:10.1186/s12874-015-0060-8

11 Petropoulou M, Nikolakopoulou A, Veroniki A-A, *et al.* Bibliographic study showed improving statistical methodology of network meta-analyses published between 1999 and 2015. *Journal of Clinical Epidemiology* 2017;**82**:20–8. doi:10.1016/j.jclinepi.2016.11.002

12 Kibret T, Richer D, Beyene J. Bias in identification of the best treatment in a Bayesian network meta-analysis for binary outcome: a simulation study. *Clinical Epidemiology* 2014;:451. doi:10.2147/CLEP.S69660

13 Mills EJ, Kanters S, Thorlund K, *et al.* The effects of excluding treatments from network meta-analyses: survey. *BMJ* 2013;**347**:f5195.

14 Trinquart L, Attiche N, Bafeta A, *et al.* Uncertainty in Treatment Rankings: Reanalysis of Network Meta-analyses of Randomized Trials. *Ann Intern Med* 2016;**164**:666–73. doi:10.7326/M15-2521

15 Veroniki AA, Straus SE, Rücker G, *et al.* Is providing uncertainty intervals in treatment ranking helpful in a network meta-analysis? *J Clin Epidemiol* 2018;**100**:122–9. doi:10.1016/j.jclinepi.2018.02.009

16 Wang Z, Carter RE. Ranking of the most effective treatments for cardiovascular disease using SUCRA: Is it as sweet as it appears? *Eur J Prev Cardiolog* 2018;**25**:842–3. doi:10.1177/2047487318767199

17 Mbuagbaw L, Rochwerg B, Jaeschke R, *et al.* Approaches to interpreting and choosing the best treatments in network meta-analyses. *Syst Rev* 2017;**6**:79. doi:10.1186/s13643-017-0473-z

18 Bafeta A, Trinquart L, Seror R, *et al.* Reporting of results from network meta-analyses: methodological systematic review. *BMJ* 2014;**348**:g1741–g1741. doi:10.1136/bmj.g1741

19 Salanti G, Nikolakopoulou A, Efthimiou O, *et al.* Introducing the Treatment Hierarchy Question in Network Meta-Analysis. *American Journal of Epidemiology* 2022;**191**:930–8. doi:10.1093/aje/kwab278

20 Hutton B, Salanti G, Caldwell DM, *et al.* The PRISMA extension statement for reporting of systematic reviews incorporating network meta-analyses of health care interventions: checklist and explanations. *Ann Intern Med* 2015;**162**:777–84. doi:10.7326/M14-2385

21 Jansen JP, Trikalinos T, Cappelleri JC, *et al.* Indirect Treatment Comparison/Network Meta-Analysis Study Questionnaire to Assess Relevance and Credibility to Inform Health Care Decision Making: An ISPOR-AMCP-NPC Good Practice Task Force Report. *Value in Health* 2014;**17**:157–73. doi:10.1016/j.jval.2014.01.004

22 Puhan MA, Singh S, Weiss CO, *et al.* A framework for organizing and selecting quantitative approaches for benefit-harm assessment. *BMC Med Res Methodol* 2012;**12**:173. doi:10.1186/1471-2288-12-173

23 Najafzadeh M, Schneeweiss S, Choudhry N, *et al.* A Unified Framework for Classification of Methods for Benefit-Risk Assessment. *Value in Health* 2015;**18**:250–9. doi:10.1016/j.jval.2014.11.001

24 van Valkenhoef G, Tervonen T, Zhao J, *et al.* Multicriteria benefit–risk assessment using network meta-analysis. *Journal of Clinical Epidemiology* 2012;**65**:394–403. doi:10.1016/j.jclinepi.2011.09.005

25 Tervonen T, Naci H, van Valkenhoef G, *et al.* Applying Multiple Criteria Decision Analysis to Comparative Benefit-Risk Assessment: Choosing among Statins in Primary Prevention. *Medical Decision Making* 2015;**35**:859–71. doi:10.1177/0272989X15587005

26 Naci H, van Valkenhoef G, Higgins JPT, *et al.* Evidence-Based Prescribing: Combining Network Meta-Analysis With Multicriteria Decision Analysis to Choose Among Multiple Drugs. *Circ Cardiovasc Qual Outcomes* 2014;**7**:787–92. doi:10.1161/CIRCOUTCOMES.114.000825

27 Mavridis D, Porcher R, Nikolakopoulou A, *et al.* Extensions of the probabilistic ranking metrics of competing treatments in network meta-analysis to reflect clinically important relative differences on many outcomes. *Biom J* 2019;:bimj.201900026. doi:10.1002/bimj.201900026

28 Chaimani A, Higgins JPT, Mavridis D, *et al.* Graphical Tools for Network Meta-Analysis in STATA. *PLoS ONE* 2013;**8**:e76654. doi:10.1371/journal.pone.0076654

29 Seo M, Furukawa TA, Veroniki AA, *et al.* The Kilim plot: A tool for visualizing network meta-analysis results for multiple outcomes. *Research Synthesis Methods* 2021;**12**:86–95. doi:10.1002/jrsm.1428

30 Ostinelli EG, Efthimiou O, Naci H, *et al.* Vitruvian plot: a visualisation tool for multiple outcomes in network meta-analysis. *Evid Based Mental Health* 2022;:ebmental-2022-300457. doi:10.1136/ebmental-2022-300457

31 Daly CH, Mbuagbaw L, Thabane L, *et al.* Spie charts for quantifying treatment effectiveness and safety in multiple outcome network meta-analysis: a proof-of-concept study. *BMC Med Res Methodol* 2020;**20**:266. doi:10.1186/s12874-020-01128-2

32 Chaimani A, Porcher R, Sbidian E, *et al.* A Markov Chain approach for ranking treatments in network meta-analysis. Epidemiology 2019. doi:10.1101/19008722

33 Nikolakopoulou A, Higgins JPT, Papakonstantinou T, *et al.* CINeMA: An approach for assessing confidence in the results of a network meta-analysis. *PLOS Medicine* 2020;**17**:e1003082. doi:10.1371/journal.pmed.1003082

34 Puhan MA, Schunemann HJ, Murad MH, *et al.* A GRADE Working Group approach for rating the quality of treatment effect estimates from network meta-analysis. *BMJ* 2014;**349**:g5630–g5630. doi:10.1136/bmj.g5630

35 Chaimani A, Caldwell DM, Li T, *et al.* Additional considerations are required when preparing a protocol for a systematic review with multiple interventions. *Journal of Clinical Epidemiology* 2017;**83**:65–74. doi:10.1016/j.jclinepi.2016.11.015

36 Page MJ, Sterne JAC, Higgins JPT, *et al.* Investigating and dealing with publication bias and other reporting biases in meta-analyses of health research: a review. *Res Syn Meth* 2020;:jrsm.1468. doi:10.1002/jrsm.1468

37 Thornton A, Lee P. Publication bias in meta-analysis: its causes and consequences. *Journal of Clinical Epidemiology* 2000;**53**:207–16. doi:10.1016/S0895-4356(99)00161-4

38 Sutton AJ, Song F, Gilbody SM, *et al.* Modelling publication bias in meta-analysis: a review. *Stat Methods Med Res* 2000;**9**:421–45. doi:10.1177/096228020000900503

39 Page MJ, Higgins JP, Sterne JA. Chapter 13: Assessing risk of bias due to missing results in a synthesis. In: *Cochrane Handbook for Systematic Reviews of Interventions*. Cochrane 2019. www.training.cochrane.org/handbook

40 Egger M, Smith GD, Schneider M, *et al.* Bias in meta-analysis detected by a simple, graphical test. *BMJ* 1997;**315**:629–34. doi:10.1136/bmj.315.7109.629

41 Sterne JAC, Egger M. Funnel plots for detecting bias in meta-analysis: Guidelines on choice of axis. *Journal of Clinical Epidemiology* 2001;**54**:1046–55. doi:10.1016/S0895-4356(01)00377-8

42 Peters JL, Sutton AJ, Jones DR, *et al.* Contour-enhanced meta-analysis funnel plots help distinguish publication bias from other causes of asymmetry. *Journal of Clinical Epidemiology* 2008;**61**:991–6. doi:10.1016/j.jclinepi.2007.11.010

43 Sterne JAC, Sutton AJ, Ioannidis JPA, *et al.* Recommendations for examining and interpreting funnel plot asymmetry in meta-analyses of randomised controlled trials. *BMJ* 2011;**343**:d4002–d4002. doi:10.1136/bmj.d4002

44 Harbord RM, Egger M, Sterne JAC. A modified test for small-study effects in meta-analyses of controlled trials with binary endpoints. *Statistics in Medicine* 2006;**25**:3443–57. doi:https://doi.org/10.1002/sim.2380

45 Peters JL. Comparison of Two Methods to Detect Publication Bias in Meta-analysis. *JAMA* 2006;**295**:676. doi:10.1001/jama.295.6.676

46 Copas JB, Shi JQ. A sensitivity analysis for publication bias in systematic reviews. *Stat Methods Med Res* 2001;**10**:251–65. doi:10.1177/096228020101000402

47 McShane BB, Böckenholt U, Hansen KT. Adjusting for Publication Bias in Meta-Analysis: An Evaluation of Selection Methods and Some Cautionary Notes. *Perspectives on Psychological Science* Published Online First: 29 September 2016. doi:10.1177/1745691616662243

48 Copas J. What Works?: Selectivity Models and Meta-Analysis. *Journal of the Royal Statistical Society Series A (Statistics in Society)* 1999;**162**:95–109.

49 Copas J, Shi JQ. Meta-analysis, funnel plots and sensitivity analysis. *Biostatistics* 2000;**1**:247–62. doi:10.1093/biostatistics/1.3.247

50 Moreno SG, Sutton AJ, Ades A, *et al.* Assessment of regression-based methods to adjust for publication bias through a comprehensive simulation study. *BMC Med Res Methodol* 2009;**9**:2. doi:10.1186/1471-2288-9-2

51 Moreno SG, Sutton AJ, Turner EH, *et al.* Novel methods to deal with publication biases: secondary analysis of antidepressant trials in the FDA trial registry database and related journal publications. *BMJ* 2009;**339**:b2981–b2981. doi:10.1136/bmj.b2981

52 Rucker G, Schwarzer G, Carpenter JR, *et al.* Treatment-effect estimates adjusted for small-study effects via a limit meta-analysis. *Biostatistics* 2011;**12**:122–42. doi:10.1093/biostatistics/kxq046

53 Rücker G, Carpenter JR, Schwarzer G. Detecting and adjusting for small-study effects in meta-analysis. *Biom J* 2011;**53**:351–68. doi:10.1002/bimj.201000151

54 Chaimani A, Salanti G. Using network meta-analysis to evaluate the existence of small-study effects in a network of interventions. *Res Syn Meth* 2012;**3**:161–76. doi:10.1002/jrsm.57

55 Moreno SG, Sutton AJ, Ades AE, *et al.* Adjusting for publication biases across similar interventions performed well when compared with gold standard data. *Journal of Clinical Epidemiology* 2011;**64**:1230–41. doi:10.1016/j.jclinepi.2011.01.009

56 Mavridis D, Sutton A, Cipriani A, *et al.* A fully Bayesian application of the Copas selection model for publication bias extended to network meta-analysis. *Statist Med* 2013;**32**:51–66. doi:10.1002/sim.5494

57 Mavridis D, Welton NJ, Sutton A, *et al.* A selection model for accounting for publication bias in a full network meta-analysis. *Statistics in Medicine* 2014;**33**:5399–412. doi:10.1002/sim.6321

58 Risk of bias tools - ROB-ME tool. https://riskofbias.info/welcome/rob-me-tool (accessed 13 Nov 2020).

Article 1: Agreement between ranking metrics in network meta-analysis: an empirical study

Virginia Chiocchia[1,2], Adriani Nikolakopoulou[1], Theodoros Papakonstantinou[1], Matthias Egger[1], Georgia Salanti[1]

[1] Institute of Social and Preventive Medicine (ISPM), University of Bern, Bern, Switzerland

[2] Graduate School of Health Sciences (GHS), University of Bern, Switzerland

My contribution:

I had the main responsibility in drafting the manuscript, performing all analyses, doing revisions after the review from co-authors and peer-review from BMJ Open.

Published: Chiocchia V, Nikolakopoulou A, Papakonstantinou T*, et al.* Agreement between ranking metrics in network meta-analysis: an empirical study. *BMJ Open* 2020;10:e037744. doi: 10.1136/bmjopen-2020-037744

ABSTRACT

Objective

To empirically explore the level of agreement of the treatment hierarchies from different ranking metrics in network meta-analysis (NMA) and to investigate how network characteristics influence the agreement.

Design

Empirical evaluation from re-analysis of network meta-analyses.

Data

232 networks of four or more interventions from randomised controlled trials, published between 1999 and 2015.

Methods

We calculated treatment hierarchies from several ranking metrics: relative treatment effects, probability of producing the best value (p_{BV}) and the surface under the cumulative ranking curve (SUCRA). We estimated the level of agreement between the treatment hierarchies using different measures: Kendall's τ and Spearman's ρ correlation; and the Yilmaz τ_{AP} and Average Overlap, to give more weight to the top of the rankings. Finally, we assessed how the amount of the information present in a network affects the agreement between treatment hierarchies, using the average variance, the relative range of variance, and the total sample size over the number of interventions of a network.

Results

Overall, the pairwise agreement was high for all treatment hierarchies obtained by the different ranking metrics. The highest agreement was observed between SUCRA and the relative treatment effect for both correlation and top-weighted measures whose medians were all equal to one. The agreement between rankings decreased for networks with less precise estimates and the hierarchies obtained from p_{BV} appeared to be the most sensitive to large differences in the variance estimates. However, such large differences were rare.

Conclusions

Different ranking metrics address different treatment hierarchy problems, however they produced similar rankings in the published networks. Researchers reporting NMA results can use the ranking metric they prefer, unless there are imprecise estimates or large imbalances

in the variance estimates. In this case treatment hierarchies based on both probabilistic and non-probabilistic ranking metrics should be presented.

Strength and limitations of this study

- To our knowledge, this is the first empirical study exploring the level of agreement of the treatment hierarchies from different ranking metrics in network meta-analysis (NMA).
- The study also explores how agreement is influenced by network characteristics.
- More than 200 published NMAs were re-analysed and three different ranking metrics calculated using both frequentist and Bayesian approaches.
- Other potential factors not investigated in this study could influence the agreement between hierarchies.

Introduction

Network meta-analysis (NMA) is being increasingly used by policy makers and clinicians to answer one of the key questions in medical decision-making: "what treatment works best for the given condition?" [1,2]. The relative treatment effects, estimated in NMA, can be used to produce ranking metrics: statistical quantities measuring the performance of an intervention on the studied outcomes, thus producing a treatment hierarchy from the most preferable to the least preferable option [3,4].

Despite the importance of treatment hierarchies in evidence-based decision making, various methodological issues related to the ranking metrics have been contested [5–7]. This ongoing methodological debate focuses on the uncertainty and bias in a single ranking metric. Hierarchies produced by different ranking metrics are not expected to agree because ranking metrics differ. For example, a *non-probabilistic ranking metric* such as the treatment effect against a common comparator considers only the mean effect (e.g. the point estimate of the odds-ratio) and ignores the uncertainty with which this is estimated. In contrast, the probability that a treatment achieves a specific rank (a *probabilistic ranking metric*) considers the entire estimated distribution of each treatment effect. However, it is important to understand why and how rankings based on different metrics differ.

There are network characteristics that are expected to influence the agreement of treatment hierarchies from different ranking metrics, such as the precision of the included studies and their distribution across treatment comparisons [4,8]. Larger imbalances in precision in the estimation of the treatment effects affects the agreement of the treatment hierarchies from probabilistic ranking metrics, but it is currently unknown whether in practice these imbalances occur and whether they should inform the choice between different ranking metrics. To our knowledge, no empirical studies have explored the level of agreement of treatment hierarchies obtained from different ranking metrics, or examined the network characteristics likely to influence the level of agreement. Here, we empirically evaluated the level of agreement between ranking metrics and examined how the agreement is affected by network features. The article first describes the methods for the calculation of ranking metrics and of specific measures to assess the agreement and to explore factors that affects it, respectively. Then, a network featuring one of the explored factors is shown as an illustrative example to display differences in treatment hierarchies from different ranking metrics.

Finally, we present the results from the empirical evaluation and discuss their implications for researchers undertaking network meta-analysis.

Methods

Data

We re-analysed networks of randomised controlled trials from a database of articles published between 1999 and 2015, including at least 4 treatments; details about the search strategy and inclusion/exclusion criteria can be found in [9,10]. We selected networks reporting arm-level data for binary or continuous outcomes. The database is accessible in the *nmadb* R package [11].

Re-analysis and calculation of ranking metrics

All networks were re-analysed using the relative treatment effect that the original publication used: odds ratio (OR), risk ratio (RR), standardised mean difference (SMD) or mean difference (MD). We estimated relative effects between treatments using a frequentist random-effects NMA model using the *netmeta* R package [12]. For the networks reporting ORs and SMDs we re-analysed them also using Bayesian models using self-programmed NMA routines in JAGS (https://github.com/esm-ispm-unibe-ch/NMAJags). To obtain probabilistic ranking metrics in a frequentist setting, we used parametric bootstrap by producing 1000 datasets from the estimated relative effects and their variance-covariance matrix. By averaging over the number of simulated relative effects we derived the *probability of treatment i to produce the best value*

$$p_{i,BV} := p_{i,1} = P\left(\mu_{ij} > 0 \ \forall j \in \mathbb{T}\right)$$

where μ_{ij} is the estimated mean relative effect of treatment i against treatment j out of a set $\mathbb{T}$ of T competing treatments. We will refer to this as p_{BV}. This ranking metric indicates how likely a treatment is to produce the largest values for an outcome (or smallest value, if the outcome is harmful). We also calculated the surface under the cumulative ranking curve ($SUCRA^F$) [3]

$$SUCRA_i = \frac{\sum_{r=1}^{T-1} c_{i,r}}{T-1}$$

where $c_{i,r} = \sum_{v=1}^{r} p_{i,v}$ are the cumulative probabilities that treatment i will produce an outcome that is among the r best values (or that it outperforms $T-r$ treatments). SUCRA, unlike p_{BV}, also considers the probability of a treatment to produce unfavourable outcome values. Therefore, the treatment with the largest SUCRA value represents the one that

outperforms the competing treatments in the network, meaning that overall it produces preferable outcomes compared to the others. We also obtained SUCRAs within a Bayesian framework ($SUCRA^B$).

To obtain the non-probabilistic ranking metric we fitted an NMA model and estimated related treatment effects. To obtain estimates for all treatments we reparametrize the NMA model so that each treatment is compared to a fictional treatment of average performance [13,14]. The estimated relative effects against a fictional treatment F of average efficacy $\hat{\mu}_{iF}$ represent the ranking metric and the corresponding hierarchy is obtained simply by ordering the effects from the largest to the smallest (or in ascending order, if the outcome is harmful). The resulting hierarchy is identical to that obtained using relative effects from the conventional NMA model, irrespective of the reference treatment. In the rest of the manuscript, we will refer to this ranking metric simply as relative treatment effect.

Agreement between ranking metrics

To estimate the level of agreement between the treatment hierarchies obtained using the three chosen ranking methods we employed several correlation and similarity measures.

To assess the correlation between ranking metrics we used Kendall's τ [15] and the Spearman's ρ [16]. Both Kendall's τ and Spearman's ρ give the same weight to each item in the ranking. In the context of treatment ranking, the top of the ranking is more important than the bottom. We therefore also used a top-weighted variant of Kendall's τ, Yilmaz τ_{AP} [17], which is based on a probabilistic interpretation of the average precision measure used in information retrieval [18] (see online supplementary Appendix).

The measures described so far can only be considered for conjoint rankings, i.e. for lists where each item in one list is also present in the other list. Rankings are non-conjoint when a ranking is truncated to a certain depth k with such lists called top-k rankings. We calculated the Average Overlap [19,20], a top-weighted measure for top-k rankings that considers the cumulative intersection (or overlap) between the two lists and averages it over a specified depth (cut-off point) k (see online supplementary Appendix for details). We calculated the Average Overlap between pairs of rankings for networks with at least six treatments (139 networks) for a depth k equal to half the number of treatments in the network, $k = {}^{T}/_{2}$ (or $((T-1))\ /\ 2$ if T is an odd number).

We calculated the four measures described above to assess the pairwise agreement between the three ranking metrics within the frequentist setting and summarised them for each pair of ranking metrics and each agreement measure using the median and the 1st and 3rd quartiles. The hierarchy according to $SUCRA^B$ was compared to that of its frequentist equivalent to check how often the two disagree.

Influence of network features on the rankings agreement

The main network characteristic considered was the amount of information in the network (reflected in the precision of the estimates). Therefore, for each network we calculated the following measures of information:

the average variance, calculated as the mean of the variances of the estimated treatment effects $mean(SE^2)$, to show how much information is present in a network altogether;

the relative range of variance, calculated as $\frac{\max SE^2 - \min SE^2}{\max SE^2}$, to describe differences in information about each intervention within the same networks;

the total sample size of a network over the number of interventions.

These measures are presented in scatter plots against the agreement measurements for pairs of ranking metrics.

All the codes for the empirical evaluation are available at https://github.com/esm-ispm-unibe-ch/rankingagreement.

Patient and public involvement

Patients and the public were not involved in this study.

Illustrative example

To illustrate the impact of the amount of information on the treatment hierarchies from different ranking metrics, we used a network of nine antihypertensive treatments for primary prevention of cardiovascular disease that presents large differences in the precision of the estimates of overall mortality [21]. The network graph and forest plot of relative treatment effects of each treatment versus placebo are presented in **Figure 1**. The relative treatment effects reported are risk ratios (RR) estimated using a random effects NMA model.

Table 1 shows the treatment hierarchies obtained using the three ranking metrics described above. The highest overall agreement is between hierarchies from the $SUCRA^F$ and the relative treatment effect as shown by both correlation (Spearman's ρ = 0.93, Kendall's τ = 0.87) and top-weighted measures (Yilmaz's τ_{AP}= 0.87; Average Overlap = 0.85). The level of

agreement decreases when $SUCRA^F$ and the relative treatment effect are compared with p_{BV} rankings (Spearman's ρ = 0.63 and ρ = 0.85 respectively). Agreement with p_{BV} especially decreases when considering top ranks only (Average Overlap is 0.48 for p_{BV} versus $SUCRA^F$ and 0.54 for p_{BV} versus relative treatment effect). All agreement measures are presented in online supplementary **Table S1**.

The reason for this disagreement is explained by the differences in precision in the estimated effects (**Figure 1**). These RRs versus placebo range from 0.82 (Diuretic/Beta-blocker versus placebo) to 0.98 (Beta-blocker versus placebo). All estimates are fairly precise except for the RR of conventional therapy versus placebo whose 95% confidence interval extends from 0.21 to 3.44. This uncertainty in the estimation is due to the fact that conventional therapy is compared only with Angiotensin Receptor Blockers (ARB) via a single study. This large difference in the precision of the estimation of the treatment effects mostly affects the p_{BV} ranking, which disagrees the most with both of the other rankings. Consequently, the Conventional therapy is in the first rank in the p_{BV} hierarchy (because of the large uncertainty) but only features in the third/fourth and sixth rank using the relative treatment effects and $SUCRA^F$ hierarchies, respectively.

To explore how the hierarchies for this network would change in case of increased precision, we reduced the standard error of the Conventional versus ARB treatment effect from the original 0.7 to a fictional value of 0.01 resulting in a confidence interval 0.77 to 0.96. The columns in the right-hand side of **Table 1** display the three equivalent rankings after the standard error reduction. The conventional treatment has moved up in the hierarchy according to $SUCRA^F$ and moved down in the one based on p_{BV}, as expected. The treatment hierarchies obtained from the $SUCRA^F$ and the relative treatment effect are now identical (Conventional and ARB share the 3.5 rank because they have the same effect estimate) and the agreement with the p_{BV} rankings also improved (p_{BV} versus $SUCRA^F$ Spearman's ρ = 0.89, Average Overlap = 0.85; p_{BV} versus relative treatment effect Spearman's ρ = 0.91, Average Overlap = 0.94; online supplementary **Table S1**).

Results

A total of 232 networks were included in our dataset. Their characteristics are shown in **Table 2**. The majority of networks (133 NMAs, 57.3%) did not report any ranking metrics in the original publication. Among those which used a ranking metric to produce a treatment

hierarchy, the probability of being the best was the most popular metric followed by the SUCRA with 35.8% and 6.9% of networks reporting them, respectively.

Table 3 presents the medians and quartiles for each similarity measures. All hierarchies showed a high level of pairwise agreement, although the hierarchies obtained from the $SUCRA^F$ and the relative treatment effect presented the highest values for both unweighted and with top-weighted measures (all measures' median equals 1). Only 4 networks (less than 2%) had a Spearman's correlation between $SUCRA^F$ and the relative treatment effect less than 90% (not reported). The correlation becomes less between the p_{BV} rankings and those obtained from the other two ranking metrics with Spearman's ρ median decreasing to 0.9 and Kendall's τ decreasing to 0.8. The Spearman's correlation between these rankings was less than 90% in about 50% of the networks (in 116 and 111 networks for p_{BV} versus $SUCRA^F$ and p_{BV} versus relative effect, respectively; results not reported). The pairwise agreement between the p_{BV} rankings and the other rankings also decreased when considering only top ranks (p_{BV} versus $SUCRA^F$ Yilmaz's τ_{AP} = 0.77, Average Overlap = 0.83; p_{BV} versus relative treatment effect Yilmaz's τ_{AP} = 0.79, Average Overlap = 0.88).

The SUCRAs from frequentist and Bayesian settings ($SUCRA^F$ and $SUCRA^B$) were compared in 126 networks (82 networks using the Average Overlap measure) as these reported OR and SMD as original measures. The relevant rankings do not differ much as shown by the median values of the agreement measures all equal to 1 and their narrow interquartile ranges (**Table 3**). Nevertheless, a few networks showed a much lower agreement between the two SUCRAs. These networks provide posterior effect estimates for which the Normal approximation is not optimal, some of which due to rare outcomes. Such cases were however uncommon as in only 6% of the networks the Spearman's correlation between $SUCRA^F$ and $SUCRA^B$ was less than 90%. Plots for the Normal distributions from the frequentist setting and the posterior distributions of the log odds-ratios (LOR) for a network with a Spearman's ρ of 0.6 between the two SUCRAs is available in online supplementary **Figure S1** [22].

Figure 2 presents how Spearman's ρ and the Average Overlap vary with the average variance of the relative treatment effect estimates in a network (scatter plots for the Kendall's τ and the Yilmaz's τ_{AP} are available in online supplementary **Figure S2**). The treatment hierarchies agree more in networks with more precise estimates (left hand side of the plots).

The association between Spearman's ρ or Average Overlap and the relative range of variance in a network (here transformed to a double logarithm of the inverse values) are displayed in

Figure 3. On the right-hand side of each plot we can find networks with smaller differences in the precision of the treatment effect estimates. Treatment hierarchies for these networks show a larger agreement than for those with larger differences in precision. The plots of the impact of the relative range of variance on all measures are available in online supplementary **Figure S3**.

The total sample size in a network over the number of interventions has a similar impact on the level of agreement between hierarchies. This confirms that the agreement between hierarchies increases for networks with a large total sample size compared to the number of treatments and, more generally, it increases with the amount of information present in a network (online supplementary **Figure S4**).

Discussion

Our empirical evaluation showed that in practice the level of agreement between treatment hierarchies is overall high for all ranking metrics used. The agreement between treatment hierarchies from $SUCRA$ and relative treatment effect was very often perfect. The agreement between the rankings from $SUCRA$ or relative treatment effect and the ranking from p_{BV} was good but decreased when the top-ranked interventions are of interest. The agreement is higher for networks with precise estimates and small imbalances in precision.

Simulation studies [6,23] using theoretical examples have shown the importance of accounting for the precision in the estimation of the treatment effects when a hierarchy is to be obtained. However, we show that cases of extreme imbalance in the precision of the treatment effects are rather uncommon.

Several factors can be responsible for imprecision in the estimation of the relative treatment effects in a network:

- large sampling error, determined by a small sample size, small number of events or a large standard deviation;
- poor connectivity of the network, when only a few links and few closed loops of evidence connect the treatments;
- residual inconsistency;
- heterogeneity in the relative treatment effects.

Random-effects models tend to provide relative treatment effects with similar precision as heterogeneity increases. In contrast, in the absence of heterogeneity when fixed-effects

models are used, the precision of the effects can vary a lot according to the amount of data available for each intervention. In the latter case, the ranking metrics are likely to disagree. Also, the role of precision in ranking disagreement is more pronounced in cases where the interventions have similar effects.

Our results also confirm that a treatment hierarchy can differ when the uncertainty in the estimation is incorporated into the ranking metric (by using, for example, a probabilistic metric rather than ranking the point estimate of the mean treatment effect) [8,24] and that rankings from the p_{BV} seem to be the most sensitive to differences in precision in the estimation of treatment effects. We showed graphically that the agreement is less in networks with more uncertainty and with larger imbalances in the variance estimates. However, we also found that such large imbalances do not occur frequently in real data and in the majority of cases the different treatment hierarchies have a relatively high agreement. We acknowledge that there could be other factors influencing the agreement between hierarchies that we did not explore, such as the chosen effect measures [25]. However, we think it is unlikely that such features play a big role in ranking agreement unless assumptions are violated or data in the network is sparse [26]. Adjustment via network meta-regression (for example, for risk of bias or small-study effects) might impact on the ranking of treatments not only by changing the point estimate but also by altering the total precision and the imbalance in the precision of the estimated treatment effects. We did not investigate the agreement between treatment hierarchies obtained from such adjusted analyses. We also did not explore non-methodological characteristics for networks with larger disagreement but we believe these characteristics are a proxy for the amount of information in a network, which is the main factor affecting the agreement between ranking metrics. For example, in some specific fields there are few or small randomised trials (e.g. surgery) and, as a consequence, the resulting networks will have less information. Also, smaller (hence more imprecise) networks might be published more often in journals with lower impact factor and get less citations than large and precise networks.

To our knowledge, this is the first empirical study assessing the level of agreement between treatment hierarchies from ranking metrics in NMA and it provides further insights into the properties of the different methods. In this context, it is important to stress that neither the objective nor the findings of this empirical evaluation imply that a hierarchy for a particular metric works better or is more accurate than one obtained from another ranking metric. The

reason why this sort of comparison cannot be made is that each ranking metric address a specific treatment hierarchy problem. For example, the $SUCRA$ ranking addresses the issue of which treatment outperforms most of the competing interventions, while the ranking based on the relative treatment effect gives an answer to the problem of which treatment is associated with the largest average effect for the outcome considered.

Our study shows that, despite theoretical differences between ranking metrics and some extreme examples, they produce very similar treatment hierarchies in published networks. In networks with large amount of data for each treatment, hierarchies based on SUCRA or the relative treatment effect will almost always agree. Large imbalances in the precision of the treatment effect estimates do not occur often enough to motivate a choice between the different ranking metrics. Therefore, our advice to researchers presenting results from NMA is the following: if the NMA estimated effects are precise, to use the ranking metric they prefer; if at least one NMA estimated effect is imprecise, to refrain from making bold statements about treatment hierarchy and present hierarchies from both probabilistic (e.g. SUCRA or rank probabilities) and non-probabilistic metrics (e.g. relative treatments effects).

Declarations

Author contributions

VC designed the study, analysed the data, interpreted the results of the empirical evaluation, and drafted the manuscript. GS designed the study, interpreted the results of the empirical evaluation and revised the manuscript. AN provided input into the study design and the data analysis, interpreted the results of the empirical evaluation and revised the manuscript. TP developed and manages the database where networks' data was accessed, provided input into the data analysis and revised the manuscript. ME provided input into the study design and revised the manuscript. All the authors approved the final version of the submitted manuscript.

Funding

This work was supported by the Swiss National Science Foundation grant/award number 179158.

Competing Interests

All authors have completed the ICMJE uniform disclosure form and declare: all authors had financial support from the Swiss National Science Foundation for the submitted work; no

financial relationships with any organizations that might have an interest in the submitted work in the previous three years; no other relationships or activities that could appear to have influenced the submitted work.

Data sharing statement

The data for the network meta-analyses included in this study are available in the database accessible using the nmadb R package.

Statement of Ethics Approval

Ethical approval was not required as human participants were not involved in this study.

Tables and Figures

Table 1: Example of treatment hierarchies from different ranking metrics for a network of nine antihypertensive treatment for primary prevention of cardiovascular disease.

Treatment	Original data			Fictional data with increased precision for Conventional treatment versus ARB		
	p_{BV} ranks	$SUCRA_F$ ranks	Relative treatment effect ranks	p_{BV} ranks	$SUCRA_F$ ranks	Relative treatment effect ranks
Conventional	1	6	3.5	3	4	3.5
Diuretic/Beta-blocker	2	1	1	1	1	1
ARB	3	3	3.5	4.5	3	3.5
CCB	4	2	2	2	2	2
Alpha-blocker	5	7	7	4.5	7	7
ACE-inhibitor	6	4	5	6.5	5	5
Diuretic	7	5	6	6.5	6	6
Placebo	8.5	9	9	8.5	9	9
Beta-Blocker	8.5	8	8	8.5	8	8

ACE=Angiotensin Converting Enzyme; CCB=Calcium Channel Blockers; ARB=Angiotensin Receptor Blockers. p_{BV}: probability of producing the best value; $SUCRA_F$: surface under the cumulative ranking curve (calculated in frequentist setting); relative treatment effect stands for the relative treatment effect against fictional treatment of average performance. The first three rankings from the left-hand side are obtained using the original data; the equivalent three rankings on the right-hand side are produced by reducing the standard error of the Conventional versus ARB treatment effect from 0.7 to a fictional value of 0.01.

Table 2: Characteristics of the 232 NMAs included in the re-analysis.

Characteristics of networks	Median	IQR
Median number of treatments compared	6	(5, 9)
Median number of studies included	19	(12, 34)
Median total sample size	6100	(2514, 17264)
	Number of NMAs	**%**
Beneficial outcome	97	41.8%
Dichotomous outcome	185	79.7%
Continuous outcome	47	20.3%
Published before 2010	42	18.1%
Ranking metric used in original publication (non-exclusive):		
Probability of producing the best value	83	35.8%
Rankograms	7	3%
Median or mean rank	3	1.3%
SUCRA	16	6.9%
Other	2	0.9%
None	133	57.3%
Published in general medicine journals†	125	53.9%
Published in health services research journals‡	3	1.3%
Published in specialty journals	104	44.8%

IQR: interquartile range; NMA: network meta-analysis; SUCRA: surface under the cumulative ranking curve.

† Includes the categories Medicine, General & Internal, Pharmacology & Pharmacy, Research & Experimental, Primary Health Care.

‡ Includes the categories Health Care Sciences & Services, Health Policy & Services.

Table 3: Pairwise agreement between treatment hierarchies obtained from the different ranking metrics measured by Spearman ρ, Kendall τ, Yilmaz τ_{AP} and Average Overlap.

	p_{BV} vs $SUCRA_F$	$SUCRA_F$ vs relative treatment effect	p_{BV} vs relative treatment effect	$SUCRA_F$ vs $SUCRA_B$
Spearman ρ	0.9 (0.8, 0.96)	1 (0.99, 1)	0.9 (0.8, 0.97)	1 (0.98, 1)
Kendall τ	0.8 (0.67, 0.91)	1 (0.95, 1)	0.8 (0.69, 0.91)	1 (0.93, 1)
Yilmaz τ_{AP}	0.78 (0.6, 0.9)	1 (0.93, 1)	0.79 (0.65, 0.9)	1 (0.93, 1)
Average Overlap	0.85 (0.72, 0.96)	1 (0.91, 1)	0.88 (0.79, 1)	1 (0.94, 1)

Medians, 1st and 3rd quartiles are reported. p_{BV}: probability of producing the best value; $SUCRA_F$: surface under the cumulative ranking curve (calculated in frequentist setting); $SUCRA_B$: surface under the cumulative ranking curve (calculated in Bayesian setting); relative treatment effect stands for the relative treatment effect against fictional treatment of average performance.

Figure 1: (left panel) Network graph of network of nine antihypertensive treatments for primary prevention of cardiovascular disease. Line width is proportional to inverse standard error of random effects model comparing two treatments. **(right panel) Forest plots of relative treatment effects of overall mortality for each treatment versus placebo.** RR: risk ratio; ACE=Angiotensin Converting Enzyme; CCB=Calcium Channel Blockers; ARB=Angiotensin Receptor Blockers; SE=standard error.

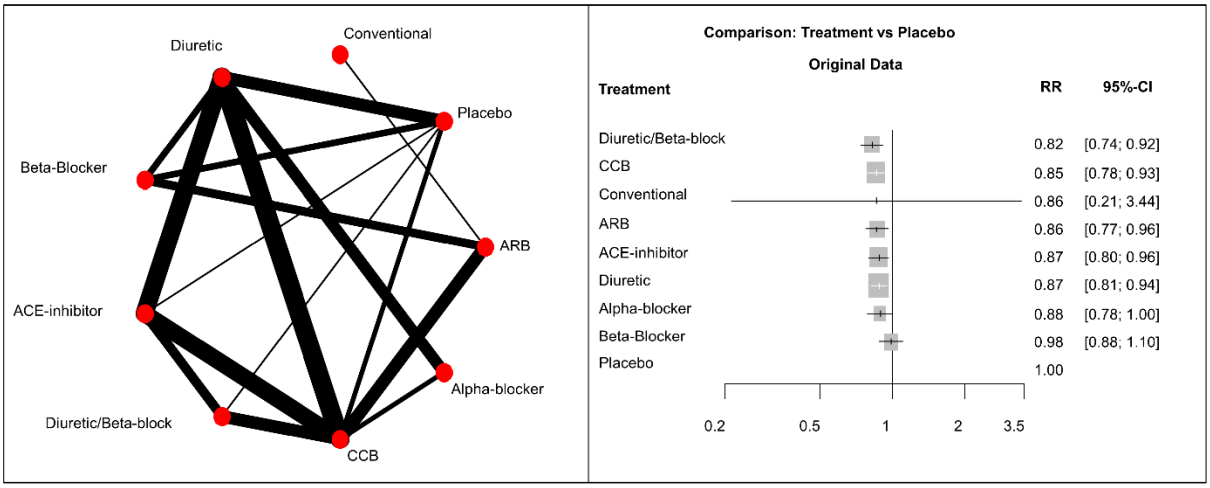


Figure 2: Scatter plots of the average variance in a network and the pairwise agreement between hierarchies from different ranking metrics. The average variance is calculated as the mean of the variances of the estimated treatment effects and describes the average information present in a network. More imprecise network are on the right-hand side of the plots. Spearman ρ (top row) and Average Overlap (bottom row) values for the pairwise agreement between p_{BV} and SUCRA (first column), SUCRA and relative treatment effect (second column), p_{BV} and relative treatment effect (third column). Purple line: cubic smoothing spline with five degrees of freedom.

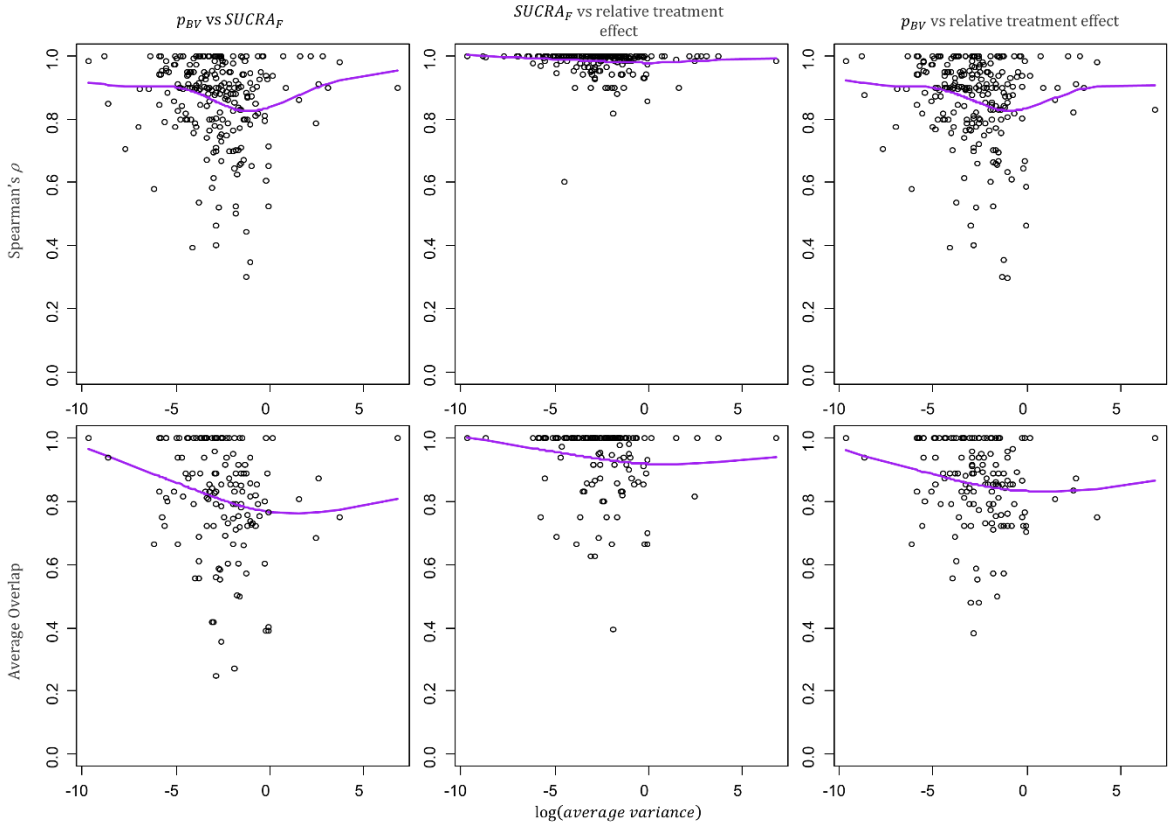
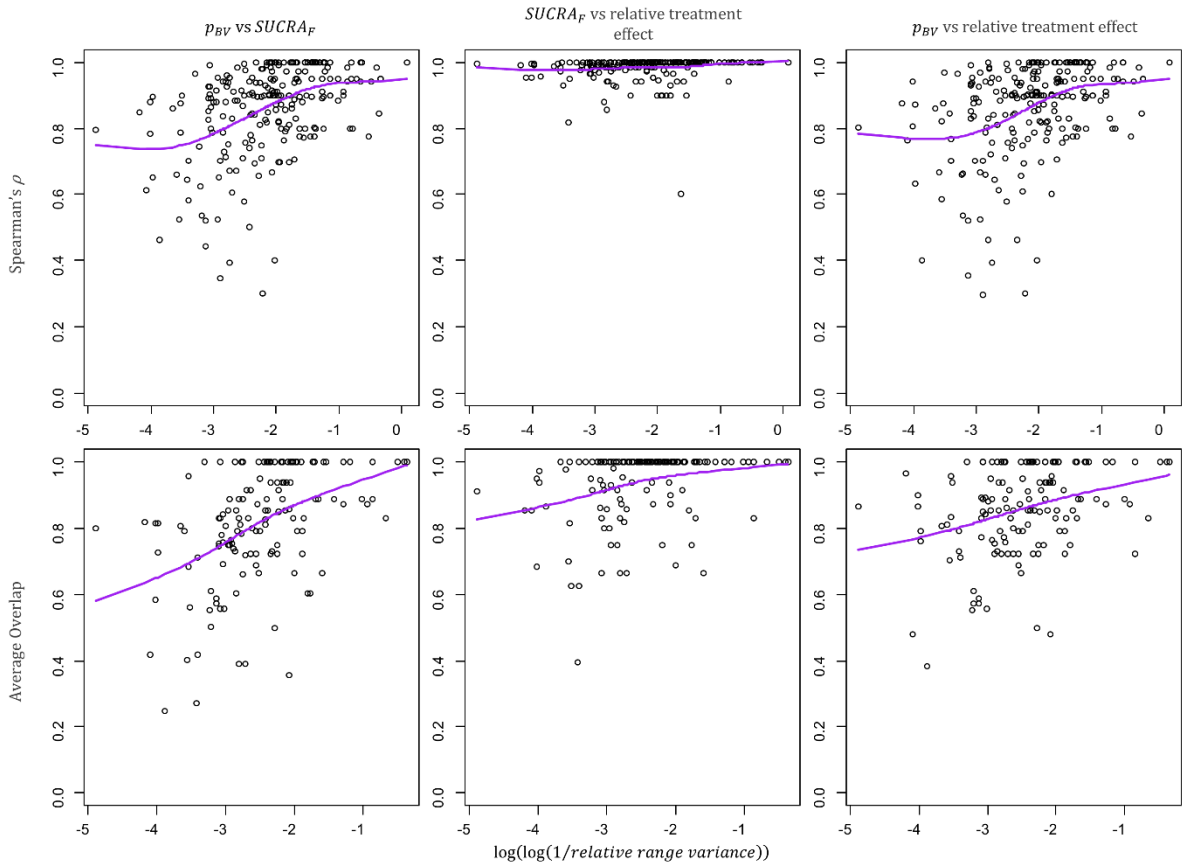


Figure 3: Scatter plots of the relative range of variance in a network and the pairwise agreement between hierarchies from different ranking metrics. The relative range of variance, calculated as $\frac{max\,SE^2 - min\,SE^2}{max\,SE^2}$, indicates how much the information differs between interventions in the same networks. Networks with larger differences in variance are on the left-hand side of the plots. Spearman ρ (top row) and Average Overlap (bottom row) values for the pairwise agreement between p_{BV} and SUCRA (first column), SUCRA and relative treatment effect (second column), p_{BV} and relative treatment effect (third column). Purple line: cubic smoothing spline with five degrees of freedom.



Supplementary material

Available at https://bmjopen.bmj.com/content/bmjopen/10/8/e037744/DC1/embed/inline-supplementary-material-1.pdf?download=true.

References Article 1

1 Efthimiou O, Debray TPA, van Valkenhoef G, *et al.* GetReal in network meta-analysis: a review of the methodology: reviewNMA. *Res Syn Meth* 2016;**7**:236–63. doi:10.1002/jrsm.1195

2 Del Fiol G, Workman TE, Gorman PN. Clinical questions raised by clinicians at the point of care: a systematic review. *JAMA Intern Med* 2014;**174**:710–8. doi:10.1001/jamainternmed.2014.368

3 Salanti G, Ades AE, Ioannidis JPA. Graphical methods and numerical summaries for presenting results from multiple-treatment meta-analysis: an overview and tutorial. *Journal of Clinical Epidemiology* 2011;**64**:163–71. doi:10.1016/j.jclinepi.2010.03.016

4 Rücker G, Schwarzer G. Ranking treatments in frequentist network meta-analysis works without resampling methods. *BMC Medical Research Methodology* 2015;**15**:58. doi:10.1186/s12874-015-0060-8

5 Trinquart L, Attiche N, Bafeta A, *et al.* Uncertainty in Treatment Rankings: Reanalysis of Network Meta-analyses of Randomized Trials. *Ann Intern Med* 2016;**164**:666–73. doi:10.7326/M15-2521

6 Kibret T, Richer D, Beyene J. Bias in identification of the best treatment in a Bayesian network meta-analysis for binary outcome: a simulation study. *Clin Epidemiol* 2014;**6**:451–60. doi:10.2147/CLEP.S69660

7 Veroniki AA, Straus SE, Rücker G, *et al.* Is providing uncertainty intervals in treatment ranking helpful in a network meta-analysis? *J Clin Epidemiol* 2018;**100**:122–9. doi:10.1016/j.jclinepi.2018.02.009

8 Jansen JP, Trikalinos T, Cappelleri JC, *et al.* Indirect Treatment Comparison/Network Meta-Analysis Study Questionnaire to Assess Relevance and Credibility to Inform Health Care Decision Making: An ISPOR-AMCP-NPC Good Practice Task Force Report. *Value in Health* 2014;**17**:157–73. doi:10.1016/j.jval.2014.01.004

9 Petropoulou M, Nikolakopoulou A, Veroniki A-A, *et al.* Bibliographic study showed improving statistical methodology of network meta-analyses published between 1999 and 2015. *Journal of Clinical Epidemiology* 2017;**82**:20–8. doi:10.1016/j.jclinepi.2016.11.002

10 Nikolakopoulou A, Chaimani A, Veroniki AA, *et al.* Characteristics of networks of interventions: a description of a database of 186 published networks. *PLoS ONE* 2014;**9**:e86754. doi:10.1371/journal.pone.0086754

11 Papakonstantinou T. *nmadb: Network Meta-Analysis Database API.* 2019. https://CRAN.R-project.org/package=nmadb

12 Rücker G, Krahn U, König J, *et al. netmeta: Network Meta-Analysis using Frequentist Methods.* 2019. https://github.com/guido-s/netmeta http://meta-analysis-with-r.org.

13 Hosmer DW, Lemeshow S. *Applied Logistic Regression: Hosmer/Applied Logistic Regression*. Hoboken, NJ, USA: : John Wiley & Sons, Inc. 2000. doi:10.1002/0471722146

14 Nikolakopoulou A, Mavridis D, Chiocchia V, *et al.* PreTA: A network meta-analysis ranking metric measuring the probability of being preferable than the average treatment. *Res Syn Meth* (submitted).

15 Kendall MG. THE TREATMENT OF TIES IN RANKING PROBLEMS. *Biometrika* 1945;**33**:239–51. doi:10.1093/biomet/33.3.239

16 Spearman C. The Proof and Measurement of Association between Two Things. *The American Journal of Psychology* 1904;**15**:72. doi:10.2307/1412159

17 Yilmaz E, Aslam JA, Robertson S. A new rank correlation coefficient for information retrieval. In: *Proceedings of the 31st annual international ACM SIGIR conference on Research and development in information retrieval - SIGIR '08*. Singapore, Singapore: : ACM Press 2008. 587. doi:10.1145/1390334.1390435

18 Yilmaz E, Aslam JA. Estimating Average Precision with Incomplete and Imperfect Judgments. In: *Proceedings of the 15th ACM International Conference on Information and Knowledge Management*. New York, NY, USA: : ACM 2006. 102–11. doi:10.1145/1183614.1183633

19 Fagin R, Kumar R, Sivakumar D. Comparing Top K Lists. In: *Proceedings of the Fourteenth Annual ACM-SIAM Symposium on Discrete Algorithms*. Philadelphia, PA, USA: : Society for Industrial and Applied Mathematics 2003. 28–36.http://dl.acm.org/citation.cfm?id=644108.644113 (accessed 15 May 2019).

20 Wu S, Crestani F. Methods for Ranking Information Retrieval Systems Without Relevance Judgments. In: *Proceedings of the 2003 ACM Symposium on Applied Computing*. New York, NY, USA: : ACM 2003. 811–6. doi:10.1145/952532.952693

21 Fretheim A, Odgaard-Jensen J, Brørs O, *et al.* Comparative effectiveness of antihypertensive medication for primary prevention of cardiovascular disease: systematic review and multiple treatments meta-analysis. *BMC Med* 2012;**10**:33. doi:10.1186/1741-7015-10-33

22 Greco T, Calabrò MG, Covello RD, *et al.* A Bayesian network meta-analysis on the effect of inodilatory agents on mortality. *British Journal of Anaesthesia* 2015;**114**:746–56. doi:10.1093/bja/aeu446

23 Davies AL, Galla T. Degree irregularity and rank probability bias in network meta-analysis. *arXiv:200307662 [cond-mat, stat]* Published Online First: 17 March 2020.http://arxiv.org/abs/2003.07662 (accessed 24 Jun 2020).

24 Chaimani A, Vasiliadis HS, Pandis N, *et al.* Effects of study precision and risk of bias in networks of interventions: a network meta-epidemiological study. *International Journal of Epidemiology* 2013;**42**:1120–31. doi:10.1093/ije/dyt074

25 Norton EC, Miller MM, Wang JJ, *et al.* Rank Reversal in Indirect Comparisons. *Value in Health* 2012;**15**:1137–40. doi:10.1016/j.jval.2012.06.001

26 van Valkenhoef G, Ades AE. Evidence Synthesis Assumes Additivity on the Scale of Measurement: Response to "Rank Reversal in Indirect Comparisons" by Norton et al. *Value in Health* 2013;**16**:449–51. doi:10.1016/j.jval.2012.11.012

Article 2: Network meta-analysis results against a fictional treatment of average performance: treatment effects and ranking metric

Adriani Nikolakopoulou[1,2], Dimitris Mavridis[3,4], **Virginia Chiocchia[1,5],** Theodoros Papakonstantinou[1], Toshi A Furukawa[5] and Georgia Salanti[1]

[1] Institute of Social and Preventive Medicine (ISPM), University of Bern, Bern, Switzerland

[2] Institute of Medical Biometry and Statistics, Faculty of Medicine and Medical Center - University of Freiburg, Germany

[3] Department of Primary Education, University of Ioannina, Ioannina, Greece

[4] Faculté de Médecine, Université Paris Descartes, Paris, France

[5] Graduate School of Health Sciences (GHS), University of Bern, Switzerland

[6] Departments of Health Promotion and Human Behavior and of Clinical Epidemiology, Kyoto University Graduate School of Medicine/School of Public Health, Kyoto, Japan

My contribution:

I contributed to the analysis by developing the code for the empirical comparison to check the agreement between the new ranking metric introduced in the paper and the existing ranking metrics. I was involved in the writing of the manuscript, and in the revision after the review from co-authors and peer-review from Research Synthesis Methods.

Published: Nikolakopoulou A, Mavridis D, Chiocchia V, Papakonstantinou T, Furukawa TA, Salanti G. Network meta-analysis results against a fictional treatment of average performance: Treatment effects and ranking metric. *Res Syn Meth.* 2021;12:161–175. doi: 10.1002/jrsm.1463

Abstract

Background: Network meta-analysis (NMA) produces complex outputs as many comparisons between interventions are of interest and a treatment ranking is often included in the aims of the evidence synthesis. The estimated relative treatment effects are usually displayed in a forest plot or in a league table and several ranking metrics are calculated and presented, such as the median and mean treatment ranks.

Methods: We estimate relative treatment effects of each competing treatment against a fictional treatment using the 'deviation from the means' coding that has been used to parameterise categorical covariates in regression models. Based on this alternative parameterisation of the NMA model, we present a new ranking metric (PreTA: Preferable Than Average) interpreted as the probability that a treatment is better than a fictional treatment of average performance.

Results: We compare PreTA with existing probabilistic ranking metrics in 232 networks of interventions. We use two networks of interventions, a network of 18 antidepressants for acute depression and a network of four interventions for heavy menstrual bleeding, to illustrate the methodology. The agreement between PreTA and existing ranking metrics depends on the precision with which relative effects are estimated.

Conclusions: PreTA is a viable alternative to existing ranking metrics which can be interpreted as the probability of being better than the 'average' treatment. It enriches the decision-making arsenal with a ranking metric which is interpreted as a probability and considers the entire ranking distributions of the involved treatments.

Keywords: Alternative parameterisation; Deviation from means; Indirect evidence; Probabilistic ranking; Treatment hierarchy

Introduction

The output that necessarily needs to be presented in a network meta-analysis (NMA) is a set of relative treatment effects between all competing treatments [1,2]. Such an output answers the primary question of NMA: to compare the performance of "all versus all" alternative treatment options for a healthcare condition. This output may be given in a forest plot against a common reference treatment or in a league table, where the names of the treatments are presented in the diagonal and each cell contains the relative treatment effect [3]. Such a table allows for the simultaneous presentation of two outcomes, or of the results from pairwise and network meta-analysis, below and above the diagonal. Additionally, by-products of relative treatment effects are often presented as ranking metrics of the included treatments. Results from NMA are often used to inform health-care decision making [4,5] and ranking metrics constitute an attempt to present such results in a coherent and understandable way. Several ranking metrics have been proposed to present NMA results, each one answering a different question. Ranking probabilities of each treatment being at each possible rank are calculated using simulation or resampling techniques either in a Bayesian or in a frequentist framework. Other ranking metrics include the surface under the cumulative ranking curve (SUCRA), that averages across all ranking probabilities for each treatment, and its frequentist analogue, P-score, which is calculated analytically [6,7]. SUCRA and P-score can be interpreted as the mean extent of certainty that a treatment is better than all the other treatments. As authors of [6] point out, however, *"it is impossible to tell what constitutes a modest or large difference in SUCRA between two treatments, either statistically or clinically"*.

In this paper, we present an alternative parameterisation of the NMA model and we use it to develop a probabilistic ranking metric that naturally incorporates uncertainty and is a viable alternative to existing ranking metrics. In section 2, we re-parameterise the NMA model to derive treatment effects against a fictional treatment of average performance using the deviation of means coding that has been used to parameterise categorical covariates in regression models [8]. In section 3, we use the derived treatment effects to compute the probability of each treatment being better than the 'average' treatment. This ranking metric aids the interpretation of NMA results in classifying treatments as superior, equivalent and inferior to an imaginary 'average' treatment.

Reparameterisation of the NMA model

Deviation from means coding in regression models

We start with a short description of the deviation from means coding in regression models as described by Hosmer and Lemeshow [8]. This is an alternative parameterisation to the most common 'reference cell coding' in order to avoid the use of a reference level. According to the reference cell coding, a categorical independent variable with C categories is expressed through $C-1$ dummy/indicator variables.

Consider, for example, that we aim to estimate the effect of a covariate with four groups on the probability of an event. We fit a logistic regression model

$$g(p(\boldsymbol{x})) = \gamma_0 + \gamma_1 \boldsymbol{x_1} + \gamma_2 \boldsymbol{x_2} + \gamma_3 \boldsymbol{x_3}$$

where $\boldsymbol{x} = (\boldsymbol{x_1}, \boldsymbol{x_2}, \boldsymbol{x_3})'$ are the dummy variables for the covariate and $g(p(\boldsymbol{x}))$ is the logit link function with $p(\boldsymbol{x})$ indicating the probability of event. According to the reference cell coding, the indicator variables are parameterised as shown in Table 1 and result into estimating logarithms of the relative odds ratios (logOR) between the categories represented by the values 0 and 1 in these indicator variables.

According to the alternative deviation from means coding, the indicator variables express effects as deviations between each category mean (here the logit of the outcome in that category) from the overall (grand) mean (here the average logit outcome over all categories) as shown in Table 1. The model results in estimating the coefficients, interpreted as the relative effects among groups versus the average effect across all groups. Note that the exponential of the coefficients are not odds ratios because in the denominator is the average odds that includes the odds of the numerator. For further information and examples on the deviation from means coding, see [8].

Notation for the NMA model

In this section, we introduce some general notation for the NMA model. Let the entire evidence base consist of $i = 1, \dots, n$ studies forming a set of treatments, denoted as $k = 1, \dots, K$. The number of treatments in study i is denoted as K_i. Index j denotes a treatment contrast. A core assumption in NMA is that of transitivity, which implies that in a network of K treatments, and subsequently $\binom{K}{2}$ possible relative treatment effects, only $K-1$ need to be estimated and the rest are derived as linear combinations of those [9,10]. The target parameter is therefore a vector $\boldsymbol{\mu}$ of $K-1$ relative treatment effects $\mu_2, \mu_3, \dots \mu_K$, called the

vector of basic parameters [11,12]. With arm-level data we can model arm level parameters, for example the event probability for a binary outcome, in study i and treatment arm k denoted as y_{ik}[13]. A link function $g(y_{ik})$ maps the parameters of interest onto a scale ranging from minus to plus infinity and u_i are the trial-specific baselines. For an overview of commonly used link functions in meta-analysis see [14]. All arm-level parameters y_{ik} across studies are collected in a vector $\boldsymbol{y^a}$ of length $\sum_{i=1}^{n} K_i$, where superscript a stands for 'arm-level'.

With contrast-level data we model trial specific summaries, for example logOR, log risk ratio, mean difference or standardized mean difference [13]. Let y_{ij} be the observed effect size for treatment contrast j in study i. The vector of the estimated contrasts across all studies is denoted as $\boldsymbol{y^c}$ and is of length $\sum_{i=1}^{n}(K_i - 1)$. The superscript $\boldsymbol{c}$ indicates the fact that 'contrast-level' data are modeled.

We will first describe the arm-level (section 2.3) and then the contrast-level (section 2.4) NMA models using reference cell coding and the equivalent alternative deviation from the means parameterisation, which allows estimation of all treatments versus a fictional treatment of average performance. Sections 2.3 and 2.4 can be read independently, i.e. the reader can skip one of the two sections. Alternatively, the reader already familiar with the NMA models that use reference cell coding can skip 2.3.1 and 2.4.1. Table 2 can be used as a reference to the four forms of the NMA model (arm-level and contrast level with reference cell and deviation from the means coding), in case parts of the remainder of section 2 are skipped.

We will exemplify the models using a hypothetical network of three treatments, A, B and C examined in four studies, one comparing A and B, one comparing A and C, one comparing B and C and one three-arm study comparing treatments A, B and C. The target vector of basic parameters is usually taken to include the relative effects of all treatments versus an arbitrary reference, here treatment A, and hence is $\boldsymbol{\mu} = \begin{pmatrix} \mu_{AB} \\ \mu_{AC} \end{pmatrix}$. The transitivity assumption implies consistency between relative treatment effects; in particular, it holds that

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

NMA with arm-level data

Reference cell coding

The model for study 1, comparing treatments A and B is shown in Table 2; $\delta_{1,AB}$ denotes the random effect of study 1 for the comparison AB and τ^2 denotes heterogeneity. It is customary

to assume that heterogeneity is common across comparisons. The model is straightforwardly generalized for the other three studies (Table 2).

In its general form, the NMA model using arm-based analysis can be written as

$$\boldsymbol{g(y^a)} = \boldsymbol{Zu} + \boldsymbol{X^a\mu} + \boldsymbol{W\delta}$$

Equation 1

where $\boldsymbol{u}$ is the vector of baselines u_i of length n, which can be assumed to be either fixed and unrelated to each other, or exchangeable drawn from a normal distribution [15]. We assume fixed and unrelated baseline effects for the remainder of this paper. Vector $\boldsymbol{\delta}$ includes the study random effects $\delta_{i,j}$ and follows the multivariate normal distribution

$$\boldsymbol{\delta} \sim \boldsymbol{N(0, \Sigma)}$$

Matrix $\boldsymbol{\Sigma}$ is a block-diagonal between-study variance-covariance matrix of dimensions $\{\sum_{i=1}^{n}(K_i - 1)\} \times \{\sum_{i=1}^{n}(K_i - 1)\}$. The matrices $\boldsymbol{Z}, \boldsymbol{X^a}, \boldsymbol{W}$ are design matrices linking the vector of baselines, basic parameters and random effects respectively with $\boldsymbol{g(y^a)}$. The construction of these design matrices depends on the modeled arm-level parameters y_{ik} and is exemplified in the following example.

For the example of Table 2, Equation 1 takes the form

$$\begin{pmatrix} g(y_{1A}) \\ g(y_{1B}) \\ g(y_{2A}) \\ g(y_{2C}) \\ g(y_{3B}) \\ g(y_{3C}) \\ g(y_{4A}) \\ g(y_{4B}) \\ g(y_{4C}) \end{pmatrix} = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} u_1 \\ u_2 \\ u_3 \\ u_4 \end{pmatrix} + \begin{pmatrix} 0 & 0 \\ 1 & 0 \\ 0 & 0 \\ 0 & 1 \\ 0 & 0 \\ -1 & 1 \\ 0 & 0 \\ 1 & 0 \\ 0 & 1 \end{pmatrix} \begin{pmatrix} \mu_{AB} \\ \mu_{AC} \end{pmatrix} + \begin{pmatrix} 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} \delta_{1,AB} \\ \delta_{2,AC} \\ \delta_{3,BC} \\ \delta_{4,AB} \\ \delta_{4,AC} \end{pmatrix}$$

with

$$\begin{pmatrix} \delta_{1,AB} \\ \delta_{2,AC} \\ \delta_{3,BC} \\ \delta_{4,AB} \\ \delta_{4,AC} \end{pmatrix} \sim N\left(\begin{pmatrix} 0 \\ 0 \\ 0 \\ 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \tau^2 & 0 & 0 & 0 & 0 \\ 0 & \tau^2 & 0 & 0 & 0 \\ 0 & 0 & \tau^2 & 0 & 0 \\ 0 & 0 & 0 & \tau^2 & \tau^2/2 \\ 0 & 0 & 0 & \tau^2/2 & \tau^2 \end{pmatrix} \right)$$

Matrix $\boldsymbol{X^a}$ indicates which elements of $\boldsymbol{\mu}$ are estimated by each $g(y_{ik})$. It contains one row per study arm and one column per basic parameter. The first row corresponds to treatment arm A of the first study taking the value 0 both for μ_{AB} and μ_{AC}. The second row indicates that μ_{AB} is estimated in treatment arm B of the first study. Similarly, the construction of the

next rows of $\boldsymbol{X^a}$, as well as that of $\boldsymbol{Z}$ and $\boldsymbol{W}$, is implied by the arm-level data included in each study and the subsequent elements of $\boldsymbol{\mu}$ to be estimated (Table 2).

Deviation from means coding

The above model in Equation 1 can be modified using the deviation from means coding [8]. The model will be parameterised in such a way to estimate the effects of each treatment versus the 'average' treatment. The target parameter of this model is a vector $\boldsymbol{b}$ that includes $K-1$ parameters b_k with $k = 2, \dots, K$ which are the effects of treatment k versus the average effect over all treatments. One of the treatments – here treatment 1 – is arbitrarily chosen to be excluded for identifiability. Results do not depend on the choice of this 'reference' treatment.

For the deviation from means coding, the model will be

$$\boldsymbol{g(y^a)} = \boldsymbol{Zu} + \boldsymbol{X^{a^*}b} + \boldsymbol{W\delta}$$

Equation 2

with $\boldsymbol{X^{a^*}}$ denoting the modified design matrix. The matrices $\boldsymbol{Z}$ and $\boldsymbol{W}$ remain unchanged. The new design matrix $\boldsymbol{X^{a^*}}$ will take values -1 for the arbitrarily chosen treatment that is not included in vector $\boldsymbol{b}$; all other entries in the matrix are as in $\boldsymbol{X^a}$.

Consider the example of Table 1 and the first two rows of the $\boldsymbol{X^a}$ matrix, $\begin{pmatrix} 0 & 0 \\ 1 & 0 \end{pmatrix}$, corresponding to the first study. According to the deviation from means coding as illustrated in Table 1, we chose a treatment (here treatment A) for which $\boldsymbol{X^{a^*}}$ will take -1 for both dummy variables (both columns of the design matrix) and the group corresponding to treatment B takes 1 and 0 for the two columns of the design matrix, as in $\boldsymbol{X^a}$. Thus, the respective part of the new design matrix will be $\begin{pmatrix} -1 & -1 \\ 1 & 0 \end{pmatrix}$. The model for study 1 with the alternative parameterisation is

$$g(y_{1A}) = u_1 - b_B - b_C$$
$$g(y_{1B}) = u_1 + b_B + \delta_{1,AB}$$
$$\delta_{1,AB} \sim N(0, \tau^2)$$

where the parameters b_B and b_C denote the effects of B versus average treatment and C versus average treatment respectively. The effect of A versus the average treatment is $-b_B - b_C$ and the relative effect of B versus A for the study 1 is derived as

$$g(y_{1B}) - g(y_{1A}) = 2b_B + b_C + \delta_{1,AB}$$

The models for all studies are given in Table 2 and the full model is written as

$$\begin{pmatrix} g(y_{1A}) \\ g(y_{1B}) \\ g(y_{2A}) \\ g(y_{2C}) \\ g(y_{3B}) \\ g(y_{3C}) \\ g(y_{4A}) \\ g(y_{4B}) \\ g(y_{4C}) \end{pmatrix} = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} u_1 \\ u_2 \\ u_3 \\ u_4 \end{pmatrix} + \begin{pmatrix} -1 & -1 \\ 1 & 0 \\ -1 & -1 \\ 0 & 1 \\ 1 & 0 \\ 0 & 1 \\ -1 & -1 \\ 1 & 0 \\ 0 & 1 \end{pmatrix} \begin{pmatrix} b_B \\ b_C \end{pmatrix} + \begin{pmatrix} 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} \delta_{1,AB} \\ \delta_{2,AC} \\ \delta_{3,BC} \\ \delta_{4,AB} \\ \delta_{4,AC} \end{pmatrix}$$

Note that the reparameterisation described using the deviation from the means coding should not be confused with different parameterisations of the NMA model to produce relative treatment effects of all treatments versus each other. We present in the Additional file 1 an example of different parameterisations for specifying the means using reference cell coding and deviation from means coding using arm-level data.

NMA with contrast-level data

Reference cell coding

In the contrast-level NMA, data from $K_i - 1$ contrasts for each study are modeled. The model for study i and treatment contrast j is written as

$$y_{ij} = \mu_j + \varepsilon_{ij} + \delta_{ij}$$
$$\varepsilon_{ij} \sim N\left(0, s_{ij}^2\right)$$
$$\delta_{ij} \sim N(0, \tau^2)$$

with ε_{ij} being the random error for study i and treatment contrast j where s_{ij}^2 is the sample variance of y_{ij}. The random effect δ_{ij} is defined as in the NMA with arm-level data. For example, for the first study the model is

$$y_{1,AB} = \mu_{AB} + \varepsilon_{1,AB} + \delta_{1,AB}$$
$$\varepsilon_{1,AB} \sim N\left(0, s_{1,AB}^2\right)$$
$$\delta_{1,AB} \sim N(0, \tau^2)$$

and, similarly, for the other studies the models are given in Table 2.

The contrast-based NMA model in its general form is then written as

$$\boldsymbol{y^c} = \boldsymbol{X^c \mu} + \boldsymbol{\delta} + \boldsymbol{\varepsilon}$$

Equation 3

with the vector of random effects $\boldsymbol{\delta}$ having the distribution given in the arm-level NMA model and the vector of random errors being distributed as

$$\boldsymbol{\varepsilon} \sim \boldsymbol{N}(\boldsymbol{0}, \boldsymbol{S})$$

where $\boldsymbol{S}$ is the block-diagonal within-study variance-covariance matrix of the same dimensions as $\boldsymbol{\Sigma}$. The design matrix $\boldsymbol{X^c}$ has dimensions $\sum_{i=1}^{n}(K_i - 1) \times (K - 1)$. The entries in each row describe the relationship between the vector of basic parameters $\boldsymbol{\mu}$ and the vector of observed contrast-level data $\boldsymbol{y^c}$.

For example, in the illustrative network of three treatments and four studies, the full model is written as

$$\begin{pmatrix} y_{1,AB} \\ y_{2,AC} \\ y_{3,BC} \\ y_{4,AB} \\ y_{4,AC} \end{pmatrix} = \begin{pmatrix} 1 & 0 \\ 0 & 1 \\ -1 & 1 \\ 1 & 0 \\ 0 & 1 \end{pmatrix} \begin{pmatrix} \mu_{AB} \\ \mu_{AC} \end{pmatrix} + \begin{pmatrix} \delta_{1,AB} \\ \delta_{2,AC} \\ \delta_{3,BC} \\ \delta_{4,AB} \\ \delta_{4,AC} \end{pmatrix} + \begin{pmatrix} \varepsilon_{1,AB} \\ \varepsilon_{2,AC} \\ \varepsilon_{3,BC} \\ \varepsilon_{4,AB} \\ \varepsilon_{4,AC} \end{pmatrix}$$

The first row of the $\boldsymbol{X^c}$ matrix indicates that the first two-arm study estimates μ_{AB}. Note that the arm-level model using reference cell coding for study 1 implies that

$$g(y_{1B}) - g(y_{1A}) = \mu_{AB} + \delta_{1,AB}$$

and, consequently, the first row of the $\boldsymbol{X^c}$ matrix results as the subtraction of the second minus the first row of $\boldsymbol{X^a}$.

Deviation from means coding

The reparameterised model will differ from that presented in Equation 3 in two ways; the target parameter to be estimated, which again are the relative effects $\boldsymbol{b}$ against an 'average' treatment, and the design matrix $\boldsymbol{X^{c^*}}$. The matrix $\boldsymbol{X^{c^*}}$ can be easily obtained from $\boldsymbol{X^{a*}}$ by subtracting its rows within each study contrast.

In its general form, the model is

$$\boldsymbol{y^c} = \boldsymbol{X^{c^*}}\boldsymbol{b} + \boldsymbol{\delta} + \boldsymbol{\varepsilon}$$

Equation 4

Consider in our example the part of $\boldsymbol{X^{a*}}$ corresponding to study 1, $\begin{pmatrix} -1 & -1 \\ 1 & 0 \end{pmatrix}$, then the row of $\boldsymbol{X^{c^*}}$ corresponding to that first study will be $(2 \quad 1)$, which is the subtraction of the two rows. This is also evident considering that

$$g(y_{1B}) - g(y_{1A}) = 2b_B + b_C + \delta_{1,AB}$$

according to the arm-based model using the deviation from means coding.

The models for studies 1 to 4 are given in Table 2 and can be written as

$$\begin{pmatrix} y_{1,AB} \\ y_{2,AC} \\ y_{3,BC} \\ y_{4,AB} \\ y_{4,AC} \end{pmatrix} = \begin{pmatrix} 2 & 1 \\ 1 & 2 \\ -1 & 1 \\ 2 & 1 \\ 1 & 2 \end{pmatrix} \begin{pmatrix} b_B \\ b_C \end{pmatrix} + \begin{pmatrix} \delta_{1,AB} \\ \delta_{2,AC} \\ \delta_{3,BC} \\ \delta_{4,AB} \\ \delta_{4,AC} \end{pmatrix} + \begin{pmatrix} \varepsilon_{1,AB} \\ \varepsilon_{2,AC} \\ \varepsilon_{3,BC} \\ \varepsilon_{4,AB} \\ \varepsilon_{4,AC} \end{pmatrix}$$

The estimation of $\boldsymbol{b}$ in the contrast-based NMA model using deviation from means coding (Equation 4) is

$$\widehat{\boldsymbol{b}} = \left(\left(\boldsymbol{X}^{\boldsymbol{c}^*} \right)' \left(\boldsymbol{S} + \widehat{\boldsymbol{\Sigma}} \right)^{-1} \boldsymbol{X}^{\boldsymbol{c}^*} \right)^{-1} \left(\boldsymbol{X}^{\boldsymbol{c}^*} \right)' \left(\boldsymbol{S} + \widehat{\boldsymbol{\Sigma}} \right)^{-1} \boldsymbol{y}^{\boldsymbol{c}}$$

with variance-covariance matrix

$$\boldsymbol{var}\left(\widehat{\boldsymbol{b}}\right) = \left(\left(\boldsymbol{X}^{\boldsymbol{c}^*} \right)' \left(\boldsymbol{S} + \widehat{\boldsymbol{\Sigma}} \right)^{-1} \boldsymbol{X}^{\boldsymbol{c}^*} \right)^{-1}$$

Vector $\widehat{\boldsymbol{b}}$ includes the estimation of the $K-1$ parameters b_k for $k = 2, \ldots, K$. The estimation of the effect of treatment $k = 1$, which was chosen to be excluded for identifiability, versus the average effect is given as

$$\hat{b}_1 = \sum_{k=2}^{K} \left(-\hat{b}_k \right)$$

with variance $\sum_{k=2}^{K} var\left(\hat{b}_k\right) + \sum_{k \neq l,\, k<l, k>1, l>1}^{K} 2cov\left(\hat{b}_k, \hat{b}_l\right)$. Note that results do not depend on the choice of reference treatment.

Network estimates $\widehat{\boldsymbol{\mu}}^{\boldsymbol{N}}$ can be derived as linear combinations of $\widehat{\boldsymbol{b}}$

$$\widehat{\boldsymbol{\mu}}^{\boldsymbol{N}} = \boldsymbol{Y}^* \widehat{\boldsymbol{b}}$$

with variance-covariance matrix

$$\boldsymbol{var}(\widehat{\boldsymbol{\mu}}^{\boldsymbol{N}}) = \boldsymbol{Y}^* \left(\left(\boldsymbol{X}^{\boldsymbol{c}^*} \right)' \left(\boldsymbol{S} + \widehat{\boldsymbol{\Sigma}} \right)^{-1} \boldsymbol{X}^{\boldsymbol{c}^*} \right)^{-1} (\boldsymbol{Y}^*)'$$

and are equivalent to the network estimates derived using reference cell coding. Matrix $\boldsymbol{Y}^*$ of dimensions $\binom{K}{2} \times (K-1)$ is constructed similarly to $\boldsymbol{X}^{\boldsymbol{c}^*}$ and connects $\widehat{\boldsymbol{b}}$ with network estimates $\widehat{\boldsymbol{\mu}}^{\boldsymbol{N}}$. We can use several methods for estimating $\boldsymbol{\Sigma}$ such as likelihood-based methods and an extension of the DerSimonian and Laird method [11,16]. For the worked example, it holds that

$$\begin{pmatrix} \hat{\mu}^N_{AB} \\ \hat{\mu}^N_{AC} \\ \hat{\mu}^N_{BC} \end{pmatrix} = \begin{pmatrix} 2 & 1 \\ 1 & 2 \\ -1 & 1 \end{pmatrix} \begin{pmatrix} \hat{b}_B \\ \hat{b}_C \end{pmatrix} = \begin{pmatrix} 2\hat{b}_B + \hat{b}_C \\ \hat{b}_B + 2\hat{b}_C \\ -\hat{b}_B + \hat{b}_C \end{pmatrix}$$

The contrast-level NMA model can be written as a two-stage model, as first described in [11,17,18], where results of separate pairwise meta-analyses are used instead of $\boldsymbol{y}^{\boldsymbol{c}}$ in the model described in Equation 3. Constructing the respective design matrix follows the logic of

constructing $\boldsymbol{X^c}$ and its modification to parameterise the model using the deviation from means coding is straightforward.

PreTA: Probability of a treatment being preferable than the average treatment

Applying the deviation from means coding in NMA models results in the derivation of the effects of each treatment against a fictional treatment of 'average' performance. In this section we use the K estimated parameters $\hat{b}_k$ to compute the probability of each treatment being better than the average treatment. To do so, we follow similar steps as those followed by Rücker and Schwarzer who derived the frequentist analogue of SUCRA, P-score [7].

Intermediate to the calculation of P-scores is the derivation of the probability that treatment k is better than treatment l, calculated as

$$P_{kl} = P(\hat{\mu}_{kl}^N > 0) = \Phi\left(\frac{\hat{\mu}_{kl}^N}{\sqrt{var(\hat{\mu}_{kl}^N)}}\right)$$

assuming that higher values represent a better outcome. Accordingly, the probability that treatment k is better than the fictional treatment of average performance (PreTA) can be derived as

$$PreTA_k = P\left(\hat{b}_k > 0\right) = \Phi\left(\frac{\hat{b}_k}{\sqrt{var\left(\hat{b}_k\right)}}\right)$$

The range of values for $PreTA_k$ is (0.5, 1) if $\hat{b}_k > 0$, and (0, 0.5) if $\hat{b}_k < 0$. As it is the case with P-scores, the mean of $PreTA_k$ across all treatments is 0.5; this means that across all treatments, the mean extent of certainty that a treatment is better than the fictional treatment of average performance is 0.5. Alternatively, the z-score $\frac{\hat{b}_k}{\sqrt{var(\hat{b}_k)}}$ can be used to classify treatments according to their 'distance' from the fictional treatment.

Of note is that the above calculations assume normality of the estimated parameters $\hat{b}_k$. However, as $\hat{b}_k$ are not effect sizes expressed for example as logOR or mean differences, using them for hypothesis testing is not meaningful. Despite that, drawing $\hat{b}_k$ along with the associated 95% confidence intervals can be useful in capturing uncertainty around the ranking produced by relative treatment effects.

Comparison of PreTAs with existing ranking metrics: theoretical considerations and empirical analysis

The, usually called, probability of being the best (pBV) is a popular ranking metric, usually calculated as the frequency that a particular treatment ranks in the first place, compared to the other alternative treatment options. pBV is interpreted as the probability of producing the best outcome value in a network of interventions (e.g. large effects for a beneficial outcome, or small effects for a harmful outcome). While its derivation might be sensible in some cases, we should not overlook the fact that it only takes into account one tail of the treatment effects' distributions; e.g. it does not account for the probability to produce a small effect on a beneficial outcome. SUCRAs and P-scores are useful summaries of the entire ranking distributions; suggested interpretations include *"the average proportion of competing treatments, which produce outcome values worse than treatment k"* and *"the mean extent of certainty that treatment k produces better values than all other treatments*" [7,19].

We performed an empirical comparison of the treatment hierarchies obtained with PreTA, pBV and SUCRA, calculated using parametric bootstrap in a frequentist framework. The agreement between ranking metrics was measured using Kendall's tau. We used a previously described database of NMAs published until 2015 including networks of four or more interventions [4]. We included networks with available outcome data in arm-level format, for which the primary outcome was analysed either as binary or as continuous. We used the effect measure used in the original review. Details about the inclusion criteria of the NMAs included in the database can be found in [4]. The empirical analysis was performed with the use of the nmadb package in R [20].

Results of the empirical analysis are presented in section 5. In the following section, we illustrate our method in two networks of interventions, for which at least some disagreements between pBV, SUCRAs and PreTAs occur.

Worked examples

Network of antidepressants

We illustrate the derivation of the method using as an example a recently published NMA comparing the effectiveness of antidepressants for major depression [21]. The primary efficacy outcome was response measured as 50% or greater reduction in the symptoms scales

between baseline and 8 weeks of follow up and results were presented as ORs. The authors aimed at comparing active antidepressants and considered the inclusion of both head-to-head and placebo-controlled trials. The network comprised 522 double-blind, parallel, RCTs comparing 21 antidepressants or placebo. In line with previous empirical evidence [22,23], the authors have found evidence that the probability of receiving placebo decreases the overall response rate in a trial and dilutes differences between active compounds [24]. Based on this ground, authors of this NMA [21] synthesized only head-to-head studies separately to estimate the relative efficacy of active interventions. Here, we will focus on the latter network that included 179 head-to-head studies comparing 18 antidepressants (Figure 1a).

Authors presented relative treatment effects between all pairs of the 18 antidepressants in a league table (figure 4 in [21]). When effect sizes are used to rank treatments, selecting a reference treatment against which to draw a forest plot of NMA effects is of particular importance. Although the choice of reference does not affect the estimates obtained, the uncertainty around NMA effects depends on the precision with which the selected reference treatment is associated. Figure 2 shows the relative treatment effects against fluoxetine and vortioxetine, the treatments that have been studied most and least respectively. While results are equivalent, choosing to present one over the other forest plot might implicitly lead to different interpretations on the similarity between the drugs based on visually inspecting the overlap of the confidence intervals.

Figure 2 also shows the derived odds of each treatment versus the odds of a fictional treatment of average response with their confidence intervals. The line of no effect is included in the graph for illustration reasons, although $e^{\hat{b}_k}$ are not suited for hypothesis testing. The amount of uncertainty around the relative effects versus the average treatment is between the amount of uncertainty around the relative effects of fluoxetine and that of vortioxetine. In fact, presenting $e^{\hat{b}_k}$ with their confidence intervals offers a solution to the ambiguity of selecting a reference treatment, in terms of the uncertainty around them and the consequent conclusions about similarity of treatments. This example shows that presenting the effects versus a fictional treatment of average performance in a forest plot, in addition to a league table presenting all relative effects, might be a viable option in networks with many treatments and in absence of a “natural” reference treatment.

Figure 3 shows the PreTAs for the 18 antidepressants; treatments around 0.5 are the treatments closest to the fictional treatment. Vortioxetine has the largest point estimate against the fictional treatment but its estimation comes with great uncertainty. Escitalopram versus fictional is more precisely estimated in favor of escitalopram and it is associated with the greatest PreTA (97%). Duloxetine and milnacipran are the treatments closest to the fictional treatment. The point estimate of nefazodone versus the average treatment is slightly larger than that of duloxetine. Due to the associated uncertainty, however, there is 34% probability that nefazodone is superior to the fictional treatment, compared to 52% of duloxetine. Fluoxetine, clomipramine, fluvoxamine, trazodone and reboxetine are among the worst treatments in the network, either because of their point estimates against the fictional treatment or because of the respective precision in the estimation. It should be noted that the hierarchy illustrated in Figure 3 refers only to one outcome and does not take into account more complex hierarchy questions.

Table 3 summarizes the ranking metrics for the network of antidepressants; pBV, the SUCRA and PreTAs are presented [6,25]. Escitalopram, which is the first treatment according to PreTA, ranks second according to SUCRA and third according to pBV. The disagreement between PreTA and pBV is explained by the fact that pBV favours vortioxetine and bupropion over escitalopram because of the mass under the right tail of the treatment effects' distribution. The small disagreement between PreTA and SUCRA reflects their different interpretations: vortioxetine, ranked first according to SUCRA, beats on average a larger proportion of treatments compared to escitalopram (0.90 versus 0.83) but escitalopram has a larger probability to be better than the fictional average treatment compared to vortioxetine (0.93 versus 0.87). Similarly, fluoxetine ranks last according to PreTA whereas it is followed by trazodone and reboxetine according to SUCRA. This disagreement arises from the fact that the smaller $var(\hat{b}_k)$ for fluoxetine leads in a greater certainty that it is worse than the fictional treatment.

Network of interventions for heavy menstrual bleeding

We use as a second example a network of interventions for the treatment of heavy menstrual bleeding. The following four interventions were compared: levenorgestel-releasing intrauterine system (Mirena), first generation endometrial destruction, second generation endometrial destruction and hysterectomy [26]. The primary outcome was patients' dissatisfaction at 12 months and the network included 20 studies (Figure 1b).

Figure 4 shows the treatment effects of the four treatments compared to a fictional average treatment and Appendix Figure 1 illustrates the relative position of each treatment according to its probability of being superior (green) or inferior (red) than the average treatment. There is a clear advantage of hysterectomy compared to the other three treatments with no treatment lying close to the 'average treatment area' (0.5 of PreTA).

In this example, hysterectomy outperforms the other three treatments and ranks first according to all ranking metrics (PreTA: 0.99, pBV: 0.97, SUCRA: 0.99, Figure 4). Similarly, all ranking metrics agree that first generation endometrial destruction is the least preferable option (PreTA: 0.01, pBV: 0.00, SUCRA: 0.17, Figure 4). The disagreement between ranking metrics occurs for the second and third position between Mirena and second generation endometrial destruction. The two interventions are similar according to the point estimates but second generation is more precise. This leads to a greater certainty that second generation is worse than the average treatment compared to Mirena, resulting in a smaller PreTA (0.12). However, second generation beats on average more treatments than Mirena does since the relative effect of second generation is larger than that of Mirena; this results in a larger SUCRA for second generation (0.47) than for Mirena (0.37).

Results of the empirical analysis

We ended up with 232 networks to be included in the empirical analysis. There was strong agreement between hierarchies obtained by PreTAs and SUCRAs, shown by a median Kendall's tau (in the following called 'correlation') of 0.94 with interquartile range (IQR) 0.86 to 1.00). Almost half of the networks (101, 44%) had correlation of 1 while only two networks (1%) had correlation less than 0.6. The network with the smallest correlation (0.4) is shown in Appendix Figure 2 [27]; it is network of five treatments, where four of them have similar treatment effects compared to the fifth one. Thus, uncertainty in the produced treatment hierarchy is high and results in disagreement between PreTA and SUCRA rankings. The agreement between PreTAs and pBV was lower with a median correlation of 0.74 (IQR 0.61 to 0.89) and 49 networks (21%) having correlation less than 0.6 (Appendix Figure 3).

As with all ranking metrics, any disagreements between PreTAs and pBV or SUCRAs are attributed to the different ways they incorporate uncertainty in the estimation. Among treatments with similar point estimates, pBV favors treatments associated with uncertainty, as the tail of the distribution of treatments with uncertain effects is larger compared to the

tail of the distribution for treatments with similar point estimate but high precision. The probability P_{kl} tends to 0.5 with increased $var(\hat{\mu}_{kl}^{N})$; consequently, the greater the uncertainty associated with a treatment, the more its P-score tends to 0.5. A recent empirical analysis investigates the role of uncertainty in the agreement between ranking metrics and a research paper describing theoretically the interpretation and the role of uncertainty in the various ranking metrics is in preparation [19,28].

Discussion

In this paper, we derived the relative treatment effects of all treatments versus a fictional treatment of average performance. To that aim, we applied the alternative deviation from means coding to the construction of design matrices in NMA models. The application of the resulted coefficients is two-fold. First, they can be used to conveniently present NMA results in large networks without an obvious reference treatment. Such a presentation would by no means substitute the presentation of a league table, or any other way of presenting all NMA relative treatment effects, in the main manuscript or in the appendix of an NMA application. On the contrary, it may only serve as a complementary presentational tool for a quick grasp of evidence. Second, we developed a new ranking metric, PreTA, interpreted as the probability of each treatment being preferable than a fictional treatment of average performance. PreTAs can be produced in all NMAs as long as the eligibility of treatments is well justified. The notion of the average treatment refers to the average absolute efficacy among the treatments included in the systematic review. Thus, as with all ranking metrics, the interpretation of PreTAs is subject to the set of treatments compared.

The usefulness of the interpretation of the $\hat{b}_k$ coefficients depends on whether the notion of an 'average' treatment makes sense. This challenge in interpreting the coefficients, and subsequently PreTA, however, may be less pronounced in NMA compared to other applications of regression models. This is because for most categorical explanatory variables (e.g. sex, ethnicity) the average category is meaningless (and this is likely the reason that the 'deviation from means' coding is very rarely used in practice). In NMA, however, the fictional treatment (a treatment with average efficacy) could in theory be developed in the future, examined in clinical trials and included in systematic reviews. A further limitation of our method is that researchers may be inclined to use hypothesis testing when interpreting the $\hat{b}_k$ coefficients, which is not suitable. Moreover, the coloring of figure 3 and Appendix figure

1 may lead to overinterpretation of the treatment hierarchy based on the dichotomy of being better or worse than the fictional average treatment. It should be noted that being better or worse than the average treatment does not necessarily mean that a treatment is good or bad; treatments may be more or less similar between them and the entire treatment effects' distributions is the only way to get all the information about all possible comparisons.

In the presence of a reference treatment, e.g. placebo, a simple and intuitive non-probabilistic ranking metric can be obtained by ranking all relative effects against placebo. Authors of NMA often present estimated treatment effects against placebo or standard care in a forest plot, providing implicitly or explicitly a treatment hierarchy. While such a hierarchy might be appropriate in many settings, they assume that treatment effects against placebo are of primary interest for the analysis. This might not be the case in other healthcare areas where one or more established therapies exists [29] or where researchers are concerned about the quality of the evidence from placebo-controlled studies [30–32] and choose to, exclusively or complementary, analyse a network without placebo. Moreover, it should be taken into account that the amount of data associated with the reference treatment might have an impact on the judgement regarding the similarity of the treatments, when such a judgement is made by visually inspecting a forest plot of NMA effects. Point estimates against the fictional average treatment provide a solution to this ambiguity. Furthermore, data from registries can be assumed to approximate the response of an average treatment, as participants may take any of the available interventions. Thus, using such external data, absolute effects can be approximated using the point estimates against the average.

Alternative methods to avoid the reference group coding have been suggested in the literature. The application of quasi-variances [33], independently proposed as 'floating absolute risks' in epidemiology [34], do avoid setting a reference group. However, the scope of their use pertains to approximating a set of variances of the model contrasts such that the variances between any linear combination of contrasts can be derived without the disposal of the covariance matrix [35]. Thus, quasi-variances approaches target a different problem from the model described in this paper and the relevance of the estimated quantities to NMA is not clear.

Producing a treatment hierarchy in NMA is popular, with 43% of published NMAs presenting at least one ranking metric [4], but also debatable. Recent developments tackle common criticisms against ranking metrics, pertaining to arguments that they are unstable [36,37],

uncertain [38], do not differentiate between clinically important and unimportant differences [2,39], do not account for multiple outcomes [40] and are not accompanied by a measure of uncertainty [41]. In particular, recent developments include extensions of P-scores for two or more outcomes [42], incorporation of clinically important values in their calculation [42], application of multiple-criteria decision analysis [43] and partial ordering of interventions according to multiple outcomes [44]. PreTAs can be easily extended to incorporate clinically important values as shown in [42]; such probabilities will then be interpreted as the probability of a treatment being better than the average by at least a certain value.

PreTA is a viable alternative to existing ranking metrics, that can be interpreted as a probability and takes into account the entire ranking distribution. As it is also the case with PreTA, all existing ranking metrics use the distribution of NMA treatment effects to produce a hierarchy of the treatments. This hierarchy can be based either on probabilities like "which is the probability that each treatment produces the best outcome value" or "which is the probability of treatment A beating treatment B" or summaries of these probabilities. Rankograms visualise the entire ranking distributions for each treatment and SUCRAs, P-scores and mean ranks summarise these probabilities in a single number for each treatment. The interpretation of these summaries is, however, not always straightforward. The development of PreTAs enriches the decision-making arsenal with a presentational and ranking tool, which can be interpreted in a clinically meaningful way.

Declarations

Data Availability Statement

Outcome data and the code for applying our methods are available in https://github.com/esm-ispm-unibe-ch/alternativenma.

Conflicts of interest

TAF reports personal fees from Mitsubishi-Tanabe, MSD and Shionogi and a grant from Mitsubishi-Tanabe, outside the submitted work; TAF has a patent 2018-177688 pending.

Funding

AN, VC, TP and GS were supported by project funding (Grant No. 179158) from the Swiss National Science Foundation.

Authors' contributions

AN conceived the idea, contributed to the modelling, produced the results and wrote the R code and the first draft of the manuscript. VC contributed to the analysis. TP contributed to the modelling and to the R code. DM, TAF and GS contributed to the modelling, reviewed the R code and contributed to the writing. All authors read and approved the final manuscript.

Tables and Figures

Table 1. Illustration of construction of dummy variables for modelling a categorical variable with four groups in regression using reference cell coding and deviation from means coding.

Reference cell coding				**Deviation from means coding**			
	Dummy variables				Dummy variables		
Covariate	x_1	x_2	x_3	Covariate	x_1	x_2	x_3
Group 1	0	0	0	Group 1	-1	-1	-1
Group 2	1	0	0	Group 2	1	0	0
Group 3	0	1	0	Group 3	0	1	0
Group 4	0	0	1	Group 4	0	0	1
				Average*	0	0	0

Table 2. Arm-level and contrast-level NMA models using reference cell coding and deviation from means coding for a fictional network of three treatments examined in four studies.

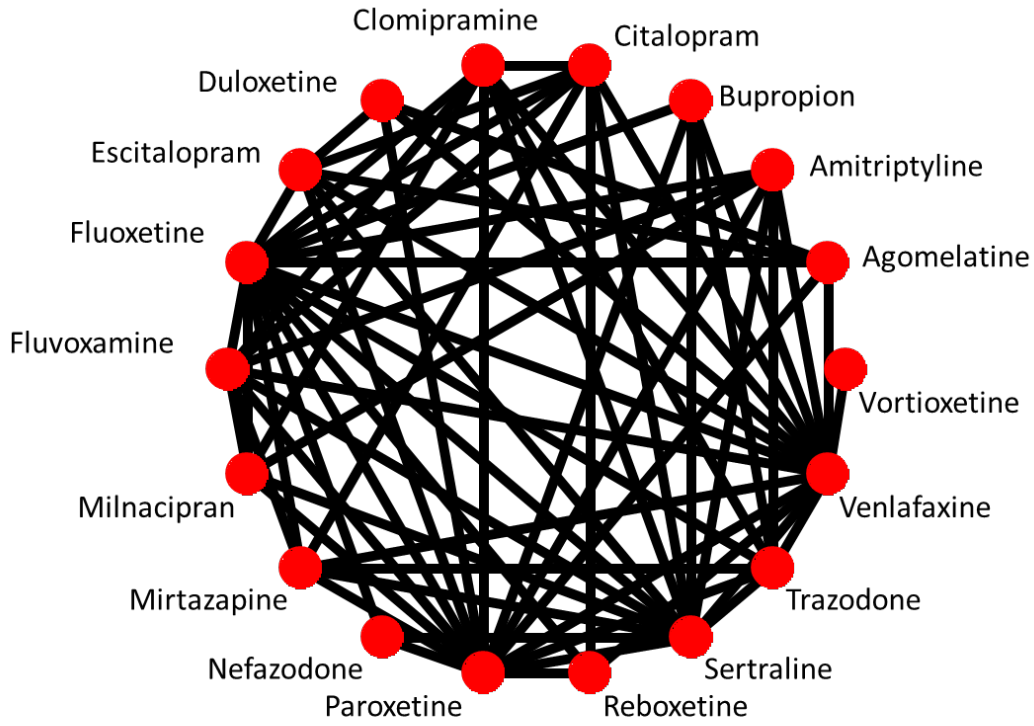
Study number, treatments compared	**Arm-based NMA**		**Contrast-based NMA**	
	Reference cell coding	Deviation from means coding	Reference cell coding	Deviation from means coding
Study 1, AB	$g(y_{1A}) = u_1$ $g(y_{1B}) = u_1 + \mu_{AB} + \delta_{1,AB}$ $\delta_{1,AB} \sim N(0, \tau^2)$	$g(y_{1A}) = u_1 - b_B - b_C$ $g(y_{1B}) = u_1 + b_B + \delta_{1,AB}$ $\delta_{1,AB} \sim N(0, \tau^2)$	$y_{1,AB} = \mu_{AB} + \varepsilon_{1,AB} + \delta_{1,AB}$ $\varepsilon_{1,AB} \sim N(0, s^2_{1,AB})$ $\delta_{1,AB} \sim N(0, \tau^2)$	$y_{1,AB} = 2b_B + b_C + \varepsilon_{1,AB} + \delta_{1,AB}$ $\varepsilon_{1,AB} \sim N(0, s^2_{1,AB})$ $\delta_{1,AB} \sim N(0, \tau^2)$
Study 2, AC	$g(y_{2A}) = u_2$ $g(y_{2C}) = u_2 + \mu_{AC} + \delta_{2,AC}$ $\delta_{2,AC} \sim N(0, \tau^2)$	$g(y_{2A}) = u_2 - b_B - b_C$ $g(y_{2C}) = u_2 + b_C + \delta_{2,AC}$ $\delta_{2,AC} \sim N(0, \tau^2)$	$y_{2,AC} = \mu_{AC} + \varepsilon_{2,AC} + \delta_{2,AC}$ $\varepsilon_{2,AB} \sim N(0, s^2_{2,AC})$ $\delta_{2,AC} \sim N(0, \tau^2)$	$y_{2,AC} = b_B + 2b_C + \varepsilon_{2,AC} + \delta_{2,AC}$ $\varepsilon_{2,AB} \sim N(0, s^2_{2,AC})$ $\delta_{2,AC} \sim N(0, \tau^2)$
Study 3, BC	$g(y_{3B}) = u_3$ $g(y_{3C}) = u_3 - \mu_{AB} + \mu_{AC} + \delta_{3,BC}$ $\delta_{3,BC} \sim N(0, \tau^2)$	$g(y_{3B}) = u_3 + b_B$ $g(y_{3C}) = u_3 + b_C + \delta_{3,BC}$ $\delta_{3,BC} \sim N(0, \tau^2)$	$y_{3,BC} = -\mu_{AB} + \mu_{AC} + \varepsilon_{3,BC} + \delta_{3,BC}$ $\varepsilon_{3,BC} \sim N(0, s^2_{3,BC})$ $\delta_{3,BC} \sim N(0, \tau^2)$	$y_{3,BC} = -b_B + b_C + \varepsilon_{3,BC} + \delta_{3,BC}$ $\varepsilon_{3,BC} \sim N(0, s^2_{3,BC})$ $\delta_{3,BC} \sim N(0, \tau^2)$
Study 4, ABC	$g(y_{4A}) = u_4$ $g(y_{4B}) = u_4 + \mu_{AB} + \delta_{4,AB}$ $g(y_{4C}) = u_4 + \mu_{AC} + \delta_{4,AC}$ $\delta_{4,AB} \sim N(0, \tau^2)$ $\delta_{4,AC} \sim N(0, \tau^2)$	$g(y_{4A}) = u_4 - b_B - b_C$ $g(y_{4B}) = u_4 + b_B + \delta_{4,AB}$ $g(y_{4C}) = u_4 + b_C + +\delta_{4,AC}$ $\delta_{4,AB} \sim N(0, \tau^2)$ $\delta_{4,AC} \sim N(0, \tau^2)$	$y_{4,AB} = \mu_{AB} + \varepsilon_{4,AB} + \delta_{4,AB}$ $y_{4,AC} = \mu_{AC} + \varepsilon_{4,AC} + \delta_{4,AC}$ $\varepsilon_{4,AB} \sim N(0, s^2_{4,AB})$ $\delta_{4,AB} \sim N(0, \tau^2)$ $\varepsilon_{4,AC} \sim N(0, s^2_{4,AC})$ $\delta_{4,AC} \sim N(0, \tau^2)$	$y_{4,AB} = 2b_B + b_C + \varepsilon_{4,AB} + \delta_{4,AB}$ $y_{4,AC} = b_B + 2b_C + \varepsilon_{4,AC} + \delta_{4,AC}$ $\varepsilon_{4,AB} \sim N(0, s^2_{4,AB})$ $\delta_{4,AB} \sim N(0, \tau^2)$ $\varepsilon_{4,AC} \sim N(0, s^2_{4,AC})$ $\delta_{4,AC} \sim N(0, \tau^2)$

Table 3. Ranking metrics for the network of antidepressants and ranks according to each ranking metric in parentheses. pBV: probability of producing the best value; SUCRA: surface under the cumulative ranking curve; PreTA: preferable than average.

	pBV	**SUCRA**	**PreTA**
Agomelatine	0.01 (6)	0.64 (6)	0.74 (8)
Amitriptyline	0.01 (7)	0.71 (5)	0.88 (4)
Bupropion	0.20 (2)	0.80 (3)	0.87 (5)
Citalopram	0.00 (17.5)	0.37 (13)	0.24 (13)
Clomipramine	0.00 (15)	0.26 (14)	0.10 (14.5)
Duloxetine	0.01 (9)	0.52 (9)	0.52 (9)
Escitalopram	0.07 (3)	0.83 (2)	0.97 (1)
Fluoxetine	0.00 (17.5)	0.23 (16)	0.01 (18)
Fluvoxamine	0.00 (12.5)	0.25 (15)	0.10 (14.5)
Milnacipran	0.01 (8)	0.48 (10)	0.46 (10)
Mirtazapine	0.03 (4)	0.75 (4)	0.91 (3)
Nefazodone	0.02 (5)	0.38 (12)	0.33 (12)
Paroxetine	0.00 (10)	0.62 (7)	0.82 (6)
Reboxetine	0.00 (15)	0.09 (18)	0.02 (16.5)
Sertraline	0.00 (11)	0.46 (11)	0.38 (11)
Trazodone	0.00 (15)	0.12 (17)	0.02 (16.5)
Venlafaxine	0.00 (12.5)	0.61 (8)	0.78 (7)
Vortioxetine	0.64 (1)	0.90 (1)	0.93 (2)

Figure 1. Panel a: Network plot of head-to-head randomized control trials comparing 18 antidepressants. Panel b: Network plot of head-to-head randomized control trials comparing 4 interventions for heavy menstrual bleeding. First and second generation interventions refer to endometrial destruction. Nodes and edges are unweighted.

Panel a



Panel b

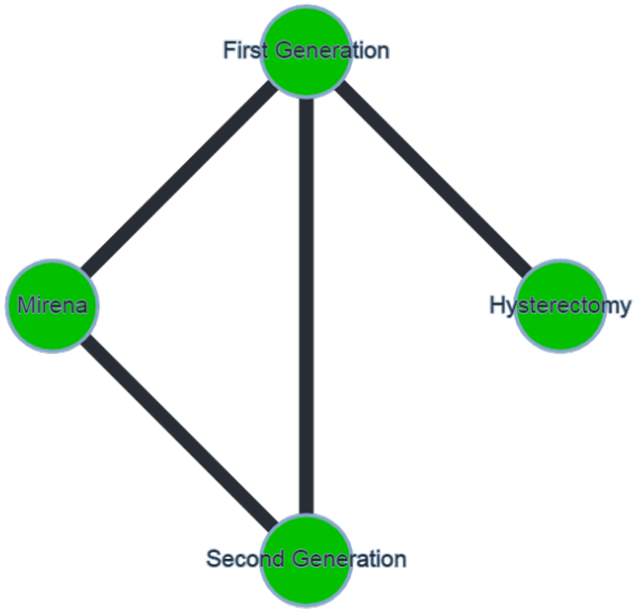


Figure 2. Odds ratios of each treatment versus fluoxetine, odds of each treatment versus odds of a fictional treatment of average response $exp(\hat{b}_k)$ and odds ratios versus vortioxetine in the network of head-to-head studies comparing 18 antidepressants. OR: odds ratio; CI: confidence interval.

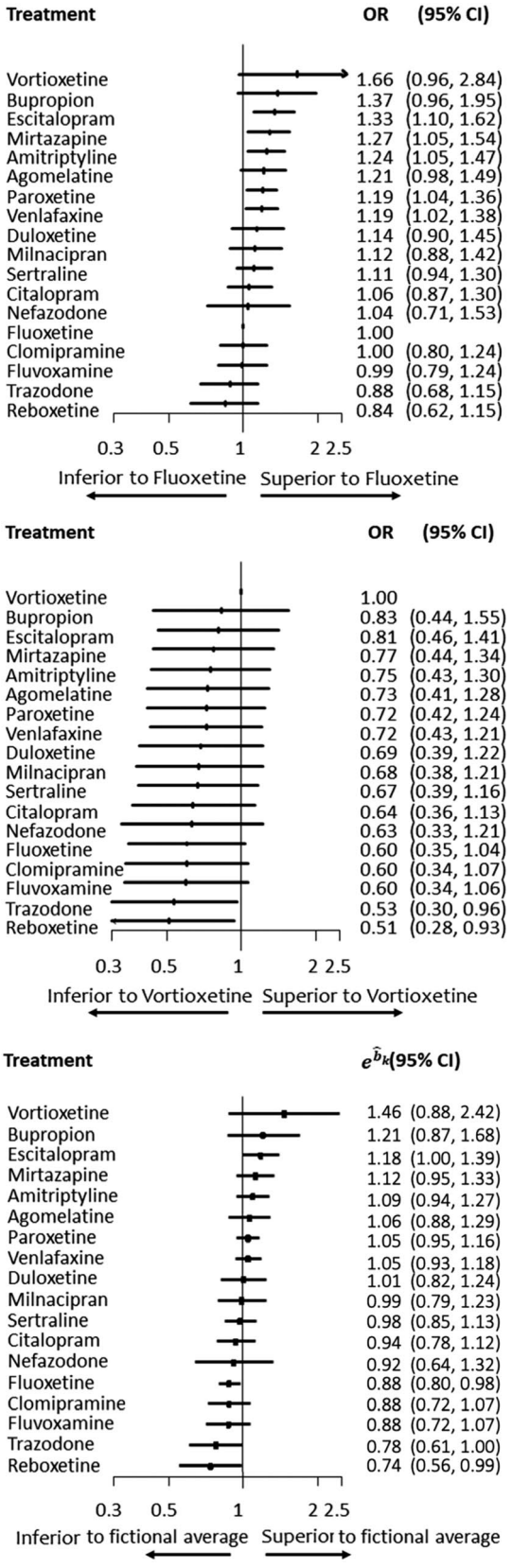


Figure 3. Classifier of interventions for the network of 18 antidepressants according to the probability of being preferable than average (PreTA).

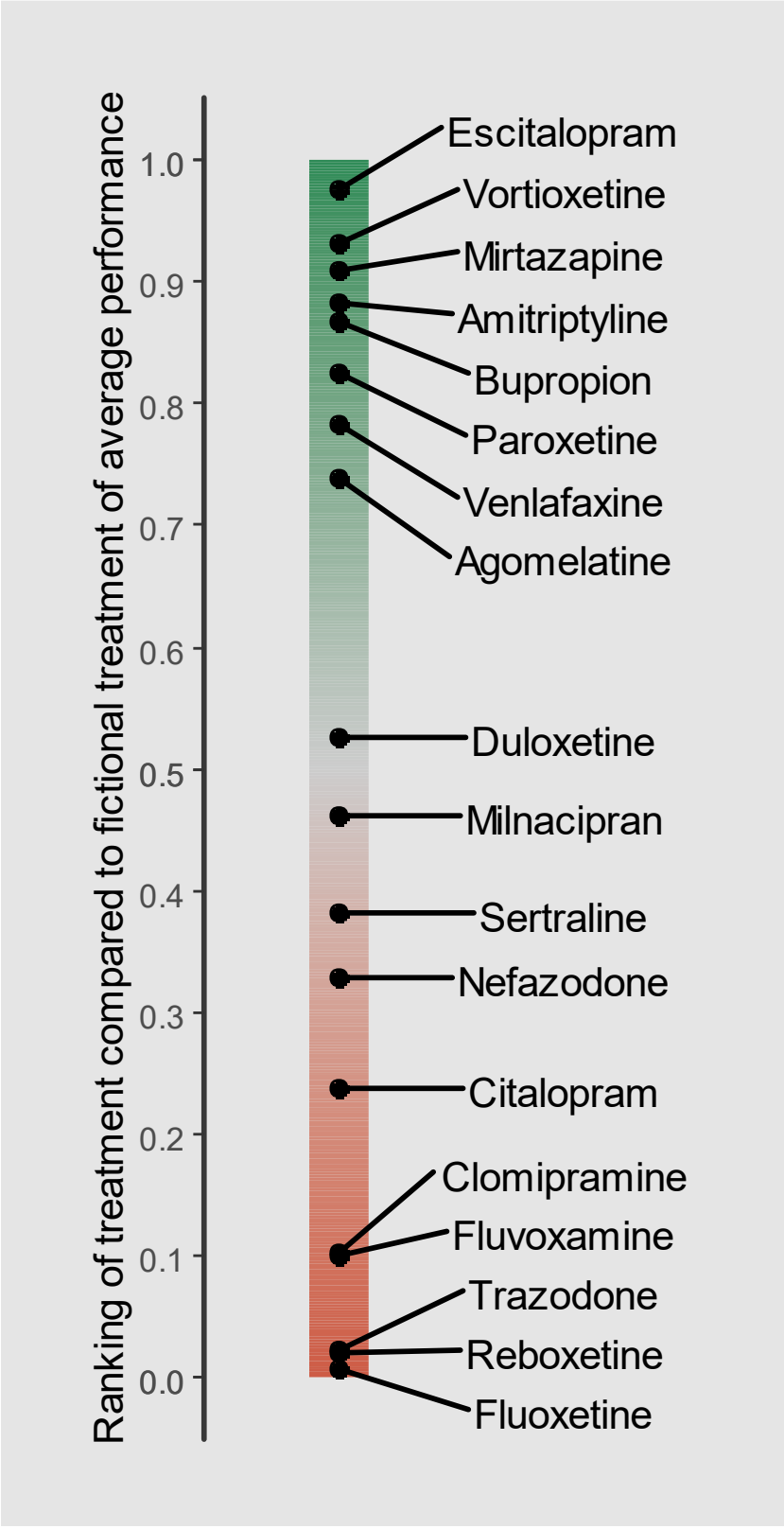
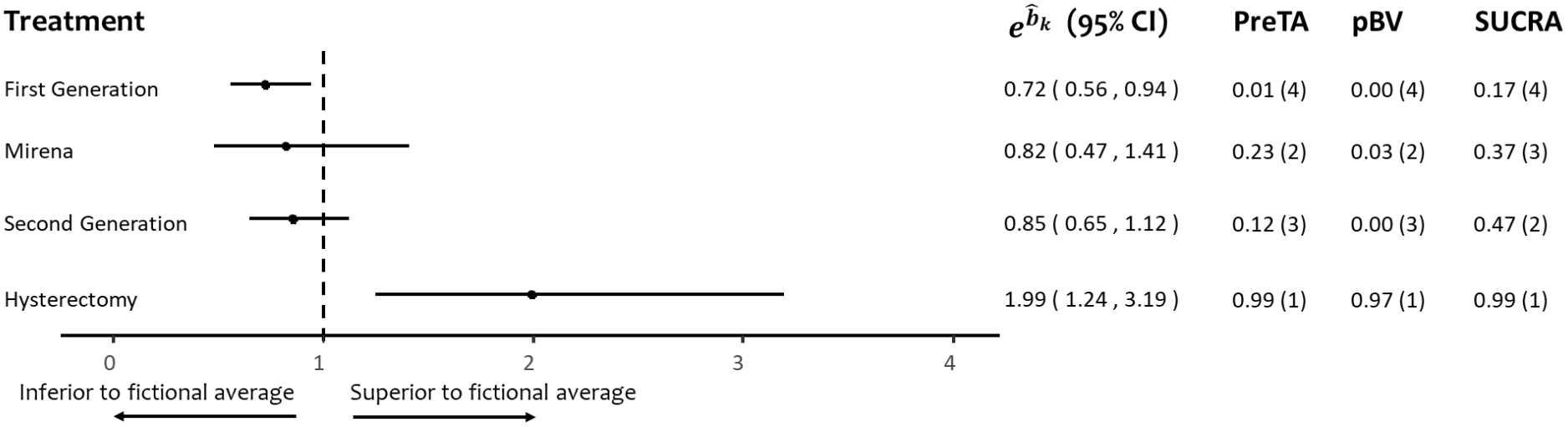


Figure 4. Odds of each treatment versus odds of a fictional treatment of average response $exp(\widehat{b}_k)$, probability of each treatment being better than the average (PreTA), probability of producing the best value (pBV) and SUCRA in the network of head-to-head studies comparing 4 interventions for heavy menstrual bleeding. Numbers in parentheses under PreTA, pBV and SUCRA represent ranks. CI: confidence interval; PreTA: preferable than average; pBV: probability of producing the best value; SUCRA: surface under the cumulative ranking curve.



Supporting Information

Available at jrsm1463-sup-0001-Figures.pdf and jrsm1463-sup-0002-supinfo.docx.

References Article 2

1 Higgins JPT, Welton NJ. Network meta-analysis: a norm for comparative effectiveness? *Lancet* 2015;**386**:628–30. doi:10.1016/S0140-6736(15)61478-7

2 Cipriani A, Higgins JPT, Geddes JR, *et al.* Conceptual and technical challenges in network meta-analysis. *Ann Intern Med* 2013;**159**:130–7. doi:10.7326/0003-4819-159-2-201307160-00008

3 Tan SH, Cooper NJ, Bujkiewicz S, *et al.* Novel presentational approaches were developed for reporting network meta-analysis. *J Clin Epidemiol* 2014;**67**:672–80. doi:10.1016/j.jclinepi.2013.11.006

4 Petropoulou M, Nikolakopoulou A, Veroniki A-A, *et al.* Bibliographic study showed improving statistical methodology of network meta-analyses published between 1999 and 2015. *J Clin Epidemiol* Published Online First: 15 November 2016. doi:10.1016/j.jclinepi.2016.11.002

5 Kanters S, Ford N, Druyts E, *et al.* Use of network meta-analysis in clinical guidelines. *Bull World Health Organ* 2016;**94**:782–4. doi:10.2471/BLT.16.174326

6 Salanti G, Ades AE, Ioannidis JPA. Graphical methods and numerical summaries for presenting results from multiple-treatment meta-analysis: an overview and tutorial. *J Clin Epidemiol* 2011;**64**:163–71. doi:10.1016/j.jclinepi.2010.03.016

7 Rücker G, Schwarzer G. Ranking treatments in frequentist network meta-analysis works without resampling methods. *BMC Med Res Methodol* 2015;**15**:58. doi:10.1186/s12874-015-0060-8

8 Hosmer DW, Lemeshow S. Interpretation of the Fitted Logistic Regression Model. In: *Applied Logistic Regression*. John Wiley & Sons, Ltd 2005. 47–90. doi:10.1002/0471722146.ch3

9 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**:80–97. doi:10.1002/jrsm.1037

10 Jansen JP, Naci H. Is network meta-analysis as valid as standard pairwise meta-analysis? It all depends on the distribution of effect modifiers. *BMC Med* 2013;**11**:159. doi:10.1186/1741-7015-11-159

11 Lu G, Welton NJ, Higgins JPT, *et al.* Linear inference for mixed treatment comparison meta-analysis: A two-stage approach. *ResSynthMeth* 2011;**2**:43–60.

12 Lu G, Ades AE. Assessing Evidence Inconsistency in Mixed Treatment Comparisons. *Journal of the American Statistical Association* 2006;**101**:447–59. doi:10.1198/016214505000001302

13 Salanti G, Higgins JPT, Ades AE, *et al.* Evaluation of networks of randomized trials. *Stat Methods Med Res* 2008;**17**:279–301. doi:10.1177/0962280207080643

14 Dias S, Sutton AJ, Ades AE, *et al.* Evidence synthesis for decision making 2: a generalized linear modeling framework for pairwise and network meta-analysis of randomized controlled trials. *Med Decis Making* 2013;**33**:607–17. doi:10.1177/0272989X12458724

15 Jackson D, Law M, Stijnen T, *et al.* A comparison of seven random-effects models for meta-analyses that estimate the summary odds ratio. *Stat Med* 2018;**37**:1059–85. doi:10.1002/sim.7588

16 van Houwelingen HC, Arends LR, Stijnen T. Advanced methods in meta-analysis: multivariate approach and meta-regression. *Stat Med* 2002;**21**:589–624.

17 Rücker G. Network meta-analysis, electrical networks and graph theory. *Res Synth Methods* 2012;**3**:312–24. doi:10.1002/jrsm.1058

18 Rücker G, Schwarzer G. Reduce dimension or reduce weights? Comparing two approaches to multi-arm studies in network meta-analysis. *Stat Med* 2014;**33**:4353–69. doi:10.1002/sim.6236

19 Salanti G, Nikolakopoulou A, Efthimiou O, *et al.* What works best? Obtaining a treatment hierarchy from network meta-analysis (in preparation).

20 Papakonstantinou T. *nmadb: Network Meta-Analysis Database API*. 2019. https://CRAN.R-project.org/package=nmadb

21 Cipriani A, Furukawa TA, Salanti G, *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* 2018;**391**:1357–66. doi:10.1016/S0140-6736(17)32802-7

22 Papakostas GI, Fava M. Does the probability of receiving placebo influence clinical trial outcome? A meta-regression of double-blind, randomized clinical trials in MDD. *Eur Neuropsychopharmacol* 2009;**19**:34–40. doi:10.1016/j.euroneuro.2008.08.009

23 Sinyor M, Levitt AJ, Cheung AH, *et al.* Does inclusion of a placebo arm influence response to active antidepressant treatment in randomized controlled trials? Results from pooled and meta-analyses. *The Journal of Clinical Psychiatry* 2010;**71**:270–9. doi:10.4088/JCP.08r04516blu

24 Salanti G, Chaimani A, Furukawa TA, *et al.* Impact of placebo arms on outcomes in antidepressant trials: systematic review and meta-regression analysis. *Int J Epidemiol* 2018;**47**:1454–64. doi:10.1093/ije/dyy076

25 Rücker G, Schwarzer G. Ranking treatments in frequentist network meta-analysis works without resampling methods. *BMC Med Res Methodol* 2015;**15**:58. doi:10.1186/s12874-015-0060-8

26 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. doi:10.1136/bmj.c3929

27 Wang L, Baser O, Kutikova L, *et al.* The impact of primary prophylaxis with granulocyte colony-stimulating factors on febrile neutropenia during chemotherapy: a systematic review and meta-analysis of randomized controlled trials. *Support Care Cancer* 2015;**23**:3131–40. doi:10.1007/s00520-015-2686-9

28 Chiocchia V, Nikolakopoulou A, Papakonstantinou T, *et al.* Empirical evaluation of the agreement between ranking metrics in network meta-analysis (in preparation).

29 Batra S, Howick J. Empirical evidence against placebo controls. *J Med Ethics* Published Online First: 9 August 2017. doi:10.1136/medethics-2016-103970

30 Turner EH, Knoepflmacher D, Shapley L. Publication Bias in Antipsychotic Trials: An Analysis of Efficacy Comparing the Published Literature to the US Food and Drug Administration Database. *PLOS Medicine* 2012;**9**:e1001189. doi:10.1371/journal.pmed.1001189

31 Ioannidis JPA, Karassa FB. The need to consider the wider agenda in systematic reviews and meta-analyses: breadth, timing, and depth of the evidence. *BMJ* 2010;**341**:c4875. doi:10.1136/bmj.c4875

32 Lee K, Bacchetti P, Sim I. Publication of clinical trials supporting successful new drug applications: a literature analysis. *PLoS Med* 2008;**5**:e191. doi:10.1371/journal.pmed.0050191

33 Ridout MS. Summarizing the Results of Fitting Generalized Linear Models to Data from Designed Experiments. In: *Statistical Modelling*. Springer, New York, NY 1989. 262–9. doi:10.1007/978-1-4612-3680-1_30

34 Easton DF, Peto J, Babiker AG a. G. Floating absolute risk: An alternative to relative risk in survival and case-control analysis avoiding an arbitrary reference group. *Statist Med* 1991;**10**:1025–35. doi:10.1002/sim.4780100703

35 Firth D, Menezes D, X R. Quasi-variances. *Biometrika* 2004;**91**:65–80. doi:10.1093/biomet/91.1.65

36 Kibret T, Richer D, Beyene J. Bias in identification of the best treatment in a Bayesian network meta-analysis for binary outcome: a simulation study. *Clin Epidemiol* 2014;**6**:451–60. doi:10.2147/CLEP.S69660

37 Mills EJ, Kanters S, Thorlund K, *et al.* The effects of excluding treatments from network meta-analyses: survey. *BMJ* 2013;**347**:f5195. doi:10.1136/bmj.f5195

38 Trinquart L, Attiche N, Bafeta A, *et al.* Uncertainty in Treatment Rankings: Reanalysis of Network Meta-analyses of Randomized Trials. *Ann Intern Med* 2016;**164**:666–73. doi:10.7326/M15-2521

39 Brignardello-Petersen R, Johnston BC, Jadad AR, *et al.* Using decision thresholds for ranking treatments in network meta-analysis results in more informative rankings. *J Clin Epidemiol* 2018;**98**:62–9. doi:10.1016/j.jclinepi.2018.02.008

40 Mbuagbaw L, Rochwerg B, Jaeschke R, *et al.* Approaches to interpreting and choosing the best treatments in network meta-analyses. *Syst Rev* 2017;**6**:79. doi:10.1186/s13643-017-0473-z

41 Veroniki AA, Straus SE, Rücker G, *et al.* Is providing uncertainty intervals in treatment ranking helpful in a network meta-analysis? *J Clin Epidemiol* 2018;**100**:122–9. doi:10.1016/j.jclinepi.2018.02.009

42 Mavridis D, Porcher R, Nikolakopoulou A, *et al.* Extensions of the probabilistic ranking metrics of competing treatments in network meta-analysis to reflect clinically important relative differences on many outcomes. *Biom J* Published Online First: 29 October 2019. doi:10.1002/bimj.201900026

43 Tervonen T, Naci H, van Valkenhoef G, *et al.* Applying Multiple Criteria Decision Analysis to Comparative Benefit-Risk Assessment: Choosing among Statins in Primary Prevention. *Med Decis Making* 2015;**35**:859–71. doi:10.1177/0272989X15587005

44 Rücker G, Schwarzer G. Resolve conflicting rankings of outcomes in network meta-analysis: Partial ordering of treatments. *Res Synth Methods* 2017;**8**:526–36. doi:10.1002/jrsm.1270

45 Lu G, Ades AE. Combination of direct and indirect evidence in mixed treatment comparisons. *Stat Med* 2004;**23**:3105–24. doi:10.1002/sim.1875

Article 3: The complexity underlying treatment rankings: how to use them and what to look at

Virginia Chiocchia[1,2], Ian R. White[3], Georgia Salanti[1]

[1] Institute of Social and Preventive Medicine, University of Bern, Bern, Switzerland

[2] Graduate School of Health Sciences, University of Bern, Bern, Switzerland

[3] Medical Research Council Clinical Trials Unit at UCL, University College London, London, UK

My contribution:

I was invited by the editors to contribute to a series of methodological and practical issues in network meta-analysis with a paper on "Understanding intervention ranking in NMA". I conceived the concept, and I had the main responsibility in drafting the manuscript and doing revisions after the review from co-authors and peer-review from BMJ Evidence-Based Medicine.

Published: Chiocchia V, White IR, Salanti G. The complexity underlying treatment rankings: how to use them and what to look at. *BMJ Evidence-Based Medicine*. doi: 10.1136/bmjebm-2021-111904

Highlights/key points

- Treatment hierarchies obtained by *SUCRA*, p_{BV}, mean ranks and mean relative effects might differ when there are large differences in the amount of data for each treatment
- Different hierarchies do not imply that one is wrong or better than the others, because the methods used to rank treatments address different "treatment hierarchy questions" based on how the "preferable treatment" is defined
- The treatment at the top of the ranking may not reflect the "best clinical choice": rankings must be considered together with relative treatment effects and quality of the evidence
- Researchers should specify in the protocol whether among the aims of the synthesis is to obtain a treatment hierarchy and, if yes, which is the "treatment hierarchy question" they aim to answer

In clinical fields where several competing treatments are available, network meta-analysis (NMA) has become an established tool to inform evidence-based decisions [1,2]. To determine which treatment is the most preferable, decision makers must account for both the quantity and the quality of the available evidence by considering both efficacy and safety outcomes as well as assessing the confidence in the obtained results [3]. It is, however, increasingly common to include in the NMA output a ranking of the competing interventions for a specific outcome of interest [4]. This article focuses on this type of rankings.

A hierarchy of treatments (or ranking) is obtained by ordering a specific ranking metric. A ranking metric is a statistic measuring the performance of an intervention and is calculated from the estimated relative treatment effects and their uncertainty in NMA [5]. A commonly used ranking metric is the point estimate of the relative treatment effects against a natural common comparator such as placebo. The rankings are unaffected by choice of comparator, so any comparator may be chosen [6]. Other commonly used metrics are the probability of producing the best outcome value, p_{BV} (sometimes called probability of being the best), and the surface under the cumulative ranking curve (*SUCRA*) or their frequentist equivalent, the P-score [7]. Treatment hierarchies are a simple and straightforward way to display the relative

performance of an intervention and aid the decision-making process, so nowadays most publications and reports present rankings [4]. Furthermore, new ranking metrics are being developed to obtain treatment hierarchies that account for important clinical and methodological aspects, such as multiple outcomes (benefits and risks), clinically important differences and the quality of the evidence.

Ranking metrics have been criticised in the literature for their lack of reliability, quoting, among other issues, limited interpretability and "instability" [8–11]. This criticism was based on the disagreement between hierarchies obtained by the different ranking metrics. Consider for example the different treatment hierarchies in **Figure 1** obtained by different ranking metrics for a network of nine antihypertensives for primary prevention of cardiovascular disease [12,13] (network graph shown in **Figure 2**). The treatment hierarchy based on p_{BV} disagrees markedly with the other hierarchies, based on relative treatment effects and *SUCRA*, particularly with respect to the top treatment. Conventional therapy, an ill-defined treatment which was evaluated in only one trial, is in the first rank in the hierarchy based on p_{BV} but only in the third/fourth and sixth rank in the hierarchies according to the relative treatment effects and *SUCRA*, respectively.

Although such examples can occur, a recent empirical study showed that they are rather rare and that in general there is a high level of agreement between the hierarchies produced by the most common ranking metrics [13]. Agreement becomes less when, as in the network of antihypertensives, there are large differences in the precision between the treatment effect estimates. These differences in precision could be produced by different data features, such as sparse or poorly connected networks, heterogeneity and inconsistency [14]. Disagreements mostly relate to hierarchies based on p_{BV}. Salanti et al also showed with theoretical examples how the uncertainty in the estimation of the relative treatment effects may affect the order of treatments in a ranking. In particular, they observed how rankings based on p_{BV} are more sensitive to differences in precision across treatment effect estimates than those based on *SUCRA*. When competing treatments have similar point estimates, p_{BV} tends to rank first the treatment with the most imprecise effect (largest confidence or credible interval); a high p_{BV} therefore tends to accompany a high probability of producing the worst value. This observation is confirmed by the empirical results in Chiocchia et al [13] and can easily be seen in the antihypertensive treatments example where the Conventional therapy drops several ranks in the hierarchy based on *SUCRA* (**Figure 1**). As displayed by the

relative treatment effects of overall mortality for each treatment versus placebo in the forest plot in **Figure 3**, the point estimates are all quite similar but the risk ratio of conventional therapy versus placebo is the only one with a large degree of uncertainty. This very imprecise effect and the large differences in the precision of the treatment effect estimates lead to the conventional therapy being the top treatment according to the p_{BV} ranking and to the disagreement between the latter and the other two rankings.

It is important to point out that all ranking metrics are statistics calculated from the data and none of them provides a "gold standard" against which each other ranking metric should be evaluated. Consequently, the criticism that some of the resulting treatment hierarchies are unreliable and unstable because they do not agree with other hierarchies is misplaced. But then, which hierarchy should one report and use to make decisions? The appropriate treatment hierarchy to use is the one resulting from the metric that answers the "treatment hierarchy question" that the systematic review is posing [14]. For example, if we are interested in "which treatment is the most likely to produce the largest positive change in the outcome" (e.g. relative drop in blood pressure or increase in quality of life) then p_{BV} will lead to the relevant treatment hierarchy. However, we think this is not the relevant treatment hierarchy question for patients. If we want to know "which treatment is likely to outperform most competitors?" then we should employ *SUCRA* rankings. Salanti et al report some examples of treatment hierarchy questions for rankings based on the most popular ranking metrics [14]. These questions and the way they are phrased are, however, not set in stone as they are suggestions based on the most common approaches and decision-making problems. Further research is needed in the field to understand what most patients and clinicians expect when they ask about the "best treatment".

Even with a careful choice of ranking metric, the treatment at the top of the resulting treatment hierarchy may not necessarily reflect the "best clinical choice". Rankings cannot be used to understand whether differences between the interventions are clinically important or not. Rankings on their own have little meaning if not presented side-by-side with measures that quantify the differences in clinical outcomes, such as mean differences or risk ratios, often presented in league tables [15]. Several choices need to be made in the full decision-making context: what outcomes are important and how do we trade-off between them? Do the observed differences reflect clinically important differences? What aspect do patients and/or clinicians value the most? How confident are we in the network meta-analysis results?

These are only some of the aspects that must be considered in the complex decision-making process. New ranking approaches have been developed to address these questions. Multi-criteria decision analysis (MCDA) is a comprehensive methodology that incorporates preference information with a benefit-risk assessment identified by explicit trade-offs across multiple outcomes [16,17]. The P-score [7] was extended to account for clinically important relative differences on more than one outcome [18] while Spie charts can be used to visualise comparative effectiveness and safety on multiple outcomes of equal or different importance to a decision-maker [19]. The Probability of Selecting a Treatment to Recommend (POST-R) incorporates important information such as the confidence in the evidence or clinical priors in the ranking algorithm [20]. A first approach to evaluate the confidence in rankings from NMA was described by Salanti et al but it has not yet been implemented into a proper framework like CINeMA [3,21]. The aim to create evidence-based guidelines also inspired the threshold analysis approach, which is not a new ranking method per se, but it informs on the robustness of treatment recommendations by quantifying how much the evidence could change before the ranking of the treatments changes [22]. In view of these new methods, NMA has the potential to provide answers to more comprehensive and complex treatment hierarchy questions and aid the decision-making process more efficiently.

If obtaining a treatment hierarchy is one of the aims of the synthesis, we recommend reviewers to specify the treatment hierarchy question a priori in the protocol, together with the appropriate ranking metric to answer that treatment hierarchy question. This is the first step to avoid misinterpreting the findings of the chosen ranking. The presented treatment hierarchy must be interpreted together with the relative treatment effects, with particular attention to the uncertainty in the estimations, as well as the quality of the synthesised evidence. More work focusing on the development of a comprehensive framework for evaluating the confidence in the rankings of treatments is needed.

Declarations

Competing Interests

IRW has received royalties as co-editor from sales of the Handbook of Meta-Analysis.

Funding

This work has been supported by the Swiss National Science Foundation (SNSF) Grant No. 179158. IRW was supported by the Medical Research Council Programme MC_UU_00004/06. The funders had no involvement in the writing of this manuscript.

Figures

Figure 1: Example of treatment hierarchies from different ranking metrics for a network of nine antihypertensive treatment for primary prevention of cardiovascular disease. ACE=Angiotensin Converting Enzyme; CCB=Calcium Channel Blockers; ARB=Angiotensin Receptor Blockers. p_{BV}: probability of producing the best value; $SUCRA$: surface under the cumulative ranking curve (calculated in frequentist setting).

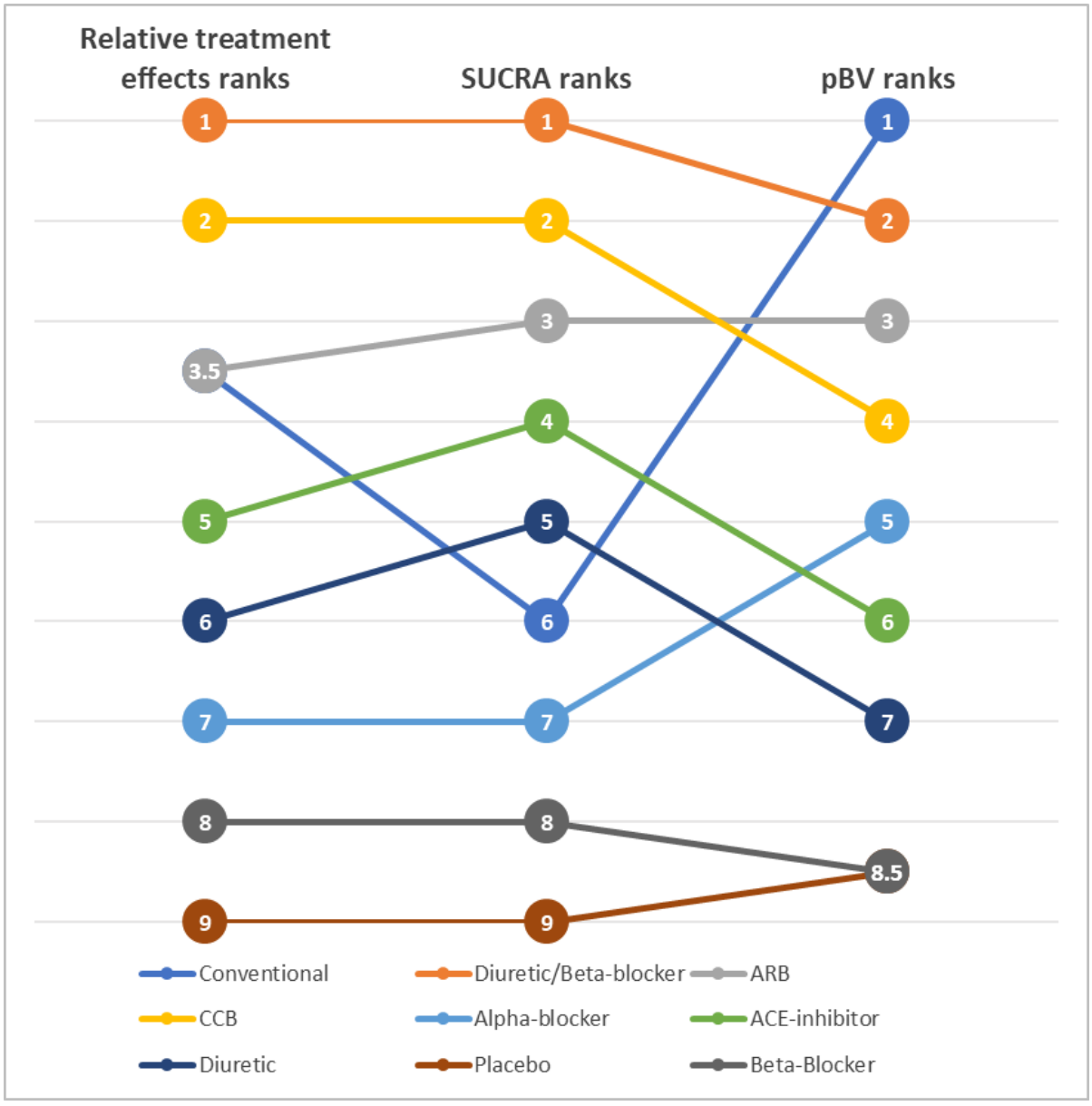


Figure 2: Graph of network of nine antihypertensive treatments for primary prevention of cardiovascular disease. Line width is proportional to inverse standard error of estimates from random effects model comparing two treatments. ACE=Angiotensin Converting Enzyme; CCB=Calcium Channel Blockers; ARB=Angiotensin Receptor Blockers.

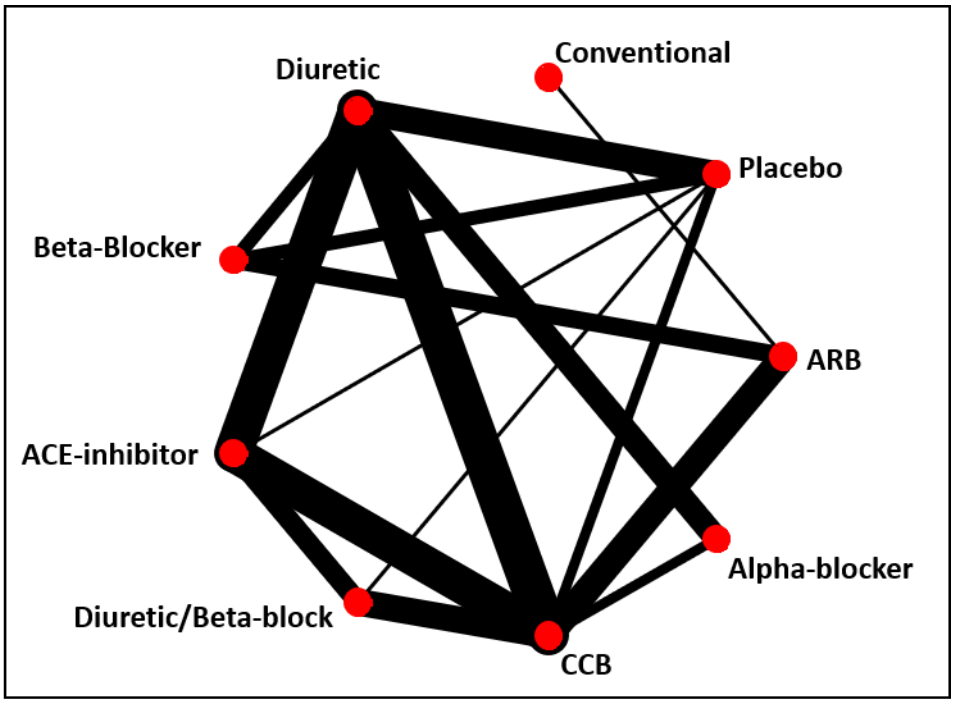
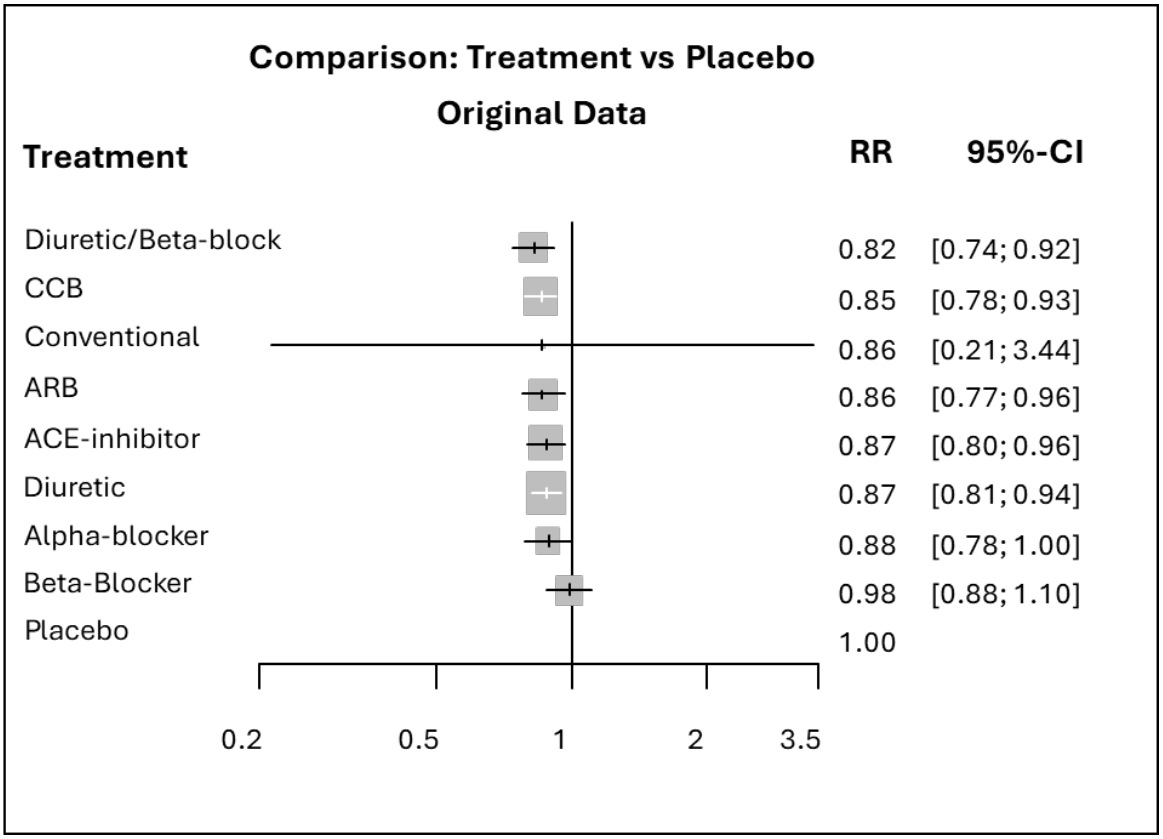


Figure 3: Forest plots of relative treatment effects of overall mortality for each treatment versus placebo. RR=risk ratio; CI=confidence interval; ACE=Angiotensin Converting Enzyme; CCB=Calcium Channel Blockers; ARB=Angiotensin Receptor Blockers.



References Article 3

1 Cipriani A, Higgins JPT, Geddes JR, *et al.* Conceptual and Technical Challenges in Network Meta-analysis. *Annals of Internal Medicine* 2013;**159**:130. doi:10.7326/0003-4819-159-2-201307160-00008

2 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**:80–97. doi:10.1002/jrsm.1037

3 Nikolakopoulou A, Higgins JPT, Papakonstantinou T, *et al.* CINeMA: An approach for assessing confidence in the results of a network meta-analysis. *PLOS Medicine* 2020;**17**:e1003082. doi:10.1371/journal.pmed.1003082

4 Petropoulou M, Nikolakopoulou A, Veroniki A-A, *et al.* Bibliographic study showed improving statistical methodology of network meta-analyses published between 1999 and 2015. *Journal of Clinical Epidemiology* 2017;**82**:20–8. doi:10.1016/j.jclinepi.2016.11.002

5 Salanti G, Ades AE, Ioannidis JPA. Graphical methods and numerical summaries for presenting results from multiple-treatment meta-analysis: an overview and tutorial. *Journal of Clinical Epidemiology* 2011;**64**:163–71. doi:10.1016/j.jclinepi.2010.03.016

6 Nikolakopoulou A, Mavridis D, Chiocchia V, *et al.* Network meta-analysis results against a fictional treatment of average performance: treatment effects and ranking metric. *Res Syn Meth* 2020;:jrsm.1463. doi:10.1002/jrsm.1463

7 Rücker G, Schwarzer G. Ranking treatments in frequentist network meta-analysis works without resampling methods. *BMC Medical Research Methodology* 2015;**15**:58. doi:10.1186/s12874-015-0060-8

8 Veroniki AA, Straus SE, Rücker G, *et al.* Is providing uncertainty intervals in treatment ranking helpful in a network meta-analysis? *J Clin Epidemiol* 2018;**100**:122–9. doi:10.1016/j.jclinepi.2018.02.009

9 Trinquart L, Attiche N, Bafeta A, *et al.* Uncertainty in Treatment Rankings: Reanalysis of Network Meta-analyses of Randomized Trials. *Annals of Internal Medicine* 2016;**164**:666. doi:10.7326/M15-2521

10 Kibret T, Richer D, Beyene J. Bias in identification of the best treatment in a Bayesian network meta-analysis for binary outcome: a simulation study. *Clinical Epidemiology* 2014;:451. doi:10.2147/CLEP.S69660

11 Mbuagbaw L, Rochwerg B, Jaeschke R, *et al.* Approaches to interpreting and choosing the best treatments in network meta-analyses. *Syst Rev* 2017;**6**:79. doi:10.1186/s13643-017-0473-z

12 Fretheim A, Odgaard-Jensen J, Brørs O, *et al.* Comparative effectiveness of antihypertensive medication for primary prevention of cardiovascular disease: systematic review and multiple treatments meta-analysis. *BMC Med* 2012;**10**:33. doi:10.1186/1741-7015-10-33

13 Chiocchia V, Nikolakopoulou A, Papakonstantinou T, *et al.* Agreement between ranking metrics in network meta-analysis: an empirical study. *BMJ Open* 2020;**10**:e037744. doi:10.1136/bmjopen-2020-037744

14 Salanti G, Nikolakopoulou A, Efthimiou O, *et al.* Introducing the treatment hierarchy question in network meta-analysis. *American Journal of Epidemiology* Published Online First: 23 November 2021. doi:10.1093/aje/kwab278

15 Hutton B, Salanti G, Caldwell DM, *et al.* The PRISMA extension statement for reporting of systematic reviews incorporating network meta-analyses of health care interventions: checklist and explanations. *Ann Intern Med* 2015;**162**:777–84. doi:10.7326/M14-2385

16 Tervonen T, Naci H, van Valkenhoef G, *et al.* Applying Multiple Criteria Decision Analysis to Comparative Benefit-Risk Assessment: Choosing among Statins in Primary Prevention. *Med Decis Making* 2015;**35**:859–71. doi:10.1177/0272989X15587005

17 van Valkenhoef G, Tervonen T, Zhao J, *et al.* Multicriteria benefit–risk assessment using network meta-analysis. *Journal of Clinical Epidemiology* 2012;**65**:394–403. doi:10.1016/j.jclinepi.2011.09.005

18 Mavridis D, Porcher R, Nikolakopoulou A, *et al.* Extensions of the probabilistic ranking metrics of competing treatments in network meta-analysis to reflect clinically important relative differences on many outcomes. *Biom J* 2019;:bimj.201900026. doi:10.1002/bimj.201900026

19 Daly CH, Mbuagbaw L, Thabane L, *et al.* Spie charts for quantifying treatment effectiveness and safety in multiple outcome network meta-analysis: a proof-of-concept study. *BMC Med Res Methodol* 2020;**20**:266. doi:10.1186/s12874-020-01128-2

20 Chaimani A, Porcher R, Sbidian E, *et al.* A Markov Chain approach for ranking treatments in network meta-analysis. Epidemiology 2019. doi:10.1101/19008722

21 Salanti G, Del Giovane C, Chaimani A, *et al.* Evaluating the Quality of Evidence from a Network Meta-Analysis. *PLoS ONE* 2014;**9**:e99682. doi:10.1371/journal.pone.0099682

22 Phillippo DM, Dias S, Welton NJ, *et al.* Threshold Analysis as an Alternative to GRADE for Assessing Confidence in Guideline Recommendations Based on Network Meta-analyses. *Annals of Internal Medicine* 2019;**170**:538. doi:10.7326/M18-3542

Article 4: Ranking competing treatments: sensitivity to the trade-off between benefits and harms on multiple clinical outcomes

Virginia Chiocchia[1,2], Toshi A. Furukawa[3], Johannes Schneider-Thoma[4], Spyridon Siafis[4], Andrea Cipriani[5,6,7], Stefan Leucht[4], Georgia Salanti[1]

[1] Institute of Social and Preventive Medicine, University of Bern, Bern, Switzerland

[2] Graduate School of Health Sciences, University of Bern, Bern, Switzerland

[3] Department of Health Promotion and Human Behavior, Kyoto University Graduate School of Medicine / School of Public Health, Kyoto, Japan

[4] Department of Psychiatry and Psychotherapy, School of Medicine, Technical University of Munich, Munich, Germany

[5] Department of Psychiatry, University of Oxford, Oxford, UK

[6] Oxford Health NHS Foundation Trust, Warneford Hospital, Oxford, UK

[7] Oxford Precision Psychiatry Lab, Oxford Health Biomedical Research Centre, Oxford, UK

My contribution:

I had the main responsibility in drafting the manuscript, performing all analyses, doing revisions after the review from co-authors.

Submitted: Chiocchia V, Furukawa T A, Schneider-Thoma J*, et al.* Ranking competing treatments: sensitivity to the trade-off between benefits and harms on multiple clinical outcomes. *Systematic Reviews*

Abstract

Background. The relative treatment effects estimated from network meta-analysis can be employed to rank treatments from the most preferable to the least preferable option. These treatment hierarchies are typically based on ranking metrics calculated from a single outcome. Some approaches have been proposed in the literature to account for multiple outcomes and individual preferences, such as the coverage area inside a spie chart, that, however, does not account for a trade-off between efficacy and safety outcomes.

We present the net-benefit standardised area within a spie chart, $SAWIS$ to explore the changes in the treatment hierarchy with different trade-offs between benefit and harms of treatments, according to a particular set of preferences.

Methods. We combine the standardised areas within spie charts for efficacy and safety/acceptability outcomes with a value λ specifying the trade-off between benefits and harms. We also describe how to derive absolute probabilities and convert outcomes on a scale between 0 and 1 for inclusion in the spie chart.

Results. We illustrate how the treatment hierarchy of three published network meta-analyses changes as the trade-off λ varies. The decrease of the $SAWIS$ quantity appears more pronounced for some drugs e.g. haloperidol. Changes in the ranking seem more frequent when SUCRA is employed as outcome measures in the spie charts.

Conclusions. $SAWIS$ should not be interpreted as a ranking metric but it is a simple approach that could help identifying which treatment is preferable when multiple outcomes are of interest and trading-off between benefits and harms is important.

Background

When multiple competing treatments are available for a specific condition, network meta-analysis (NMA) is used to identify which one is preferable by estimating relative treatment effects between each pair of treatments for a given outcome of interest (1,2). From these relative treatment effects, one can obtain summary statistics to measure the performance of an intervention. Such quantitative measures are called ranking metrics and are used to produce treatment hierarchies from the most preferable to the least preferable option. Among the most commonly reported treatment hierarchies we find those based on the probability of producing the best value, the SUCRA or its frequentist equivalent, the P-score, as well as the relative treatment effects against a reference or control treatment (e.g. placebo) (3–5). These ranking metrics are typically calculated for a single outcome, so network meta-analyses often present several treatment hierarchies for all harmful and beneficial outcomes. Extension of the existing ranking metrics that involve multiple outcomes have been recently presented. Mavridis et al. extended the idea of the P-score for multiple outcomes (6) and Daly et al. introduced a new framework, the spie charts, to measure the effectiveness or safety of each treatment on multiple outcomes (7). In a spie chart, the importance of each outcome is represented by the angle of an outcome-specific sector composing the spie chart, whose coverage area represents the quantity by which to rank the treatments. Efficacy and safety outcomes should not, however, be plotted on the same spie chart since a single value for the resulting area inside could mask important information on safety, so the authors suggest producing two separate spie charts, one for benefit and one for harmful outcomes.

The aim of this paper is to combine the areas of the two spie charts to produce a visual and numerical way to explore the sensitivity of a treatment hierarchy to the different perceptions of the trade-off between benefit and harms of treatments, subject to a particular set of preferences in terms of outcome importance. We illustrate if and how the treatment hierarchy of three published network meta-analyses changes for varying levels of the trade-off.

Motivating examples

The first example is a network of head-to-head studies investigating 18 antidepressants for the acute treatment of adults with major depressive disorder (8). We consider two efficacy dichotomous outcomes: response to treatment (defined as a reduction of at least 50% in the

score between baseline and week 8 on a standardised rating scale for depression) and remission. We also consider two outcomes about harms: acceptability (dropout due to any cause) and tolerability (dropout due to adverse events). The presentation of both beneficial and harmful outcomes is important for the clinical decision-making process, because some antidepressants, like amitriptyline, have a good efficacy profile, particularly in terms of response to treatment, but perform poorly in terms of acceptability and tolerability.

The second example is a network of placebo-controlled studies of 32 antipsychotics for the acute treatment of adults with multi-episode schizophrenia (9). We consider one efficacy outcome, overall symptoms of schizophrenia as measured by rating scales, and four safety outcomes: use of antiparkinson medication (as a proxy of extra-pyramidal symptoms), weight gain, prolactin elevation, and QTc prolongation (as a proxy of cardiac risk). Some antipsychotics show a clear distinction between their own efficacy and safety profiles. For instance, haloperidol and olanzapine are among the most effective antipsychotics but they are associated with high rates of antiparkinson medication use and weight gain, respectively.

The third example is a network of pharmacological and dietary-supplement treatments for autism spectrum disorder (10). We consider two efficacy outcomes: changes in core symptoms for social-communication difficulties and repetitive behaviours as measured by any validated scale, and a safety outcome i.e. number of patients reporting any adverse event.

Methods

We first introduce the standardised area within a spie chart as reported by Daly et al. (7) and then we illustrate how we combine it with a trade-off value. Let us consider *j=1, …, J* outcomes for *i=1, …, N* treatments. The outcome measures y_{ij} range between 0 and 1 and have weights w_j reflecting the importance of each outcome.

The standardised area within a spie chart (SAWIS)

For a specific treatment *i*, the formula for the standardised area within a spie chart (A_i^{std}) for J outcomes measures y_{ij} with weights w_j is the following.:

$$A_i^{std} = \frac{1}{2\pi}\sum_{j=1}^{J} w_j y_{ij}^2$$

The weights w_j represent the angles of the *J* sectors composing the spie chart and range between 0 and 2π, where $w_j = 0$ implies outcome *j* does not contribute to the area i.e., it is irrelevant for the purpose of ranking the treatments, and $w_j = 2\pi$ implies outcome *j* is the

sole contributor to the area i.e., it is the only outcome considered important to rank the treatments. Daly et al illustrate various methods to derive the contribution of the outcomes in terms of weights. These include but are not limited to preference elicitation from decision makers or experts, coefficients from prognostic models and, more generally, evidence in the literature (7).

The outcome measures y_{ij} must be on the same scale, such as SUCRA or an absolute probability. However, this is challenging, as dichotomous, continuous or time-to-event outcomes are often relevant to the same treatment hierarchy. Below, we describe some existing methods to convert treatment effects of dichotomous and continuous outcomes to y_i scaled between 0 and 1. When it is not possible to perform these conversion methods and/or obtain absolute probabilities, the alternative of using SUCRA as the outcome measures y_i can always be employed, as shown by Daly et al (7). Another important aspect to note is that the chosen outcomes to be included in the same spie chart must also be measured in the same direction. That means, that higher values for the efficacy outcomes indicate a higher benefit, while higher values for the safety outcomes indicate a higher harm. Therefore, to outperform its competitors, a treatment would achieve the largest area within the efficacy spie chart but the smallest area within the safety spie chart.

Transforming outcome measures on a 0 to 1 scale: Dichotomous outcomes

For each treatment *i* the absolute probabilities y_i for dichotomous outcomes can be calculated by using the odds ratios OR_i versus control from the outcome-specific network meta-analyses and the odds in the control group $odds_{control}$ which can be estimated by meta-analysing the reference arms.

$$y_i = \frac{odds_i}{1 + odds_i}$$

$$odds_i = OR_i \times odds_{control}$$

Transforming outcome measures on a 0 to 1 scale: Continuous outcomes

If the continuous outcome is measured using different scales (e.g. symptoms scores or rating scales), the mean differences MD_i of each treatment *i* versus the control group can be calculated using the relative standardised mean differences SMD_i versus the control group from the outcome-specific network meta-analysis and the pooled standard deviation

$SDpooled_{rep.study}$ from a representative study in the field reporting the outcome in the chosen scale.

$$MD_i = SMD_i * SDpooled_{rep.study}$$

The absolute mean effects M_i for each treatment *i* can then be calculated from the MD_i and the absolute mean for the control group $mean_{rep.study}$ from the chosen representative study

$$M_i = MD_i + mean_{rep.study}$$

If the continuous outcome is measured on a scale where a defined minimum and maximum value exists, the obtained absolute mean effects M_i for each treatment *i* are standardised using the minimum and maximum values, min and max, for the relevant outcome scale

$$y_i = \frac{M_i - min}{max - min}$$

If the continuous outcome is not defined within a specific range, it can be converted into a "response/risk" probabilities y_i from the control group probability $p_{control}$ and the SMD_i of group *i* versus the control group using Furukawa's method (11–13)

$$y_i = \Phi(SMD_i - \Phi^{-1}(1 - p_{control}))$$

where Φ is the cumulative standard normal distribution and Φ^{-1} its inverse. The control group probabilities $p_{control}$ represents the probability of scores of patients beyond the cut-off value *C* used to distinguish between those with and without treatment response (11)

$$p_{control} = \Phi\left(\frac{mean_{control} - C}{SD_{control}}\right)$$

If dichotomous variables defining how many patients reach the specific cut-off *C* are available, $p_{control}$ can also be estimated from a meta-analysis of proportions.

Another way to transform outcomes measured on the same scale is with the use of a *partial value function* as described by Tervonen et al (14). The idea is to bound the region in which the outcome values are likely to fall by setting two points c'_k and c''_kas the least and most preferable values, respectively. A (linear) partial value function could then be defined as $u_k(c_k) = (c_k - c'_k)/(c''_k - c'_k)$, for an outcome where larger values are preferable, and it is normalized by $u_k(c'_k) = 0$ and $u_k(c'_k) = 1$.

Benefit and harms trade-off: the net-benefit standardised area within a spie chart ($SAWIS$)

To trade-off between benefits and harms, we introduce a value λ to combine the standardised areas within the spie charts for efficacy and safety/acceptability outcomes within the same

formula. The latter will produce a numerical quantity, the net-benefit standardised area within a spie chart ($SAWIS$) that could be interpreted as the "net benefit" with the treatment, measured on a SAWIS difference scale

$$SAWIS_i = A_i^B - \lambda * A_i^H.$$

A_i^B is the (standardised) area within the spie chart for benefit from efficacy outcomes, while A_i^H is the (standardised) area within the spie chart for "harm", that includes safety and acceptability outcomes. We set $\lambda = {}^1/_u$, where *u* can vary between 1 and infinity to reflect the amount of harms we are willing to accept for an increase in benefit, measured on SAWIS scale.

$SAWIS_i$ cannot easily be interpreted as a ranking metric. In our case, the coverage area within a spie chart is a weighted sum of the efficacy or safety and acceptability outcomes measured on a 0-1 scale, and the obtained value is a difference adjusted for a specific willingness-to-pay in terms of negative outcomes (λ). In addition, it might be difficult to elicit plausible values for *u*, as the unit of measure is not a probability, change in scores or a specific outcome that the patients would be able to trade-off with harms, but the area inside a spie chart. Therefore, our approach should be employed as a sensitivity analysis to show whether and how the treatment hierarchy changes according to different trade-offs between benefits and harms, i.e. for the whole range of λ values.

Implementation

We developed an R function to implement the methods described above, incorporating the Spie chart R code provided by its authors. The user must specify the efficacy outcomes and safety/acceptability outcomes as separate vectors and, similarly, the relative vectors of weights as values between 0 and 1. The weights values must add up to 1 for the efficacy and safety/acceptability outcomes separately. The user is also required to specify outcome labels as string vectors for the two set of outcomes separately and a value, or a series of values, for the trade-off λ. As a default, the function calculates the quantity, $SAWIS_i$, for λ ranging from 0 to 1 by increment of 0.05. The function returns three objects: the benefit spie charts and the harms spie charts, both containing the plot and the value of the area within the spie chart for each treatment; and a table showing the values for the treatments at each value of λ. The R code for the function and to reproduce the plots in this paper are freely available on GitHub (https://github.com/esm-ispm-unibe-ch/nb-spie).

Results

Ranking antidepressants

We present the variability in treatment hierarchy obtained with our approach for the network of 18 antidepressants for the acute treatment of adults with major depressive disorder. We show how the hierarchy varies for different values of the trade-off λ.

We first estimated for each treatment the absolute probabilities of response or risk for the outcomes as described in the methods section. Fluoxetine was chosen as the reference drug, so we estimated the odds in this control group by meta-analysing the Fluoxetine arms.

The probability $p_{Fluoxetine}$ for response, remission, dropouts for any cause and due to side effects were 0.569, 0.347, 0.236 and 0.078, respectively, as reported in Table 1.

After consulting with clinicians, we gave a weight of 0.3 and 0.7 to the response and remission outcomes respectively; and weights of 0.7 and 0.3 to dropout due to side effects and dropout due to any cause outcomes, respectively.

The $SAWIS_i$ values for different λ are shown in Additional File 1 and illustrated in Figure 1. The $SAWIS_i$ for all drugs decreases with increasing values of λ. However, this decrease is less pronounced for some treatments, such as vortioxetine which, for the chosen weights, seems to retain its high performance for any trade-off between beneficial and harmful outcomes. Whatever the trade-off between benefits and risks, reboxetine remains the worst-performing drug, while vortioxetine, bupropion and escitalopram are consistently the best options.

Ranking antipsychotics

To transform the efficacy outcome, overall symptoms of schizophrenia, to the same scale, we selected a representative study that measures change in symptoms on the PANSS scale (15). The mean endpoint $mean_{rep.study}$ and $SDpooled_{rep.study}$ for Placebo were 98.4 and 21.4, respectively.

Since the outcomes must be between 0 and 1, we have standardised the absolute mean score using the formula $y_i = \frac{M_i - 30}{210 - 30}$ (PANSS score can range between 30 and 210). Then, the obtained value was reversed so that higher values equate to better outcomes ($1 - y_i$).

The absolute probabilities of risk for antiparkinson medication use were obtained by estimating the odds for placebo by meta-analysing the reference arms; $p_{Placebo}$ was estimated to be 0.093 as reported in Additional File 2.

The absolute risk probabilities for the remaining harmful outcomes, weight gain, prolactin elevation, and QTc prolongation, were converted from the corresponding continuous outcomes. To derive the control group probabilities $p_{Placebo}$ we used the dichotomous variables to distinguish patients with and without the response based on a cut-off *C* of at least 7% for weight gain and study-specific thresholds for prolactin elevation and QTc prolongation. The estimated $p_{Placebo}$values were 0.034, 0.019 and 0.006, for weight gain, prolactin elevation, and QTc prolongation, respectively. The obtained probabilities and corresponding SMDs for each treatment are available in Additional File 2. Due to missing data for one or more outcomes, 18 antipsychotics were not included in the calculation of the $SAWIS_i$.

After consulting with clinicians, we gave a weight of 0.4 and 0.3 to antiparkinson medication use and weight gain, respectively, to reflect the importance of these safety outcomes compared to the other two which were both given a weight of 0.15. The $SAWIS_i$ values for different λ values are shown in Additional File 3 and illustrated in Figure 2.

The $SAWIS_i$ for all drugs decreases with increasing values of λ. However, this decrease is particularly evident for haloperidol which, for the chosen weights, goes from being among the best treatments to being the worst active drug when the willingness to tolerate harms decreases. Whatever the trade-off between benefits and harms, placebo remains at the bottom of the hierarchy, while amisulpride, olanzapine and risperidone remain in the first three ranks.

Ranking pharmacological and dietary-supplement treatments for autism spectrum disorder

For this example, it was not possible to estimate absolute probabilities or use any of the conversion methods described previously due to the variety of scales used to calculate this score and the lack of a specific cut-off to define responders using these scales. Therefore, for all outcomes we first estimated the relative treatment effects (SMD for continuous outcomes and OR for the safety outcome) of each intervention versus placebo and then, produced the relative SUCRA that we employed as the outcome measures to plot in the spie charts. For the safety outcome, any adverse event, we calculated the SUCRA to reflect the fact that in the spie chart framework higher values for the safety outcomes must indicate a higher harm, as previously explained. Therefore, the corresponding SUCRA ranking is reversed compared to what the ordinary SUCRA ranking for a safety outcome would look like, i.e. the best-performing treatment in terms of rate of adverse events will have the lowest SUCRA value in

our case, instead of the highest value, as it would usually be. The calculated SUCRA values are reported in Additional File 4.

We calculated $SAWIS_i$ by giving a weight of 0.5 to both efficacy outcomes, i.e. social-communication difficulties and repetitive behaviours. Due to missing data for one or two outcome, 16 interventions were not included in the calculation of the $SAWIS_i$.

The $SAWIS_i$ values for different trade-off values are reported in Additional File 5 and illustrated in Figure 3. Again, the $SAWIS_i$ decreases with increasing values of λ, even though this decrease is nearly null for some treatments, such as folinic acid, sapropterin and sertraline. While, for some treatments the decrease in the $SAWIS_i$ values is so large that they drop several ranks in the hierarchy, moving from being the most efficacious treatments to being among the least beneficial ones, particularly risperidone (ranked 1st at λ = 0 and 15th at λ = 1) and guanfacine (ranked 6th at λ = 0 and last at λ = 1).

Unlike the previous examples, the range of $SAWIS_i$ for this example is quite broad, also including negative values. This is because the quantities here are calculated from SUCRA which estimates the probability that a treatment X outperforms its competitors Y, Z etc, and hence depends on the performance of the competitors Y, Z, etc. Consequently, SUCRA values can be large (e.g. above 0.7) for "high-performing" interventions, even when the outcome is rare (as in safety outcomes). Therefore, the SUCRA values for harms outcomes, when included in the spie charts and, in turn, in the $SAWIS_i$ formula, could then produce negative values if the corresponding SUCRAs for positive outcome are not of the same magnitude.

Discussion

In this manuscript we presented an innovative and easy-to-implement method to show if a ranking varies for different trade-offs between efficacy and safety/acceptability outcomes by combining the area within a spie chart for both benefit and harms. We tested this approach with three different datasets (and different outcomes) from network meta-analyses published in different fields of mental health; however, this approach can be generalised to other fields of medicine.

Being an extension of the spie chart approach, our method shares the same limitations. First, this is only applicable when all treatments have data available for all outcomes of interest. As the author of the original spie chart paper reports, one option is to include the outcome by assigning it a value of zero in the calculation of the areas within the spie charts for those

treatments where it is unavailable. In this way, however, such treatments are penalised for having missing outcomes. For this reason, in the antipsychotics and autism examples, we decided to exclude the treatments with missing outcome data, following implicitly the assumption that these are not irrelevant for the ranking purpose. Second, the choice of outcomes to be included and their relative importance should be based on clinical grounds. Furthermore, the choice of outcomes must be given careful consideration not only in relation to the availability of data, but also with regards to the correlation between the specific outcomes and the information they provide. For example, for antidepressants we decided to exclude the continuous outcome from rating scales as we thought they would essentially duplicate the information already given by the response rate. Then, the values produced by this method depends greatly on the choice of the outcome measures to include in the spie charts which must also be on the same scale. We described how we obtained the absolute probabilities for binary outcomes and how we converted the continuous outcomes into a dichotomous scale (11–13) but reviewers should consult available guidance in the literature to obtain absolute probabilities for other type of outcomes e.g., rate or time-to-event data (16). Also, the absolute probabilities of response (or risk) do not account for the uncertainty of the relative treatment effects, which is instead encompassed by a measure such as the SUCRA, or the P-score, which is still between 0 and 1. We have, however, shown in the third illustrative example how the use of SUCRA affect the results in our approach, creating more variation in the rankings for the range of the trade-odd values. As explained in the Results section, this is due to the nature of the SUCRA, whose value depends on the performance of a treatment compared to all of its competitors. We, therefore, recommend the use of absolute probabilities or absolute mean effects scaled between 0 and 1 over the use of SUCRA values for our approach, whenever possible.

We want to draw attention to the fact that the quantities produced by our method should not be regarded as a new ranking metric. The interpretation of the area within a spie chart is not straightforward itself as it depends on the outcome measures plotted in the charts. Additionally, the final quantity we obtain from our approach is a difference between the two coverage areas adjusted for a specified trade-off value which adds further complexity to the interpretation. Specifically, we made our trade-off equal to the ratio 1 / u, where u could be set by the answer to questions such as "how many harms could you tolerate for an increase in benefit?" so that λ ranges between 0 and 1. However, u should theoretically be in the same

unit as A_i^H , the value of the area within the harms spie chart, to make the final value of our method interpreted as a proper ranking metric. Future research could expand this method and focus on the interpretation of the trade-off λ and the produced value. We recommend instead to use our method as a sensitivity analysis to check if a specific ranking produced in NMA stays consistent when multiple outcomes and different preferences are considered. These preferences are expressed by the importance of the different outcomes, represented by the weights in the spie chart formula and by the trade-off λ that allows to indicate "how much" one is willing to tolerate to see an increase in benefit. As this trade-off is bound to remain very subjective, it is even more important to examine the variability of the ranking for all plausible values of λ.

In the broader context of comparing treatments, health economic modelling and, specifically, cost-effectiveness analysis is sometimes used on top of NMA results to assess the performance of competing treatments accounting for costs. Our proposed approach draws from the cost-effectiveness methodology and uses the net-benefit concept to trade-off between harms and benefits. However, costs cannot be considered the same as harms as they can also vary substantially by country and reimbursement systems. When the interest lies in ranking treatments according to the cost-effectiveness profile, a proper economic evaluation approach would be required.

Various visualization tools have been proposed lately to facilitate communication of results for multiple outcomes (17,18). However, unlike our new approach, they all assume the different benefits and harms outcomes have the same importance. We believe that the method described in this paper could help clinicians, patients and policy makers to make decisions on which treatment is preferable when multiple outcomes are of interest and trading-off between benefits and harms is important.

List of abbreviations

NMA: Network meta-analysis

SUCRA: Surface under the cumulative ranking curve

OR: Odds ratio

MD: Mean difference

SMD: Standardised mean difference

SD: Standard deviation

SAWIS: Net-benefit standardised area within a spie chart

Declarations

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Availability of data and materials

The datasets generated and analysed, and the code to replicate the results study are available in the GitHub repository https://github.com/esm-ispm-unibe-ch/nb-spie.

Competing Interests

TAF reports personal fees from Kyoto University Original, grants and personal fees from Mitsubishi-Tanabe, personal fees from SONY, grants and personal fees from Shionogi, outside the submitted work, patents 2020-548587 and 2022-082495 pending, and intellectual properties for Kokoro-app licensed to Mitsubishi-Tanabe. TAF is Deputy Editor for the Evidence-Based Mental Health journal. SL is Associate Editor for the Evidence-Based Mental Health journal. GS is Section Editor for the Evidence-Based Mental Health journal. AC is Editor for the Evidence-Based Mental Health journal.

Funding

This work was supported by the Swiss National Science Foundation (SNSF) Grant No. 179158.

Authors' contributions

VC and GS conceived the idea and designed the study. VC performed all analysis, produced the results and wrote the R code and the first draft of the manuscript. TA, JST, SS, AC and SL provided the data and clinical input. All authors contributed to the interpretation of the results, read and approved the final manuscript.

Tables and Figures

Table 1: Probabilities of response to treament, remission, dropout due to any cause and dropout due to side effects, estimated from the network of 18 antidepressants for the acute treatment of adults with major depressive disorder

	Response	Remission	Dropout for any cause	Dropout for side effects
agomelatine	0.613	0.362	0.208	0.054
amitriptyline	0.621	0.381	0.268	0.120
bupropion	0.645	0.459	0.249	0.095
citalopram	0.583	0.333	0.228	0.074
clomipramine	0.569	0.378	0.315	0.171
duloxetine	0.601	0.383	0.295	0.150
escitalopram	0.638	0.393	0.212	0.066
fluoxetine	0.569	0.347	0.236	0.078
fluvoxamine	0.569	0.365	0.276	0.098
milnacipran	0.596	0.352	0.249	0.070
mirtazapine	0.629	0.371	0.246	0.096
nefazodone	0.579	0.342	0.273	0.117
paroxetine	0.611	0.375	0.244	0.092
reboxetine	0.524	0.295	0.329	0.162
sertraline	0.596	0.357	0.234	0.067
trazodone	0.540	0.340	0.278	0.109
venlafaxine	0.610	0.377	0.262	0.124
vortioxetine	0.682	0.413	0.176	0.062

Figure 1: $SAWIS_i$ for the network of 18 antidepressants for the acute treatment of major depressive disorder. The benefit spie chart included response and remission with weights 0.3 and 0.7, respectively, and the harm spie chart included dropout due to side effects and due to any cause with weights 0.7 and 0.3, respectively.

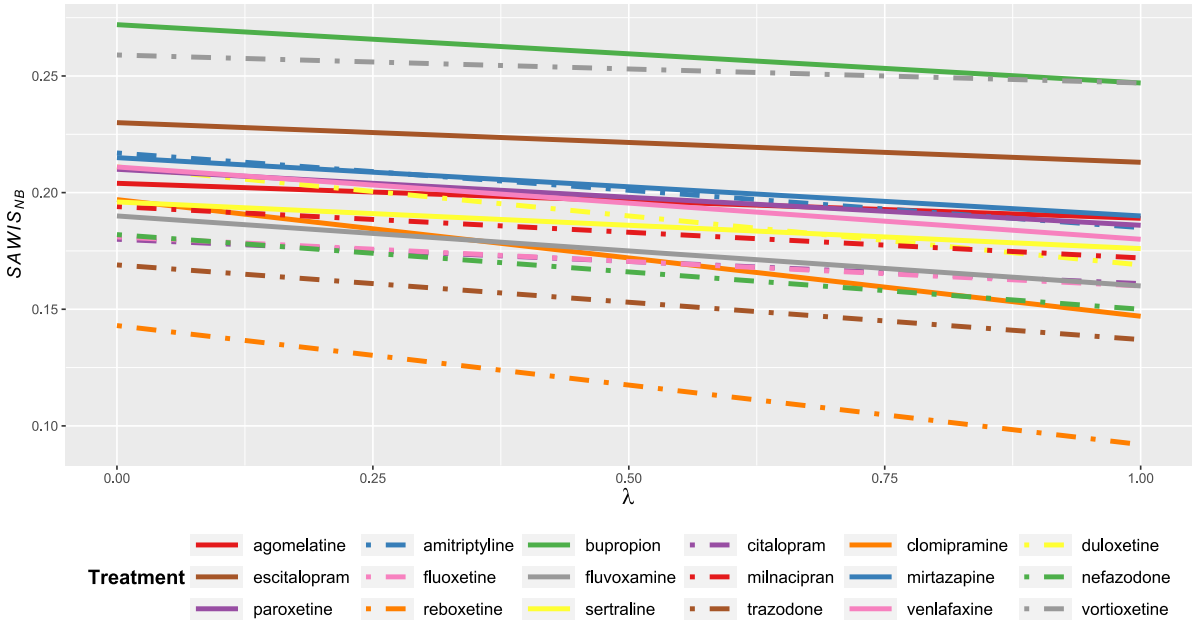


Figure 2: $SAWIS_i$ quantity for the network of antipsychotics for the acute treatment of multi-episode schizophrenia. The benefit spie chart included only one efficacy outcome, overall symptoms of schizophrenia, and the harm spie chart included antiparkinson medication use, weight gain, prolactin elevation, and QTc prolongation with weights 0.4, 0.3, 0.15 and 0.15, respectively.

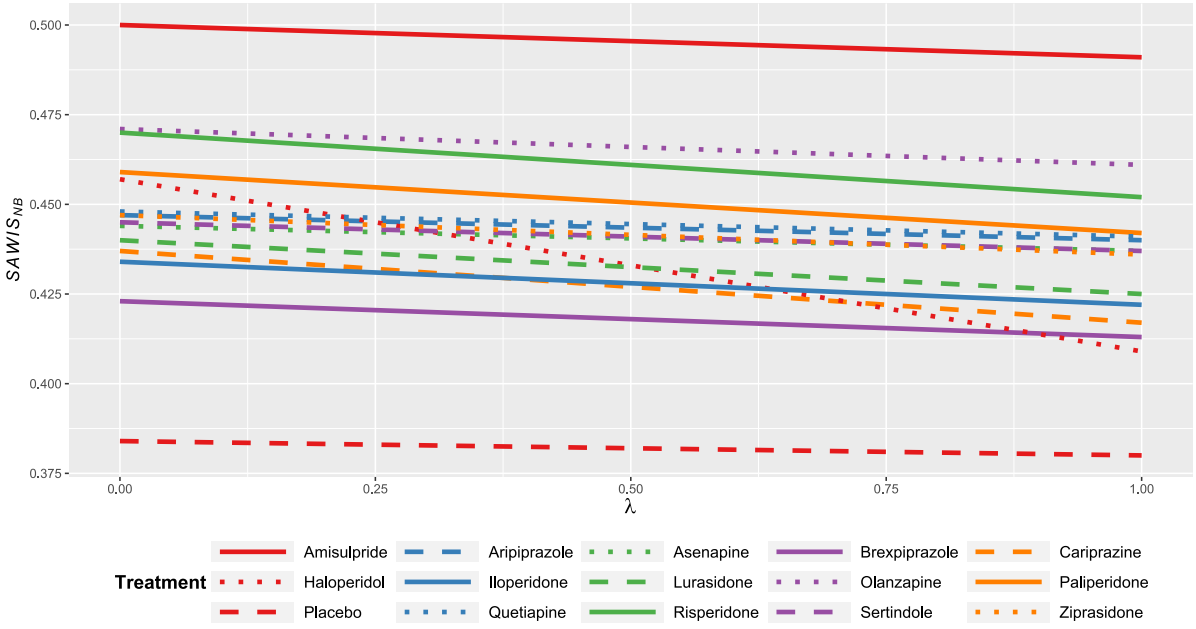
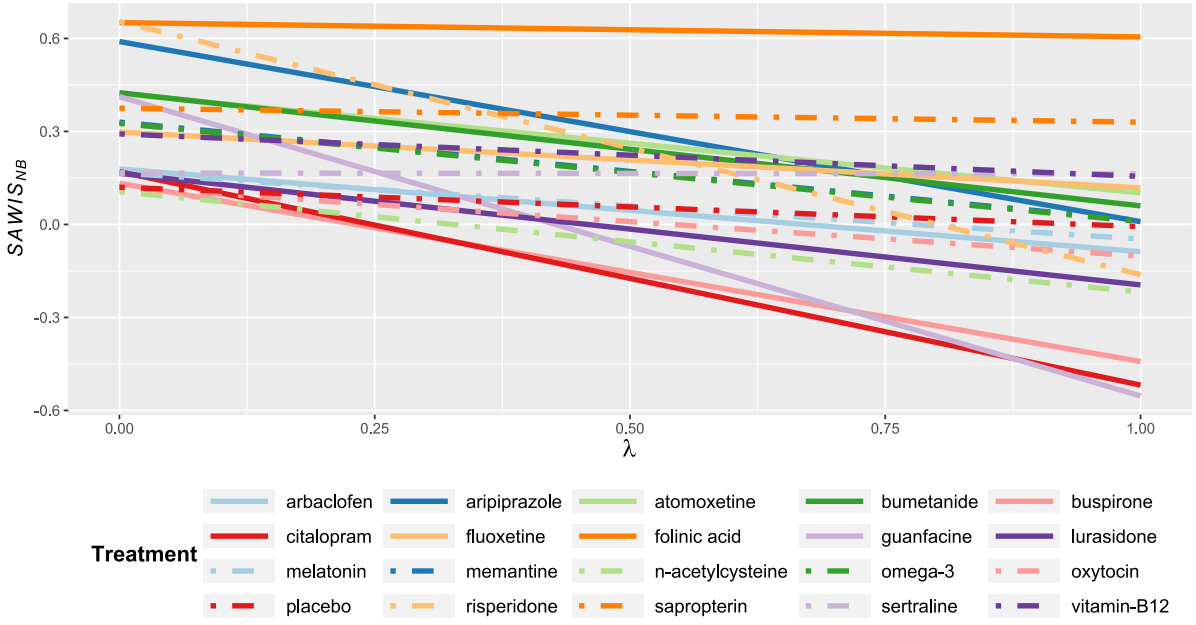


Figure 3: $SAWIS_i$ quantity for the network of pharmacological and dietary-supplement treatments for autism spectrum disorder. The benefit spie chart included two efficacy outcomes, changes in core symptoms for social-communication difficulties, and repetitive behaviours with weights 0.5 each, and the safety spie chart included one outcome, any adverse event.



Additional files

Additional file 1: Values of the $SAWIS$ for the network of 18 antidepressants

	agom	amit	bupr	cita	clom	dulo	esci	fluo	fluv	miln	mirt	nefa	paro	rebo	sert	traz	venl	vort
0	0.204	0.217	0.272	0.18	0.197	0.211	0.23	0.181	0.19	0.194	0.215	0.182	0.21	0.143	0.196	0.169	0.211	0.259
0.5	0.20325	0.2154	0.27075	0.17905	0.1945	0.2089	0.22915	0.17995	0.1885	0.1929	0.21375	0.1804	0.2088	0.14045	0.195	0.1674	0.20945	0.2584
0.1	0.2025	0.2138	0.2695	0.1781	0.192	0.2068	0.2283	0.1789	0.187	0.1918	0.2125	0.1788	0.2076	0.1379	0.194	0.1658	0.2079	0.2578
0.15	0.20175	0.2122	0.26825	0.17715	0.1895	0.2047	0.22745	0.17785	0.1855	0.1907	0.21125	0.1772	0.2064	0.13535	0.193	0.1642	0.20635	0.2572
0.2	0.201	0.2106	0.267	0.1762	0.187	0.2026	0.2266	0.1768	0.184	0.1896	0.21	0.1756	0.2052	0.1328	0.192	0.1626	0.2048	0.2566
0.25	0.20025	0.209	0.26575	0.17525	0.1845	0.2005	0.22575	0.17575	0.1825	0.1885	0.20875	0.174	0.204	0.13025	0.191	0.161	0.20325	0.256
0.3	0.1995	0.2074	0.2645	0.1743	0.182	0.1984	0.2249	0.1747	0.181	0.1874	0.2075	0.1724	0.2028	0.1277	0.19	0.1594	0.2017	0.2554
0.35	0.19875	0.2058	0.26325	0.17335	0.1795	0.1963	0.22405	0.17365	0.1795	0.1863	0.20625	0.1708	0.2016	0.12515	0.189	0.1578	0.20015	0.2548
0.4	0.198	0.2042	0.262	0.1724	0.177	0.1942	0.2232	0.1726	0.178	0.1852	0.205	0.1692	0.2004	0.1226	0.188	0.1562	0.1986	0.2542
0.45	0.19725	0.2026	0.26075	0.17145	0.1745	0.1921	0.22235	0.17155	0.1765	0.1841	0.20375	0.1676	0.1992	0.12005	0.187	0.1546	0.19705	0.2536
0.5	0.1965	0.201	0.2595	0.1705	0.172	0.19	0.2215	0.1705	0.175	0.183	0.2025	0.166	0.198	0.1175	0.186	0.153	0.1955	0.253
0.55	0.19575	0.1994	0.25825	0.16955	0.1695	0.1879	0.22065	0.16945	0.1735	0.1819	0.20125	0.1644	0.1968	0.11495	0.185	0.1514	0.19395	0.2524
0.6	0.195	0.1978	0.257	0.1686	0.167	0.1858	0.2198	0.1684	0.172	0.1808	0.2	0.1628	0.1956	0.1124	0.184	0.1498	0.1924	0.2518
0.65	0.19425	0.1962	0.25575	0.16765	0.1645	0.1837	0.21895	0.16735	0.1705	0.1797	0.19875	0.1612	0.1944	0.10985	0.183	0.1482	0.19085	0.2512
0.7	0.1935	0.1946	0.2545	0.1667	0.162	0.1816	0.2181	0.1663	0.169	0.1786	0.1975	0.1596	0.1932	0.1073	0.182	0.1466	0.1893	0.2506
0.75	0.19275	0.193	0.25325	0.16575	0.1595	0.1795	0.21725	0.16525	0.1675	0.1775	0.19625	0.158	0.192	0.10475	0.181	0.145	0.18775	0.25
0.8	0.192	0.1914	0.252	0.1648	0.157	0.1774	0.2164	0.1642	0.166	0.1764	0.195	0.1564	0.1908	0.1022	0.18	0.1434	0.1862	0.2494
0.85	0.19125	0.1898	0.25075	0.16385	0.1545	0.1753	0.21555	0.16315	0.1645	0.1753	0.19375	0.1548	0.1896	0.09965	0.179	0.1418	0.18465	0.2488
0.9	0.1905	0.1882	0.2495	0.1629	0.152	0.1732	0.2147	0.1621	0.163	0.1742	0.1925	0.1532	0.1884	0.0971	0.178	0.1402	0.1831	0.2482
0.95	0.18975	0.1866	0.24825	0.16195	0.1495	0.1711	0.21385	0.16105	0.1615	0.1731	0.19125	0.1516	0.1872	0.09455	0.177	0.1386	0.18155	0.2476
1	0.189	0.185	0.247	0.161	0.147	0.169	0.213	0.16	0.16	0.172	0.19	0.15	0.186	0.092	0.176	0.137	0.18	0.247

Additional file 2: Outcomes for the network of 32 antipsychotics for the treatment of schizophrenia

	id	Prob efficacy	Prob antiparkinson	Prob weight gain	SMD weight	Prob prolactin	SMD prolactin	Prob QTc Prolongation	SMD qtc
Amisulpride	1	0.707	0.141	0.036	0.031	0.055	0.480	0.038	0.759
Aripiprazole	2	0.669	0.127	0.050	0.180	0.011	-0.203	0.006	0.000
Asenapine	3	0.666	0.113	0.076	0.392	0.028	0.168	0.012	0.269
Brexpiprazole	4	0.650	0.152	0.051	0.194	0.026	0.134	0.004	-0.086
Cariprazine	5	0.661	0.219	0.054	0.217	0.015	-0.102	0.005	-0.048
Chlorpromazine	6	0.673	0.212	0.125	0.678	0.026	0.127	#N/A	#N/A
Clopenthixol	7	0.669	0.516	0.032	-0.021	#N/A	#N/A	#N/A	#N/A
Clozapine	8	0.726	0.041	0.131	0.704	#N/A	#N/A	#N/A	#N/A
Flupentixol	9	0.671	0.331	0.043	0.114	0.016	-0.057	#N/A	#N/A
Fluphenazine	10	0.650	0.404	#N/A	#N/A	#N/A	#N/A	#N/A	#N/A
Haloperidol	11	0.676	0.340	0.047	0.156	0.084	0.698	0.007	0.099
Iloperidone	12	0.659	0.145	0.109	0.594	0.029	0.185	0.014	0.337
Levomepromazine	13	0.623	0.131	0.118	0.642	#N/A	#N/A	#N/A	#N/A
Loxapine	14	0.673	0.324	0.067	0.331	#N/A	#N/A	#N/A	#N/A
Lurasidone	15	0.663	0.192	0.041	0.087	0.037	0.287	0.004	-0.104
Molindone	16	0.669	0.297	0.008	-0.586	#N/A	#N/A	#N/A	#N/A
Olanzapine	17	0.686	0.097	0.138	0.738	0.027	0.147	0.011	0.228
Paliperidone	18	0.678	0.156	0.082	0.434	0.186	1.182	0.007	0.049
Penfluridol	19	0.667	0.344	#N/A	#N/A	#N/A	#N/A	#N/A	#N/A
Perazine	20	0.655	0.068	0.030	-0.054	0.004	-0.592	#N/A	#N/A
Perphenazine	21	0.687	0.263	#N/A	#N/A	#N/A	#N/A	#N/A	#N/A
Pimozide	22	0.656	0.804	#N/A	#N/A	0.024	0.095	#N/A	#N/A
Placebo	23	0.620	0.093	0.034	0.000	0.019	0.000	0.006	0.000
Quetiapine	24	0.669	0.101	0.098	0.534	0.015	-0.084	0.009	0.162
Risperidone	25	0.685	0.170	0.083	0.439	0.177	1.148	0.011	0.225
Sertindole	26	0.667	0.090	0.119	0.649	0.052	0.451	0.053	0.917
Sulpiride	27	0.677	0.249	0.025	-0.126	#N/A	#N/A	#N/A	#N/A

Thioridazine	28	0.685	0.103	#N/A	#N/A	#N/A	#N/A	#N/A	#N/A
Thiothixene	29	0.695	0.420	#N/A	#N/A	#N/A	#N/A	#N/A	#N/A
Trifluoperazine	30	0.649	0.288	0.026	-0.112	#N/A	#N/A	#N/A	#N/A
Ziprasidone	31	0.669	0.163	0.039	0.063	0.028	0.167	0.017	0.412
Zotepine	32	0.692	0.198	0.170	0.873	0.000	-1.782	#N/A	#N/A
Zuclopenthixol	33	0.681	0.286	0.100	0.544	#N/A	#N/A	#N/A	#N/A

Additional file 3: Values of the $SAWIS$ for the network of antipsychotics

λ	Amisulpride	Aripiprazole	Asenapine	Brexpiprazole	Cariprazine	Haloperidol	Iloperidone	Lurasidone	Olanzapine	Paliperidone	Placebo	Quetiapine	Risperidone	Sertindole	Ziprasidone
0	0.5	0.447	0.444	0.423	0.437	0.457	0.434	0.44	0.471	0.459	0.384	0.448	0.47	0.445	0.447
0.05	0.49955	0.44665	0.44365	0.4225	0.436	0.4546	0.4334	0.43925	0.4705	0.45815	0.3838	0.44765	0.4691	0.4446	0.44645
0.1	0.4991	0.4463	0.4433	0.422	0.435	0.4522	0.4328	0.4385	0.47	0.4573	0.3836	0.4473	0.4682	0.4442	0.4459
0.15	0.49865	0.44595	0.44295	0.4215	0.434	0.4498	0.4322	0.43775	0.4695	0.45645	0.3834	0.44695	0.4673	0.4438	0.44535
0.2	0.4982	0.4456	0.4426	0.421	0.433	0.4474	0.4316	0.437	0.469	0.4556	0.3832	0.4466	0.4664	0.4434	0.4448
0.25	0.49775	0.44525	0.44225	0.4205	0.432	0.445	0.431	0.43625	0.4685	0.45475	0.383	0.44625	0.4655	0.443	0.44425
0.3	0.4973	0.4449	0.4419	0.42	0.431	0.4426	0.4304	0.4355	0.468	0.4539	0.3828	0.4459	0.4646	0.4426	0.4437
0.35	0.49685	0.44455	0.44155	0.4195	0.43	0.4402	0.4298	0.43475	0.4675	0.45305	0.3826	0.44555	0.4637	0.4422	0.44315
0.4	0.4964	0.4442	0.4412	0.419	0.429	0.4378	0.4292	0.434	0.467	0.4522	0.3824	0.4452	0.4628	0.4418	0.4426
0.45	0.49595	0.44385	0.44085	0.4185	0.428	0.4354	0.4286	0.43325	0.4665	0.45135	0.3822	0.44485	0.4619	0.4414	0.44205
0.5	0.4955	0.4435	0.4405	0.418	0.427	0.433	0.428	0.4325	0.466	0.4505	0.382	0.4445	0.461	0.441	0.4415
0.55	0.49505	0.44315	0.44015	0.4175	0.426	0.4306	0.4274	0.43175	0.4655	0.44965	0.3818	0.44415	0.4601	0.4406	0.44095
0.6	0.4946	0.4428	0.4398	0.417	0.425	0.4282	0.4268	0.431	0.465	0.4488	0.3816	0.4438	0.4592	0.4402	0.4404
0.65	0.49415	0.44245	0.43945	0.4165	0.424	0.4258	0.4262	0.43025	0.4645	0.44795	0.3814	0.44345	0.4583	0.4398	0.43985
0.7	0.4937	0.4421	0.4391	0.416	0.423	0.4234	0.4256	0.4295	0.464	0.4471	0.3812	0.4431	0.4574	0.4394	0.4393
0.75	0.49325	0.44175	0.43875	0.4155	0.422	0.421	0.425	0.42875	0.4635	0.44625	0.381	0.44275	0.4565	0.439	0.43875
0.8	0.4928	0.4414	0.4384	0.415	0.421	0.4186	0.4244	0.428	0.463	0.4454	0.3808	0.4424	0.4556	0.4386	0.4382
0.85	0.49235	0.44105	0.43805	0.4145	0.42	0.4162	0.4238	0.42725	0.4625	0.44455	0.3806	0.44205	0.4547	0.4382	0.43765
0.9	0.4919	0.4407	0.4377	0.414	0.419	0.4138	0.4232	0.4265	0.462	0.4437	0.3804	0.4417	0.4538	0.4378	0.4371
0.95	0.49145	0.44035	0.43735	0.4135	0.418	0.4114	0.4226	0.42575	0.4615	0.44285	0.3802	0.44135	0.4529	0.4374	0.43655
1	0.491	0.44	0.437	0.413	0.417	0.409	0.422	0.425	0.461	0.442	0.38	0.441	0.452	0.437	0.436

Additional file 4: Outcomes for the network of 36 treatments for autism spectrum disorder

	id	social_SUCRA	repbehav_SUCRA	ae_SUCRA
arbaclofen	1	0.5591	0.2111	0.5171
aripiprazole	2	0.7387	0.7966	0.7625
atomoxetine	3	0.4593	0.7951	0.5657
balovaptan	4	0.3452	#N/A	0.4071
bumetanide	5	0.5984	0.7014	0.6043
buspirone	6	0.3401	0.3870	0.7586
carnosine	7	0.4933	0.3512	#N/A
cholesterol	8	0.6743	#N/A	#N/A
citalopram	9	0.4472	0.3731	0.8296
dimethylglycine	10	0.8216	#N/A	0.2075
donepezil	11	0.2807	#N/A	#N/A
fluoxetine	12	0.4549	0.6236	0.4239
folinic acid	13	0.8491	0.7624	0.2148
guanfacine	14	0.4408	0.7937	0.9825
L1-79	15	0.6872	0.8622	#N/A
lamotrigine	16	0.3145	0.3591	#N/A
lurasidone	17	0.4942	0.2912	0.5999
mecamylamine	18	0.1853	0.3729	#N/A
melatonin	19	0.4632	0.3431	0.4617
memantine	20	0.3828	0.7163	0.5642
n-acetylcysteine	21	0.1884	0.4205	0.5692
omega-3	22	0.6539	0.4747	0.5626
oxytocin	23	0.3773	0.3126	0.4710
placebo	24	0.3660	0.3269	0.3557
probiotics	25	0.6682	0.3031	#N/A
riluzole	26	0.5580	#N/A	0.0960
risperidone	27	0.7519	0.8624	0.9041
sapropterin	28	0.5988	0.6256	0.2125
sertraline	29	0.4131	0.4012	0.0589
simvastatin	30	0.1946	0.1861	#N/A
sulforaphane	31	0.2934	0.3938	#N/A
tianeptine	32	0.5555	#N/A	#N/A
tideglusib	33	0.7778	0.6461	#N/A
vitamin-B12	34	0.5330	0.5486	0.3706
vitamin-D	35	0.4229	0.2583	#N/A
whey-protein	36	0.6174	#N/A	#N/A

Additional file 5: Values of the $SAWIS$ for the network of treatments for autism spectrum disorder

λ	arbaclofen	aripiprazole	atomoxetine	bumetanide	buspirone	citalopram	fluoxetine	folinic acid	guanfacine	lurasidone	melatonin	memantine	n-acetylcysteine	omega 3	oxytocin	placebo	risperidone	sapropterin	sertraline	vitamin B12
0	0.179	0.590	0.422	0.425	0.133	0.170	0.298	0.651	0.412	0.165	0.166	0.330	0.106	0.326	0.120	0.120	0.655	0.375	0.166	0.292
0.05	0.166	0.561	0.406	0.407	0.104	0.136	0.289	0.649	0.364	0.147	0.155	0.314	0.090	0.310	0.109	0.114	0.614	0.373	0.166	0.285
0.1	0.152	0.532	0.390	0.389	0.076	0.101	0.280	0.646	0.316	0.129	0.145	0.298	0.074	0.294	0.098	0.107	0.573	0.371	0.166	0.278
0.15	0.139	0.503	0.374	0.370	0.047	0.067	0.271	0.644	0.267	0.111	0.134	0.282	0.057	0.278	0.087	0.101	0.532	0.368	0.166	0.271
0.2	0.126	0.474	0.358	0.352	0.018	0.032	0.262	0.642	0.219	0.093	0.123	0.266	0.041	0.263	0.076	0.095	0.492	0.366	0.165	0.265
0.25	0.112	0.445	0.342	0.334	-0.011	-0.002	0.253	0.640	0.171	0.075	0.113	0.251	0.025	0.247	0.065	0.088	0.451	0.364	0.165	0.258
0.3	0.099	0.416	0.326	0.316	-0.040	-0.036	0.244	0.637	0.123	0.057	0.102	0.235	0.009	0.231	0.053	0.082	0.410	0.362	0.165	0.251
0.35	0.086	0.387	0.310	0.297	-0.068	-0.071	0.235	0.635	0.074	0.039	0.091	0.219	-0.007	0.215	0.042	0.076	0.369	0.359	0.165	0.244
0.4	0.072	0.358	0.294	0.279	-0.097	-0.105	0.226	0.633	0.026	0.021	0.081	0.203	-0.024	0.199	0.031	0.069	0.328	0.357	0.165	0.237
0.45	0.059	0.329	0.278	0.261	-0.126	-0.140	0.217	0.630	-0.022	0.003	0.070	0.187	-0.040	0.183	0.020	0.063	0.287	0.355	0.165	0.230
0.5	0.046	0.300	0.262	0.243	-0.155	-0.174	0.208	0.628	-0.071	-0.015	0.060	0.171	-0.056	0.168	0.009	0.057	0.247	0.353	0.165	0.224
0.55	0.032	0.270	0.246	0.224	-0.183	-0.208	0.199	0.626	-0.119	-0.033	0.049	0.155	-0.072	0.152	-0.002	0.050	0.206	0.350	0.164	0.217
0.6	0.019	0.241	0.230	0.206	-0.212	-0.243	0.190	0.623	-0.167	-0.051	0.038	0.139	-0.088	0.136	-0.013	0.044	0.165	0.348	0.164	0.210
0.65	0.005	0.212	0.214	0.188	-0.241	-0.277	0.181	0.621	-0.215	-0.069	0.028	0.123	-0.105	0.120	-0.024	0.037	0.124	0.346	0.164	0.203
0.7	-0.008	0.183	0.198	0.170	-0.270	-0.312	0.172	0.619	-0.264	-0.087	0.017	0.107	-0.121	0.104	-0.035	0.031	0.083	0.344	0.164	0.196
0.75	-0.021	0.154	0.182	0.151	-0.298	-0.346	0.163	0.617	-0.312	-0.105	0.006	0.092	-0.137	0.088	-0.047	0.025	0.042	0.341	0.164	0.189
0.8	-0.035	0.125	0.166	0.133	-0.327	-0.380	0.154	0.614	-0.360	-0.123	-0.004	0.076	-0.153	0.072	-0.058	0.018	0.001	0.339	0.164	0.182
0.85	-0.048	0.096	0.150	0.115	-0.356	-0.415	0.145	0.612	-0.408	-0.141	-0.015	0.060	-0.169	0.057	-0.069	0.012	-0.039	0.337	0.163	0.176
0.9	-0.061	0.067	0.134	0.097	-0.385	-0.449	0.136	0.610	-0.457	-0.159	-0.026	0.044	-0.186	0.041	-0.080	0.006	-0.080	0.335	0.163	0.169
0.95	-0.075	0.038	0.118	0.078	-0.413	-0.484	0.127	0.607	-0.505	-0.177	-0.036	0.028	-0.202	0.025	-0.091	-0.001	-0.121	0.332	0.163	0.162
1	-0.088	0.009	0.102	0.060	-0.442	-0.518	0.118	0.605	-0.553	-0.195	-0.047	0.012	-0.218	0.009	-0.102	-0.007	-0.162	0.330	0.163	0.155

References Article 4

1 Caldwell DM, Ades AE, Higgins JP. Simultaneous comparison of multiple treatments: combining direct and indirect evidence. *BMJ* 2005;**331**:897–900. doi:10.1136/bmj.331.7521.897

2 Lu G, Ades AE. Combination of direct and indirect evidence in mixed treatment comparisons. *Statistics in Medicine* 2004;**23**:3105–24. doi:10.1002/sim.1875

3 Salanti G, Ades AE, Ioannidis JP. Graphical methods and numerical summaries for presenting results from multiple-treatment meta-analysis: an overview and tutorial. *JClinEpidemiol* 2011;**64**:163–71. doi:10.1016/j.jclinepi.2010.03.016

4 Rücker G, Schwarzer G. Ranking treatments in frequentist network meta-analysis works without resampling methods. *BMC Med Res Methodol* 2015;**15**:58. doi:10.1186/s12874-015-0060-8

5 Nikolakopoulou A, Mavridis D, Chiocchia V, *et al.* Network meta-analysis results against a fictional treatment of average performance: treatment effects and ranking metric. *Res Syn Meth* 2020;:jrsm.1463. doi:10.1002/jrsm.1463

6 Mavridis D, Porcher R, Nikolakopoulou A, *et al.* Extensions of the probabilistic ranking metrics of competing treatments in network meta-analysis to reflect clinically important relative differences on many outcomes. *Biom J* 2019;:bimj.201900026. doi:10.1002/bimj.201900026

7 Daly CH, Mbuagbaw L, Thabane L, *et al.* Spie charts for quantifying treatment effectiveness and safety in multiple outcome network meta-analysis: a proof-of-concept study. *BMC Med Res Methodol* 2020;**20**:266. doi:10.1186/s12874-020-01128-2

8 Cipriani A, Furukawa TA, Salanti G, *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. *The Lancet* 2018;**391**:1357–66. doi:10.1016/S0140-6736(17)32802-7

9 Huhn M, Nikolakopoulou A, Schneider-Thoma J, *et al.* Comparative efficacy and tolerability of 32 oral antipsychotics for the acute treatment of adults with multi-episode schizophrenia: a systematic review and network meta-analysis. *The Lancet* 2019;**394**:939–51. doi:10.1016/S0140-6736(19)31135-3

10 Siafis S, Çıray O, Wu H, *et al.* Pharmacological and dietary-supplement treatments for autism spectrum disorder: a systematic review and network meta-analysis. *Molecular Autism* 2022;**13**:10. doi:10.1186/s13229-022-00488-4

11 da Costa BR, Rutjes AW, Johnston BC, *et al.* Methods to convert continuous outcomes into odds ratios of treatment response and numbers needed to treat: meta-epidemiological study. *International Journal of Epidemiology* 2012;**41**:1445–59. doi:10.1093/ije/dys124

12 Furukawa TA, Cipriani A, Barbui C, *et al.* Imputing response rates from means and standard deviations in meta-analyses. *International Clinical Psychopharmacology* 2005;**20**:49–52.

13 Furukawa TA, Leucht S. How to Obtain NNT from Cohen's d: Comparison of Two Methods. *PLOS ONE* 2011;**6**:e19070. doi:10.1371/journal.pone.0019070

14 Tervonen T, van Valkenhoef G, Buskens E, *et al.* A stochastic multicriteria model for evidence-based decision making in drug benefit-risk analysis. *Statist Med* 2011;**30**:1419–28. doi:10.1002/sim.4194

15 Beasley CM, Sanger T, Satterlee W, *et al.* Olanzapine versus placebo: results of a double-blind, fixed-dose olanzapine trial. *Psychopharmacology* 1996;**124**:159–67. doi:10.1007/BF02245617

16 Dias S, Sutton AJ, Ades AE, *et al.* Evidence synthesis for decision making 2: a generalized linear modeling framework for pairwise and network meta-analysis of randomized controlled trials. *MedDecisMaking* 2013;**33**:607–17. doi:10.1177/0272989X12458724

17 Seo M, Furukawa TA, Veroniki AA, *et al.* The Kilim plot: A tool for visualizing network meta-analysis results for multiple outcomes. *Research Synthesis Methods* 2021;**12**:86–95. doi:10.1002/jrsm.1428

18 Ostinelli EG, Efthimiou O, Naci H, *et al.* Vitruvian plot: a visualisation tool for multiple outcomes in network meta-analysis. *Evid Based Mental Health* 2022;:ebmental-2022-300457. doi:10.1136/ebmental-2022-300457

Article 5: ROB-MEN: a tool to assess risk of bias due to missing evidence in network meta-analysis

Virginia Chiocchia[1,2], Adriani Nikolakopoulou[1,3], Julian PT Higgins[4], Matthew J Page[5], Theodoros Papakonstantinou[1,3], Andrea Cipriani[6,7], Toshi A Furukawa[8], George CM Siontis[9], Matthias Egger[1,4,10], Georgia Salanti[1]

[1]Institute of Social and Preventive Medicine, University of Bern, Bern, Switzerland

[2]Graduate School for Health Sciences, University of Bern, Switzerland

[3]Institute of Medical Biometry and Statistics, Faculty of Medicine and Medical Center, University of Freiburg, Germany

[4]Population Health Sciences, Bristol Medical School, University of Bristol, Bristol, United Kingdom

[5]School of Public Health and Preventive Medicine, Monash University, Melbourne, Australia

[6]Department of Psychiatry, University of Oxford, Oxford, UK

[7]Oxford Health NHS Foundation Trust, Warneford Hospital, Oxford, UK

[8]Department of Health Promotion and Human Behavior, Kyoto University Graduate School of Medicine and School of Public Health, Kyoto, Japan

[9]Department of Cardiology, Bern University Hospital, Inselspital, Bern, Switzerland

[10]Centre for Infectious Disease Epidemiology and Research, University of Cape Town, Cape Town, South Africa

My contribution:

I had the main responsibility in drafting the manuscript and doing revisions after the review from co-authors and peer-review from BMC Medicine. I conceived the idea, oversaw the project, and developed the web application to implement the risk of bias tool.

Published: Chiocchia V, Nikolakopoulou A, Higgins J P T, *et al*. ROB-MEN: a tool to assess risk of bias due to missing evidence in network meta-analysis. *BMC Medicine* (2021) 19:304

doi: 10.1186/s12916-021-02166-3

Abstract

Background

Selective outcome reporting and publication bias threaten the validity of systematic reviews and meta-analyses and can affect clinical decision-making. A rigorous method to evaluate the impact of this bias on the results of network meta-analyses of interventions is lacking. We present a tool to assess the Risk Of Bias due to Missing Evidence in Network meta-analysis (ROB-MEN).

Methods

ROB-MEN first evaluates the risk of bias due to missing evidence for each of the possible pairwise comparison that can be made between the interventions in the network. This step considers possible bias due to the presence of studies with unavailable results (within-study assessment of bias) and the potential for unpublished studies (across-study assessment of bias). The second step combines the judgements about the risk of bias due to missing evidence in pairwise comparisons with (i) the contribution of direct comparisons to the network meta-analysis estimates, (ii) possible small-study effects evaluated by network meta-regression, and (iii) any bias from unobserved comparisons. Then, a level of "low risk", "some concerns" or "high risk" for the bias due to missing evidence is assigned to each estimate, which is our tool's final output.

Results

We describe the methodology of ROB-MEN step-by-step using an illustrative example from a published NMA of non-diagnostic modalities for the detection of coronary artery disease in patients with low risk acute coronary syndrome. We also report a full application of the tool on a larger and more complex published network of 18 drugs from head-to-head studies for the acute treatment of adults with major depressive disorder.

Conclusions

ROB-MEN is the first tool for evaluating the risk of bias due to missing evidence in network meta-analysis and applies to networks of all sizes and geometry. The use of ROB-MEN is facilitated by an R Shiny web application that produces the Pairwise Comparisons and ROB-MEN Table and is incorporated in the reporting bias domain of the CINeMA framework and software.

Keywords: risk of bias, missing evidence, network meta-analysis, evidence synthesis, publication bias, selective outcome reporting, reporting bias.

Background

A challenging issue in evidence-based medicine is the bias introduced by the selective non-reporting of primary studies or results. Failure to report all findings can lead to results being missing from a meta-analysis. Either a whole study may remain unpublished, commonly referred to as 'publication bias', or specific results may not be reported in a publication, usually referred to as 'selective outcome reporting bias' or 'selective non-reporting of results'.

Several methods are available to investigate such bias in pairwise meta-analysis [1]. These include generic approaches, for example, comparisons of study protocols with published reports and comparison of results obtained from published versus unpublished sources, as well as statistical methods (e.g. funnel plots [2–4], tests for small-study effects [2,5–7] and selection models [8,9]). Recently, a tool to evaluate Risk Of Bias due to Missing Evidence (ROB-ME) integrated these approaches into an overall assessment of the risk of bias due to missing evidence in pairwise meta-analysis [10].

Network meta-analysis extends pairwise meta-analysis to enable multiple treatments comparison by combining direct and indirect evidence within a network of randomised trials or other comparative studies. Several of the numerical approaches to evaluate bias developed for pairwise meta-analysis have been adapted to the network meta-analysis setting [11–15]. Still, a rigorous methodology for assessing the risk of bias due to missing results in network meta-analysis estimates is currently lacking.

To address this gap, we developed the Risk Of Bias due to Missing Evidence in Network meta-analysis (ROB-MEN) tool, which incorporates qualitative and quantitative methods. We assume that investigators assembled studies into a coherent network according to a pre-specified protocol, checked the assumptions and deemed them plausible, and used appropriate statistical methods to obtain relative treatment effects for pairs of interventions. Then, ROB-MEN can be used to assess the risk of bias due to missing evidence in each of the relative treatment effects estimated in network meta-analysis. We illustrate the ROB-MEN approach step by step using a network meta-analysis of non-invasive diagnostic tests for coronary artery disease [16]. We also report an application of the tool to a network of 18 antidepressants from head-to-head studies [17].

Methods

The ROB-MEN tool was developed between April and November 2020 within the CINeMA framework to evaluate confidence in results from network meta-analysis [18,19]. The authors are epidemiologists, statisticians, systematic reviewers, trialists, and health services researchers, many of whom are involved with Cochrane systematic reviews, methods groups, and training events. The initial proposal drew on existing methods for assessing selective outcome reporting bias [20] and publication bias [2,5,8] in pairwise meta-analysis, as summarised in the Cochrane Handbook for Systematic Reviews of Interventions [1]. A draft tool was developed in line with the preliminary version of the ROB-ME tool [10] and presented to all co-authors. Improvements and modifications were informed by relevant methodological literature, previously published tools for assessing methodological quality of meta-analyses and by the authors' experience of developing tools to assess the risk of bias in randomised and non-randomised studies, and systematic reviews [21,22]. The group met several times to discuss the approach and agreed on the tool's structure, content and step-wise application. An R Shiny web application to facilitate the implementation of ROB-MEN for the users was developed alongside the tool's conceptual framework by two of the co-authors and checked by the whole group. Refinements were made following feedback received also from training and research events.

We outline the methodology using the example of a network of randomised controlled trials comparing non-invasive diagnostic strategies for the detection of coronary artery disease in patients presenting with symptoms suggestive of an acute coronary syndrome [16]. The outcome of interest is referral to coronary angiography, for which the network included 18 trials comparing exercise electrocardiogram (ECG), single-photon emission computed tomography-myocardial perfusion imaging (SPECT-MPI), coronary computed tomographic angiography (CCTA), cardiovascular magnetic resonance (CMR), stress echocardiography (Stress echo), and standard care. Standard care was based on the discretion of the clinicians or local diagnostic strategies. The network graph is shown in Fig. 1a, and a summary of the network meta-analysis methods and results is available in Additional file 1.

Overview of ROB-MEN

In ROB-MEN, 'bias due to missing evidence' refers to bias arising when some study results are unavailable because of their results. This situation may, for example, arise because of non-

significant p-values, small magnitudes of effect, or harmful effects. It can be due to two types of missing evidence, as described in the recently developed ROB-ME tool [10]: i) the selective reporting of results within studies that are published or otherwise known to exist, called "*within study assessment of bias*" in the tool; ii) studies that remain entirely unpublished and are not known to exist, referred to as *"across-study assessment of bias"* (see also the glossary in Box 1).

In network meta-analysis, estimates of treatment effects are derived by combining direct and indirect evidence. Direct evidence refers to evidence about pairs of treatments that have been directly compared within studies (e.g. the 8 pairwise comparisons with data shown in Fig. 1a). Indirect evidence refers to evidence on pairs of treatments that is "indirectly" derived from the sources of direct evidence via a common comparator or chain of comparisons (Box 1). In ROB-MEN, we first evaluate the likely risk of bias due to missing evidence for each pairwise comparison between the interventions of interest, irrespective of the availability of direct evidence (Fig. 1b). We then consider the risk of bias from pairwise comparisons and their contribution to each estimate [23] with the additional risk of bias from indirect comparisons and any evidence of small-study effects to evaluate the overall risk of bias due to missing evidence in each network meta-analysis estimate.

Two tables that record the assessments for each pairwise comparison and each estimate are at the tool's core: the *Pairwise Comparisons Table* and the *ROB-MEN Table* (see Tables 1 and 2 for examples). Both tables are completed separately for each outcome in the review. The Pairwise Comparisons Table facilitates the assessments in the ROB-MEN Table. The output of the Pairwise Comparisons Table provides judgement on possible bias due to missing evidence for each of the possible comparisons made from the interventions in the network. The ROB-MEN Table is the main output of the tool. It combines the information from the Pairwise Comparisons Table with (i) information about the structure and the amount of data in the network and (ii) the potential impact of missing evidence on the network meta-analysis results to reach an overall judgement about the risk of bias for each estimate. Fig. 2 summarises the process. An R Shiny web application (https://cinema.ispm.unibe.ch/rob-men/) facilitates the ROB-MEN process, including creating the two core tables, as described in Additional file 2 and Additional file 3 [24].

Box 1: Glossary of terms.

Pairwise comparisons: all treatment comparisons in the network irrespective of the availability of data. A network with T treatments has T(T-1)/2 pairwise comparisons. Depending on whether there are studies reporting the studied outcome, the pairwise comparisons can be distinguished into *observed for this outcome, observed for other outcomes,* and *unobserved*.

Direct evidence: The evidence available (statistical information derived from data) about a pairwise comparison that is available from direct, within-study information about that comparison.

Indirect evidence: The evidence available (statistical information derived from data) about a pairwise comparison that is *not* available from within-study information, i.e. is obtained indirectly via a common comparator or chain of comparisons.

'Only direct' estimate: Relative treatment effect estimated in an network meta-analysis that is derived only from direct evidence.

'Only indirect' estimate: Relative treatment effect estimated in an network meta-analysis that is derived only from indirect evidence.

Mixed estimate: Relative treatment effect estimated in an network meta-analysis that is derived from both direct and indirect evidence.

Network meta-analysis estimate: estimates of relative treatment effects derived from network meta-analysis; these can be distinguished into 'Only direct', 'Only indirect' and Mixed estimates.

Within-study assessment of bias due to missing evidence: bias arising from missing results due to selective outcome reporting i.e. results being reported, but not others, within studies published or otherwise known to exist.

Across -study assessment of bias due to missing evidence: bias introduced from missing studies because they are entirely unpublished i.e. not known to exist.

Risk of bias due to missing evidence in pairwise comparisons

The assessment of bias due to missing evidence in all possible pairwise comparisons follows the ROB-ME tool for pairwise meta-analysis [10]. Like ROB-ME, we consider the studies contributing to the network meta-analysis of the outcome of interest and the studies contributing to networks of other outcomes in a systematic review. Such studies are informative about possible selective non-reporting of the outcome being addressed in the current network meta-analysis. ROB-MEN differs from ROB-ME by considering all possible pairwise comparisons between the interventions in the network. There may be missing evidence for any directly observed comparisons and missing evidence for the indirect comparisons that were *not* observed among the included studies. The possible pairwise

comparisons between the interventions involved in the network, that is, all combinations of two treatments, are organised into three groups:

A. *"observed for this outcome":* the comparisons for which there is direct evidence contributing to the network meta-analysis for the current outcome
B. *"observed for other outcomes":* the pairwise comparisons for which there is direct evidence only for other outcomes in the systematic review
C. *"unobserved":* the pairwise comparisons that have not been investigated in any of the identified studies in the systematic review.

These groups constitute the rows of the Pairwise Comparisons Table for a specific outcome. Instructions for filling in the table are summarised in Additional file 2.

For each comparison, the first two columns report the total number of studies with results for the current outcome or any outcome, respectively. In brackets, we enter the total sample size by adding up all participants randomised in the studies investigating the specific comparison for that outcome. By definition, the unobserved comparisons will have zero in both columns. In contrast, those observed for other outcomes will have zero in the first column.

The groups of comparisons are presented in Table 1 for the example of non-invasive diagnostic modalities for the detection of coronary artery disease. Of the possible 15 comparisons, 8 were observed for the outcome of interest. The remaining 7 were unobserved, i.e. not observed for the outcome of interest or any other outcomes.

Within-study assessment of bias due to missing evidence. The evaluation of bias due to selective non-reporting of results within studies concerns studies identified for the review but missing from the synthesis. They are known to exist, but the results are unavailable: the studies report on other outcomes than the outcome of interest. The presence of selective non-reporting of results in each study is assessed using study-specific tools such as Step 2 of the ROB-ME tool [10,20]. Then, the likely impact of the missing results across all studies may be assessed using two signalling questions to reach an overall judgement of *no bias detected* or *suspected bias favouring X* for each comparison (Table 3). The preliminary version of the ROB-ME tool describes various approaches to evaluate the within-study assessment of bias by considering the plausibility of scenarios where study results are or are *not* unavailable because of the p-value, magnitude or direction of the treatment effects [1,10].

A thorough within-study assessment of bias due to missing evidence is labour intensive but particularly valuable as the impact of selective non-reporting or under-reporting of results can be quantified more easily than the impact of selective non-publication of an unknown number of studies [1]. However, suppose the number of studies (or the sample size) not reporting the outcome of interest (i.e. the difference between the first two columns in Table 1) is small compared to the number of studies (or the total sample size) reporting the outcome (the first column in Table 1). In that case, the assessment of these few studies is unlikely to affect the judgment from the within-study assessment significantly. Reviewers may then decide to assign *no bias detected* to the relevant comparison without carrying out the assessment. *No bias detected* is also assigned when no study is suspected of selective non-reporting or under-reporting of results for a specific comparison (i.e. the numbers in the first two columns are equal). For the unobserved comparisons, the assessment is not applicable ("NA", Table 1).

In the example of the non-invasive diagnosis of coronary artery disease, there were no additional studies that did not report results for the outcome of interest. Therefore, we assume that there is no selective outcome reporting bias, and we assign *no bias detected* for the within-study assessment of bias to all observed comparisons. In the 'Application of ROB-MEN to a network of antidepressants' the within-study assessment of bias is completed using the signalling questions for additional studies not reporting the outcome of interest.

Across-study assessment of bias due to missing evidence. This situation refers to studies undertaken but not published, so reviewers are unaware of them. Each comparison is assessed for risk of publication bias using qualitative and quantitative considerations. First, a qualitative judgement is made to assign a level of *no bias detected* or *suspected bias*. Conditions that may indicate bias include:

- Failure to search for unpublished studies and grey literature.
- The meta-analysis may be based on a few positive findings on a newly introduced drug as the early evidence likely overestimates efficacy [25].
- Previous evidence may have shown the presence of publication bias for that comparison [26].

Conditions suggesting no bias include data from unpublished studies and agreement of their findings with those of published studies or a tradition of prospective trial registration in the field.

For comparisons with at least 10 studies (in the first column in Table 1), judgements can additionally consider statistical techniques such as contour-enhanced funnel plots [4], meta-regression models and statistical tests for small-study effects [2,6,7,27–29] or selection models for pairwise meta-analysis (e.g. Copas [8]). These can be useful when it is difficult to assess publication bias reliably, e.g. when protocols and records from trial registries were unavailable. The direction of any bias should be noted: it will generally reflect the larger benefits observed in smaller studies.

We implemented the across-study assessment of bias in the network meta-analysis of non-invasive diagnostic tests of coronary artery disease using qualitative considerations (see Additional file 4). None of the comparisons included 10 or more studies and no assessment using graphical or statistical methods was therefore performed. The judgements for all comparisons are reported in Table 1.

Overall risk of bias for pairwise comparisons. The last step in the Pairwise Comparisons Table is to combine the levels of risk assigned in the previous steps into a final judgement of *no bias detected* or *suspected bias*. In case of *suspected bias*, the predicted direction of the bias, i.e. which treatment the bias is likely to favour, should also be specified (see Figure 1). For the unobserved comparisons (group C), the overall risk of bias will be the same as the judgement made for the across-study assessment of bias, as this is the only assessment applicable to these comparisons.

For the comparisons observed for the outcome of interest or other outcomes (group A and B), the overall judgement will consider qualitative assessments for both the within-study and the across-study assessment of bias. The assessment of selective outcome reporting bias ("within-study assessment of bias") is likely to be the most valuable because its impact can be quantified more easily than that of publication bias ("across-study assessment of bias"). The process of forming a final judgement for each pairwise comparison is illustrated in the flowchart in Additional file 5.

Since there was no within-study assessment of bias for the example of non-invasive diagnosis of coronary artery disease, the overall bias judgement will only consider the across-study assessment of bias. The final overall risk of bias judgements is reported in the Pairwise Comparison Table (Table 1).

Risk of bias due to missing evidence in network meta-analysis estimates

Once the assessments of overall bias for each pairwise comparison are complete, we integrate them in the assessment of risk of bias for each network estimate in the ROB-MEN Table. We organise the estimates into two groups, "mixed/only direct" and "only indirect", depending on the type of evidence contributing to each estimate (see Box 1). Here, we describe the detailed steps for filling in the relevant column in the ROB-MEN Table and illustrate them using the network of trials of non-invasive coronary artery disease diagnosis. Instructions are summarised in Additional file 3.

Contribution of comparisons with suspected bias to network meta-analysis estimates. The first step is to consider the contribution matrix of the network. The cells of this matrix provide the percentage contribution that each comparison with direct evidence (columns of the matrix) makes to the calculation of the corresponding network meta-analysis relative treatment effect (rows of the matrix) [23]. Additional file 6 shows the contribution matrix for the network of non-invasive diagnosis of coronary artery disease. Each comparison with direct evidence is combined with the risk of bias as judged in the Pairwise Comparisons Table (Table 1). This way, the percentage contribution from direct evidence with suspected bias (reported in the first and second column of the ROB-MEN Table, see Table 2 for example) can be estimated. The evaluation of the contribution from comparisons with suspected bias is reported in the third column. Specifically, the possible levels are:

- *No substantial contribution from bias*: there is no substantial contribution from evidence with bias favouring one of the two treatments;
- *Substantial contribution from bias balanced*: there is a substantial contribution from evidence with suspected bias, but the biases favouring one or the other treatment are balanced and cancel each other out;
- *Substantial contribution from bias favouring X*: there is a substantial contribution from evidence with bias favouring one of the two treatments (say X).

In the non-invasive diagnosis of coronary artery disease network meta-analysis, we considered the contribution from biased evidence as substantially in favour of one treatment if the relative difference between treatments was at least 15%. Among the mixed estimates, five of them have a clear separation of high contribution coming from biased evidence between the two treatments (e.g. CCTA vs SPECT-MPI). Among the indirect estimates, only

three estimates showed such clear separation (e.g. CMR vs SPECT-MPI). The relevant bias judgements for this step are in column 3 of the ROB-MEN Table (Table 2).

Additional risk of bias for indirect estimates. Indirect relative effects are calculated from sources of direct evidence in the Pairwise Comparisons Table with contributions as shown in the contribution matrix. The absence of direct evidence for these indirect comparisons may lead to bias if any studies are missing for reasons associated with their results. Therefore, for the indirect estimates, we need to account for this potential source of bias, which is represented by the final judgement of the overall bias for pairwise comparisons *observed for other outcomes* or completely *unobserved* in the Pairwise Comparisons Table. We copy the final judgements from column 5 of the Pairwise Comparisons Table (see Table 1 for example) into column 4 of the ROB-MEN Table (see Table 2) of our illustrative example, and we consider only those of the indirect estimates. Three estimates were at suspected bias favouring CCTA, CMR and SPECT-MPI.

Small-study effects in network meta-analysis. To evaluate small-study effects, we run a network meta-regression model with a measure of precision (e.g. variance or standard error) as the covariate. This model generates an adjusted relative effect by extrapolating the regression line to the smallest observed variance (the 'largest' study) independently for each comparison. To assess the presence of small-study effects, we compare the obtained adjusted estimates with the original (unadjusted) estimates by looking at the overlap of their corresponding confidence (or credible) intervals. A lack of overlap between the two intervals (or between one estimate and the interval for the other estimate) is an indication that effect estimates differ between smaller and larger studies. Note that this approach assumes there is no other explanation for the difference between the original, and the adjusted estimates, i.e. other covariates do not explain it. The evaluation of small-study effects is reported in the penultimate column of the ROB-MEN Table (Table 2), with levels indicating whether there is evidence of small-study effects and, if so, which treatment is favoured by the small studies.

For the example of non-invasive diagnostic modalities, we ran a network meta-regression model using the variance of the estimate (pooled variance for multi-arm studies) as a covariate to investigate small-study effects in the whole network. The adjusted estimates via extrapolation to the smallest observed variance are reported in column 6 of the ROB-MEN Table next to the original network meta-analysis summary effect (column 5 in Table 2). None of the network meta-regression estimates are markedly different from their unadjusted

counterparts, and the credible intervals for estimates overlap. Therefore, "No evidence of small-study effects" is reported in column 7 for all the estimates.

Overall risk of bias for network meta-analysis estimates. We propose rules for assigning a final judgement on the overall risk of bias due to missing evidence for estimates which are described in Table 4. If there is a substantial contribution from evidence with suspected bias (column 3), we have concerns regarding the risk of bias for that estimate. Suppose this contribution is split between evidence with bias favouring one of the treatments and evidence with bias favouring the other treatment. In that case, the biases may cancel out, assuming the bias is about the same in the two directions. Concerns about the risk of bias are then defined by the overall bias of unobserved comparisons in column 4 (for indirect estimates) and the evidence about small-study effects (column 7). The final judgements for the overall risk of bias are reported in column 8 (see Table 2). The reviewer can decide to follow our proposed rules to assign the overall risk of bias level but, if “stricter” or “more relaxed” approaches are preferred, they can also reach their final judgement based on their own reasoning. Whatever their reasoning, every choice and assessment should be justified and clearly described.

Given that most of the mixed estimates have substantial contributions from biased evidence favouring one of the two treatments. Still, there was no evidence of small-study effects for any of the estimates, we have some concerns about the risk of bias due to missing evidence. The exceptions are exercise ECG vs standard care, Exercise ECG vs stress echo and SPECT-MPI vs standard care. There, the level was decreased to "Low risk" due to lack of substantial contribution from biased evidence favouring either one of the two treatments. Similarly, we assign "Some concerns" to indirect estimates, where a substantial contribution from biased evidence was favouring one of the two treatments. For CMR vs stress echo, the level was increased to "High risk" because of the additional bias from the corresponding indirect comparison assessed in the Pairwise Comparisons Table (Table 1), despite the fact that there is no evidence of small-study effects. The other indirect estimates were assigned a level of "Low risk" of bias because i) there was no substantial contribution from biased evidence or it canceled each other out; ii) there was no additional bias from the indirect comparison assessed in the Pairwise Comparisons Table (Table 1); iii) there was no evidence of small-study effects. No estimate was judged to be at high risk of bias due to missing evidence. The final judgements on the overall risk of bias due to missing evidence are reported in column 8 of Table 2.

Results

Application of ROB-MEN to a network of antidepressants

We applied the ROB-MEN tool to a network of head-to-head studies (i.e. trials of active interventions) of 18 antidepressants [17]. The outcome of interest is the response to treatment defined as a reduction of at least 50% in the score between baseline and week 8 on a standardised rating scale for depression [30].

Pairwise Comparisons Table

There are 153 possible comparisons between the 18 drugs. Seventy compared the response to the antidepressant (group A) and 2 (amitriptyline vs bupropion and amitriptyline vs nefazodone, group B) compared other outcomes (dropouts and remission). The remaining 82 possible comparisons were not covered in any of the studies ("unobserved", group C) (see Additional file 7).

We *carried* out the within-study assessment of bias due to missing evidence for the two comparisons in the "observed for other outcomes" group (*no bias detected*) and for the comparisons in the group "observed for this outcome" for which extra studies were identified that did not report the outcome of interest. We judged four of these to be potentially biased because the extra studies did not report the full results and were sponsored by the company manufacturing the drug favoured by the bias. We judged the other four comparisons as *no bias detected*: the unavailable results were unlikely to be missing due to non-significant p-values or the directions of the results and unlikely to affect the overall results. For example, selective outcome reporting bias was suspected for an additional study of fluoxetine versus paroxetine but unlikely to affect the synthesised results given its small sample size (21 participants) relative to the total sample size (1364 participants). We assigned all other comparisons observed for this outcome a level of *no bias detected* in this step. The withn-study assessment of bias was not applicable to the 82 unobserved comparisons.

The across-study assessment of bias was carried out for all comparisons. We considered that bias, when suspected, would favour the newest drug, following the novel agent bias principle. The exceptions were comparisons where agomelatine, paroxetine, bupropion and vortioxetine were the newest drug because the authors obtained all unpublished data from the manufacturers. This qualitative consideration took priority over findings from contour-enhanced funnel plots and tests for small-study effects for comparisons with at least 10

studies. Based on the findings from these statistical techniques, neither amitriptyline versus fluoxetine nor citalopram versus escitalopram would be judged at suspected bias. We nevertheless agreed our judgement from the across-study assessment of bias for both comparisons as *suspected bias favouring the newest drug* because the review authors could not exclude the possibility of hidden studies with unfavourable results towards the newer drug in the comparison (fluoxetine and escitalopram).

Considering the previous assessments, most of the pairwise comparisons were considered at *suspected bias favouring the newest drug*. The only ones judged with *no bias detected* were all comparisons involving agomelatine and vortioxetine, as well as other 12 comparisons involving other drugs. The judgements for all pairwise comparisons are reported in the last column of the Pairwise Comparisons Table (Additional file 7).

ROB-MEN Table

Once the Pairwise Comparison Table is complete with all judgements, we integrate them in the ROB-MEN Table. First, the overall risk of bias judgements for comparisons with direct evidence are combined with the results from the contribution matrix to calculate for each network meta-analysis estimate the contribution coming from direct evidence at suspected bias favouring either of the two treatments, and in total. We considered an estimate to have substantial contribution from evidence at suspected bias favouring one of the two treatments in the contrast if the difference between the first and second column (contribution from evidence at suspected bias favouring first and favouring second treatment, respectively) was at least 15 percentage points.

The bias assessment for indirect evidence is only considered for the "only indirect" estimates and is copied from the last column of the Pairwise Comparison Table. This potential risk for "missing studies" is particularly important for the indirect estimates because it drives the bias evaluation to a "high risk" level in case there is also substantial contribution from direct evidence with suspected bias in the same direction.

The last part of the risk of bias assessment for the network estimate involves running a network meta-regression model to evaluate the presence (or absence) of small-study effects. We run the model using the smallest observed variance as a covariate and assuming unrelated coefficients. All estimates and their adjusted counterpart were similar, and their credible intervals had a good level of overlap, providing no evidence of small-study effects.

Following the rules set out in Table 3 we assign the final judgements on the overall risk of bias due to missing evidence to the estimates and report it in the last column of the ROB-MEN Table (Additional file 8). Overall, the risk of bias for most estimates was classified as *some concerns* or *low risk*. In particular, none of the comparisons involving agomelatine, paroxetine, venlafaxine or vortioxetine were at high risk of bias. All 153 network meta-analysis estimates with their relative ROB-MEN levels are reported in Table 5.

Discussion

To our knowledge, ROB-MEN is the first tool for assessing the risk of bias due to missing evidence in network meta-analysis. ROB-MEN builds on an approach recently proposed for pairwise meta-analysis [1,10] and adapts it to the network setting . Specifically, the assessments for selective outcome reporting and publication bias in pairwise comparisons are combined with (i) the percentage contribution of direct evidence for each pairwise comparison to the network meta-analysis estimates, (ii) evidence about the presence of small-study effects and (iii) any bias arising from unobserved comparisons.

Our examples demonstrate that the tool applies to different network meta-analyses, including very large and complex networks, for which assessing the risk of bias can be lengthy and labour-intensive. We developed an R Shiny web application [24] to facilitate the ROB-MEN use. Once the user has evaluated the risk of bias for all pairwise comparisons and estimates, the app produces the Pairwise Comparisons and ROB-MEN Table. The ROB-MEN tool is also incorporated in the reporting bias domain of the CINeMA framework and software [18,19].

ROB-MEN is not applicable in situations where an intervention of interest is disconnected from the network. It was not designed to cover comparisons involving disconnected interventions. In case of disconnected networks, we recommend to evaluate each subnetwork separately. Like for any other evaluation of results' credibility in evidence synthesis, many of the judgements in the ROB-MEN process involve subjective decisions. Judging bias due to missing evidence is challenging, particularly for publication bias, as reviewers will often not know about unpublished studies. However, the subjectivity of our approach, specifically in the pairwise comparisons step, is shared by other approaches, as described in the Cochrane Handbook and ROB-ME tool [1,10]. Also, the novel quantitative

methods, the contribution matrix [23] and network meta-regression that we integrated into the assessment rely less on the reviewer's subjectivity.

Conclusions

We encourage the evidence-synthesis community to conduct studies of the reliability and reproducibility of the ROB-MEN tool. We recommend reviewers specify the criteria used and explain the reasoning behind the judgements to enhance transparency. We believe that ROB-MEN will help those performing network meta-analyses reach better-informed conclusions and enhance the toolbox of available methods for evaluating the credibility of network meta-analysis results.

Declarations

Ethics approval and consent to participate. Not applicable.

Consent for publication. Not applicable.

Availability of data and materials. Data sharing not applicable as no new datasets were generated for this study.

Funding. The development of the ROB-MEN web application and part of the presented work was supported by the Cochrane Collaboration. GS, VC, TP, AN and ME are supported by project funding (Grant No. 179158, 189498) from the Swiss National Science Foundation (SNSF). AN is supported by an SNSF personal fellowship (P400PM_186723). JPTH is a National Institute for Health Research (NIHR) Senior Investigator (NF-SI-0617-10145) and is supported by the National Institute for Health Research (NIHR) Bristol Biomedical Research Centre at University Hospitals Bristol and Weston NHS Foundation Trust and the University of Bristol, NIHR Applied Research Collaboration West (ARC West) at University Hospitals Bristol and Weston NHS Foundation Trust and NIHR Health Protection Research Unit in Evaluation of Interventions at the University of Bristol in partnership with Public Health England. MJP is supported by an Australian Research Council Discovery Early Career Researcher Award (DE200101618). AC is supported by the National Institute for Health Research (NIHR) Oxford Cognitive Health Clinical Research Facility, by an NIHR Research Professorship (grant RP-2017-08-ST2-006), by the NIHR Oxford and Thames Valley Applied Research Collaboration and by the NIHR Oxford Health Biomedical Research Centre (grant BRC-1215-20005). The views expressed in this article are those of the authors and do not necessarily represent those of the SNSF, NHS, the NIHR, MRC, or the Department of Health and Social Care.

Competing interests. All authors have completed the ICMJE uniform disclosure form at www.icmje.org/coi_disclosure.pdf and declare: AC has received research and consultancy fees from INCiPiT (Italian Network for Paediatric Trials), CARIPLO Foundation and Angelini Pharma; TAF reports personal fees from MSD, grants and personal fees from Mitsubishi-Tanabe, grants and personal fees from Shionogi, outside the submitted work; TAF has a patent 2018-177688 pending, and a patent Kokoro-app issued; no other relationships or activities that could appear to have influenced the submitted work.

Authors' contributions. VC and GS conceived and oversaw the project. VC, AN, JPTH, MJP, TP, ME and GS contributed to development of ROB-MEN. GCMS, AC and TAF contributed to the application of ROB-MEN on the network examples. VC wrote the first draft of the manuscript. All authors reviewed and commented on drafts of the manuscript. VC and GS will act as guarantors. The corresponding author attests that all listed authors meet authorship criteria and that no others meeting the criteria have been omitted. All authors read and approved the final manuscript.

Acknowledgements. We thank Tianjing Li for her contribution and suggestions to the development of the ROB-MEN methodology and manuscript.

Tables and Figures

Table 1: Pairwise Comparisons Table for the network of non-invasive diagnostic modalities for detecting coronary artery disease.

Column No.	1	2	3	4	5
Pairwise comparisons	**No. of studies in each comparison**		**Within-study assessment of bias**	**Across-study assessment of bias**	**Overall bias**
	Reporting this outcome (sample size)	Total identified in the SR (total sample size)	Evaluation of selective reporting within studies using signalling questions	Qualitative and quantitative assessment of publication bias	Overall judgement
Group A: observed for this outcome					
CCTA vs Exercise ECG	1 (562)	1 (562)	No bias detected	No bias detected	No bias detected
CCTA vs SPECT-MPI	2 (1149)	2 (1149)	No bias detected	Suspected bias favouring CCTA	Suspected bias favouring CCTA
CCTA vs Standard care	7 (4015)	7 (4015)	No bias detected	Suspected bias favouring CCTA	Suspected bias favouring CCTA
CMR vs Standard care	2 (214)	2 (214)	No bias detected	Suspected bias favouring CMR	Suspected bias favouring CMR
Exercise ECG vs Standard care	1 (130)	1 (130)	No bias detected	No bias detected	No bias detected
Exercise ECG vs Stress Echo	4 (1086)	4 (1086)	No bias detected	No bias detected	No bias detected
SPECT-MPI vs Standard care	2 (4165)	2 (4165)	No bias detected	No bias detected	No bias detected
Standard care vs Stress Echo	1 (132)	1 (132)	No bias detected	Suspected bias favouring Stress Echo	Suspected bias favouring Stress Echo
Group B: observed for other outcomes (no studies)					
Group C: Unobserved					
CCTA vs CMR	0	0	NA	No bias detected	No bias detected
CCTA vs Stress Echo	0	0	NA	Suspected bias favouring CCTA	Suspected bias favouring CCTA
CMR vs Exercise ECG	0	0	NA	No bias detected	No bias detected
CMR vs SPECT-MPI	0	0	NA	No bias detected	No bias detected
CMR vs Stress Echo	0	0	NA	Suspected bias favouring CMR	Suspected bias favouring CMR
Exercise ECG vs SPECT-MPI	0	0	NA	Suspected bias favouring SPECT-MPI	Suspected bias favouring SPECT-MPI
SPECT-MPI vs Stress Echo	0	0	NA	No bias detected	No bias detected

CCTA: coronary computed tomographic angiography; CMR: cardiovascular magnetic resonance; ECG: electrocardiogram; Echo: echocardiography; SPECT-MPI: single-photon emission computed tomography-myocardial perfusion imaging; SR: systematic review.

Table 2: ROB-MEN Table for the network of non-invasive diagnostic modalities for detection of coronary artery disease in patients with low risk acute coronary syndrome.

	1	2	3	4	5	6	7	8
NMA estimate	**% contribution of evidence from pairwise comparisons with suspected bias**		**Evaluation of contribution from evidence with suspected bias**	**Bias assessment for indirect evidence**	**NMA treatment effect**	**NMR treatment effect at the smallest observed variance**	**Evaluation of small-study effects**	**Overall risk of bias**
	Favouring first treatment	**Favouring second treatment**						
Mixed/ only direct								
CCTA vs Exercise ECG	20.2%	0%	Substantial contribution from bias favouring CCTA		1.97 (1.06, 3.79)	1.74 (0.82, 3.66)	No evidence of small-study effects	Some concerns
CCTA vs SPECT-MPI	66.0%	0%	Substantial contribution from bias favouring CCTA		1.29 (0.93, 1.78)	1.30 (0.88, 2.04)	No evidence of small-study effects	Some concerns
CCTA vs Standard care	89.2%	0%	Substantial contribution from bias favouring CCTA		1.17 (0.93, 1.50)	1.18 (0.89, 1.58)	No evidence of small-study effects	Some concerns
CMR vs Standard care	100%	0%	Substantial contribution from bias favouring CMR		0.37 (0.17, 0.81)	0.35 (0.08, 1.37)	No evidence of small-study effects	Some concerns
Exercise ECG vs Standard care	0%	0%	No substantial contribution from bias		0.59 (0.31, 1.12)	0.68 (0.33, 1.39)	No evidence of small-study effects	Low risk
Exercise ECG vs Stress Echo	0%	2.2%	No substantial contribution from bias		1.89 (1.25, 2.81)	2.03 (1.23, 3.35)	No evidence of small-study effects	Low risk
SPECT-MPI vs Standard care	0%	0%	No substantial contribution from bias		0.91 (0.68, 1.24)	0.91 (0.62, 1.25)	No evidence of small-study effects	Low risk
Standard care vs Stress Echo	0%	55.2%	Substantial contribution from bias favouring Stress Echo		3.15 (1.49, 6.37)	2.99 (1.41, 6.50)	No evidence of small-study effects	Some concerns

Only indirect								
CCTA vs CMR	45.9%	46.8%	Substantial contribution from bias balanced	No bias detected	3.15 (1.40, 7.20)	3.40 (0.81, 15.70)	No evidence of small-study effects	Low risk
CCTA vs Stress Echo	20.9%	12.8%	Substantial contribution from bias balanced	Suspected bias favouring CCTA	3.71 (1.83, 7.92)	3.53 (1.61, 7.72)	No evidence of small-study effects	Low risk
CMR vs Exercise ECG	37.7%	0%	Substantial contribution from bias favouring CMR	No bias detected	0.62 (0.22, 1.77)	0.51 (0.10, 2.47)	No evidence of small-study effects	Some concerns
CMR vs SPECT-MPI	47.0%	0%	Substantial contribution from comparisons with suspected bias favouring CMR	No bias detected	0.41 (0.18, 0.93)	0.39 (0.08, 1.60)	No evidence of small-study effects	Some concerns
CMR vs Stress Echo	33.6%	13.4%	Substantial contribution from bias favouring CMR	Suspected bias favouring CMR	1.17 (0.40, 3.51)	1.04 (0.19, 5.17)	No evidence of small-study effects	High risk
Exercise ECG vs SPECT-MPI	0%	0%	No substantial contribution from bias	Suspected bias favouring SPECT-MPI	0.65 (0.32, 1.28)	0.75 (0.35, 1.67)	No evidence of small-study effects	Low risk
SPECT-MPI vs Stress Echo	0%	13.3%	No substantial contribution from bias	No bias detected	2.87 (1.37, 6.45)	2.68 (1.18, 6.16)	No evidence of small-study effects	Low risk

CCTA: coronary computed tomographic angiography; CMR: cardiovascular magnetic resonance; ECG: electrocardiogram; Echo: echocardiography; NMA: network meta-analysis; NMR: network meta-regression; SPECT-MPI: single photon emission computed tomography-myocardial perfusion imaging. Effects in column 5 and 6 are odds ratios and 95% credible intervals.

Table 3: Signalling questions for the within-study bias assessment of comparisons observed for the outcome of interest or other outcomes.

Signalling question	**Responses for each comparison (group A and B only)**		
1. Was there any eligible study for which results for the outcome of interest were unavailable, likely because of the P-value, magnitude or direction of the result generated?	Yes	Yes	No
2. (If Yes to the previous question) Was the amount of information omitted from the synthesis sufficient to have a notable effect on the magnitude of the synthesised result?	Yes	No	-
Overall judgment	*Suspected bias (favouring X)*	*No bias detected*	*No bias detected*

Table 4: Proposed rules for judging the overall risk of bias due to missing evidence for network meta-analysis estimates

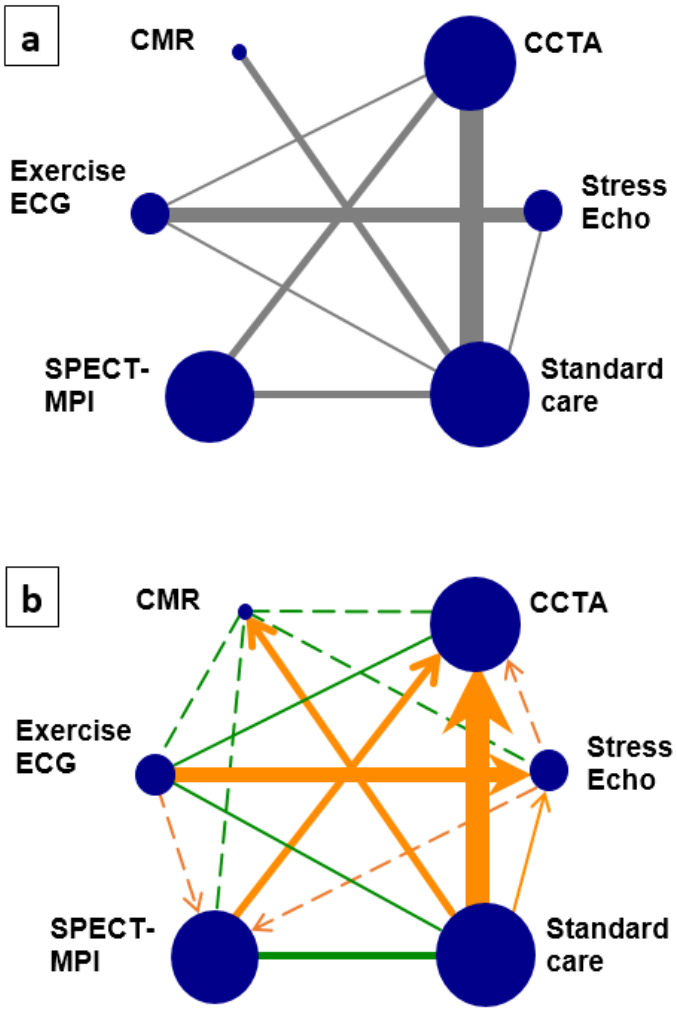
Low risk	There is no substantial contribution from evidence with suspected bias favouring one of the two treatments, OR There is substantial contribution from evidence at suspected bias but it is split more or less equally between evidence with bias favouring one of the treatments and evidence with bias favouring the other treatment **AND** There is no evidence of small-study effects favouring one of the two treatments OR [For indirect estimates only] There is no suspected bias favouring one of the two treatments from the assessment of indirect evidence.
Some concerns	All other combinations
High risk	There is substantial contribution from evidence with suspected bias favouring one of the two treatments, say X **AND** There is evidence of small-study effects favouring the same treatment X OR [For indirect estimates only] There is suspected bias favouring that treatment X from the assessment of indirect evidence.

Table 5: League table of the network estimates and corresponding risk of bias due to missing evidence for the network of 18 antidepressants.

Ago																	
0.96 (0.75, 1.22)	Ami		Cit		Dul	Esc				Mir	Nef		Reb				
0.87 (0.58, 1.30)	0.91 (0.62, 1.33)	Bup	Cit		Dul	Esc			Mil	Mir	Nef		Reb				
1.13 (0.87, 1.47)	1.18 (0.93, 1.48)	1.29 (0.88, 1.93)	Cit		Dul				Mil		Nef						
1.20 (0.92, 1.57)	1.25 (0.99, 1.59)	1.38 (0.92, 2.07)	1.06 (0.83, 1.38)	Clo	Dul	Esc				Mir	Nef		Reb				
1.05 (0.81, 1.37)	1.10 (0.85, 1.42)	1.21 (0.81, 1.82)	0.93 (0.71, 1.23)	0.88 (0.66, 1.16)	Dul		Dul	Dul						Dul	Dul		
0.90 (0.71, 1.14)	0.94 (0.74, 1.18)	1.03 (0.70, 1.53)	0.80 (0.65, 0.97)	0.75 (0.58, 0.96)	0.85 (0.67, 1.08)	Esc		Esc							Esc		
1.20 (0.98, 1.47)	1.25 (1.06, 1.47)	1.37 (0.96, 1.96)	1.06 (0.87, 1.29)	1.00 (0.82, 1.22)	1.14 (0.91, 1.44)	1.33 (1.11, 1.60)	Fluo										
1.20 (0.91, 1.61)	1.26 (0.99, 1.60)	1.38 (0.92, 2.08)	1.07 (0.82, 1.39)	1.01 (0.76, 1.32)	1.14 (0.85, 1.55)	1.34 (1.02, 1.75)	1.01 (0.81, 1.26)	Fluvo			Nef		Reb				
1.07 (0.80, 1.45)	1.12 (0.87, 1.44)	1.23 (0.81, 1.88)	0.95 (0.72, 1.26)	0.90 (0.67, 1.19)	1.02 (0.75, 1.40)	1.20 (0.91, 1.58)	0.90 (0.70, 1.13)	0.89 (0.66, 1.18)	Mil								
0.93 (0.72, 1.21)	0.98 (0.78, 1.21)	1.07 (0.72, 1.59)	0.83 (0.65, 1.06)	0.78 (0.60, 1.01)	0.89 (0.67, 1.17)	1.04 (0.81, 1.33)	0.78 (0.64, 0.94)	0.77 (0.60, 1.00)	0.87 (0.66, 1.15)	Mir							
1.15 (0.76, 1.74)	1.20 (0.81, 1.78)	1.32 (0.80, 2.23)	1.02 (0.67, 1.55)	0.96 (0.63, 1.47)	1.09 (0.71, 1.68)	1.28 (0.84, 1.92)	0.96 (0.66, 1.40)	0.95 (0.62, 1.46)	1.07 (0.70, 1.63)	1.23 (0.81, 1.85)	Nef				Nef		
1.01 (0.82, 1.24)	1.05 (0.90, 1.23)	1.15 (0.80, 1.67)	0.89 (0.73, 1.10)	0.84 (0.68, 1.02)	0.96 (0.76, 1.20)	1.12 (0.93, 1.34)	0.84 (0.74, 0.96)	0.84 (0.67, 1.04)	0.94 (0.75, 1.18)	1.08 (0.90, 1.30)	0.88 (0.60, 1.28)	Par					
1.44 (1.02, 2.05)	1.50 (1.07, 2.09)	1.65 (1.03, 2.64)	1.28 (0.92, 1.75)	1.20 (0.84, 1.72)	1.37 (0.94, 1.97)	1.61 (1.15, 2.22)	1.20 (0.88, 1.63)	1.20 (0.83, 1.70)	1.34 (0.93, 1.93)	1.54 (1.09, 2.17)	1.25 (0.79, 2.00)	1.43 (1.05, 1.95)	Reb	Reb	Reb		
1.07 (0.85, 1.37)	1.12 (0.93, 1.35)	1.23 (0.84, 1.80)	0.95 (0.77, 1.19)	0.89 (0.71, 1.12)	1.02 (0.79, 1.31)	1.20 (0.96, 1.47)	0.90 (0.76, 1.06)	0.89 (0.70, 1.14)	1.00 (0.77, 1.30)	1.15 (0.93, 1.43)	0.93 (0.63, 1.38)	1.06 (0.91, 1.26)	0.75 (0.54, 1.04)	Ser			
1.35 (0.98, 1.86)	1.41 (1.07, 1.85)	1.54 (1.03, 2.31)	1.19 (0.88, 1.63)	1.13 (0.81, 1.52)	1.28 (0.92, 1.78)	1.51 (1.10, 2.03)	1.13 (0.86, 1.46)	1.12 (0.81, 1.54)	1.26 (0.90, 1.74)	1.45 (1.08, 1.93)	1.17 (0.75, 1.85)	1.34 (1.03, 1.73)	0.94 (0.63, 1.39)	1.26 (0.95, 1.65)	Tra		
1.01 (0.81, 1.25)	1.06 (0.87, 1.27)	1.16 (0.81, 1.66)	0.90 (0.73, 1.11)	0.84 (0.67, 1.06)	0.96 (0.77, 1.21)	1.13 (0.92, 1.37)	0.85 (0.73, 0.97)	0.84 (0.66, 1.07)	0.94 (0.73, 1.21)	1.08 (0.88, 1.32)	0.88 (0.59, 1.31)	1.00 (0.86, 1.17)	0.70 (0.52, 0.96)	0.94 (0.79, 1.13)	0.75 (0.57, 0.98)	Ven	
0.72 (0.42, 1.25)	0.76 (0.44, 1.28)	0.83 (0.45, 1.53)	0.64 (0.37, 1.12)	0.60 (0.34, 1.04)	0.69 (0.39, 1.20)	0.81 (0.46, 1.39)	0.60 (0.35, 1.01)	0.60 (0.34, 1.05)	0.67 (0.38, 1.17)	0.77 (0.45, 1.33)	0.63 (0.33, 1.20)	0.72 (0.42, 1.21)	0.50 (0.27, 0.92)	0.67 (0.39, 1.15)	0.53 (0.30, 0.95)	0.72 (0.43, 1.19)	Vor

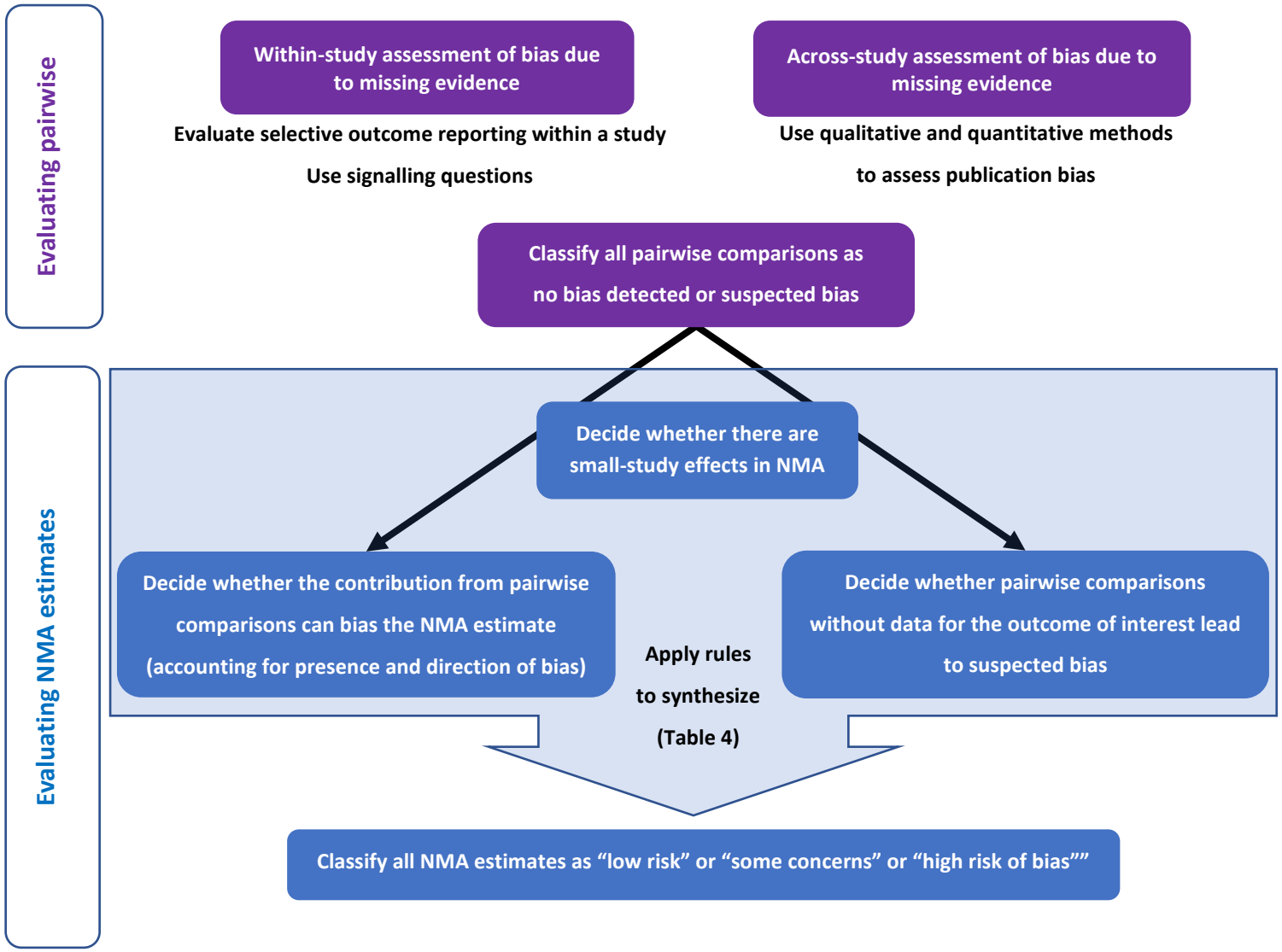
The values in the lower triangle represent the relative treatment effect (odds ratios and 95% credible intervals) of the treatment on the top (column) versus the treatment on the row. Colours indicate the ROB-MEN levels: green = Low risk; yellow: Some concerns; red = High risk. Names in the upper triangle indicate the treatment favoured by the bias in the high risk estimates (red cells). Risk of bias assessments were obtained using the Shiny app. Ago = agomelatine; Ami = amitriptyline; Bup = bupropion; Cit = citalopram; Clo = clomipramine; Dul = duloxetine; Esc = escitalopram; Fluo = fluoxetine; Fluvo = fluvoxamine; Mil = milnacipran; Mir = mrtazapine; Nef = nefazodone; Par = paroxetine; Reb = reboxetine; Ser = sertraline; Tra = trazodone; Ven = venlafaxine; Vor = vortioxetine

Figure 1: Network plots of network meta-analysis of non-invasive diagnostic modalities for detecting coronary artery disease. (a) Standard network plot. (b) Network graph showing risk of bias assessment for pairwise comparisons. Sizes of solid lines and nodes are proportional to number of studies in each comparison and total sample size for each treatment, respectively. Solid lines represent the observed direct comparisons, dotted lines represent unobserved comparisons between interventions. Green indicates *no bias detected*, orange indicates *suspected bias favouring the treatment indicated by the arrow*.



ECG: electrocardiogram; CCTA: coronary computed tomographic angiography; CMR: cardiovascular magnetic resonance; SPECT-MPI: single-photon emission computed tomography-myocardial perfusion imaging; Stress Echo: stress echocardiography.

Figure 2: Overview of the ROB-MEN process.



List of abbreviations

ROB-ME	Risk Of Bias due to Missing Evidence
ROB-MEN	Risk Of Bias due to Missing Evidence in Network meta-analysis
CINeMA	Confidence In Network Meta-Analysis
ECG	Electrocardiogram
SPECT-MPI	Single-Photon Emission Computed Tomography-Myocardial Perfusion Imaging
CCTA	Coronary Computed Tomographic Angiography
CMR	Cardiovascular Magnetic Resonance
Stress echo	Stress Echocardiography

Additional files

Available in the links below and online under Supplementary Information.

Additional file 1: Network graph, methods and forest plot for the network meta-analysis of non-invasive diagnostic modalities for the detection of coronary artery disease in patients with low risk acute coronary syndromes.

Additional file 2: Instructions for filling in the Pairwise Comparisons Table.

Additional file 3: Instructions for filling in the ROB-MEN Table.

Additional file 4: Description of the judgements from the across-study assessment of bias for the example of non-invasive diagnostic modalities for detection of coronary artery disease in patients with low risk acute coronary syndromes.

Additional file 5: Flow chart for assessing overall risk of bias due to missing evidence in pairwise comparisons.

Additional file 6: Contribution matrix for the network of non-invasive diagnostic modalities for coronary artery disease in patients with low risk acute coronary syndrome.

Additional file 7: Pairwise Comparisons Table for the network of 18 antidepressants.

Additional file 8: ROB-MEN Table for the network of 18 antidepressants.

References Article 5

1 Page MJ, Higgins JP, Sterne JA. Chapter 13: Assessing risk of bias due to missing results in a synthesis. In: *Cochrane Handbook for Systematic Reviews of Interventions*. Cochrane 2019. www.training.cochrane.org/handbook

2 Egger M, Smith GD, Schneider M, *et al.* Bias in meta-analysis detected by a simple, graphical test. *BMJ* 1997;**315**:629–34. doi:10.1136/bmj.315.7109.629

3 Sterne JAC, Egger M. Funnel plots for detecting bias in meta-analysis: Guidelines on choice of axis. *Journal of Clinical Epidemiology* 2001;**54**:1046–55. doi:10.1016/S0895-4356(01)00377-8

4 Peters JL, Sutton AJ, Jones DR, *et al.* Contour-enhanced meta-analysis funnel plots help distinguish publication bias from other causes of asymmetry. *Journal of Clinical Epidemiology* 2008;**61**:991–6. doi:10.1016/j.jclinepi.2007.11.010

5 Sterne JAC, Sutton AJ, Ioannidis JPA, *et al.* Recommendations for examining and interpreting funnel plot asymmetry in meta-analyses of randomised controlled trials. *BMJ* 2011;**343**:d4002–d4002. doi:10.1136/bmj.d4002

6 Harbord RM, Egger M, Sterne JAC. A modified test for small-study effects in meta-analyses of controlled trials with binary endpoints. *Statistics in Medicine* 2006;**25**:3443–57. doi:https://doi.org/10.1002/sim.2380

7 Peters JL. Comparison of Two Methods to Detect Publication Bias in Meta-analysis. *JAMA* 2006;**295**:676. doi:10.1001/jama.295.6.676

8 Copas JB, Shi JQ. A sensitivity analysis for publication bias in systematic reviews. *Stat Methods Med Res* 2001;**10**:251–65. doi:10.1177/096228020101000402

9 McShane BB, Böckenholt U, Hansen KT. Adjusting for Publication Bias in Meta-Analysis: An Evaluation of Selection Methods and Some Cautionary Notes. *Perspectives on Psychological Science* Published Online First: 29 September 2016. doi:10.1177/1745691616662243

10 Risk of bias tools - ROB-ME tool. https://riskofbias.info/welcome/rob-me-tool (accessed 13 Nov 2020).

11 Chaimani A, Salanti G. Using network meta-analysis to evaluate the existence of small-study effects in a network of interventions. *Res Syn Meth* 2012;**3**:161–76. doi:10.1002/jrsm.57

12 Mavridis D, Sutton A, Cipriani A, *et al.* A fully Bayesian application of the Copas selection model for publication bias extended to network meta-analysis. *Statist Med* 2013;**32**:51–66. doi:10.1002/sim.5494

13 Mavridis D, Welton NJ, Sutton A, *et al.* A selection model for accounting for publication bias in a full network meta-analysis. *Statistics in Medicine* 2014;**33**:5399–412. doi:10.1002/sim.6321

14 Chaimani A, Higgins JPT, Mavridis D, *et al.* Graphical Tools for Network Meta-Analysis in STATA. *PLoS ONE* 2013;**8**:e76654. doi:10.1371/journal.pone.0076654

15 Moreno SG, Sutton AJ, Ades AE, *et al.* Adjusting for publication biases across similar interventions performed well when compared with gold standard data. *Journal of Clinical Epidemiology* 2011;**64**:1230–41. doi:10.1016/j.jclinepi.2011.01.009

16 Siontis GC, Mavridis D, Greenwood JP, *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;:k504. doi:10.1136/bmj.k504

17 Cipriani A, Furukawa TA, Salanti G, *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. *The Lancet* 2018;**391**:1357–66. doi:10.1016/S0140-6736(17)32802-7

18 Nikolakopoulou A, Higgins JPT, Papakonstantinou T, *et al.* CINeMA: An approach for assessing confidence in the results of a network meta-analysis. *PLOS Medicine* 2020;**17**:e1003082. doi:10.1371/journal.pmed.1003082

19 *CINeMA: Confidence in Network Meta-Analysis*. Institute of Social and Preventive Medicine, University of Bern 2017. https://cinema.ispm.unibe.ch/ (accessed 11 Oct 2021).

20 Kirkham JJ, Altman DG, Chan A-W, *et al.* Outcome reporting bias in trials: a methodological approach for assessment and adjustment in systematic reviews. *BMJ* 2018;:k3802. doi:10.1136/bmj.k3802

21 Sterne JAC, Savović J, Page MJ, *et al.* RoB 2: a revised tool for assessing risk of bias in randomised trials. *BMJ* 2019;:l4898. doi:10.1136/bmj.l4898

22 Sterne JA, Hernán MA, Reeves BC, *et al.* ROBINS-I: a tool for assessing risk of bias in non-randomised studies of interventions. *BMJ* 2016;:i4919. doi:10.1136/bmj.i4919

23 Papakonstantinou T, Nikolakopoulou A, Rücker G, *et al.* Estimating the contribution of studies in network meta-analysis: paths, flows and streams. *F1000Res* 2018;**7**:610. doi:10.12688/f1000research.14770.3

24 *ROB-MEN: Risk Of Bias due to Missing Evidence in Network meta-analysis*. Institute of Social and Preventive Medicine, University of Bern 2021. https://cinema.ispm.unibe.ch/rob-men/ (accessed 11 Oct 2021).

25 Guyatt GH, Oxman AD, Montori V, *et al.* GRADE guidelines: 5. Rating the quality of evidence—publication bias. *Journal of Clinical Epidemiology* 2011;**64**:1277–82. doi:10.1016/j.jclinepi.2011.01.011

26 Turner EH, Matthews AM, Linardatos E, *et al.* Selective Publication of Antidepressant Trials and Its Influence on Apparent Efficacy. *N Engl J Med* 2008;**358**:252–60. doi:10.1056/NEJMsa065779

27 Moreno SG, Sutton AJ, Ades A, *et al.* Assessment of regression-based methods to adjust for publication bias through a comprehensive simulation study. *BMC Med Res Methodol* 2009;**9**:2. doi:10.1186/1471-2288-9-2

28 Moreno SG, Sutton AJ, Turner EH, *et al.* Novel methods to deal with publication biases: secondary analysis of antidepressant trials in the FDA trial registry database and related journal publications. *BMJ* 2009;**339**:b2981–b2981. doi:10.1136/bmj.b2981

29 Moreno SG, Sutton AJ, Thompson JR, *et al.* A generalized weighting regression-derived meta-analysis estimator robust to small-study effects and heterogeneity: A REGRESSION-DERIVED META-ANALYSIS MODEL ROBUST TO SMALL-STUDY EFFECTS. *Statist Med* 2012;**31**:1407–17. doi:10.1002/sim.4488

30 Furukawa TA, Salanti G, Atkinson LZ, *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. doi:10.1136/bmjopen-2015-010919

Overall discussion and outlook

Main findings

The work conducted in this thesis contributes to two main topics in the context of network meta-analysis: the evaluation and extension of existing techniques used to rank competing treatments, and the assessment of the risk of bias due to missing evidence.

In Article 1 we show how the level of agreement between treatment hierarchies obtained by different ranking metrics is affected by the amount of information present in a network. These differences in level of agreement are particularly evident when there are large imbalances in the precision of the estimates, though we find that such imbalances are rare in practice.

As part of this empirical evaluation, we employed rankings based on relative treatment effects against a fictional treatment of average performance, estimated through an alternative parameterisation of the network meta-analysis model, as described in Article 2. Such relative treatment effects and the corresponding probabilistic metric – interpreted as probability that a treatment is preferable than an average-performing treatment – are useful in networks of interventions where a natural reference (or control) treatment does not exist.

In Article 3 we provide guidance on the use and choice of ranking metrics, and how to properly interpret the resulting treatment hierarchies together with the relative treatment effects and quality of the evidence. We reiterated their advantages and limits in the decision-making process. We also emphasised that differences between rankings must not be interpreted as one ranking metric being preferable over another, as they address different treatment hierarchy questions which must be defined in advance.

We then expanded on the existing ranking methodology by combining a recently developed ranking metric, accounting for both multiple outcomes and individual preferences, with different trade-offs between benefits and harms. In Article 4 we show the usefulness of the obtained quantity as a sensitivity analysis to explore variation in the ranking for a whole range of trade-off values and a specific set of individual preferences.

In Article 5 we described the development of the first framework and tool for evaluating the risk of bias due to missing evidence in network meta-analysis. We also provide a user-friendly web application, part of the more comprehensive CINeMA framework, to make the tool more accessible and semi-automate some of the assessments.

Limitations and implications for future research

In Article 1, we acknowledge that we did not explore other potential factors that could influence the agreement between treatment hierarchies from different ranking metrics. These could be, for example, the effect measure chosen to synthesise the results, adjustment via network meta-regression, as well as non-methodological network characteristics e.g. clinical settings and/or fields. However, most of these factors are either unlikely to affect the agreement substantially, or are likely to be associated with the amount of information available in the network [1,2]. Furthermore, our empirical results are in line with results from simulation studies and theoretical examples [3,4].

In Article 2, the interpretation of the estimated coefficients (i.e. the relative effects of all treatments versus a fictional treatment of average performance) is straightforward only in situations where the notion of a *fictional treatment of average performance* is in some way meaningful. Such a fictional treatment does not (or may not) exist in practice, and it refers to the average absolute efficacy among the treatments included in the systematic review. Therefore, the interpretation of the obtained coefficients and of PreTAs depends on the set of compared treatments, though this is the case for all ranking metrics.

The quantity *SAWIS* that we propose in Article 4 is useful to show the variability of rankings for different individual preferences and trade-offs between benefits and harms. However, we are unsure that it can be presented as a ranking metric due to the complexity of its interpretation. Specifically, this quantity depends highly on the choice of outcome measures used to calculate the standardised area within a spie chart [5]. Daly *et al.* state that the interpretation of the standardised area within a spie chart is comparable to the interpretation of SUCRA as it represents the probability that a treatment ranks best overall. However, the formula involves a sum of probabilities – either SUCRA values or absolute probabilities – and it is unclear whether the resulting value would be a probability itself. Besides, the formula of our quantity adds further complexity as it is a difference between two areas within spie charts which also includes the trade-off value λ. Specifically, the challenge is to express the latter in the same unit as the standardised area within the spie chart, and it is difficult to give a value to *u* that can meaningfully be interpreted by patients as “how much in terms of side effects/harms they would be willing to experience for an increase in benefit”. Therefore, in its current formulation, our net-benefit standardised area within a spie chart cannot be used

to address a specific treatment hierarchy question, and further research is necessary to expand and improve upon interpretation of the trade-off value and the quantity as a whole.

The ROB-MEN tool presented in Article 5 was developed in the context of the CINeMA framework [6] and is, therefore, aimed at connected networks of aggregated data from randomised controlled trials. While the assessment of reporting bias may not be entirely relevant or fully applicable for non-randomised and observational studies, as described in the Risk of Bias In Non-randomized Studies – of Interventions (ROBINS-I) tool [7], the CINeMA and ROB-MEN web applications [8,9] cannot currently be used on network meta-analyses which make use of individual participant data (IPD) or component network meta-analyses (CNMA). Also, the ROB-MEN web application currently requires the user to load data in arm-based format for dichotomous and continuous outcomes and, like CINeMA, it does not support other types of outcomes such as time-to-event or count data. Future research could, therefore, look at updating the frameworks and relevant web applications to allow more flexibility in terms of types of data and analyses supported, such as dose-response NMAs or NMAs of diagnostic test accuracy.

Finally, as for all risk of bias assessments and tools, ROB-MEN also partly involves subjective judgements, but we tried to limit the reviewers' subjectivity by implementing quantitative elements and specific instructions. A certain degree of subjectivity is, however, inevitable in such evaluations. Consequently, the interrater agreement may also be affected but whether and how this subjectivity plays a role is currently unknown. Future work could focus on the reproducibility of assessments made by reviewers for ROB-MEN and, more generally, the CINeMA framework.

One of the main points of criticism regarding the use of ranking metrics is the fact that they do not incorporate an assessment for the quality of evidence [10]. Treatments at the top ranks of a hierarchy may be based on biased evidence, or heterogeneous or inconsistent data, and so they should not be recommended without examining the credibility of results first. A first attempt to evaluate the credibility of rankings was made by Salanti *et al.* [11] but, unlike the CINeMA framework for evaluating the confidence in the NMA effect estimates, this strategy for evaluating confidence in the ranking has not been developed further. Similarly, as a spin-off of CINeMA, the ROB-MEN tool focuses on the evaluation of bias due to missing evidence in the NMA estimates without incorporating or making judgements on the rankings. Future improvements of CINeMA, and consequently ROB-MEN, could also put emphasis on this

aspect and include extensions for a more specific tool on the credibility of ranking. However, it is unclear if and how a specific framework to evaluate confidence in a ranking is truly useful and necessary per se, given that treatment hierarchies are already interpreted together with the relative treatment effects and their confidence evaluations (as described in Article 3).

As previously described, new ranking methods and visualisation techniques for multiple outcomes like the *POST-R* and the Vitruvian plot [12,13] now integrate CINeMA confidence ratings. An implementation like the one in the Vitruvian plot (i.e. colouring the outcome sectors according to the confidence ratings) could potentially be applied also to spie charts [5] and, consequently, to our *SAWIS* approach. Another option that would involve integrating the confidence ratings within the ranking metric, is to down- or up-rank treatments in a fashion similar to how relative treatment effects are downgraded in the CINeMA or GRADE frameworks [6,14].

Recently, a new approach to evaluate the confidence in treatment recommendations based on NMA results has also been proposed and presented as an alternative to CINeMA and GRADE [15,16]. Threshold analysis is a sensitivity analysis that produces a threshold quantifying the amount by which the evidence could change before the recommendation changes (i.e. a different treatment or set of treatments is recommended). The robustness of the recommendation is judged by assessing whether it is plausible that the evidence could change (e.g. due to bias) by more than the determined threshold. Thus, this approach also involves some subjectivity like the other frameworks but, unlike them, does not properly tackle heterogeneity, inconsistency, or indirectedness.

The suggestions above for incorporating the quality of evidence may provide a better and more comprehensive picture of the whole output of an NMA, without having to look at confidence in the effects estimates and confidence in the rankings separately. However, developing ranking metrics to answer complex treatment hierarchy questions remains challenging, as the metrics aim to incorporate qualitative information (such as the confidence in the evidence) into a set of numerical quantities [4]. It is therefore not surprising that new developments in ranking metrics may remain difficult to use and interpret in practice.

Outlook and concluding remarks

Disagreement between treatment hierarchies obtained from different ranking metrics is not unusual, since each one addresses a distinct treatment hierarchy question and has, therefore, a specific interpretation. More complex hierarchy questions that are not linked to any of the available ranking metrics may be relevant in healthcare decision-making, hence triggering the need for extension of these methods [4]. However, instead of focusing on producing a specific ranking metric, a different and potentially more meaningful approach is to translate decision questions into hierarchy questions and calculate their uncertainty.

In this context, Papakonstantinou *et al.* [17] recently proposed a method that uses simulations to calculate the relative frequencies of each possible ranking obtainable in an NMA. They express the hierarchy question as a criterion, or set of criteria, and quantify the certainty around it by summing the relative frequencies of all hierarchies satisfying such criteria. The method is described for a single outcome and cannot handle benefit-risk assessments, so future research could look at extensions of this approach which adapt it to decision questions involving both efficacy and harms.

Despite the plethora of methods available to present results from evidence synthesis and aid treatment choice, their use in clinical practice often remains problematic. Barriers and challenges to knowledge translation include a lack of skills from end-users to appraise, comprehend, and implement the methods, among others [18–21]. A practical example of an initiative that tried to overcome such issues is a recent ongoing project in the mental health field that is combining evidence synthesis techniques and prediction modelling to build a web-based decision support algorithm [22]. With this tool, patients and clinicians' preferences are entered to produce personalised treatment recommendations tailored to the individual-patient level. The tool will then be tested in a randomised controlled trial to assess its feasibility, effectiveness, and acceptability in a real-world clinical setting.

We developed a user-friendly web application that facilitates the application of the framework we presented for assessing bias due to missing evidence in NMA. The tool has been adopted in the field and is already being used in practice. The code is freely available online and we accept feedback and contributions from the community. We also provide training material and documentation which is key when providing tools tailored for use by individual users without expert guidance.

In this thesis we made significant contributions to the evidence synthesis field. We conducted the first empirical study assessing the level of agreement between rankings and we showed that this is usually very high, unless there are large imbalances in the precision of the estimates, which are however rare in practice. We extended the existing ranking methodology by producing a quantity that can be used to explore if rankings vary for a specific set of individual preferences across a range of different trade-off between benefits and harms. We developed the first tool to evaluate the risk of bias due to missing evidence in a network of interventions and we provided a user-friendly web application to facilitate the assessment and automate some of the required steps. These contributions provide knowledge and tools that can support clinicians, policy makers and patients to choose the most preferable treatment for a specific condition. The development of new methods must be conceived with the aim of translating them into practice and delivering them to the end-users, not to purely produce research output to be confined within the scientific and academic community. In this sense, our results represent a move forward in the direction of translating knowledge into practical use and actively implementing evidence synthesis methodology. In general, more implementation research in clinical practice to guide decision-making is needed, focusing on facilitating the adoption of new methods and ensuring the perspectives of all end-users, including patients, are considered in the process.

References overall discussion and outlook

1 Norton EC, Miller MM, Wang JJ, *et al.* Rank Reversal in Indirect Comparisons. *Value in Health* 2012;**15**:1137–40. doi:10.1016/j.jval.2012.06.001

2 van Valkenhoef G, Ades AE. Evidence Synthesis Assumes Additivity on the Scale of Measurement: Response to “Rank Reversal in Indirect Comparisons” by Norton et al. *Value in Health* 2013;**16**:449–51. doi:10.1016/j.jval.2012.11.012

3 Davies AL, Galla T. Degree irregularity and rank probability bias in network meta-analysis. *Research Synthesis Methods* 2021;**12**:316–32. doi:10.1002/jrsm.1454

4 Salanti G, Nikolakopoulou A, Efthimiou O, *et al.* Introducing the Treatment Hierarchy Question in Network Meta-Analysis. *American Journal of Epidemiology* 2022;**191**:930–8. doi:10.1093/aje/kwab278

5 Daly CH, Mbuagbaw L, Thabane L, *et al.* Spie charts for quantifying treatment effectiveness and safety in multiple outcome network meta-analysis: a proof-of-concept study. *BMC Med Res Methodol* 2020;**20**:266. doi:10.1186/s12874-020-01128-2

6 Nikolakopoulou A, Higgins JPT, Papakonstantinou T, *et al.* CINeMA: An approach for assessing confidence in the results of a network meta-analysis. *PLOS Medicine* 2020;**17**:e1003082. doi:10.1371/journal.pmed.1003082

7 Sterne JA, Hernán MA, Reeves BC, *et al.* ROBINS-I: a tool for assessing risk of bias in non-randomised studies of interventions. *BMJ* 2016;:i4919. doi:10.1136/bmj.i4919

8 CINeMA: Confidence in Network Meta-Analysis. 2017.https://cinema.ispm.unibe.ch/ (accessed 11 Oct 2021).

9 ROB-MEN: Risk Of Bias due to Missing Evidence in Network meta-analysis. 2021.https://cinema.ispm.unibe.ch/rob-men/ (accessed 11 Oct 2021).

10 Mbuagbaw L, Rochwerg B, Jaeschke R, *et al.* Approaches to interpreting and choosing the best treatments in network meta-analyses. *Syst Rev* 2017;**6**:79. doi:10.1186/s13643-017-0473-z

11 Salanti G, Del Giovane C, Chaimani A, *et al.* Evaluating the Quality of Evidence from a Network Meta-Analysis. *PLoS ONE* 2014;**9**:e99682. doi:10.1371/journal.pone.0099682

12 Chaimani A, Porcher R, Sbidian E, *et al.* A Markov Chain approach for ranking treatments in network meta-analysis. Epidemiology 2019. doi:10.1101/19008722

13 Ostinelli EG, Efthimiou O, Naci H, *et al.* Vitruvian plot: a visualisation tool for multiple outcomes in network meta-analysis. *Evid Based Mental Health* 2022;:ebmental-2022-300457. doi:10.1136/ebmental-2022-300457

14 Puhan MA, Schunemann HJ, Murad MH, *et al.* A GRADE Working Group approach for rating the quality of treatment effect estimates from network meta-analysis. *BMJ* 2014;**349**:g5630–g5630. doi:10.1136/bmj.g5630

15 Phillippo DM, Dias S, Ades AE, *et al.* Sensitivity of treatment recommendations to bias in network meta-analysis. *J R Stat Soc A* 2018;**181**:843–67. doi:10.1111/rssa.12341

16 Phillippo DM, Dias S, Welton NJ, *et al.* Threshold Analysis as an Alternative to GRADE for Assessing Confidence in Guideline Recommendations Based on Network Meta-analyses. *Ann Intern Med* 2019;**170**:538. doi:10.7326/M18-3542

17 Papakonstantinou T, Salanti G, Mavridis D, *et al.* Answering complex hierarchy questions in network meta-analysis. *BMC Med Res Methodol* 2022;**22**:47. doi:10.1186/s12874-021-01488-3

18 Straus SE, Tetroe J, Graham I. Defining knowledge translation. *Canadian Medical Association Journal* 2009;**181**:165–8. doi:10.1503/cmaj.081229

19 Gravel K, Légaré F, Graham ID. Barriers and facilitators to implementing shared decision-making in clinical practice: a systematic review of health professionals' perceptions. *Implementation Sci* 2006;**1**:16. doi:10.1186/1748-5908-1-16

20 Cabana MD, Rand CS, Powe NR, *et al.* Why Don't Physicians Follow Clinical Practice Guidelines?: A Framework for Improvement. *JAMA* 1999;**282**:1458. doi:10.1001/jama.282.15.1458

21 Milner M, Estabrooks CA, Myrick F. Research utilization and clinical nurse educators: a systematic review. *J Eval Clin Pract* 2006;**12**:639–55. doi:10.1111/j.1365-2753.2006.00632.x

22 Tomlinson A, Furukawa TA, Efthimiou O, *et al.* Personalise antidepressant treatment for unipolar depression combining individual choices, risks and big data (PETRUSHKA): rationale and protocol. *Evid Based Mental Health* 2020;**23**:52–6. doi:10.1136/ebmental-2019-300118

Acknowledgements

I would like to start by thanking my supervisor Prof. Georgia Salanti for giving me the opportunity to join her research group and for her guidance and constructive criticism. I am grateful to her for creating a supportive work environment and for her understanding during personal difficult times. I wish to thank my co-referee Prof. Tianjing Li for her support and feedback and for always being accommodating and willing to help. Despite we were never able to meet in person, I really enjoyed our chats and discussions during the supervisory meetings. I would also like to thank all my co-authors and collaborators as this PhD project would have not been possible without their invaluable input and contributions. I would particularly like to recognise the helpful advice and assistance of Dr Adriani Nikolakopoulou and Dr Theodoros Papakonstantinou.

I had the great pleasure and privilege to be part of the Evidence Synthesis Methods team and learn a lot from its past and current members. Thanks to my fellow PhD colleagues, Mike, Konstantina and Tasnim, for the mutual support and their availability to help at any time. I cannot begin to express my thanks to Alex Holloway for his work on the ROB-MEN app, for his help on all sorts of software issues, for proof-reading this thesis, and for always encouraging me to improve my programming (as well as my skateboarding) skills. Special thanks go to Chiara Gastaldon for her professional insights and assistance on both work and personal matters, and for never saying no whenever I asked for help.

Thanks to my colleagues and officemates, many of whom I have the pleasure of calling friends. Thanks for the chats, laughs, coffee breaks, afternoon teas & cakes, gelato breaks and drinks, which made this journey even more enjoyable.

Thanks to Dr Cinzia Del Giovane for unofficially being my mentor for the past four years and for always listening, particularly when I needed it most. I will never forget her valuable advice and our fun times in Bern, Bologna and San Benedetto.

Thanks to my close friends in Bern as they have been like a second family to me. I cannot imagine how I would have managed the past few years without their love and support.

Thanks to my friends abroad who, despite the distance, have always been present and are more siblings than friends to me. They are and have been my anchor, especially in the past two years.

Infine, il mio più sentito ringraziamento va ai miei genitori per avermi dato le opportunità di conseguire i miei obiettivi e per i loro insegnamenti, che mi hanno resa quello che sono. Grazie per avermi trasmesso la passione per i viaggi, l'arte, la scienza e la cultura, per aver creduto in me e per avermi spinto a guardare oltre e a crescere intellettualmente sempre. Ringrazio mia madre per avermi sostenuto in ogni momento, anche a distanza, mettendo da parte tutte le sue paure e debolezze. Ringrazio mio padre per rimanere al suo fianco in questo momento particolarmente difficile, facendo ogni giorno del suo meglio per strapparle un sorriso, e per la straordinaria forza d'animo che dimostra in ogni momento.

Declaration of Originality

Last name, first name: Chiocchia Virginia

Matriculation number: 18-130-625

I hereby declare that this thesis represents my original work and that I have used no other sources except as noted by citations.

All data, tables, figures and text citations which have been reproduced from any other source, including the internet, have been explicitly acknowledged as such.

I am aware that in case of non-compliance, the Senate is entitled to withdraw the doctorate degree awarded to me on the basis of the present thesis, in accordance with the "Statut der Universität Bern (Universitätsstatut; UniSt)", Art. 69, of 7 June 2011.

Place, date

Bern, 19/10/2022

Signature