OPEN ACCESS

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]

10.1136/bmjebm-2021-111904

[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, University College London, London, UK

Correspondence to:
Virginia Chiocchia, Institute of Social and Preventive Medicine, University of Bern, 3012 Bern, Switzerland; virginia.chiocchia@ispm.unibe.ch

To cite: Chiocchia V, White IR, Salanti G. *BMJ Evidence-Based Medicine* 2023;**28**:180–182.

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

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.

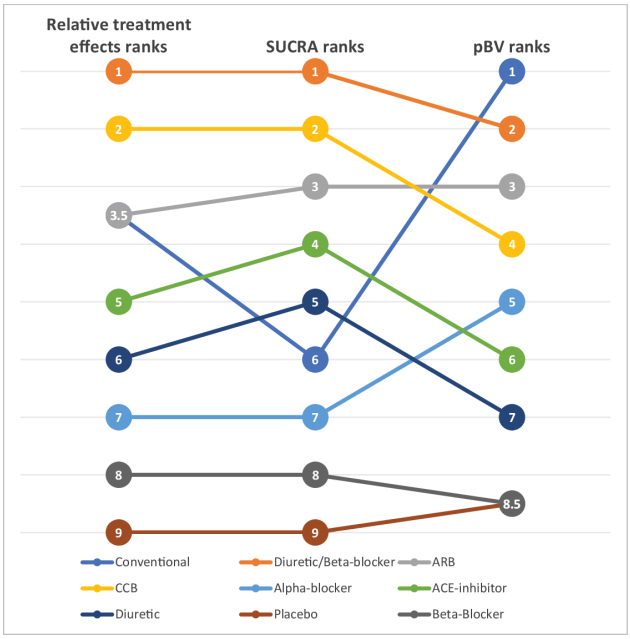

Figure 1 Example of treatment hierarchies from different ranking metrics for a network of nine antihypertensive treatment for primary prevention of cardiovascular disease. ARB, angiotensin receptor blockers; CCB, calcium channel blockers; p_{BV}, probability of producing the best value; $SUCRA$, surface under the cumulative ranking curve (calculated in frequentist setting).

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 vs 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

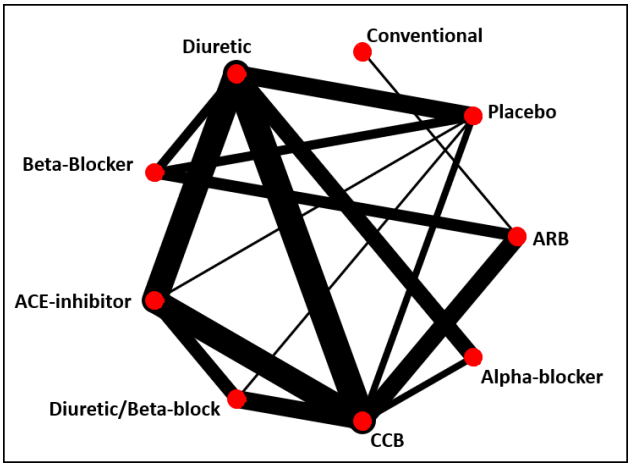

Figure 2 Graph of network of nine antihypertensive treatments for primary prevention of cardiovascular disease. line width is proportional to inverse SE of estimates from random effects model comparing two treatments. ARB, angiotensin receptor blockers; CCB, calcium channel blockers.

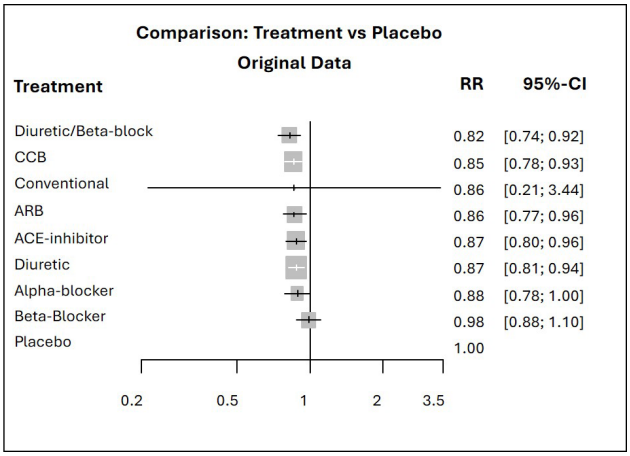

Figure 3 Forest plots of relative treatment effects of overall mortality for each treatment vs placebo. ARB, angiotensin receptor blockers; CCB, calcium channel blockers; RR, risk ratio.

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' (eg, 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 NMA 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. Multicriteria decision analysis 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 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.

Contributors VC drafted the manuscript. All authors reviewed and commented on drafts and on the final version 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.

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.

Disclaimer The funders had no involvement in the writing of this manuscript.

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

Patient consent for publication Not applicable.

Provenance and peer review Not commissioned; externally peer reviewed.

ORCID iDs
Virginia Chiocchia http://orcid.org/0000-0002-6196-3308
Georgia Salanti http://orcid.org/0000-0002-3830-8508

References

1 **Cipriani A**, Higgins JPT, Geddes JR, *et al*. Conceptual and technical challenges in network meta-analysis. *Ann Intern Med* 2013;**159**:130.
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. *Res Synth Methods* 2012;**3**:80–97.
3 **Nikolakopoulou A**, Higgins JPT, Papakonstantinou T, *et al*. Cinema: an approach for assessing confidence in the results of a network meta-analysis. *PLoS Med* 2020;**17**:e1003082.
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* 2017;**82**:20–8.
5 **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.
6 **Nikolakopoulou A**, Mavridis D, Chiocchia V. Network meta-analysis results against a fictional treatment of average performance: treatment effects and ranking metric.. *Res Syn Meth* 2020.
7 **Rücker G**, Schwarzer G. Ranking treatments in frequentist network meta-analysis works without resampling methods. *BMC Med Res Methodol* 2015;**15**:58.
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.
9 **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.
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. *Clin Epidemiol* 2014;**451**.
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.
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.
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.
14 **Salanti G**, Nikolakopoulou A, Efthimiou O, *et al*. Introducing the treatment hierarchy question in network meta-analysis. *Am J Epidemiol* 2021. doi:10.1093/aje/kwab278. [Epub ahead of print: 23 Nov 2021].
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.
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.
17 **van Valkenhoef G**, Tervonen T, Zhao J, *et al*. Multicriteria Benefit-risk assessment using network meta-analysis. *J Clin Epidemiol* 2012;**65**:394–403.
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* 2020;**62**:375–385.
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.
20 **Chaimani A**, Porcher R, Sbidian E. A Markov chain approach for ranking treatments in network meta-analysis. *Epidemiology* 2019.
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.
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. *Ann Intern Med* 2019;**170**:538.