RANKING PROBABILITIES IN NETWORK META-ANALYSIS

METHODOLOGICAL AND ANALYTICAL CONSIDERATIONS ON RANKING PROBABILITIES IN NETWORK META-ANALYSIS: EVALUATING COMPARATIVE EFFECTIVENESS AND SAFETY OF INTERVENTIONS

By CAITLIN HELEN DALY, Hons B.Sc., M.Sc.

A Thesis Submitted to the School of Graduate Studies in Partial Fulfilment of the Requirements for the Degree of Doctor of Philosophy

Doctor of Philosophy (2020) McMaster University

Health Research Methods, Evidence, & Impact Hamilton, Ontario, Canada

TITLE: Methodological and analytical considerations on ranking probabilities in network meta-analysis: Evaluating comparative effectiveness and safety of interventions

AUTHOR: Caitlin Helen Daly
B.Sc. (Honours), McMaster University
M.Sc. (Statistics), McMaster University

SUPERVISOR: Dr. Jemila S. Hamid

NUMBER OF PAGES: xvi, 165

Lay abstract

Decisions on how to best treat a patient should be informed by all relevant evidence comparing the benefits and harms of available options. Network meta-analysis (NMA) is a statistical method for combining evidence on at least three treatments and produces a coherent set of results. Nevertheless, NMA results are typically presented separately for each health outcome (e.g., length of hospital stay, mortality) and the volume of results can be overwhelming to a knowledge user. Moreover, the results can be contradictory across multiple outcomes. Statistics that facilitate the ranking of treatments may aid in easing this interpretative burden while limiting subjectivity. This thesis aims to address methodological gaps and limitations in current ranking approaches by providing alternative methods for evaluating the robustness of treatment ranks, establishing comparative rankings, and integrating ranking probabilities across multiple outcomes. These contributions provide objective means to improve the use of comparative treatment rankings in NMA.

Abstract

Network meta-analysis (NMA) synthesizes all available direct (head-to-head) and indirect evidence on the comparative effectiveness of at least three treatments and provides coherent estimates of their relative effects. Ranking probabilities are commonly used to summarize these estimates and provide comparative rankings of treatments. However, the reliability of ranking probabilities as summary measures has not been formally established and treatments are often ranked for each outcome separately. This thesis aims to address methodological gaps and limitations in current literature by providing alternative methods for evaluating the robustness of treatment ranks, establishing comparative rankings, and integrating ranking probabilities across multiple outcomes. These novel tools, addressing three specific objectives, are developed in three papers.

The first paper presents a conceptual framework for quantifying the robustness of treatments ranks and for elucidating potential sources of lack of robustness. Cohen's kappa is proposed for quantifying the agreement between two sets of ranks based on NMAs of the full data and a subset of the data. A leave one-study-out strategy was used to illustrate the framework with empirical data from published NMAs, where ranks based on the surface under the cumulative ranking curve (SUCRA) were considered. Recommendations for using this strategy to evaluate sensitivity or robustness to concerning evidence are given.

When two or more cumulative ranking curves cross, treatments with large probabilities of ranking the best, second best, third best, etc. may rank worse than treatments with smaller corresponding probabilities based on SUCRA. This limitation of

SUCRA is addressed in the second paper through the proposal of partial SUCRA (pSUCRA) as an alternative measure for ranking treatments. pSUCRA is adopted from the partial area under the receiver operating characteristic curve in diagnostic medicine and is derived to summarize relevant regions of the cumulative ranking curve.

Knowledge users are often faced with the challenge of making sense of large volumes of NMA results presented across multiple outcomes. This may be further complicated if the comparative rankings on each outcome contradict each other, leading to subjective final decisions. The third paper addresses this limitation through a comprehensive methodological framework for integrating treatments' ranking probabilities across multiple outcomes. The framework relies on the area inside spie charts representing treatments' performances on all outcomes, while also incorporating the outcomes' relative importance. This approach not only provides an objective measure of the comparative ranking of treatments across multiple outcomes, but also allows graphical presentation of the results, thereby facilitating straightforward interpretation.

All contributions in this thesis provide objective means to improve the use of comparative treatment rankings in NMA. Further extensive evaluations of these tools are required to assess their validity in empirical and simulated networks of different size and sparseness.

Publications related to thesis

Daly, C. H., Neupane, B., Beyene, J., Thabane, L., Straus, S. E., & Hamid, J. S. (2019). Empirical evaluation of SUCRA-based treatment ranks in network meta-analysis: quantifying robustness using Cohen's kappa. *BMJ Open, 9*(9), e024625. https://doi.org/10.1136/bmjopen-2018-024625

Daly, C. H., Mbuagbaw, L., Thabane, L., Straus, S. E., & Hamid, J. S. (2020a). Partial surface under the cumulative ranking curve (pSUCRA) as an alternative measure for ranking treatments in network meta-analysis. *Submitted to Clinical Epidemiology.*

Daly, C. H., Mbuagbaw, L., Thabane, L., Straus, S. E., & Hamid, J. S. (2020b). Spie charts for quantifying treatment effectiveness and safety in multiple outcome network meta-analysis: A proof-of-concept study. *Submitted to BMC Medical Research Methodology*. https://doi.org/10.21203/rs.3.rs-36139/v1

Acknowledgements

I am thankful to have been blessed with so many kind and amazing people who have supported and uplifted me throughout this journey.

I wish to thank my incredible supervisor, Dr. Jemila Hamid, for her invaluable guidance, wisdom, encouragement, and patience. There are not enough words to express how grateful I am for the opportunities and knowledge you have given me. Thank you for listening and believing in me.

I would like to express my deepest gratitude to my supervisory committee members Dr. Lawrence Mbuagbaw, Dr. Sharon Straus, and Dr. Lehana Thabane for their support and constructive feedback. Thank you for taking the time to meet with me, teach me, and help me progress. I also greatly appreciate the several research and teaching opportunities you have provided me.

In addition, I would like to extend my sincere thanks to my external examiner, Dr. George Wells, for taking the time to review this thesis and for providing very helpful feedback.

I would also like to recognize the support I have received from Lorraine Carroll, Dr. Steven Hanna, Sophia Piro, and Kristina Vukelic at McMaster University during my PhD studies. I have also been supported by Dr. Andrea Tricco and Dr. Areti Angeliki Veroniki at the Li Ka Shing Knowledge Institute. Thank you for sharing your knowledge with me and for encouraging me. I am also indebted to my colleagues at the University of Bristol,

especially Prof. Tony Ades, Prof. Sofia Dias, and Prof. Nicky Welton; I have learned so much from you.

I wish to express my sincere appreciation and gratitude to my friends and family in Canada and the United Kingdom for keeping me afloat. A special thank you to my friends in graduate school, Charlene, Erik, Joycelyne, Sayantee, and Thuva. Jarred, I cannot emphasize enough how much I appreciate everything you have done for me, thank you for being incredibly patient with me. To my parents, thank you for your unconditional love and support. I appreciate how you have always been there for me. Finally, Emma, thank you for being the brightest part of my life.

Table of contents

List of figures

Chapter 2

Chapter 3

Chapter 4

Chapter 5

List of tables

Chapter 3

Chapter 4

Chapter 5

List of abbreviations

AUC	area under the receiver operating characteristic curve
CHD	coronary heart disease
CI	confidence interval
COPD	chronic obstructive pulmonary disease
CrI	credible interval
CVD	cardiovascular disease
DLQI	dermatology life quality index
HDL-c	high-density lipoprotein cholesterol
IPD	individual patient data
LDL-c	low-density lipoprotein cholesterol
MCMC	Markov chain Monte Carlo
NICE	National Institute for Health and Care Excellence
NMA	network meta-analysis
OR	odds ratio
PASI	psoriasis area severity index
pAUC	partial area under the receiver operating characteristic curve
pSUCRA	partial surface under the cumulative ranking curve
RCT	randomized controlled trial
ROC	receiver operating characteristic
SUCRA	surface under the cumulative ranking curve
TC	total cholesterol
TG	triglycerides
Trt	treatment

Declaration of academic achievement

This thesis is structured as a "sandwich thesis", comprised of three manuscripts prepared for journal publication. The first manuscript, Paper I presented in Chapter 3, has been published in *BMJ Open*, while Paper II (Chapter 4) and Paper III (Chapter 5) have been submitted to *Clinical Epidemiology* and *BMC Medical Research Methodology*, respectively. Caitlin Daly contributed to the following components of the manuscripts: led the conceptualization and design of the studies; conducted all the statistical analyses; provided interpretation of the results; drafted and revised all the manuscripts as the first author; submitted the manuscripts to journals and took leadership in addressing and responding to reviewers' comments. This scholarly work was conducted between Fall 2014 and Summer 2020.

Chapter 1
Introduction and objectives

1.1 Introduction

Individuals make decisions every day. ‘Decision’ has been defined as “a choice or judgment that you make after thinking and talking about what is the best thing to do” (“Decision”, 2020). Making a choice implies there are at least two options to select from. Reasons for choosing one option over another include chance (e.g., flipping a coin) and preferences that naturally formulate an ordered list of the options. Selecting the best option of such lists adheres to the theory of rational choice (Osborne, 2004). In the context of health care, when recommending or selecting a treatment, the overarching preferences may be defined in terms of efficacy and safety, and if relevant to the decision maker, financial costs (Dias et al., 2018). Ideally, the best intervention will be the most efficacious and tolerable, with the least side effects and costs.

Randomized controlled trials (RCTs) are the optimal study design for comparing the efficacy of interventions and thereby inform treatment decisions (Djulbegovic & Guyatt, 2017). When there are multiple RCTs addressing the same research question defined in terms of the population, intervention(s), comparator, and outcome(s) of interest, a systematic review may be conducted to collect and synthesize their results (Higgins et al., 2019a; Liberati et al., 2009; Moher et al., 2015a). Doing so provides a means to summarize multiple results presented across trials, which may have not been comprehendible to a decision maker otherwise. Meta-analysis is a statistical approach for combining quantitative comparative evidence on two interventions, while network meta-analysis (NMA) combines direct and indirect evidence on three or more interventions (McKenzie et al., 2019). Meta-analysis helps determine whether one intervention is better than another, while NMA may determine which intervention is best among several possible options. If there are no biases present among the included RCTs, both in terms of their conduct and generalizability, these synthesis methods should deliver more precise estimates of the true intervention effects defined by the research question. Meta-analysis and NMA also provide an objective means of quantifying and examining heterogeneity across studies and offer a comprehensive framework for elucidating sources of heterogeneity. Researchers may assess and adjust for known effect modifiers through subgroup analyses or meta-regression (Berkey et al., 1995; Cooper et al., 2009; Oxman & Guyatt, 1992). If individual patient data are available, within-study variation may be explored as well (Higgins et al., 2001).

Although well-intentioned researchers strive to reduce or eliminate the presence of bias within RCTs and the systematic reviews that include them, bias still arises and should nevertheless be of concern to a decision maker. The amount of uncertainty should also play a role since estimates can only be derived with a certain level of precision. Rigorous quality and credibility assessment tools for RCTs, e.g., Cochrane Risk-of-Bias (RoB) Tool, and the NMAs that include them, e.g., the Grading of Recommendations Assessment, Development and Evaluation (GRADE) approach and Confidence in Network Meta-Analysis (CINeMA), aim to assist a decision maker in this regard (Brignardello-Petersen et al., 2018; Nikolakopoulou et al., 2020; Puhan et al., 2014; Sterne et al., 2019). However, these tools currently only describe the credibility of evidence; how this should translate to a hierarchy of interventions remains unclear. For example, an estimate may not be biased enough to impact a rational decision maker’s choice (Phillippo et al., 2019). It may be more helpful to outline thresholds indicating how much an intervention effect estimate must change such that an alternative intervention would be recommended based on a predefined decision rule (Phillippo et al., 2017). Regardless of the approach, balancing bias and uncertainty alongside multiple measures of efficacy and safety in an objective and transparent way is not an easy task.

Information overload becomes more likely as the numbers of interventions and outcomes increase, which can hinder a decision maker’s ability to make a rational choice (Buchanan & Kock, 2001). Reducing the volume of information, while still communicating important criteria that should inform a decision maker’s preferences is, therefore, important. Since the ultimate task of a decision maker is to develop an ordinal list of the options, rankings statistics may aid in this venture and lead to evidence-informed policies and decisions.

1.2 Motivation and objectives

Ranks are increasingly being included in published NMAs to summarize the comparative effectiveness and safety of interventions, and yet their use has been met with criticism (Petropoulou et al., 2017; Veroniki et al., 2018). One main and understandable concern is that differences between ranks implies there are differences between the relative effects of interventions, when there may be no relevant difference at all (Mbuagbaw et al., 2017; Mills et al., 2012; Trinquart et al., 2016). Ranks also do not acknowledge potential biases in the evidence (Mbuagbaw et al., 2017). In general, ranks are ordinal data and thus do not provide as much information as interval or ratio data such as relative effects (Stevens, 1946).

While it is clear that decisions should not solely be based on ranking statistics in their current form, they are still being used and even recommended in NMAs (Jansen et al., 2011; Petropoulou et al., 2017). Psychologists have long noted the human tendency to take shortcuts when making decisions (Korteling et al., 2018; Tversky & Kahneman, 1974). This might mean decision makers place an undue reliance on rankings if they are presented in NMA results. It is, therefore, of interest to study the behaviour of treatment ranks in NMA and see if they can be adapted to better inform and meet the needs of decision makers. Ranking probabilities, that is, the probability of an intervention ranking best, second best, third best, etc., and their summaries may serve as a starting point, as they incorporate the uncertainty in the relative effects estimated in an NMA (Salanti et al., 2011). Ranking probabilities are also advantageous for decision making as they translate comparative effectiveness and safety measures to a scale between 0 and 1, thereby providing a mechanism to compare interventions across multiple outcomes.

This thesis aims to address some of the concerns over the use of ranks to summarize NMA results by improving our understanding and use of ranking probabilities. Novel frameworks and summary measures are developed with the aim of providing tools for 1) assessing the impact questionable evidence (e.g., a study at high risk of bias) may have on a hierarchy (ranking) of interventions, 2) uncovering important information masked by an existing ranking measure, and 3) combining intervention hierarchies over multiple outcomes, including situations where the outcomes may not be of equal importance to knowledge users. These contributions may assist a decision maker in incorporating important aspects of NMA results into a single ordinal list representing their preferences.

1.3 Scope of the thesis

This sandwich-structured thesis has been motivated by an introduction to the challenges of using NMA results to develop a hierarchy of interventions, followed by a plan to address these challenges through ranking probabilities. Chapter 2 expands on this by providing a comprehensive literature review of systematic review methodology, evidence synthesis including meta-analysis and NMA, ranking probabilities, and the area under their cumulative curves, known as the surface under the cumulative ranking curve (SUCRA) (Salanti et al., 2008). Chapters 3, 4, and 5 then aim to provide tools for answering the following research questions:

1. How robust are SUCRA-based treatment ranks to the inclusion or exclusion of RCTs in the network?
2. Are there any study characteristics (e.g., sample size) associated with the robustness of SUCRA-based treatment ranks?
3. Is it helpful to consider a portion of the cumulative ranking curve, rather than the full curve, when ranking treatments?
4. Which treatment is best overall across multiple outcomes?

SUCRA objectively balances the estimated relative effects as well as their uncertainty and is becoming one the most increasingly used measures to rank treatments in NMA (Petropoulou et al., 2017). Nevertheless, there is limited information on the robustness or sensitivity of SUCRA and its corresponding ranks. This motivates research questions 1 and 2, which are addressed in Chapter 3. Chapter 3 corresponds to Paper I of this sandwich thesis, where we develop a novel framework for quantifying the robustness of SUCRA-based treatment ranks in determining intervention hierarchies. The framework also allows investigation of potential factors associated with lack of robustness of SUCRA-based ranks. In developing the framework, Cohen's kappa is proposed as a measure to quantify the agreement between SUCRA-based ranks calculated in an NMA with the full evidence and an NMA of a subset of evidence (Cohen, 1960, 1968). Thus, this provides a measure of the robustness of the ranks to the evidence removed in the latter network. This permits an assessment of how concerning or questionable evidence (e.g., because of bias) impacts the intervention hierarchy. The advantages of Cohen's kappa are discussed, including the option to apply weights to different changes in rank, which may range in their degree of importance to a stakeholder.

Although SUCRA is favoured among ranking measures as it incorporates the uncertainty of the estimated treatment effects (Salanti et al., 2014), it has some limitations, particularly in situations where the cumulative ranking curves of two or more treatments cross. Interventions with similar values of SUCRA do not necessarily imply they are equally effective, since SUCRA is a trade-off of the magnitude of the effect estimates and their uncertainty. For this reason, interventions that have higher probabilities of ranking among the best may have lower SUCRA values than interventions with smaller probabilities of ranking among the best. This limitation motivates research question 3, which is addressed in Paper II (Chapter 4) of this thesis. In this paper, we adapt the conceptual and methodological framework behind the partial area under the receiver operating characteristic curve (pAUC) from diagnostic medicine (McClish, 1989). We provide an alternative method for ranking treatments or interventions in NMA, which we refer to as the partial surface under the cumulative ranking curve (pSUCRA).

Systematic reviews and their corresponding evidence synthesis often involve multiple outcomes. Traditionally, NMAs are conducted outcome by outcome and knowledge users are consequently inundated with large amounts of information. Moreover, conflicting ranking in terms of the relative efficacy and/or safety of interventions might arise in NMAs involving multiple outcomes, and hence lead to challenges in appropriately interpreting and utilizing the results of NMA. This motivates research question 4, and this is addressed in Chapter 5, which corresponds to Paper III of this thesis. In this paper, we present a comprehensive methodological framework for integrating ranking probabilities across multiple outcomes. We use the area inside a spie chart (Stafoggia et al., 2011) to integrate SUCRA values across multiple outcomes and provide an objective measure for ranking interventions. We illustrate mechanisms to calculate weights to reflect the desired contributions of each outcome and discuss some evidence-based sources for informing the weights. This paper also illustrates that the spie chart is more favourable than a radar plot, an alternative method used in a recently published NMAs.

Finally, a summary of the thesis and some concluding remarks are provided in Chapter 6, where we also highlight key contributions and potential limitations. We also discuss potential extensions and areas for future research.

Chapter 2
Background

2.1 Evidence synthesis

Research studies are typically performed on a sample of the population, and thus deliver estimates of the true answer to a research question. How well the estimate represents the truth depends on the nature of data (e.g., variability, distribution) collected by the study investigators as well as the sample of the population. Studies may be repeated because of unawareness or scepticism of estimates from previous studies, or a need to extend the generalizability or applicability of the results to a different context (Cooper & Hedges, 2009). Even if a study was replicated in the exact same way, a different estimate may be obtained because of random variation. As such, there may be multiple estimates of the true parameter of interest.

Evidence synthesis is the practice of combining knowledge on a topic from multiple sources (Donnelly et al., 2018). Researchers may adopt this practice prior to conducting primary research to gain an understanding of what evidence already exists and to identify any gaps (Chalmers et al., 2014; Macleod et al., 2014). Results from an evidence synthesis are also used by stakeholders to make evidence-informed policies and decisions, and to provide evidence-informed patient care (Mays et al., 2005; Moat et al., 2013). Several researchers have classified existing synthesis methods based on theoretical grounds (Grant & Booth, 2009; Perrier et al., 2016; Royal Pharmaceutical Society, 2011; Tricco et al., 2016a), and have collectively identified nearly 30 approaches (Littell, 2018). These approaches may be utilized to synthesize evidence qualitatively, quantitatively, or both, and may be used in combination with each other e.g., systematic reviews and meta-analysis, as described shortly (Liberati et al., 2009).

Reasons for selecting an evidence synthesis approach may include its purpose, scope, the type of data to be synthesized (i.e., qualitative or quantitative), and the resources available (Kastner et al., 2016; Littell, 2018). For example, meta-synthesis is a common approach used to gain perceptions of a population from qualitative data (Tricco et al., 2016b). Overviews of reviews, rapid reviews, scoping reviews, and systematic reviews are used to synthesize evidence on the comparative effectiveness and safety of interventions in health care (Tricco et al., 2018). When an urgent decision must be made in a short timeframe, rapid reviews may be the preferred choice to synthesize such evidence (Tricco et al., 2018). However, a full systematic review is considered to be the gold standard and is thus preferable when the required resources are available (Khangura et al., 2012).

2.2 Systematic reviews

Systematic reviews aim to provide a comprehensive review of all available evidence on a pre-specified research question (Lasserson et al., 2019). They were introduced in the social sciences in the 1970s and have since been adopted by health researchers (Moher et al., 2015b). In addition to evidence on health care interventions, they may summarize evidence on other health care topics such as diagnostic tests and prognostic factors. What sets systematic reviews apart from other types of reviews is the rigorous methodology they employ to minimize any potential biases that may arise from the conduct of the review

(Liberati et al., 2009). For example, a plan or protocol for a systematic review should be developed in advance, where the scope, research question, search strategy, inclusion/exclusion criteria, and synthesis methods are explicitly set out (Lasserson et al., 2019). The Preferred Reporting Items for Systematic Review and Meta-Analysis Protocols (PRISMA-P) statement provides guidance on developing sufficient protocols (Moher et al., 2015a). Similar to randomized controlled trials (RCTs), systematic reviewers are encouraged to register their review on the International Prospective Register of Systematic Reviews (PROSPERO) (Stewart et al., 2012). Specifying the aforementioned building blocks of a systematic review on this platform increases transparency and reduces research waste (Moher et al., 2014).

Research questions regarding the comparative effectiveness of interventions typically take on a PICO format, where the Population, Interventions, Comparators, and Outcomes are specified (Richardson et al., 1995). Both the Joanna Briggs Institute Reviewer's Manual and the Cochrane Handbook for Systematic Reviews of Interventions recognize the importance of specifying PICO questions to guide the development of a systematic review (Aromataris & Munn, 2020; Higgins et al., 2019a). These questions will inform the literature search terms, as well as the criteria for including or excluding studies (Eriksen & Frandsen, 2018). To avoid missing relevant studies, databases should be strategically selected bearing in mind the discipline (topic) of the review, and sources of unpublished studies should be identified as well (Lefebvre et al., 2019). The literature search strategy may also be improved using the Peer Review of Electronic Search Strategies (PRESS) checklist (McGowan et al., 2016). Titles, abstracts, and full text of potentially relevant articles should be screened in duplicate against a pre-determined list of inclusion and exclusion criteria that aims to deliver a collection of evidence that meets the scope of the review (Edwards et al., 2002). Similarly, data may be extracted in duplicate or verified by a second reviewer to reduce possible extraction errors (Shamseer et al., 2015). In the end, these thorough measures should help make a systematic review reproducible (Lasserson et al., 2019). Each study included in the systematic review should be assessed for risk of bias in order to get a sense of the robustness of any conclusions drawn from the evidence (Shamseer et al., 2015). Authors may improve the transparency of a systematic review by following PRISMA (Liberati et al., 2009), and the quality of a systematic review may be critically assessed using A MeaSurement Tool to Assess systematic Reviews (AMSTAR) 2 (Shea et al., 2017).

Evidence on the comparative effectiveness and safety of interventions collected in a systematic review may be synthesized through narrative or statistical means (McKenzie et al., 2019). When a narrative approach is adopted, the individual study estimates may be summarized in text, tabular, or graphical form. Statistical approaches can be used to summarize the individual study estimates numerically. For example, the distribution of the effect sizes may be summarized by their range and p-values may be combined to compute a chi-squared test statistic (McKenzie et al., 2019). However, if the aim is to estimate the true relative effectiveness of interventions, and the included studies provide estimates of these, along with the corresponding measures of variability, then these estimates should be statistically synthesized, provided that the included studies are sufficiently similar to combine (Borenstein et al., 2009a; McGough & Faraone, 2009; Sullivan & Feinn, 2012).

Meta-analysis is used to quantitatively combine estimates of the relative effectiveness between two interventions (Borenstein et al., 2010). When there is evidence on the relative effects between three or more interventions forming a connected network, this evidence may be pooled using network meta-analysis (NMA) (Higgins & Welton, 2015).

2.3 Comparative effectiveness

In theory, RCTs are the best source of evidence on the comparative effectiveness of interventions (Devereaux & Yusuf, 2003). This is because randomization should eliminate confounders (i.e., determinants of whether a participant receives one intervention over another) and balance any known and unknown prognostic factors (i.e., predictors of the outcome) across the intervention groups (Moher et al., 2010). If, for example, patients of a more severely diseased population were purposely assigned to one intervention group over another, this could skew the comparisons between interventions. Thus, in a well-conducted and properly designed RCT, any difference between the outcomes of the intervention and comparator groups can be attributed to the causal effect of the intervention (Greenland, 1990). Although this thesis focuses on synthesis of evidence from RCTs, approaches for including observational or quasi-experimental studies exist (Metelli & Chaimani, 2020; Reeves et al., 2019).

When combining evidence on the comparative effectiveness of interventions across RCTs, relative effects, rather than absolute effects in each intervention arm, are synthesized to preserve the benefits of randomization (da Costa & Jüni, 2014). This is because prognostic factors will not be balanced across the intervention arms if the absolute effects for each arm are synthesized separately, from which a summary relative effect is calculated. 'Comparative effectiveness research' concerns the differences in both the benefits and harms of interventions (Sox, 2010). Measures for quantifying such differences, often referred to as effect measures, depend on the type of outcome (e.g., continuous, binary, time-to-event, count) and should be computed in an analysis that accounts for the study design (e.g., parallel, crossover, cluster) (Higgins et al., 2019b).

Effect measures for continuous data include the 'mean difference', which may be used to measure the absolute difference in two groups' mean outcomes on the same scale (e.g., systolic blood pressure), as well as the 'standardized mean difference', which may be used when outcomes are reported on different scales (e.g., depression scores) (Fu et al., 2008). Differences in the mean change from baseline may also be of interest when individual-level changes are important; for example, to summarize the relative effect of a weight loss intervention compared to another. Risk differences, risk ratios and odds ratios are common effect measures for binary outcomes such as mortality, stroke, or discontinuation of treatment for any reason (Borenstein et al., 2009b; Sekhon et al., 2017). Time-to-event outcomes such as progression-free survival are typically summarized as hazard ratios (Tierney et al., 2007), while count outcomes such as number of dental cavities may be summarized as rate ratios (Dias et al., 2018).

2.4 Frequentist and Bayesian inference

Statistical inference is the process of drawing conclusions for a population based on a sample of it (Casella & Berger, 2002). In meta-analysis or NMA, we are usually concerned with the point (e.g., the sample mean) and interval (e.g., 95% confidence interval) estimation of a parameter, θ. For example, a parameter of interest may be the true difference in mean systolic blood pressure levels between senior patients receiving diuretics versus those receiving beta-blockers. There are two main approaches to estimate the parameters of interest and to conduct inference: frequentist and Bayesian.

In a frequentist setting, a parameter is considered fixed and the data are generated from a random process characterized by this parameter (Casella & Berger, 2002). A point estimate alone does not provide any indication of how well it reflects the truth. A confidence interval may be constructed around this point estimate to provide a range of plausible values that may include the true parameter value. The interpretation of this interval relies on large numbers of hypothetical repetitions of the sampling process (Dobson & Barnett, 2008). For example, if the sampling process were repeated 1000 times, 950 of the estimated 95% confidence intervals for each sample would contain the true value.

In a Bayesian framework, the observed data, $\mathbf{y} = y_1, y_2, ..., y_n$, are considered to be fixed and the parameter depends on the data (Dobson & Barnett, 2008). A posterior distribution for a parameter, $P(\theta \mid \mathbf{y})$, depends on the likelihood of the data, $P(\mathbf{y} \mid \theta)$, and a prior belief about the parameter, expressed by a probability distribution, $P(\theta)$ (Lunn et al., 2013):

$$P(\theta \mid \mathbf{y}) \propto P(\mathbf{y} \mid \theta) P(\theta).$$

The mean or median of this posterior distribution is often taken to be a point estimate of the parameter, and an interval is usually defined by percentiles of the posterior distribution.

For example, the 2.5th and 97.5th percentiles may define the limits of a 95% credible interval. Note that a credible interval is different from a confidence interval. Since the parameter is not fixed, Bayesian estimates may be interpreted in probabilistic terms. For example, there is a 95% chance that the true parameter falls within a 95% credible interval.

The principles behind Bayesian inference reflect a decision maker's thought process (Dias et al., 2018). Usually a decision maker will have a pre-existing belief (prior), which is updated by new data (likelihood) to inform a new belief (posterior distribution). For example, consider the density plots of the simulated distributions presented in Figure 1A. The prior suggests θ is likely to be greater than zero, while the data captured by the likelihood suggests θ is more likely to be less than zero. The posterior distribution falls somewhere in between the prior and likelihood. It is slightly more influenced by the prior distribution, which has more certainty than the likelihood.

One criticism of Bayesian inference concerns the subjectivity that may be introduced through the specification of the prior distribution (Lambert et al., 2005), as frequentist inference solely relies on the data for parameter estimation. To mitigate this criticism, non-informative priors may be used (e.g, from the Gaussian distribution with mean 0 and large variance), so that the data largely inform the posterior distributions. For example, a vaguer prior is specified in Figure 1B, compared to the prior used in Figure 1A, as reflected by the spread of their distributions. Since there is more uncertainty in the prior distribution compared to the likelihood, the posterior distribution is influenced more by the likelihood, i.e., the data.

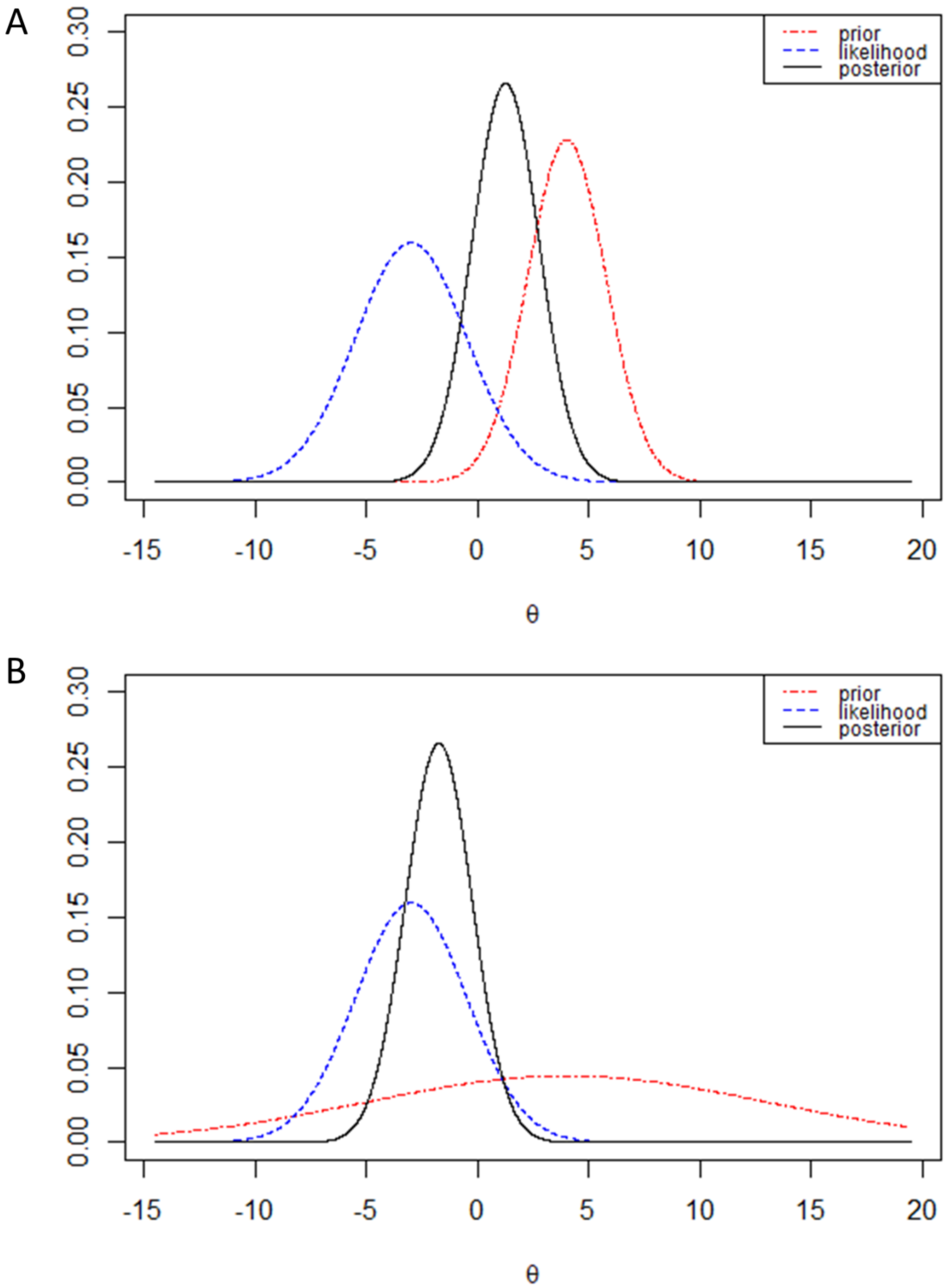


Figure 1: *Comparison of influence of the prior distribution and the likelihood on the posterior distribution in Bayesian inference. In (A), an informative prior distribution is specified; in (B) a less informative prior is specified.*

Until the invention of computers, Bayesian inference was largely hindered by the need to evaluate difficult integrals to derive summaries of the posterior distribution. For example, the mean is calculated as the expected value of the posterior distribution: $E\left(\theta \mid \mathbf{y}\right) = \int_{\theta} \theta P\left(\theta \mid \mathbf{y}\right) d\theta$ (Lunn et al., 2013). As such, conjugate priors are often used for mathematical simplicity. However, this becomes even more difficult when there are multiple parameter models from which a joint posterior distribution must be derived. For example, if $P\left(\boldsymbol{\theta} \mid \mathbf{y}\right)$ is the joint posterior distribution of parameters $\boldsymbol{\theta} = \left\{\theta_1, \theta_2, ..., \theta_K\right\}$, the marginal posterior distribution of θ_1 is $P\left(\theta_1 \mid \mathbf{y}\right) = \int_{\theta_2} \ldots \int_{\theta_K} P\left(\boldsymbol{\theta} \mid \mathbf{y}\right) d\theta_2 \ldots d\theta_K$ (Lunn et al., 2013). Such integrals may be approximated using Markov chain Monte Carlo (MCMC) simulation methods. These iterative algorithms, such as the Gibbs sampler (Casella & George, 1992), are used to simulate Markov chains which will eventually converge to a stationary distribution which should capture the joint posterior distribution of interest (Lunn et al., 2013). By sampling from this stationary distribution, we can approximate summaries of the marginal distributions using the principles of Monte Carlo integration (Lunn et al., 2013).

2.5 Meta-analysis

In a frequentist framework, a generic method for conducting meta-analysis is known as the inverse-variance approach (Deeks et al., 2019). In a meta-analysis involving n studies, the parameter of interest, θ, is first estimated in each study, θ_i, $i = 1, ..., n$, along with the corresponding variances of the estimates, V_i. Then, these study-specific estimates are pooled as a weighted average:

$$\theta = \sum_{i=1}^{n} \frac{w_i \theta_i}{w_i}$$

where w_i correspond to the individual study weights. A fixed effect meta-analysis assumes each study estimates the same parameter and $w_i = \frac{1}{V_i}$. Hence, larger studies will receive more weight than smaller studies in terms of their contribution to the pooled relative effect (Borenstein & Higgins, 2013).

It may be unreasonable to assume that each study estimates the same parameter, as they may differ in terms of potential effect modifiers such as the age of the patients or length of follow up (Riley et al., 2011). Effect modifiers lead to heterogeneity between the study estimates. Instead, a random effects meta-analysis may be conducted, where each study estimates its own parameter θ_i, but these parameters are related and follow a common a distribution (DerSimonian & Laird, 1986). The between-study variance in the parameters, τ^2, is captured in the study weights, $w_i = \frac{1}{V_i + \tau^2}$, and the study estimates are combined as a weighted average as described above (Borenstein et al., 2010). Note that, if $\tau^2 = 0$, a random effects meta-analysis produces similar results as a fixed effect meta-analysis.

Evidence of heterogeneity may be assessed statistically through the comparison of the fit of models assuming fixed or random effects, or through the chi-square test for heterogeneity (Deeks et al., 2019; Dias et al., 2011). In addition to the value of τ^2 itself, heterogeneity is sometimes assessed using the I^2 statistic, which measures the proportion of observed variation that is attributable to the difference in the true effects estimated by each study (Borenstein et al., 2017). Wherever possible, heterogeneity should be explained by accounting for differences in the distributions of known treatment effect modifiers across trials using subgroup analysis or meta-regression (Berkey et al., 1995; Oxman & Guyatt, 1992).

In addition to combining measures of relative effect, the inverse-variance approach is generic enough to combine estimates of a variety of parameters such as prevalence or incidence rates. However, since this approach assumes estimates are normally distributed around the true parameter (fixed effects) or the average of the true parameters (random effects), the parameters and their estimates should be transformed to a continuous scale. This may bias estimates of non-normal outcomes. Other meta-analytical approaches that do not require normal approximations include Peto's odds ratio, the arcsine difference, the Mantel-Haenszel approach, generalized linear (mixed) models, and Bayesian hierarchical models, which have been contrasted in (Efthimiou, 2018).

2.6 Network meta-analysis (NMA)

Network meta-analysis (NMA) is generalization of meta-analysis that allows the synthesis of relative effect estimates on three or more treatments connected in a network of evidence (Caldwell et al., 2005; Lu & Ades, 2004). Consider the simplest case, where the relative effects between treatments A, B, and C form a triangle network of evidence (Figure 2). Let

d_{AB}, d_{AC}, d_{BC} be the true relative effects of the treatments: B vs. A, C vs. A, and C vs. B, respectively. Each trial directly estimates the relative effects between the treatments it compares. For example, trials comparing treatments B and C provide direct evidence on the relative effect of C vs. B, d_{BC}^{direct}, where the variance of this estimate is denoted as $V\left(d_{BC}^{direct}\right)$. Indirect evidence may also be obtained from trials that compare one of these treatments to a common comparator by making use of the constraint

$$d_{BC} = d_{AC} - d_{AB}.$$

The above equation is known as a consistency equation (Lu & Ades, 2004). The parameters on the right side of this equation are called basic parameters (i.e., all treatment effects relative to a common comparator; in this example, treatment A). All other relative effects are simply a function of the basic parameters and are thus called functional parameters.

As long as the above constraint holds, direct estimates, d_{BC}^{direct}, and indirect estimates, $d_{BC}^{indirect} = d_{AC}^{direct} - d_{AB}^{direct}$, $V\left(d_{BC}^{indirect}\right) = V\left(d_{AB}^{direct}\right) + V\left(d_{AC}^{direct}\right)$, may be combined and should deliver a more precise estimate of the relative effect, d_{BC}^{pooled}. The simplest method to combine the direct and indirect estimates in a network of 3 treatments is known as the Bucher method (Bucher et al., 1997). Using this approach, the pooled estimate is a weighted average of the direct and indirect estimates,

$$d_{BC}^{pooled} = \frac{w_{BC}^{direct} d_{BC}^{direct} + w_{BC}^{indirect} d_{BC}^{indirect}}{w_{BC}^{direct} + w_{BC}^{indirect}}$$

where

$$w_{BC}^{direct} = \frac{1}{Var\left(d_{BC}^{direct}\right)}, w_{BC}^{indirect} = \frac{1}{Var\left(d_{BC}^{indirect}\right)}.$$

Thus, the contribution of the direct or indirect estimate to the pooled estimate depends on its precision. More weight is given to the more precise estimate.

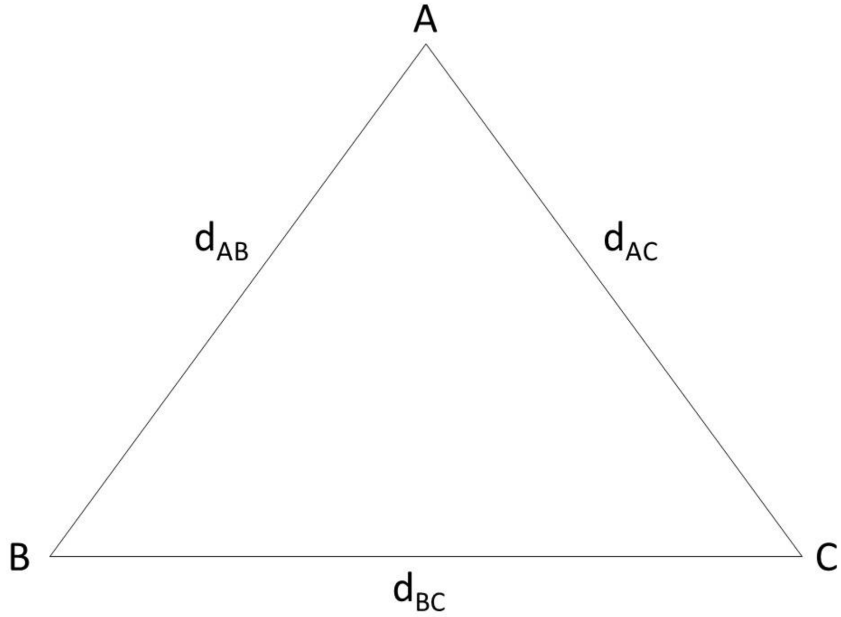


Figure 2: *Example network of evidence between treatments A, B, and C. The parameters of interest are the relative effects of B vs. A, d_{AB}, C vs. A, d_{AC}, and C vs. B, d_{BC}.*

While the Bucher method is simple to implement in a network of three treatments, the calculation of the indirect estimates increases in complexity as the number of treatments increase. For example, consider the four networks presented in Figure 3. To estimate d_{CD} in the star network of evidence (Figure 3A), we would only be able to use the indirect evidence derived from $d_{CD}^{indirect} = d_{AD}^{direct} - d_{AC}^{direct}$. In Figure 3B, there is direct evidence on C vs. B, creating another pathway to derive indirect evidence, $d_{CD}^{indirect} = d_{AD}^{direct} - d_{AB}^{direct} - d_{BC}^{direct}$, which would have to be combined with $d_{CD}^{indirect} = d_{AD}^{direct} - d_{AC}^{direct}$. In Figure 3C, the network expands through additional comparisons involving treatments E and F. Indirect evidence on D vs. C may then be obtained by $d_{CD}^{indirect} = d_{ED}^{direct} - d_{EF}^{direct} - d_{FC}^{direct}$, which would have to be combined with the indirect evidence from the pathways mentioned previously. In Figure 3D, all treatments have been directly compared in head to head RCTs, creating multiple pathways to derive indirect evidence. As such, more advanced techniques are required to obtain indirect estimates.

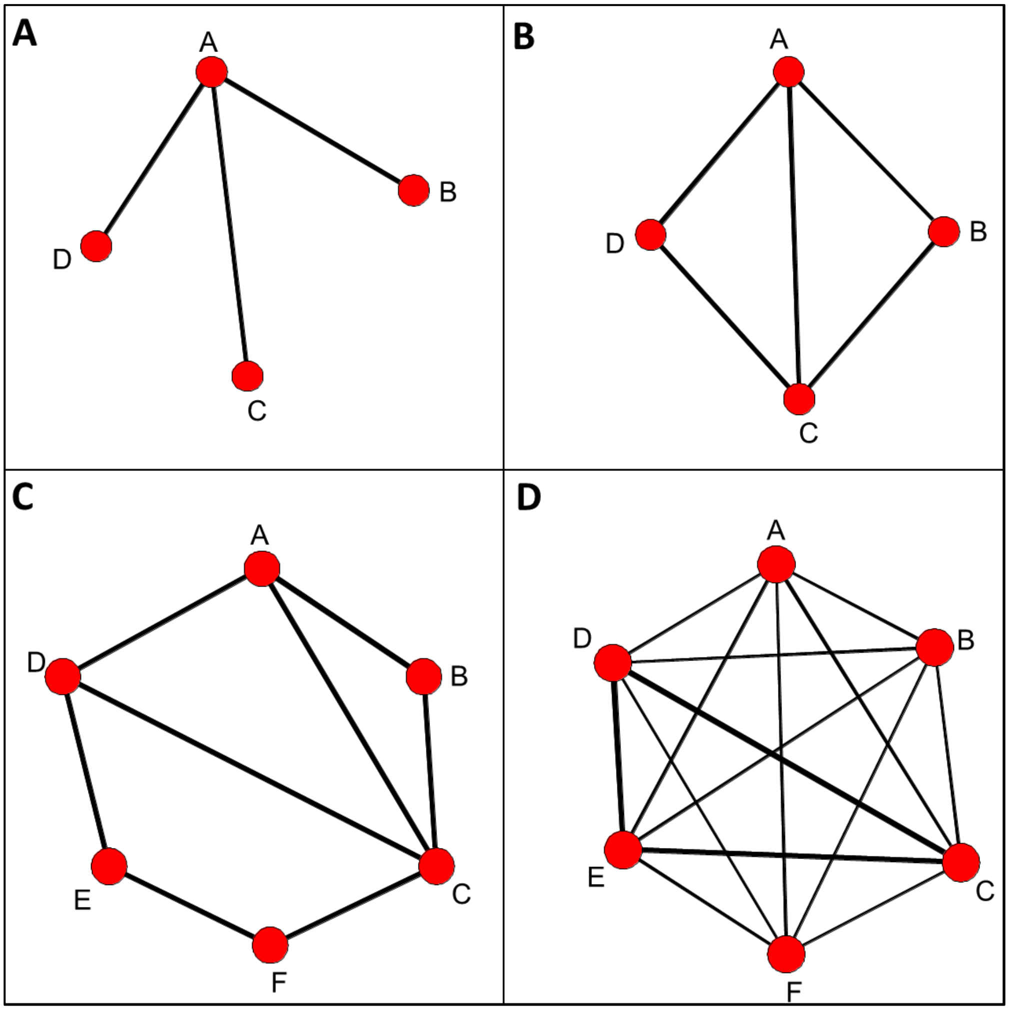


Figure 3: *Examples of networks of varying geometry. Nodes (circles) represent treatments while solid lines represent head to head comparisons made in randomized controlled trials.*

2.6.1 Statistical approaches

In addition to the Bucher method, there are four main approaches to conduct an NMA (Efthimiou et al., 2016). NMA may be conducted using meta-regression, where the relative effects are the dependent variables and the treatment comparisons are covariates in a regression model (Lumley, 2002). Another approach based on graph-theory and adopted from the analysis of electrical networks has been shown to produce identical estimates to the meta-regression approach (Rücker, 2012; Rücker & Schwarzer, 2014). Alternatively, the basic parameters may be modelled using a multivariate approach (White et al., 2012). Note these three approaches are commonly implemented in frequentist software that require the study-specific relative effects to be calculated before synthesis and thus make normal approximations. Bayesian hierarchical models can fit an exact likelihood to the data available on each treatment arm, thus negating the need for normal approximations (Lu & Ades, 2004). In addition, a review of 456 NMAs published between 1999 and 2015 noted Bayesian hierarchical models were the most commonly applied approach (Petropoulou et al., 2017), and ranking measures were initially proposed in this framework (Salanti et al., 2011). Thus, in this thesis, Bayesian hierarchical models are fitted and are described in more detail below.

2.6.1.1 Bayesian hierarchical model

Let θ_{ik} be a parameter characterizing an outcome in arm k of trial i. For example, if the observed data follow a binomial distribution,

$$r_{ik} \sim Binomial\left(\theta_{ik}, n_{ik}\right)$$

where r_{ik} and n_{ik} are the observed number of events and participants in arm k in trial i, respectively, θ_{ik} would then be the probability of an event among the population receiving treatment k. If the observed data follow a normal distribution,

$$\hat{\theta}_{ik} \sim Normal\left(\theta_{ik}, V_{ik}\right)$$

where $\hat{\theta}_{ik}$ and V_{ik} are the observed mean and variance of the outcome in arm k in trial i, respectively, θ_{ik} would then be the mean outcome in the population receiving treatment k (Dias et al., 2011).

Let $g\left(.\right)$ be an appropriate link function that transforms the parameter to a linear scale. For example, the logit link would transform a probability ranging between 0 and 1 to a log-odds ranging between -∞ and ∞. Then, the estimated parameter in each treatment arm may be modelled using a generalized linear model (Dias et al., 2011; Lu & Ades, 2004),

$$g\left(\theta_{ik}\right) = \mu_i + \delta_{ik}$$

where μ_i is the transformed outcome for the treatment group in arm 1 of trial i and is a nuisance parameter and δ_{ik} is the study-specific relative effect of treatment k vs. the treatment in arm 1 of trial i such that

$$\begin{array}{ll} \delta_{ik} = d_{t_{ik}} - d_{t_{i1}} & \text{fixed effect model} \\ \delta_{ik} \sim Normal\left(d_{t_{ik}} - d_{t_{i1}}, \tau^2_{t_{i1},t_{ik}}\right) & \text{random effects model} \end{array},$$

where t_{ik} is the treatment in arm k of trial i, d_j are the relative effects of treatment $j = 1,...,T$ compared to a reference treatment in a network of T treatments (i.e., the basic parameters). Finally, $\tau^2_{t_{i1},t_{ik}}$ is the between-study variance of the t_{ik} vs. t_{i1} relative effects.

There is rarely enough evidence in a network to estimate the between-study variance for each relative effect, and so a common between-study variance τ^2 is often assumed (Lu & Ades, 2009). The correlation between the $K-1$ multiple random effects, $\delta_{i2},...,\delta_{iK}$, in a K-arm trial is accounted for by assuming that the vector of random effects follows a multivariate normal distribution. When a common between-study heterogeneity is assumed (Lu & Ades, 2004), the multivariate normal distribution is

$$\boldsymbol{\delta}_{\mathbf{i}} = \begin{pmatrix} \delta_{i2} \\ \vdots \\ \delta_{iK} \end{pmatrix} \sim MVN\left(\begin{pmatrix} d_{t_{i2}} - d_{t_{i1}} \\ \vdots \\ d_{t_{iK}} - d_{t_{i1}} \end{pmatrix}, \begin{pmatrix} \tau^2 & \tau^2/2 & \cdots & \tau^2/2 \\ \tau^2/2 & \tau^2 & \cdots & \tau^2/2 \\ \vdots & \vdots & \ddots & \vdots \\ \tau^2/2 & \tau^2/2 & \cdots & \tau^2 \end{pmatrix} \right).$$

Sometimes trials will only report relative effects, rather than summary outcome data in each arm. In these trials, the relative effects on a continuous scale (e.g., mean difference or log-odds ratios) compared to the treatment in arm 1 may be modelled using a normal or multivariate normal likelihood, where the latter is used for multi-arm trials (Dias et al., 2018).

2.6.2 Assumptions

In NMA, it is assumed that all included trials are similar in terms of treatment effect modifiers (Efthimiou et al., 2016). All trials could have hypothetically compared every treatment in the network, estimating the same (fixed effects) or similar (random effects) true relative effects, but only a subset of the relative effects are reported by each trial and the others are missing at random (Dias et al., 2018; Lu & Ades, 2006). This assumption is formulated in the consistency equations described earlier.

Although trials are assumed to be similar in terms of treatment effect modifiers, the presence of treatment effect modifiers must be explored. This may be assessed through epidemiological means by comparing the distribution of known treatment effect modifiers across trials (Efthimiou et al., 2016). Alternatively, this may be assessed statistically by assessing the agreement between direct and indirect evidence, also known as consistency assessments (Dias et al., 2010a; Veroniki et al., 2014). Since large degrees of between-study heterogeneity may mask inconsistency, and differences in effect modifiers contribute to heterogeneity, it is important to assess the magnitude of heterogeneity as well (Efthimiou et al., 2016).

2.7 Treatment ranks

As described previously, when an NMA is implemented in a Bayesian framework, samples of the posterior distributions of the relative effects may be obtained using MCMC methods and compared iteratively. For example, in a network of 4 treatments A, B, C, and D, we may sample from the posterior distributions of the treatment effects relative to A. In one iteration, the sampled effects for A, B, C, and D may be 0, 0.5, 1, and 1.4 respectively. Note that the relative effects of A vs. itself will always be 0. Assuming that higher values are preferred, in this sample D ranked best, followed by C, then B, then A. Another iteration may sample 0, 0.25, -0.5, 0.1 as the relative effects for A, B, C, and D, respectively, in which case B ranked best, followed by D, then A, then C. This iterative process may be repeated to obtain a large number of samples, which provides a natural means to estimate ranking measures and their uncertainty.

2.7.1 Ranking probabilities

Consider a network of T treatments. Let $P(k,h)$ denote the probability that treatment $k=1,\ldots,T$ ranks h^{th} best, $h=1,\ldots,T$, where T is the worst rank, i.e., last. These ranking probabilities may be estimated as the proportion of MCMC samples where treatment k's sampled relative effect was the h^{th} best (Salanti et al., 2011). Note that these ranking probabilities will form a $T\times T$ matrix, where the marginal sums must add to 1. The ranking probabilities may be depicted graphically using rank-o-grams (Figure 4).

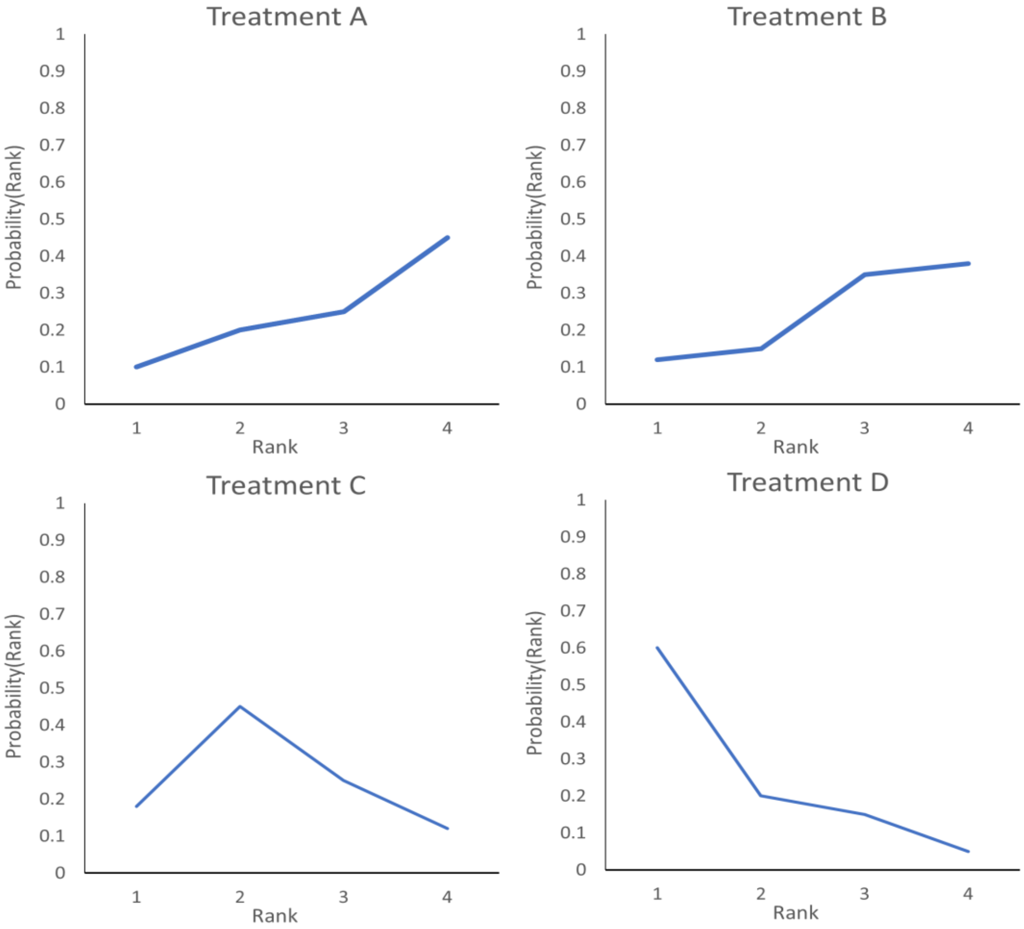


Figure 4: *Simulated example of rank-o-grams based on a network of four treatments.*

2.7.2 Surface under the cumulative ranking curve (SUCRA)

Alternatively, the cumulative ranking probabilities, $F\left(k,h\right)=\sum_{j=1}^{h}P\left(k,j\right)$, may be calculated and plotted (Figure 5). The motivation for doing so is to enable a comparison of the area or surface under the cumulative ranking curve (SUCRA) (Salanti et al., 2011). The larger the area, the greater the certainty that a treatment is the best. This may be numerically calculated as

$$SUCRA\left(k\right)=\frac{1}{T-1}\sum_{j=1}^{T-1}F\left(k,j\right).$$

Iteratively sampling from the posterior distributions of the relative effects and recording the ranks at each iteration will produce a posterior distribution for each treatment's rank. This posterior distribution may be summarized by its mean or median, as well as its percentiles to outline a range of credible values (i.e., credible intervals). SUCRA been shown to be an inversely scaled transformation of the mean rank (Rücker & Schwarzer, 2015),

$$SUCRA\left(k\right)=\frac{T-E\left(rank\left(k\right)\right)}{T-1}.$$

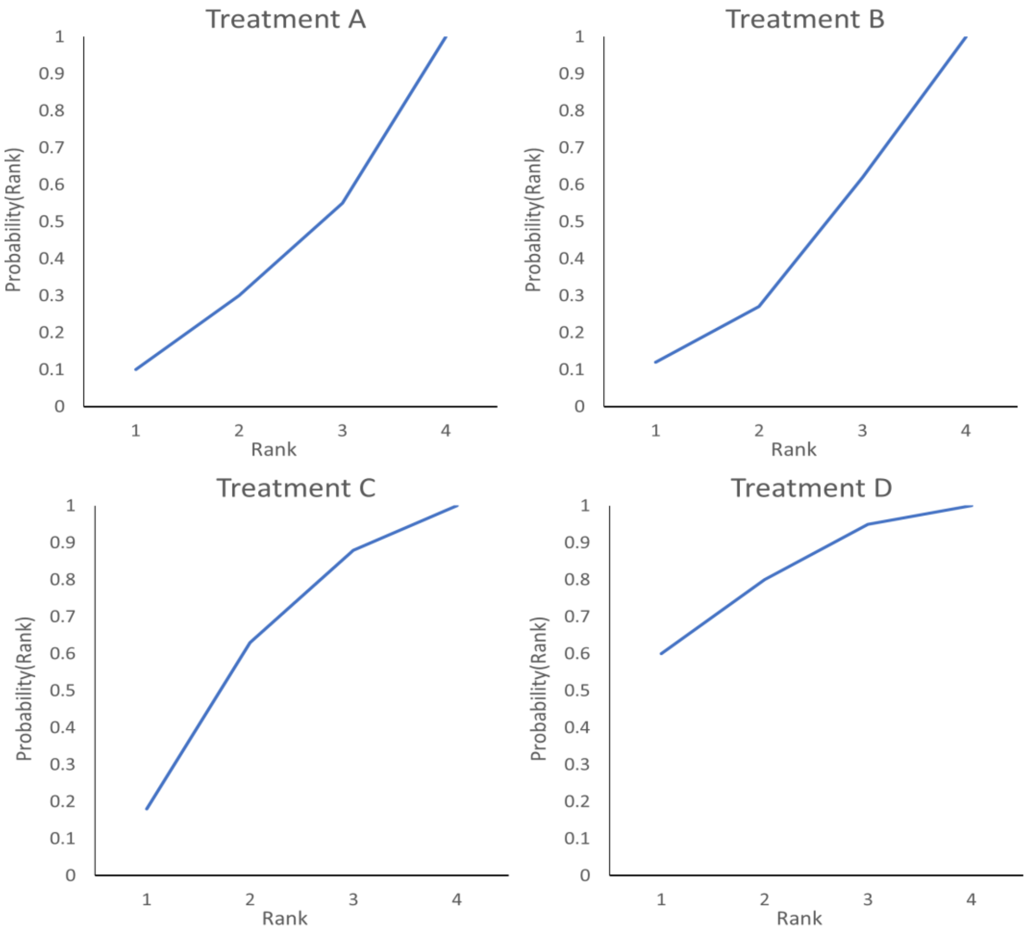


Figure 5: *Plots of cumulative ranking probabilities from simulated example presented in Figure 4.*

2.7.3 Ranking in a frequentist framework

Although we have presented ranking probabilities in a Bayesian context, ranking measures such as the probability of a treatment ranking best and SUCRA may be estimated through parametric bootstrapping of the relative effects, which are assumed to be normally distributed (White et al., 2012). Alternatively, the P-score may also be calculated in a frequentist framework without resampling techniques, and its estimates and interpretation are similar to that of SUCRA (Rücker & Schwarzer, 2015).

2.8 Summary

This chapter has provided an overview of the knowledge this thesis draws upon. Tools for synthesizing evidence on the comparative effectiveness and safety of interventions have been described. The next three chapters expand on the use of ranking probabilities and SUCRA for comparing treatments based on the results of an NMA conducted in a Bayesian framework.

Chapter 3

Empirical evaluation of SUCRA-based treatment ranks in network meta-analysis: quantifying robustness using Cohen’s kappa

Caitlin H Daly[1], Binod Neupane[1], Joseph Beyene[1,2], Lehana Thabane[1,3],
Sharon E Straus[4,5], Jemila S Hamid[1,6]

[1]Department of Health Research Methods, Evidence, and Impact, McMaster University, Hamilton, Ontario, Canada

[2]Department of Mathematics and Statistics, McMaster University, Hamilton, Ontario, Canada

[3]Biostatistics Unit, Father Sean O’Sullivan Research Centre, St Joseph’s Healthcare Hamilton, Hamilton, Ontario, Canada

[4]Knowledge Translation Program, Li Ka Shing Knowledge Institute, St. Michael’s Hospital, Toronto, Ontario, Canada

[5]Department of Medicine, University of Toronto, Toronto, Ontario, Canada

[6]Clinical Research Unit, Children’s Hospital of Eastern Ontario, Ottawa, Ontario, Canada

Citation: Daly, C. H., Neupane, B., Beyene, J., Thabane, L., Straus, S. E., & Hamid, J. S. (2019). Empirical evaluation of SUCRA-based treatment ranks in network meta-analysis: quantifying robustness using Cohen’s kappa. BMJ Open, 9(9), e024625. https://doi.org/10.1136/bmjopen-2018-024625

Abstract

Objective: To provide a framework for quantifying the robustness of treatment ranks based on Surface Under the Cumulative RAnking curve (SUCRA) in network meta-analysis (NMA) and investigating potential factors associated with lack of robustness.

Methods: We propose the use of Cohen's kappa to quantify the agreement between SUCRA-based treatment ranks estimated through NMA of a complete data set and a subset of it. We illustrate our approach using five published NMA data sets, where robustness was assessed by removing studies one at a time.

Results: Overall, SUCRA-based treatment ranks were robust to individual studies in the five data sets we considered. We observed more incidences of disagreement between ranks in the networks with larger numbers of treatments. Most treatments moved only one or two ranks up or down. The lowest quadratic weighted kappa estimate observed across all networks was in the network with the smallest number of treatments (4), where weighted kappa=40%. In the network with the largest number of treatments (12), the lowest observed quadratic weighted kappa=89%, reflecting a small shift in this network's treatment ranks overall. Preliminary observations suggest that a study's size, the number of studies making a treatment comparison, and the agreement of a study's estimated treatment effect(s) with those estimated by other studies making the same comparison(s) may explain the overall robustness of treatment ranks to studies.

Conclusions: Investigating robustness or sensitivity in an NMA may reveal outlying rank changes that are clinically or policy-relevant. Cohen's kappa is a useful measure that permits investigation into study characteristics that may explain varying sensitivity to individual studies. However, this study presents a framework as a proof of concept and further investigation is required to identify potential factors associated with the robustness of treatment ranks using more extensive empirical evaluations.

Strengths and limitations of this study

- To the best of our knowledge, robustness of Surface Under the Cumulative RAnking curve (SUCRA)-based treatment ranks has not been formally assessed in the literature, despite the controversy surrounding its use.
- The adoption of Cohen's kappa as a means to quantify the robustness of SUCRA-based treatment ranks to individual studies in network meta-analysis allows one to empirically investigate reasons for robustness.
- This is a proof-of-concept study; any observations made in the five illustrations are limited to these data sets and are mainly hypothesis-generating; more extensive empirical evaluation is needed to investigate reasons for robustness to studies.
- Simulation studies are ultimately needed to establish the validity and generalisability of the methodological framework to examine robustness of SUCRA-based treatment ranks.

Introduction

Network meta-analysis (NMA) simultaneously compares the efficacy or safety of three or more treatments by synthesising evidence directly and indirectly contributed by studies, including randomised controlled trials (RCTs).[1–3] This helps answer questions such as 'which treatment is best?' in addressing a clinical problem. Ideally, all studies providing information that will assist in answering a carefully defined research question will inform the NMA. A well-thought-out systematic review will aim to produce a collection of such studies.[4] This is done by identifying potentially relevant studies in an extensive literature search and vetting them against inclusion and exclusion criteria that have been designed to ensure the question of interest is being addressed by each study.

Despite the desire to provide a holistic body of evidence in attempt to determine a hierarchy of the efficacy or safety of all available treatments, individual studies within an NMA are understandably subjected to further scrutiny often in the form of risk-of-bias assessments.[5] Studies that considerably increase the between-study heterogeneity because of differences in treatment effect estimates beyond chance (eg, poor overlap of confidence intervals (CIs)) may also be flagged for further investigation.[6] It is not surprising then for those interpreting NMAs to raise concerns about the inclusion or contribution of a particular study or subset of studies to the pooled treatment effect estimates, even if they passed strict inclusion and exclusion criteria.

Identifying a sensible hierarchy of treatments based on the results of an NMA is not straightforward. The interpretation of several relative treatment effect estimates (eg, 6 in the case of 4 treatments and 45 in the case of 10 treatments) for each outcome can be overwhelming. To draw a knowledge user's attention to the most efficacious or safest treatments for a particular outcome, a ranking system for each outcome can be presented alongside the treatment effect estimates. In a Bayesian framework, ranks may be determined based on the mean or median of the posterior distribution of the ranks, the probability of a treatment ranking best or the Surface Under the Cumulative RAnking curve (SUCRA).[7–10] Alternatively, in a frequentist framework, ranks may be based on a measure similar to SUCRA, referred to as the P-score.[11] The probability of a treatment ranking best is appealing in terms of the ease of its interpretation, and a large value (eg, >0.90) may reflect that treatment is quite certain to be the most efficacious or safest. However, treatments that have large uncertainty around their estimated effects are more likely to have higher probabilities of ranking best.[10] When there is a lot of overlap and uncertainty in the treatment effects, this will be reflected across all ranking probabilities (ie, probability of ranking best, second best, etc), and SUCRA summarises this.[7 12] An overview of the characteristics of these different ranking measures is provided by Veroniki et al.[12]

Ranking treatments, in general, is not without controversy. For example, even if there are no clinically or statistically relevant differences between the efficacy of treatments, the difference in their ranks will imply there is one.[9 13] Recently proposed methodology explores how much an estimated treatment effect (in a study or synthesis of studies making the same comparison) must change to impact treatment recommendations.[14 15] Further, treatment ranks that are based on the probability of ranking best may be biased and influenced by the removal of treatments from an NMA.[16 17] In addition, the removal of a study can impact ranking probabilities and ranks based on the probability of ranking best.[18 19] Since ranking probabilities contribute to the calculation of SUCRA, which is often estimated with large uncertainty,[12] yet is increasingly being used in published NMAs,[20] it is of interest to examine the robustness of SUCRA-based treatment ranks and to quantify sensitivity with respect to evidence contributed by individual studies. It is also imperative to make knowledge users aware of factors in an NMA that may influence how well a relative effect may be estimated (eg, the structure of the network or heterogeneity between study estimates), which, in turn, impacts the treatment ranks.

To the best of our knowledge, no study has specifically looked at the robustness of SUCRA-based treatment ranks and quantified their sensitivity. Within published NMAs, it is not uncommon to find authors investigating the robustness of their conclusions regarding the hierarchy of treatments in general through subgroup or sensitivity analyses. They may then narratively compare the hierarchies in these additional analyses to the one produced in the base-case analysis. We aim to adopt this idea to investigate the robustness of SUCRA-based treatment ranks. This paper serves as a first step to do this. Here, we present a framework that makes use of an appropriate measure to quantify changes in treatment hierarchies (or ranks), which further enables a more rigorous investigation to understand why certain studies may impact conclusions made in an NMA. Our objectives are to (1) provide an objective measure to quantify robustness or sensitivity of SUCRA-based treatment ranks through Cohen's kappa and (2) illustrate how we may use the aforementioned measure to examine what features of the evidence might explain why the removal of some studies change the rank of treatments more than other studies.

Methods

Description of illustrative data

To illustrate our approach, we selected five NMAs from an internal collection of data extracted from published NMAs that reported the trial outcome data. Our proposed approach described below can only be applied to networks where outcome data on each treatment are provided by at least two studies. Of the 15 data sets available to us, 5 were excluded from consideration as they did not meet this requirement. We selected 5 of the remaining 10 NMA data sets because they contained the largest number of treatments and studies, and varied in terms of their network connectivity and size of information (eg, number of patients per treatment and number of RCTs per comparison) which we planned to investigate as potential reasons for variation in rank sensitivity. We refer to these data sets as the 'chronic obstructive pulmonary disease' ('COPD'),[21] 'depression',[22] 'diabetes',[23] 'heavy menstrual bleeding'[24] and 'stroke'[25] networks as these NMAs compare treatments addressing these medical conditions. Network diagrams and the SUCRA values for each treatment, produced using complete data, are shown in figure 1, while characteristics of the evidence within the networks are presented in online supplementary table S1.

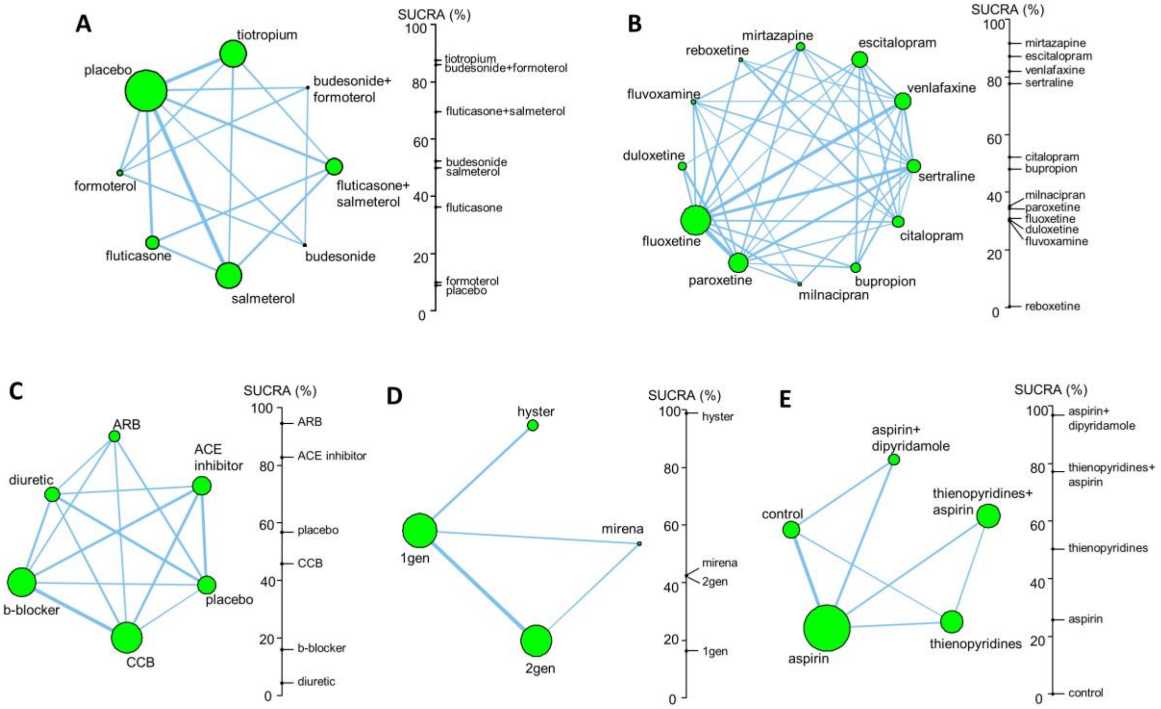


Figure 1 *Network diagrams (left) and SUCRA (right) for (A) chronic obstructive pulmonary disease network,[21] (B) depression network,[22] (C) diabetes network,[23] (D) heavy menstrual bleeding network and[24] (E) stroke network.[25] The sizes of the nodes are proportional to the number of patients randomised to the treatments, and the widths of the edges are proportional to the number of studies comparing two nodes. 1gen, 1st generation endometrial destruction; 2gen, 2nd generation endometrial destruction; ACE, angiotensin-converting-enzyme; ARB, angiotensin-receptor blockers; b-blocker, β blocker; CCB, calcium channel blocker; hyster, Hysterectomy; SUCRA, Surface Under the Cumulative RAnking curve.*

The COPD network consisted of evidence on 8 treatments from 39 RCTs, and it had the least direct evidence on all possible treatment comparisons (57.1% out of a total of 28 possible comparisons) (online supplementary table S1). Tiotropium was ranked the best treatment in this network based on SUCRA, followed closely by budesonide+formoterol (figure 1A). Despite containing evidence from the largest number of trials (111) comparing the largest number of treatments (12), the depression network had the second least number of patients (24 595) of the five networks (figure 1B, online supplementary table S1). The diabetes network, on the other hand, contained evidence from the largest number of patients (154 176) and most of the 15 possible comparisons between the 6 treatments were made in at least 1 trial, making it the most well-connected network (figure 1C, online supplementary table S1). The heavy menstrual bleeding network is the smallest of the five networks in terms of number of treatments (4), RCTs (20, 2-arm only) and patients (2886) (figure 1D, online supplementary table S1). The stroke network had the second smallest number of treatments (5), but had the second largest number of patients (55 463). All direct comparisons were made in at least two RCTs in the stroke network, and the ranking of treatments based on their SUCRA values is well-established, as exemplified by the distance between them (figure 1E, online supplementary table S1).

Empirical evaluation

For each data set, we selected and proceeded with a model that was appropriate for the data type, as our purpose was to use the networks for illustration and not for clinical interpretation or generalisability. For the interested reader, we have included details on the selected model, model fit statistics and results of inconsistency checks in online supplementary table S2.

An NMA was initially conducted with all studies included and the ranks of treatments based on the SUCRA results of this NMA were recorded. Sensitivity analyses were subsequently conducted, where for each sensitivity analysis, a single study was removed, an NMA was conducted based on the data set excluding this single study, and the SUCRA-based treatment ranks were documented. This was repeated for all studies, removing them one at a time. This procedure is similar to those used in influence analysis in regression, where the influence of an observation on a regression model is investigated through comparison of regression models fitted with and without the observation in the data set.[26] The motivation for this was to enable exploratory analysis, provided there is sufficient variability in the impact of trials and potential explanatory variables of interest (eg, number of patients).

For each NMA, the analysis was performed in a Bayesian framework using the gemtc package (V.0.8-2) in R.[27 28] Vague priors were used for all model parameters (Normal(0, 10 000) for baseline and treatment effects, and Uniform(0, 5) for common between-study SD). Results were based on 100000 samples with a thinning rate of 10 after an adaption phase of 20000 samples in each of three chains of Markov chain Monte Carlo simulations. Convergence was assessed using trace plots as well as the Gelman-Brooks-Rubin diagnostic test.[29 30]

We ranked treatments based on their SUCRA values.[7] To calculate SUCRA in a Bayesian framework, the ranking probabilities, $P(i, j)$ — the probability that treatment i ranks j^{th} best for a particular outcome — were calculated for each treatment. The cumulative distribution function of a treatment's ranking probabilities — the probability that treatment i ranks k^{th} best or better — was subsequently calculated as

$$F(i,k) = \sum_{j=1}^{k} P(i, j).$$

The SUCRA value for treatment i was then taken to be the surface under the curve defined by this cumulative distribution function. Mathematically, it was calculated as

$$SUCRA(i) = \frac{\sum_{k=1}^{n-1} F(i,k)}{n-1}$$

where n was the number of treatments in the network. The treatment with the largest SUCRA value was ranked the best, the treatment with the second-largest SUCRA value was ranked second best, and so on, such that the treatment with the smallest SUCRA value was ranked n^{th} (the worst) for the outcome.

Quantifying, presenting and elucidating robustness of treatment ranks

To quantify the influence a study had on all SUCRA-based treatment ranks, we used Cohen's kappa,[31] which measured the agreement between the treatment ranks produced with the complete data and the ranks produced when a study was removed. We use the term robustness in reference to the sensitivity of the treatment ranks with respect to individual studies, indicated by departure from the ranks produced with the complete data. The kappa statistic offers flexibility in the assessment of the robustness or sensitivity of treatment ranks in the sense that different weighting schemes will allow one to focus on different questions regarding the difference in treatment ranks. For example, the unweighted (simple) kappa separated studies based on the number of treatments that changed rank and considered any change to be the same, regardless of the size of the rank displacement. In this sense, unweighted kappa provides information similar to the percentage of treatments whose rank remains unchanged and serves as an overall indicator of rank robustness or sensitivity. A more appropriate weighting scheme may be quadratic weights,[32] where the weights between two disagreeing ranks are differences between the original and new ranks squared (eg, if a treatment's rank changed from 2 to 5, the corresponding disagreement weight would be $(2-5)^2$). This would distinguish, for example, the importance of a change in treatment rank of two places from a change in treatment rank by one place. The quadratic weighted kappa is equivalent to Pearson's correlation coefficient applied to the SUCRA-based ranks, as well as Spearman's rank correlation if it was applied to the SUCRA-based ranks or SUCRA values themselves.[32] Other weighting schemes may incorporate distances in SUCRA, or may be designed to reflect changes in the top three ranks. However, in this paper, we are providing a general framework for the proposed approach, and we illustrate it by holding interest in changes across all treatment ranks, quantified by kappa with no weights and quadratic weights.

In order to investigate rank robustness or sensitivity with respect to study characteristics, we compared the distributions of study characteristics between groups of studies with a similar impact on treatment ranks via density plots and descriptive statistics. In particular, we looked at characteristics that may highlight the contribution of studies to the direct evidence in the network. A study's contribution to a network is a factor of its own characteristics, as well as those of other studies included in the network. In a frequentist setting, a study's contribution has been previously summarised as a single quantity.[33 34] The contribution of evidence within a direct comparison to an NMA has also been quantified.[35] However, given the limited information available on the study characteristics in the selected publicly available data sets, we explored only trial size (ie, total subjects) and the amount of information available (ie, number of studies) on treatment comparisons. In this empirical evaluation, we initially considered these characteristics using univariate analysis. Since both networks contained multi-arm RCTs, we considered the number of studies per treatment comparison, N_s, for a given trial s, under two scenarios: (1) as an average across all comparisons made by a single trial s:

$$N_s = \frac{1}{n_s}\sum_{i=1}^{n_s}\left(\text{number of RCTs that made direct comparison } i\right)$$

where $n_s = \frac{k_s\left(k_s - 1\right)}{2}$ is the number of unique direct comparisons within a k-arm RCT s and (2) at a comparison level. In the latter scenario, multi-arm RCTs had multiple values characterising the number of studies that made each comparison, whereas two-arm RCTs had only one value. Finally, we considered the change in between-study variance after the removal of each study. This characterised the heterogeneity between the treatment effect(s) estimated by the removed study and those estimated in other studies making the same comparison(s). A large relative change would suggest a large difference in the treatment effect(s) observed in a particular study, compared with the treatment effect(s) observed in other studies.

As the rank of a specific (eg, locally available or cheaper) treatment may be of interest to knowledge users, we also explored how often and how much each of the treatments' ranks changed after the removal of a study. We quantified robustness of a treatment's rank by the proportion of studies whose removal resulted in a change in its rank and compared it with the width of the 95% credible interval (CrI) for its rank. This was done to assess the relationship between the uncertainty and robustness of a treatment's rank, the former of which is a cause of recent concern.[9] To calculate the 95% CrI for each rank, we made use of the relationship between SUCRA and the expected rank ($\bar{r}$)[11]:

$$SUCRA = \frac{n-\bar{r}}{n-1}$$

where n is the number of treatments in the network. In our illustrative examples, we computed the 95% CrIs for SUCRA based on the 2.5th and 97.5th percentiles of the posterior distribution of SUCRA. We then transformed the CrIs for each SUCRA $\left(LL_s, UL_s\right)$ into CrIs for the expected rank $\left(LL_r, UL_r\right)$ using this relation:

$$\left(LL_r, UL_r\right) = \left(n-\left(n-1\right)LL_s, n-\left(n-1\right)UL_s\right).$$

Patient and public involvement

Patients and the public were not involved in this study.

Results

Apart from the depression network, the majority of RCTs within each network did not individually impact the SUCRA-based treatment ranks (table 1). In the stroke network, the removal of an individual RCT did not impact any of the SUCRA-based treatment ranks across all RCTs, and thus the observed agreement beyond chance was universally perfect in this network (unweighted kappa (κ_{UW}) = weighted kappa (κ_W = 1)). The smallest beyond chance agreement was observed in the heavy menstrual bleeding network (κ_{UW} = 0%). In this case, the removal of an RCT displaced three of the four treatments' ranks, and the corresponding weighted agreement, where the importance of disagreement increases as the change in rank increases, was κ_W = 40%.

The largest absolute change in a treatment's rank after the removal of an RCT was observed in the depression network (table 1). In one instance, the removal of one RCT resulted in milnacipran and fluvoxamine exchanging ranks. In the complete data set, they had ranked 7th and 11th best, respectively, and so each treatments' rank changed by 4 places. The observed agreement beyond chance between the ranks based on the complete data set and subset of data with this RCT removed was κ_{UW} = 82%. This observed agreement is equal to cases in the depression network where the removal of an RCT resulted in two treatments exchanging neighbouring ranks (eg, seventh and eighth), highlighting an important change that the unweighted agreement measure does not capture. This illustrates the usefulness of the weighted agreement measure in terms of distinguishing the qualitatively different impacts of RCTs. In the former situation, the weighted agreement was κ_W = 89%, while in the latter situation, the weighted agreement was κ_W = 99%.

Table 1 *Summary of impact of individual studies on SUCRA-based treatment ranks in five NMAs.*

	Number of trials that did not change any treatment rank (%)	Among studies that impacted treatment's ranks…						
		Median of Observed kappas (κ) (Minimum, Maximum)			Number (%) of Studies Displacing Treatment's Rank By…			
Network		Unweighted	Weighted	Treatment*	1 Rank	2 Ranks	3 Ranks	4 Ranks
COPD[21]	30 (76.9%)	71% (14%, 71%)	98% (81%, 98%)	tiotropium	3 (33%)	0 (0%)	0 (0%)	0 (0%)
				budesonide+ formoterol	3 (33%)	0 (0%)	0 (0%)	0 (0%)
				fluticasone+ salmeterol	1 (11%)	0 (0%)	0 (0%)	0 (0%)
				budesonide	4 (44%)	0 (0%)	2 (22%)	0 (0%)
				salmeterol	5 (56%)	0 (0%)	0 (0%)	0 (0%)
				fluticasone	2 (22%)	0 (0%)	0 (0%)	0 (0%)
				formoterol	4 (44%)	0 (0%)	0 (0%)	0 (0%)
				placebo	2 (22%)	1 (11%)	0 (0%)	0 (0%)
Depression[22]	36 (32.4%)	82% (45%, 82%)	99% (89%, 99%)	mirtazapine	1 (1%)	0 (0%)	0 (0%)	0 (0%)
				escitalopram	2 (3%)	0 (0%)	0 (0%)	0 (0%)
				venlafaxine	2 (3%)	0 (0%)	0 (0%)	0 (0%)
				sertraline	1 (1%)	0 (0%)	0 (0%)	0 (0%)
				citalopram	3 (4%)	0 (0%)	0 (0%)	0 (0%)
				bupropion	3 (4%)	0 (0%)	0 (0%)	0 (0%)
				milnacipran	13 (17%)	1 (1%)	1 (1%)	1 (1%)
				paroxetine	19 (25%)	0 (0%)	0 (0%)	0 (0%)
				fluoxetine	37 (49%)	21 (28%)	0 (0%)	0 (0%)
				duloxetine	61 (81%)	3 (4%)	2 (3%)	0 (0%)
				fluvoxamine	30 (40%)	7 (9%)	1 (1%)	2 (3%)
				reboxetine	0 (0%)	0 (0%)	0 (0%)	0 (0%)
Diabetes[23]	21 (95.5%)	60% (60%, 60%)	94% (94%, 94%)	ARB	0 (0%)	0 (0%)	0 (0%)	0 (0%)
				ACE-inhibitor	0 (0%)	0 (0%)	0 (0%)	0 (0%)
				placebo	1 (100%)	0 (0%)	0 (0%)	0 (0%)
				CCB	1 (100%)	0 (0%)	0 (0%)	0 (0%)
				b-blocker	0 (0%)	0 (0%)	0 (0%)	0 (0%)
				diuretic	0 (0%)	0 (0%)	0 (0%)	0 (0%)
Heavy menstrual bleeding[24]	16 (80%)	33% (0%, 33%)	80% (40%, 80%)	hyster	0 (0%)	0 (0%)	0 (0%)	0 (0%)
				mirena	3 (75%)	1 (0%)	0 (0%)	0 (0%)
				2gen	4 (100%)	0 (0%)	0 (0%)	0 (0%)
				1gen	1 (25%)	0 (0%)	0 (0%)	0 (0%)
Stroke[25]	26 (100%)	N/A	N/A	aspirin+dipyrida-mole	N/A			
				thieno-pyridines+ aspirin				
				thieno-pyridines				
				aspirin				
				control				

[*] Treatments listed in order of SUCRA-based rank computed from NMA of complete data set.

COPD, chronic obstructive pulmonary disease; N/A, Not applicable; NMA, network meta-analysis; SUCRA, Surface Under the Cumulative RAnking curve.

In most cases, when the removal of an RCT impacted the treatment ranks, treatments exchanged ranks with a neighbouring treatment (eg, tiotropium and budesonide+formoterol in the COPD network (figure 1A)). Changes between neighbouring treatments' ranks are more common between treatments with small differences in SUCRA, compared with treatments that have larger differences between their SUCRA values. For example, in the depression network, milnacipran and paroxetine have SUCRA values of 35.2% and 34.3%, respectively, and fluoxetine, duloxetine and fluvoxamine have SUCRA values of 30.9%, 30.5% and 30.0%, respectively (figure 1B). These treatments' ranks changed by one place after the removal of a relatively higher number of RCTs, compared with other treatments in the network (table 1). The treatment ranking best according to SUCRA in the diabetes, heavy menstrual bleeding and stroke networks was never affected by the removal of an RCT in each network (table 1), and we note the considerable difference between the SUCRA values between the best and second best ranking treatments in all three networks (figure 1C–E).

Since there was substantial variability in the impact of RCTs to the SUCRA-based treatment ranks in COPD and depression networks, compared with the other networks, we explored potential reasons to explain why some RCTs in these networks impacted treatment ranks more than others.

Results of further investigation into ranks in the COPD NMA

The largest changes in rank were observed for two RCTs, identified as study 13 (κ_W = 83%) and study 18 (κ_W = 81%) (figure 2). Both of these studies compared four treatments:

budesonide, formoterol, budesonide–formoterol and placebo (online supplementary table S3). The removal of study 13 resulted in budesonide, originally ranked fourth best, to become the best, while the removal of study 18 resulted in the same treatment ranking seventh (just better than the worst treatment). Apart from study 10, the remaining RCTs for which changes in treatment rank were observed had the same level of weighted rank agreement (figure 2).

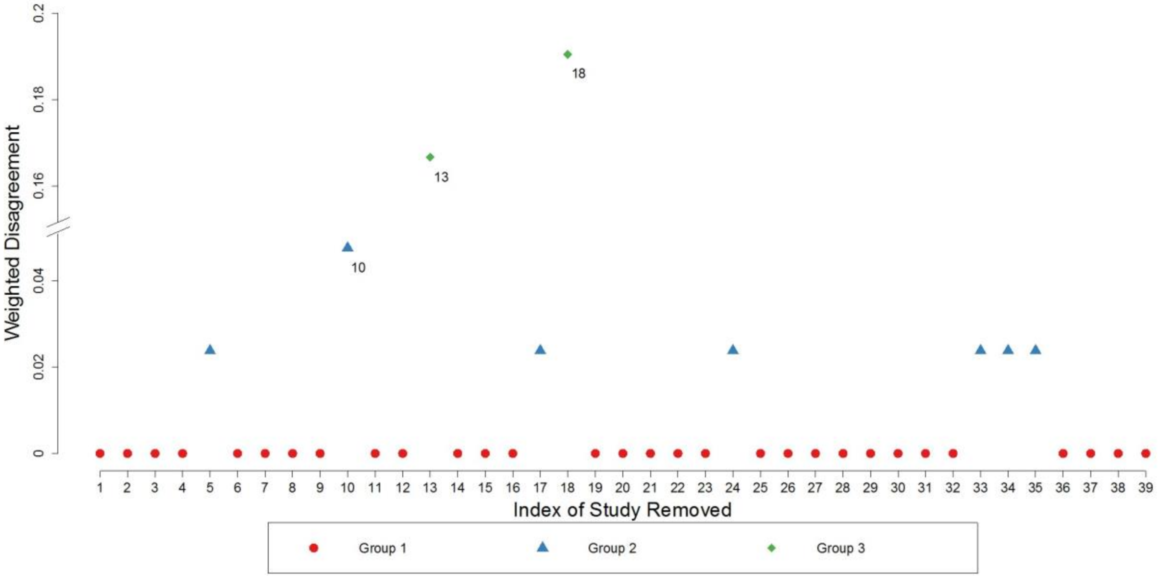
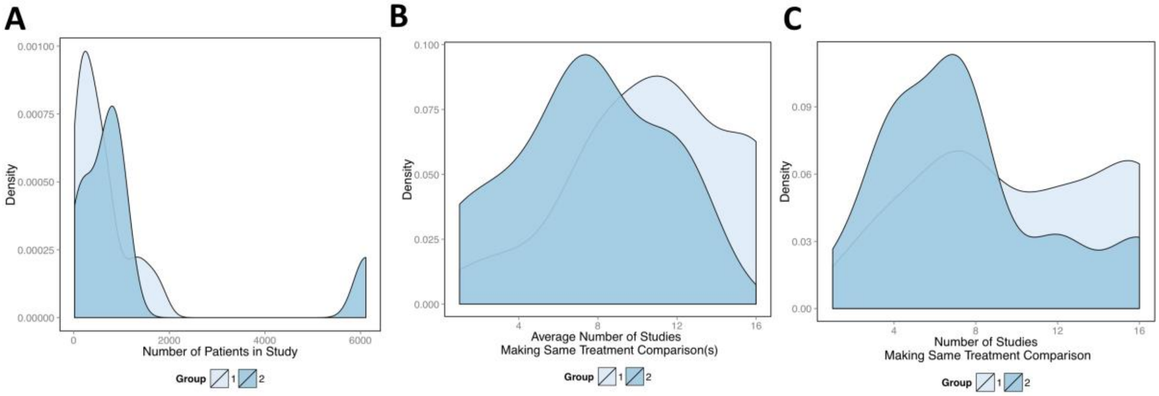


Figure 2 *Observed 1—quadratic weighted kappa (κ_W) in the chronic obstructive pulmonary disease network,*[21] *which quantifies the weighted disagreement between the treatment ranks produced from the complete data set and the ranks produced from a sub-data set where one RCT (indexed on the x-axis) was removed. Studies are grouped by a similar impact on rankings, as indicated by markers described in the legend, for further investigation. RCT, randomised controlled trial.*

Excluding outlying studies 13 and 18, we examined and compared the number of patients in RCTs that changed treatment ranks (Group 2 of figure 2) and those that did not (Group 1 of figure 2). The group of RCTs whose removal did not result in a change in treatment ranks included two clusters of studies (Group 1 of figure 3A), one containing some of the smallest numbers of patients in the network, and the other containing relatively larger numbers of patients. The size of the RCTs that displaced treatment ranks fell into three clusters; the first of which also include some of the smallest numbers of patients in the network, but most of these RCTs contained relatively larger numbers of patients (Group 2 of figure 3A). There was also an exceptionally large RCT that shifted treatment ranks. Compared with the RCTs that did not change treatment ranks, there was a slight shift in the distribution of the size of studies that impacted ranks, indicating that they tended to be larger than the majority of RCTs that did not impact ranks (figure 3A). In terms of the average number of RCTs per comparison, there is a shift in the mode of the distributions between the two groups, suggesting RCTs that displaced treatment ranks tended to make less common comparisons on average (figure 3B). At a comparison level, RCTs that changed treatment ranks were more often than not making infrequently studied treatment comparisons (Group 2 of figure 3C). However, the bimodal distribution belonging to the group of RCTs that did not change ranks is mostly, in part, driven by multi-arm RCTs that made common comparisons, as well as uncommon comparisons (Group 1 of figure 3C).



Group	Median Number of Patients (Min – Max)	Group	Median Information Available for Comparisons, Averaged across All Comparisons (Min – Max)	Group	Median Information Available for Comparisons, by Comparison (Min – Max)
1	420 (16 – 1829)	1	12 (1 – 16)	1	10 (1 – 16)
2	723 (139 – 6112)	2	7.7 (1 – 12)	2	7 (1 – 16)

Figure 3 *Study characteristics between two groupings of studies in chronic obstructive pulmonary disease network[21]: Group 1, where the individual removal of these RCTs had no impact on treatment rankings ($\kappa_W = 1$), and group 2, where the individual removal of these RCTs had a small impact on treatment rankings (0.95 < κ_W < 0.98) (identified in figure 2). Density plots, as well as descriptive statistics, of (A) the number of patients within studies, (B) information available for comparisons (ie, number of studies in the network making each comparison), averaged across all comparisons made within a study, and (C) information available for comparisons across all comparisons made by a study, are displayed. RCTs, randomised controlled trials.*

Further investigation as to why studies 13 and 18 produced extreme rank changes revealed that these four-arm RCTs provide the only direct evidence on five out of the six possible comparisons between four treatments (budesonide, formoterol, budesonide–formoterol and placebo) in the network. Furthermore, these studies provided conflicting evidence on the placebo versus budesonide comparison (study 13: OR (95%CI) = 0.81 (0.57 to 1.16); study 18: OR (95%CI) = 2.31 (1.37 to 3.87)). This conflicting evidence drives the magnitude of the between-study variance, as the between-study variance decreased after the removal of each of these two RCTs, and the magnitude of the change in between-study variance was much larger for study 13, compared with the changes observed after the removal of all other RCTs. In addition, the conflicting evidence is reflected by the large uncertainty in budesonide's rank. Its 95% CrI, 1–8, indicates that there is a 95% probability that budesonide's rank in terms of reducing exacerbations could be as high as 1 (ie, best treatment in terms of efficacy) or as low as 8 (ie, worst treatment). Based on the limited number of treatments and hence datapoints, we were not able to conclude the existence or non-existence of a relationship between the CrIs for each treatment's rank and the number of RCTs that impacted their rankings (online supplementary figure S1).

Results of further investigation into ranks in the depression NMA

Among the 75 RCTs whose removal resulted in a change in treatment ranks, 41 (54.7%) only affected the ranks of 2 treatments, hence the high value of weighted kappa estimates (κ_W), suggesting very good weighted agreement among the ranks (median κ_W = 99% (minimum 89%, maximum 99%)) (figure 4). Only the removal of two RCTs, studies 31 and 55, resulted in the observed maximum rank change of four places (table 1, figure 4). For instance, the removal of study 31 resulted in the exchange of ranks between milnacipran (7th best) and fluvoxamine (11th best). This study provided the only direct comparison of these two agents (online supplementary table S4), suggesting that sparseness of a network may influence the robustness of SUCRA-based ranks. The removal of study 55 resulted in fluvoxamine's rank increasing from 11th best to 7th best, and three other treatments' ranks subsequently decreased by one or two ranks. Study 55 provided the only direct evidence between fluvoxamine and venlafaxine, but venlafaxine's rank was unaffected by the removal of this RCT. Although these RCTs had the largest impact on treatment ranks, the change in between-study variance was minimal in comparison to the changes observed after the removal of other RCTs.

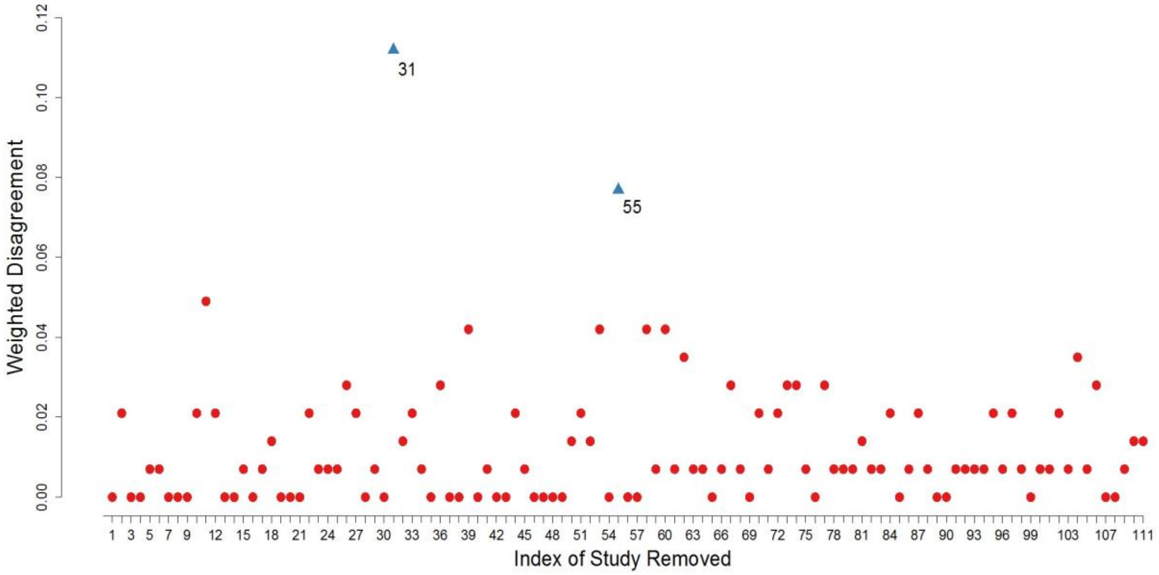


Figure 4 *Observed 1—quadratic weighted kappa (κ_W) in the depression network,[22] which quantifies the weighted disagreement between the treatment ranks produced from the complete data set and the ranks produced from a sub-data set where one RCT (indexed on the x-axis) was removed. RCT, randomised controlled trial.*

Finally, we investigated the relationship between the robustness of individual treatment ranks with their precision as measured by the width of the 95% CrIs when the SUCRA-based ranks were calculated using the complete data set. Similar to the COPD NMA, the small number of datapoints did not reveal any conclusive relationship (online supplementary figure S2).

Discussion

This study proposes a novel approach for quantifying robustness or sensitivity of treatment ranks using Cohen's kappa in NMA. We illustrated the approach using five publicly available NMAs and the results show that SUCRA-based ranks in most of these NMAs are in general robust with respect to the exclusion of individual studies. However, we have observed even a single study can change the pooled evidence enough (ie, relative effects) to influence SUCRA-based treatment ranks. When this occurs, this should serve as a flag for further investigation as to whether the change is important enough to impact how confident a knowledge user may be in terms of the hierarchy of the efficacy or safety of treatments. As such, rigorous scrutiny of such studies is important when conducting an NMA; this might be particularly crucial in a sparse network where direct evidence on some treatment comparisons is limited. Note that the results and conclusions drawn from the five networks are for illustrative purposes only and are not intended for clinical interpretation and use. The observations made regarding the robustness of the treatment ranks are limited to the five networks evaluated and may not be true for all networks.

Most changes in treatment ranks were observed between treatments in close proximity of each other's SUCRA values. SUCRA summarises the relative strength and precision of the estimated treatment effects, and similar SUCRA values might truly reflect treatments that are equally efficacious (or safe), where the small differences observed might be because of random error. On the other hand, similar SUCRA values might reflect true but small (and sometimes clinically important) differences in the efficacy or safety between the treatments. This highlights why it is important to interpret treatment ranks alongside point estimates and confidence or credible intervals of relative effects, to assess the relevance of any differences between treatments.[36] In terms of investigating SUCRA-based treatment ranks using Cohen's kappa, a weighting scheme that incorporates differences in SUCRA would highlight studies that have a meaningful impact on SUCRA-based treatment ranks. Alternatively, a different ranking measure may be used to distinguish the relative efficacy of these treatments and should be investigated in future work.

The use of weighted kappa to quantify rank sensitivity as opposed to other rank agreement measures offers the advantage of incorporating a weighting scheme that distinguishes trials or subgroups based on the importance of their influence. For example, if investigators were only concerned about changes in the top-ranked treatments, a weighted kappa that gives more weight to disagreements among the top three ranks may highlight which trials impact the top-ranked treatments more than others. A weighting scheme could also incorporate changes in relative effects within treatments, or differences in relative effects between treatments, to reflect clinically important changes. Nevertheless, one may explore other agreement measures to assess the robustness of ranks, including the prevalence-adjusted bias-adjusted kappa.[37] In addition, while the motivation for using kappa or other agreement measures is to quantify the robustness of SUCRA-based treatment ranks, users may be able to accompany the agreement measures with CIs or assess their significance using established tests (eg,[32]) provided the sample size, which is equal to the number of treatments in the network, is sufficient.

In practice, we note that knowledge users are encouraged to examine the uncertainty of the SUCRA values from NMA through their CrIs to assess whether there is a relevant difference in the efficacy and safety of treatments.[9] In the COPD and depression networks we explored, we were not able to make any conclusions regarding a relationship between the width of CrIs for the ranks (from which CrI of SUCRA may be derived through a transformation)[11] and robustness or sensitivity of the treatment's rank. More extensive empirical evaluation, as well as simulation studies, is required to explore this further and establish a relationship, if any.

Investigation of the robustness or sensitivity of treatment ranks with respect to study characteristics is possible by identifying clusters of studies with similar kappa values. For example, further investigation of five distinct groups (clusters) of RCTs in terms of unweighted kappa in the depression network (online supplementary figure S3) revealed the median number of patients in these clusters increased as the unweighted rank agreement decreased (online supplementary figure S4A). However, there was no association between the amount of information available on treatment comparisons and rank agreement measured by unweighted kappa (online supplementary figure S4B,C). Due to limited study-level data from the selected publicly available NMAs, further exploration of identified clusters was not possible in this manuscript, but is of interest in a more rich data set.

A study's size and the number of trials comparing the treatments it compared could be offered as explanations for a studies' outlying impact on treatment ranks. Heterogeneity between evidence provided by studies within direct comparisons or between direct and indirect evidence (ie, inconsistency) might also explain why some studies are more influential in networks than others, leading to rank differences/disagreements. Exploring the robustness of treatments ranks may, thus, help to pinpoint the sources (ie, studies) of important heterogeneity or inconsistency. However, small sample size or small number of studies per treatment comparison may mask potentially important heterogeneity between studies, which would be reflected in the overlap of wide confidence or credible intervals of treatment effects estimated by individual studies. Furthermore, sparseness in the network, that is, a single study or no sources of direct evidence on many treatment comparisons, would limit evaluation of heterogeneity. This gives credibility to a common criticism of NMA, that knowledge users should interpret the results for the treatment comparisons with little direct evidence with caution. Thus, a combination of these factors may explain why some studies influence treatment ranks more than others, and should be considered together in a multivariate setting.[33] [34] However, an investigation into clinical characteristics of a study or its quality (eg, patient population, treatment administration and risk of bias) may be more informative and helpful to knowledge users concerned about the potential influence of studies.

Bootstrapping techniques could serve as an option for assessing the robustness of NMA results, but this could lead to disconnected subnetworks in some bootstrap samples. Moreover, leaving one or more studies out follows the practice of sensitivity or subgroup analyses commonly employed in NMA, and quantifying changes in rank with kappa provides an objective summary of robustness. We would like to highlight that the approach used in this paper is not meant to identify outlying studies that should be excluded from an NMA. This approach follows the same principles that guide influence analysis across a variety of modelling situations. Outlying observations may or may not impact the model, whereas influential observations do.[38] In deviance-based analyses, if there is a concern regarding the contribution of a particular datapoint, a common practice is to present the results with and without the datapoint. It is then up to the knowledge users to decide which data set is most representative of the problem at hand. For example, in the context of NMA, if a study includes a population that was not included in other studies and is not relevant to the research question, a knowledge user may choose to interpret the results without that particular study. Alternatively, provided there is enough information available, meta-regression may be used to adjust a study's contribution to an NMA based on a known effect modifier. At a minimum, investigating studies' influence on the treatment ranks may highlight studies that require a secondary check against the inclusion and exclusion criteria, or for data extraction errors.

Finally, we note that the magnitude of Cohen's kappa is often categorised into levels of agreement for interpretation (eg, poor (<0%), slight (0%–20%), fair (21%–40%), moderate (41%–60%), substantial (61%–80%) and almost perfect (81%–100%) agreement).[39] This is an ad-hoc procedure and ultimately depends on the context of the area it is applied to. Knowledge users should carefully consider whether a kappa value of 90%, for example, is indeed indicative of almost perfect rank agreement based on their expertise in the clinical area.

Conclusion

Motivated by the concerns surrounding the stability of treatment ranks in NMA, this study provides a framework for investigating the robustness of SUCRA-based treatment ranks and reasons for varying sensitivity to individual studies in NMAs. It lays the groundwork for quantifying, visualising and elucidating the robustness or sensitivity of SUCRA-based treatment ranks with respect to direct evidence provided in individual studies. Similar to deviance-based analyses done to investigate outlying studies, we recommend that future NMAs should include sensitivity analyses to assist knowledge users in assessing the robustness of treatment ranks to individual studies. This will also help knowledge users to understand how the robustness of treatment ranks may depend on the contribution and features of the studies making up the network. The approach described in this paper will draw a knowledge user's attention to a study or groups of studies that have undue influence on the treatment ranks, which may prompt them to adjust the ranks, if certain aspects of the studies makes it necessary to do so (eg, because of an inclusion of a poorly conducted study, or large uncertainty in evidence resulting from very heterogeneous results).

References

1. 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.

2. Higgins JPT, Welton NJ. Network meta-analysis: a norm for comparative effectiveness? Lancet 2015;386:628–30.

3. Efthimiou O, Debray TPA, van Valkenhoef G, et al. GetReal in network meta-analysis: a review of the methodology. Res Synth Methods 2016;7:236–63.

4. Higgins JPT, Green S. Cochrane handbook for systematic reviews of interventions. Version 5.1.0. The Cochrane Collaboration, 2011.

5. Higgins JPT, Altman DG, Gøtzsche PC, et al. The Cochrane collaboration's tool for assessing risk of bias in randomised trials. BMJ 2011;343:d5928.

6. Chan L, Macdonald ME, Carnevale FA, et al. Reconciling disparate data to determine the right answer: a grounded theory of meta analysts' reasoning in meta-analysis. Res Synth Methods 2018;9:25–40.

7. Salanti G, Ades AE, Ioannidis JPA. Graphical methods and numerical summaries for presenting results from multiple-treatment metaanalysis: an overview and tutorial. J Clin Epidemiol 2011;64:163–71.

8. Trinquart L, Abbé A, Ravaud P. Impact of reporting bias in network meta-analysis of antidepressant placebo-controlled trials. PLoS One 2012;7:e35219.

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;164:666–73.

10. 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 Health 2014;17:157–73.

11. Rücker G, Schwarzer G. Ranking treatments in frequentist network meta-analysis works without resampling methods. BMC Med Res Methodol 2015;15:28.

12. 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.

13. Mills EJ, Ioannidis JPA, Thorlund K, et al. How to use an article reporting a multiple treatment comparison meta-analysis. JAMA 2012;308:1246–53.

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

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

16. 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.

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

18. Brignardello-Petersen R. Should network meta-analysis become the standard in evidence-based clinical practice? Toronto, Ontario: University of Toronto, 2016.

19. Zhang J, Yuan Y, Chu H. The impact of excluding trials from network meta-analyses – an empirical study. PLoS One 2016;11:e0165889.

20. Petropoulou M, Nikolakopoulou A, Veroniki A-A, et al. Bibliographic study showed improving statistical methodology of network metaanalyses published between 1999 and 2015. J Clin Epidemiol 2017;82:20–8.

21. Baker WL, Baker EL, Coleman CI. Pharmacologic treatments for chronic obstructive pulmonary disease: a mixed-treatment comparison meta-analysis. Pharmacotherapy 2009;29:891–905.

22. Cipriani A, Furukawa TA, Salanti G, et al. Comparative efficacy and acceptability of 12 new-generation antidepressants: a multipletreatments meta-analysis. Lancet 2009;373:746–58.

23. Elliott WJ, Meyer PM. Incident diabetes in clinical trials of antihypertensive drugs: a network meta-analysis. Lancet 2007;369:201–7.

24. 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.

25. Thijs V, Lemmens R, Fieuws S. Network meta-analysis: simultaneous meta-analysis of common antiplatelet regimens after transient ischaemic attack or stroke. Eur Heart J 2008;29:1086–92.

26. Weisberg S. Applied linear regression. 3rd edn. Hoboken, New Jersey: John Wiley & Sons, Inc, 2005.

27. van Valkenhoef G, Kuiper J. gemtc: GeMTC network meta-analysis. R package version 0.6-1, 2014. Available: https://cran.r-project.org/ web/packages/gemtc/index.html [Accessed 4 Jul 2017].

28. R Core Team. R: A language and environment for statistical computing. Vienna, Austria: R Foundation for Statistical Computing, 2014. (accessed 4 July 2017).

29. Brooks S, Gelman A. General methods for monitoring convergence of iterative simulations. J Comput Graph Stat 1998;7:434–55.

30. Gelman A, Rubin DB. Inference from iterative simulation using multiple sequences. Statist Sci 1992;7:457–72.

31. Cohen J. A coefficient of agreement for nominal scales. Educ Psychol Meas 1960;20:37–46.

32. Cohen J. Weighted kappa: nominal scale agreement provision for scaled disagreement or partial credit. Psychol Bull 1968;70:213–20.

33. König J, Krahn U, Binder H. Visualizing the flow of evidence in network meta-analysis and characterizing mixed treatment comparisons. Stat Med 2013;32:5414–29.

34. Jackson D, White IR, Price M, et al. Borrowing of strength and study weights in multivariate and network meta-analysis. Stat Methods Med Res 2017;26:2853–68.

35. Salanti G, Del Giovane C, Chaimani A, et al. Evaluating the quality of evidence from a network meta-analysis. PLoS One 2014;9:e99682.

36. Mbuagbaw L, Rochwerg B, Jaeschke R, et al. Approaches to interpreting and choosing the best treatments in network metaanalyses. Syst Rev 2017;6:79–83.

37. Byrt T, Bishop J, Carlin JB. Bias, prevalence and kappa. J Clin Epidemiol 1993;46:423–9.

38. Stevens JP. Outliers and influential data points in regression analysis. Psychol Bull 1984;95:334–44.

39. Landis JR, Koch GG. The measurement of observer agreement for categorical data. Biometrics 1977;33:159–74.

Supplementary Material

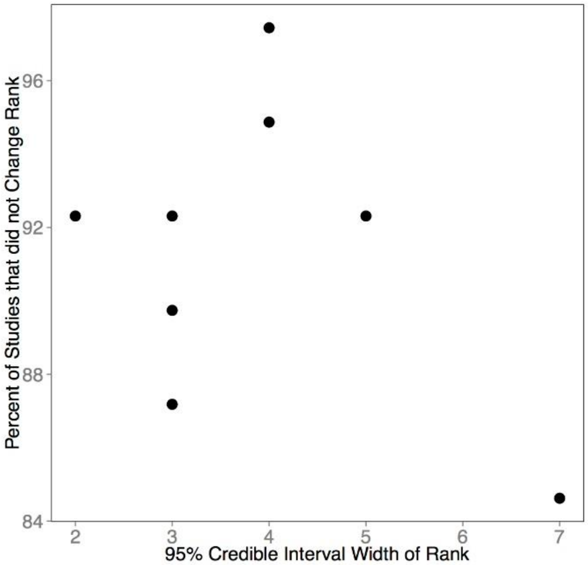


Figure S1 *Comparison of treatment sensitivity, reflected by the proportion of studies for which their removal did not change a treatment's rank, and rank uncertainty, reflected by the width of the 95% credible intervals of the treatment's rank, for treatments in the COPD network [1].*

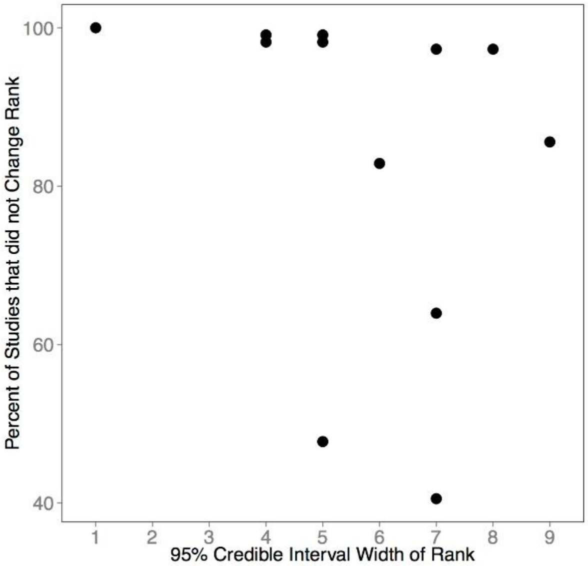


Figure S2 *Comparison of treatment sensitivity, reflected by the proportion of studies for which their removal did not change a treatment's rank, and rank uncertainty, reflected by the width of the 95% credible intervals of the treatment's rank, for treatments in the depression network [2].*

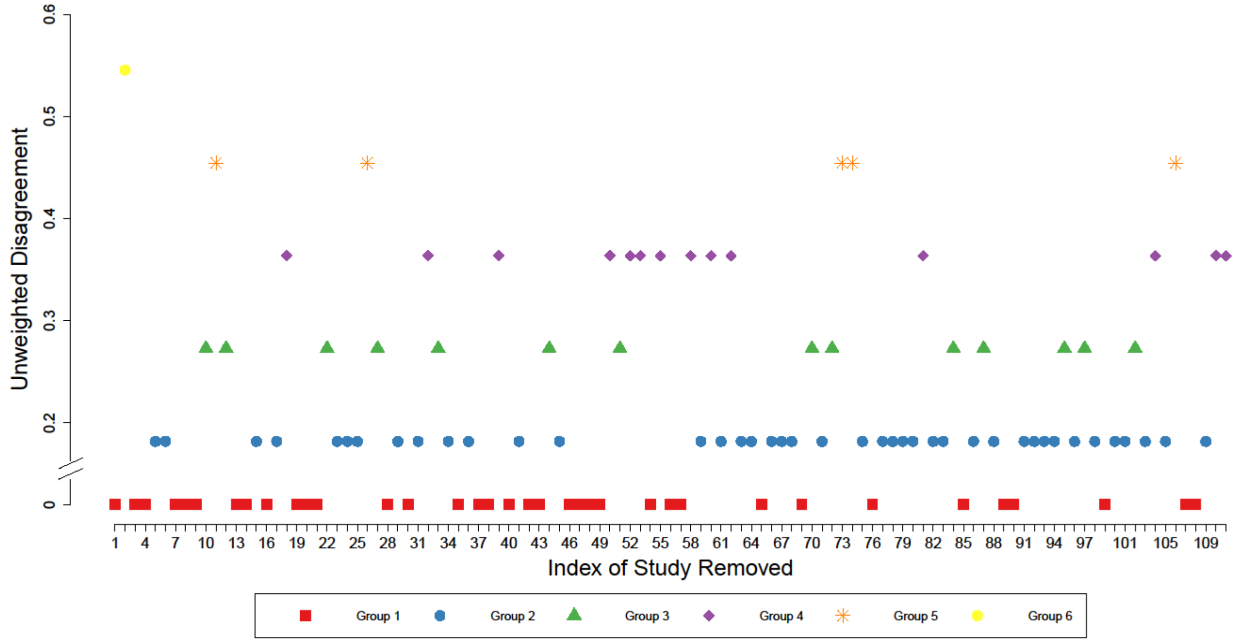
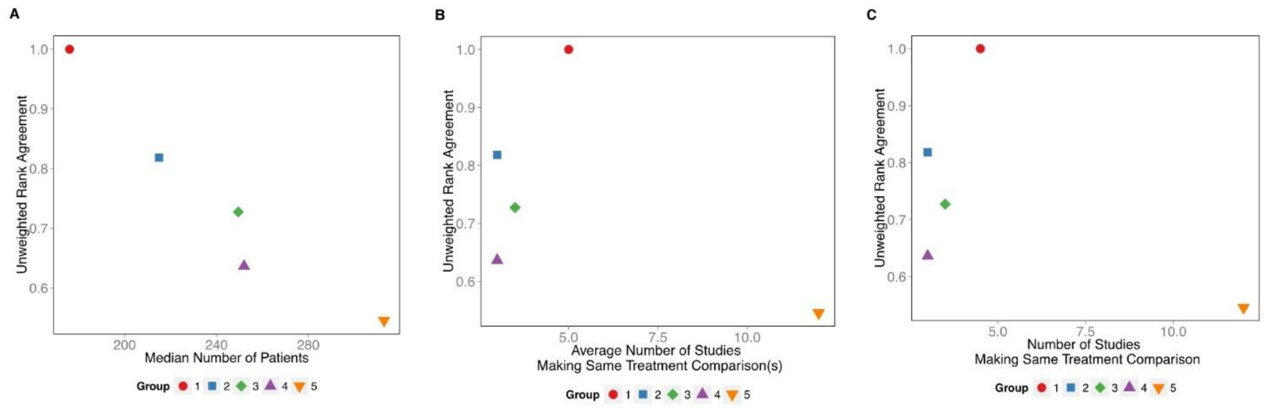


***Figure S3** Observed 1 – unweighted kappa ($\hat{\kappa}_{UW}$) in the depression network,[2] which quantifies the disagreement between the treatment ranks produced from the complete dataset and the ranks produced from a sub-dataset where one RCT (indexed on the x-axis) is removed. Studies are grouped by similar impact on rankings, as indicated by markers described in legend, for further investigation.*



Group	Median Number of Patients (Min – Max)	Group	Median Information Available for Comparisons, Averaged across All Comparisons (Min - Max)	Group	Median Information Available for Comparisons, by Comparison (Min – Max)
1	176 (22 – 382)	1	5 (1 – 12)	1	4.5 (1 – 12)
2	215 (52 – 547)	2	3 (1 – 12)	2	3 (1 – 12)
3	249.5 (43 – 400)	3	3.5 (1 – 12)	3	3.5 (1 – 12)
4	252 (46 – 300)	4	3 (1 – 12)	4	3 (1 – 12)
5	313 (190 – 708)	5	12 (3 – 12)	5	12 (3 – 12)

Figure S4 *Study characteristics between five groupings of studies in depression network [2]: Groups 1-5 correspond to the Groups 1-5 identified in Figure S3. Scatter plots, as well as descriptive statistics, of A) the number of patients within studies, B) information available for comparisons (ie, number of studies making each comparison), averaged across all comparisons made by a study, and C) information available for comparisons across all comparisons made by a study, are displayed.*

Table S1 *Characteristics of each network.*

Network	Outcome	Number of 2- / 3- / 4-arm Trials	Number of Treatments	Number of Patients	Number of Patients/ Treatment			Proportion of Possible Comparisons Made in RCTs	Number of Trials/ Direct Comparison		
					Median	Min	Max		Median	Min	Max
COPD [1]	Proportion who experienced at least one exacerbation	29/ 4/ 6	8	28,235	3,165.5	455	8,745	0.571	3.5	1	16
Depression [2]	Proportion who responded to treatment	109/ 2/ 0	12	24,595	1,779.5	569	5,061	0.636	2	1	12
Diabetes [3]	Proportion who developed diabetes	18/ 4/ 0	6	154,176	23,166.5	14,185	38,809	0.933	2	1	5
Heavy Menstrual Bleeding [4]	Proportion who were dissatisfied	20/ 0/ 0	4	2,886	759	128	1,240	0.667	4	1	11
Stroke [5]	Proportion who developed incident vascular events	23/ 3/ 0	5	55,463	10,311	5,138	21,402	0.70	4	2	12

Table S2 *Description of model fitted to each NMA dataset.*

		Model			Model Fit		Evidence of Potential Inconsistency?	
NMA	Outcome	Fixed or Random	Likelihood	Link	Data points	Total Residual Deviance	Global assessment	Loop assessment
COPD [1]	Proportion who experienced at least one exacerbation	Random	Binomial	logit	94	106.6	Yes	Yes
Depression [2]	Proportion who responded to treatment	Random	Binomial	logit	224	230.1	No	Yes
Diabetes [3]	Proportion who developed diabetes	Random	Binomial	cloglog	48	53.6	Yes	Yes
Heavy Menstrual Bleeding [4]	Proportion who were dissatisfied	Fixed	Binomial	logit	40	39.9	No	No
Stroke [5]	Proportion who developed incident vascular events	Fixed	Binomial	logit	55	47.0	No	No

In practice, the fit of an appropriate base-case model should be assessed, and checks for inconsistency should be conducted before any sensitivity or subgroup analysis. We selected either a fixed- or random-effects model with a common between study variance assumed, where the best fitting model based on a meaningfully lower total residual deviance or DIC (about 3-5 points) was chosen (Table S2).[6] Inconsistency was assessed in this base-case NMA by comparing the fit of an unrelated mean model to the NMA model on the complete datasets, as well as node-splitting.[7-9]

Table S3 *Study indices by treatment comparison (COPD network).*

Treatment Comparison	**Study Index**
budesonide vs. budesonide-formoterol	13, 18
budesonide vs. formoterol	13, 18
budesonide vs. placebo	13, 18
budesonide-formoterol vs. formoterol	13, 18
budesonide-formoterol vs. placebo	13, 18
fluticasone vs. placebo	1, 8, 9, 14, 17, 22, 33, 39
fluticasone vs. fluticasone-salmeterol	9, 14, 17, 33
fluticasone vs. salmeterol	9, 14, 17, 33
fluticasone-salmeterol vs. placebo	9, 14, 16, 17, 25, 28, 33
fluticasone-salmeterol vs. salmeterol	9, 14, 16, 17, 28, 33, 34
fluticasone-salmeterol vs. tiotropium	36
formoterol vs. placebo	10, 13, 18, 21
formoterol vs. tiotropium	24
salmeterol vs. placebo	3, 4, 6, 7, 9, 11, 12, 14, 15, 16, 17, 28, 29, 31, 33, 37
salmeterol vs. tiotropium	7, 11, 20
tiotropium vs. placebo	2, 5, 7, 11, 19, 23, 26, 27, 30, 32, 35, 38

Table S4 *Study indices by treatment comparison (Depression network).*

Treatment Comparison	**Study Index**
bupropion vs. escitalopram	28, 29
bupropion vs. fluoxetine	33, 48, 106
bupropion vs. paroxetine	105
bupropion vs. sertraline	32, 36, 61
bupropion vs. venlafaxine	5, 96, 109
citalopram vs. escitalopram	24, 34, 68, 72
citalopram vs. fluoxetine	21, 22, 63
citalopram vs. fluvoxamine	56
citalopram vs. mirtazapine	67
citalopram vs. paroxetine	2
citalopram vs. reboxetine	17, 65
citalopram vs. sertraline	44
citalopram vs. venlafaxine	7
duloxetine vs. escitalopram	62, 78, 104
duloxetine vs. fluoxetine	51
duloxetine vs. paroxetine	39, 52, 58, 81
escitalopram vs. fluoxetine	59, 87
escitalopram vs. paroxetine	12, 23
escitalopram vs. sertraline	88, 101
escitalopram vs. venlafaxine	18, 71
fluoxetine vs. fluvoxamine	37, 82
fluoxetine vs. milnacipran	11, 53, 66
fluoxetine vs. mirtazapine	9, 57, 102, 107, 108
fluoxetine vs. paroxetine	1, 26, 40, 45, 46, 47, 49, 50, 73, 74, 79, 97
fluoxetine vs. reboxetine	19, 20, 27, 95
fluoxetine vs. sertraline	4, 16, 46, 47, 77, 89, 94, 100
fluoxetine vs. venlafaxine	8, 30, 35, 38, 41, 42, 76, 84, 92, 98, 99
fluvoxamine vs. milnacipran	31
fluvoxamine vs. mirtazapine	86
fluvoxamine vs. paroxetine	10, 60, 64
fluvoxamine vs. sertraline	75, 83
fluvoxamine vs. venlafaxine	55
milnacipran vs. paroxetine	90
milnacipran vs. sertraline	110
mirtazapine vs. paroxetine	14, 85, 103
mirtazapine vs. sertraline	13
mirtazapine vs. venlafaxine	15, 54
paroxetine vs. sertraline	3, 46, 47, 111
paroxetine vs. venlafaxine	69
reboxetine vs. sertraline	43
reboxetine vs. venlafaxine	6
sertraline vs. venlafaxine	25, 70, 80, 91, 93

References

1. Baker W, Baker E, Coleman C. Pharmacologic treatments for chronic obstructive pulmonary disease: a mixed-treatment comparison meta-analysis. Pharmacotherapy 2009;29(8):891-905.

2. Cipriani A, Furukawa T, Salanti G, et al. Comparative efficacy and acceptability of 12 new-generation antidepressants: a multiple-treatments meta-analysis. Lancet 2009;373(9665):746-58.

3. Elliott W, Meyer P. Incident diabetes in clinical trials of antihypertensive drugs: a network meta-analysis. Lancet 2007;369(9557):201-7.

4. Middleton L, 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.

5. Thijs V, Lemmens R, Fieuws S. Network meta-analysis: simultaneous meta-analysis of common antiplatelet regimens after transient ischaemic attack or stroke. Eur Heart J 2008;29(9):1086-92.6.

6. Spiegelhalter D., et al., Bayesian measures of model complexity and fit. J R Stat Soc Series B Stat Methodol 2002;64(4):583-639.

7. Dias S, et al. NICE DSU Technical Support Document 4: Inconsistency in networks of evidence based on randomised controlled trials, in Technical Support Document. 2011.

8. Dias S, et al., Evidence Synthesis for Decision Making 4: Inconsistency in networks of evidence based on randomized controlled trials. Medical Decision Making 2013;33:641-56.

9. van Valkenhoef G, et al., Automated generation of node-splitting models for assessment of inconsistency in network meta-analysis. Res Synth Meth 2016;7:80-93.

Chapter 4

Partial surface under the cumulative ranking curve (pSUCRA) as an alternative measure for ranking treatments in network meta-analysis

Caitlin H Daly[1,2], Lawrence Mbuagbaw[1,3], Lehana Thabane[1,3],
Sharon E Straus[4,5], Jemila S Hamid[1,6]

[1] Department of Health Research Methods, Evidence, and Impact, McMaster University

[2] Population Health Sciences, Bristol Medical School, University of Bristol

[3] Biostatistics Unit, Father Sean O'Sullivan Research Centre, St Joseph's Healthcare Hamilton

[4] Knowledge Translation Program, Li Ka Shing Knowledge Institute, St. Michael's Hospital

[5] Department of Medicine, Faculty of Medicine, University of Toronto

[6] Department of Mathematics and Statistics, University of Ottawa

Citation: Daly, C. H., Mbuagbaw, L., Thabane, L., Straus, S. E., & Hamid, J. S. (2020a). Partial surface under the cumulative ranking curve (pSUCRA) as an alternative measure for ranking treatments in network meta-analysis. *Submitted to Clinical Epidemiology*.

Abstract

Background: The surface under the cumulative ranking curve (SUCRA) in network meta-analysis (NMA) is commonly used to determine the comparative rankings of treatments. When the cumulative ranking curves of two or more treatments cross, treatments with the highest probabilities of ranking the best, 2nd best, 3rd best, etc may rank further down an ordered list of SUCRA. This situation is similar to the areas under two crossing receiver operating characteristic (ROC) curves in diagnostic medicine.

Methods: This is a proof of concept study, where we adapt the concept of partial area under the ROC curve, used to evaluate and compare performances of diagnostic tests, and extend its use to the context of ranking treatments in NMA. The partial surface under the cumulative ranking curve (pSUCRA) is proposed as an alternative measure for ranking treatments in NMA. We use simulated data as well as data from a published NMA on acute mania treatments to illustrate pSUCRA, its interpretation, and investigate its optimality.

Results: In both illustrative datasets presented in this paper, pSUCRA is shown to favour treatments with the strongest estimated effects and highest probabilities of being ranked in the top. This is because higher weights are assigned, by design, to treatments with more favourable ranking probabilities (eg ranks 1, 2, 3).

Conclusion: pSUCRA has the potential to highlight treatments that outperform others in a relevant region of the cumulative ranking curve. This is important when considering large numbers of treatments. Further investigation using extensive simulations and empirical evaluations are required.

Introduction

Network meta-analysis (NMA) simultaneously compares and synthesizes direct and indirect evidence on multiple treatments to produce a coherent list of relative effects.[1,2] Combining evidence in such a way permits the ranking of treatments. The surface under the cumulative ranking curve (SUCRA) is an increasingly common measure for ranking treatments.[3] It has been argued that this ranking measure is more favourable than the probability of a treatment ranking best as it incorporates the full distribution of the ranking probabilities and hence the uncertainty in the evidence.[3,4] However, there have been concerns that SUCRA does not convey important information on the relative effects and other information (eg bias) that may be relevant to the decision maker.[5]

SUCRA is simply a measure of how certain a treatment is to be the best among the alternative treatments in the network.[3] A treatment can still have a high value of SUCRA even if one cannot statistically conclude it is more effective or safe than a placebo. For example, the confidence or credible intervals of the relative effects vs. placebo include the null effect. To mitigate this misleading feature, analysts may rank a subset of the treatments in a Bayesian framework, where the subset only includes treatments a decision maker is willing to consider for recommendation. For example, in a National Institute for Health and Care Excellence (NICE) guideline on post-traumatic stress disorder management, treatments that had been studied on at least 100 patients were ranked against each other, as this was considered to be the minimum amount of evidence to draw conclusions.[6] This approach still allows all relevant evidence to contribute to the NMA while the distributions of the ranking probabilities are only estimated based on the relative effects of the treatments eligible for a decision.[7]

Another potentially misleading situation may arise when two or more treatments have similar SUCRA, giving the impression their estimated relative effects are similar in terms of magnitude and uncertainty. However, this will not always be the case, as SUCRA is a balance of the comparative location and scale (variation) of the posterior distributions of the estimated treatment effects.[3,8] The trade-off between the location and scale of these distributions could lead to the crossing of cumulative ranking curves, which could result in them having the same area under the curve. In addition, there are also situations where crossing cumulative ranking curves lead to a treatment (say A) with the highest ranking probabilities of being among the best to have a lower SUCRA-based rank than another treatment (say B) with much lower probabilities of ranking among the best. These two situations are similar to the crossing of two receiver operating characteristic (ROC) curves in diagnostic medicine, where the ROC curve represents the trade-off between the sensitivity and specificity of a test, and the area under the ROC curve (AUC) is used as a summary of a test's performance.[9-11]

When two ROC curves cross and the corresponding AUCs are similar, important differences in the sensitivity and specificity are masked. In some cases, the test with the highest sensitivity within a range of acceptable specificity (ie left side of the ROC curves), may not correspond to the highest AUC value. For example, consider the curves in Figure 1A. A test may be moderately sensitive when highly specific (solid black line in Figure 1), while another test may be minimally sensitive when highly specific, but its sensitivity improves at a greater rate compared to the other test at a cost of specificity (dashed red line in Figure 1). The partial AUC (pAUC) may be used to compare and distinguish the performances of these diagnostic tests in a more important region of sensitivity or specificity.[12,13]

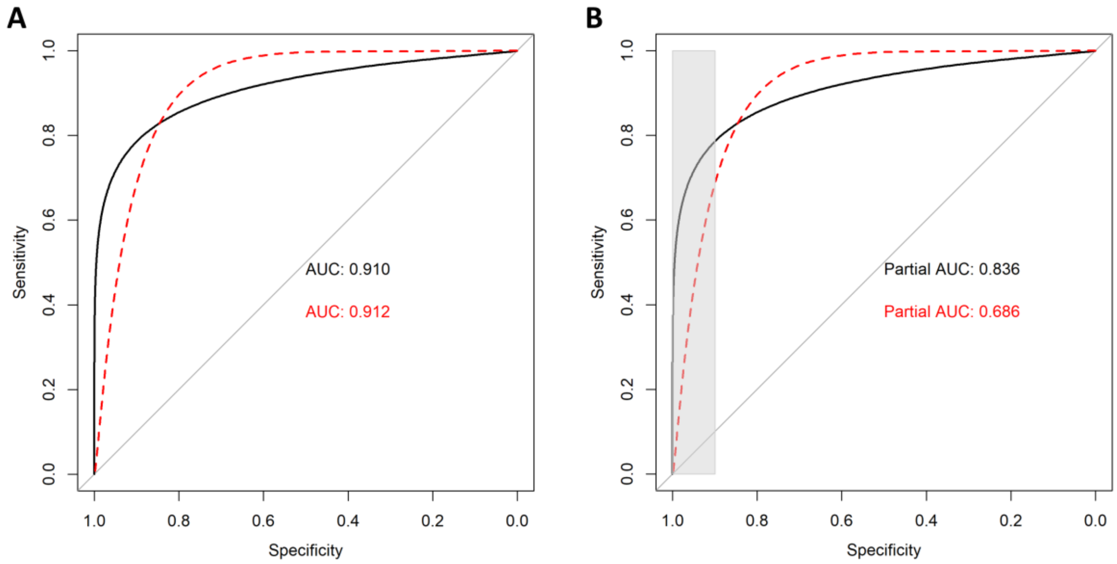


Figure 1: *Two simulated receiver operating characteristic curves with similar areas under the curves (A). The same curves are plotted in (B), however, the partial areas under the curves for specificities between 0.9 and 1 suggest the solid curve is associated with the superior diagnostic test.*

If, for instance, a false positive rate greater than 10% would make a test useless or not applicable in practice, then the tests' pAUCs would have distinguished their performance in a clinically relevant range of specificities (Figure 1B). Using a similar argument, considering a portion of the cumulative ranking curve in NMA may help distinguish the performances of treatments among the top ranks.

In this paper, we adapt the concept of pAUC and extend its use to provide an alternative measure for ranking treatments in NMA. We refer to this measure as the partial surface under the cumulative ranking curve (pSUCRA). In this proof of concept study, we also seek to improve the understanding of ranking probabilities and their summary measures through the development and implementation of pSUCRA. We illustrate an application of pSUCRA using a simulated dataset and a published NMA. We provide comparative evaluation, where we show that pSUCRA may be a more suitable alternative to SUCRA in some situations and that SUCRA may not capture a decision maker's aversity to uncertainty.

Material and methods

Surface under the cumulative ranking curve (SUCRA)

SUCRA, the estimate for the area under the cumulative ranking curve, can be viewed as the area under a cumulative step function.[3] This is demonstrated in Figure 2 and we provide a step-by-step illustration of how the measure is calculated. To simplify its presentation, we consider an NMA involving $n = 5$ treatments, where the cumulative ranking curve for treatment i is represented by a plot in Figure 2.

The formula used to calculate SUCRA for any treatment is equivalent to that of a left hand Riemann sum.[14] That is, rectangles of equal width are plotted along the curve so that the point of each curve intersects with the top left corner of each rectangle (Figure 2A). The sum of the area of the four rectangles (length l multiplied by width w) is used to estimate SUCRA:

$$\begin{aligned} SUCRA(i) &= A(i) \\ &= \sum_{j=1}^{4} A(i,j) \\ &= l_1 w_1 + l_2 w_2 + l_3 w_3 + l_4 w_4 \\ &= F(i,1)w + F(i,2)w + F(i,3)w + F(i,4)w \\ &= \left(F(i,1) + F(i,2) + F(i,3) + F(i,4)\right)w \\ &= w\sum_{j=1}^{4} F(i,j) \end{aligned}$$

where the length, $l_i = F(i,j)$, is taken to be the cumulative ranking probability of treatment i at rank j.

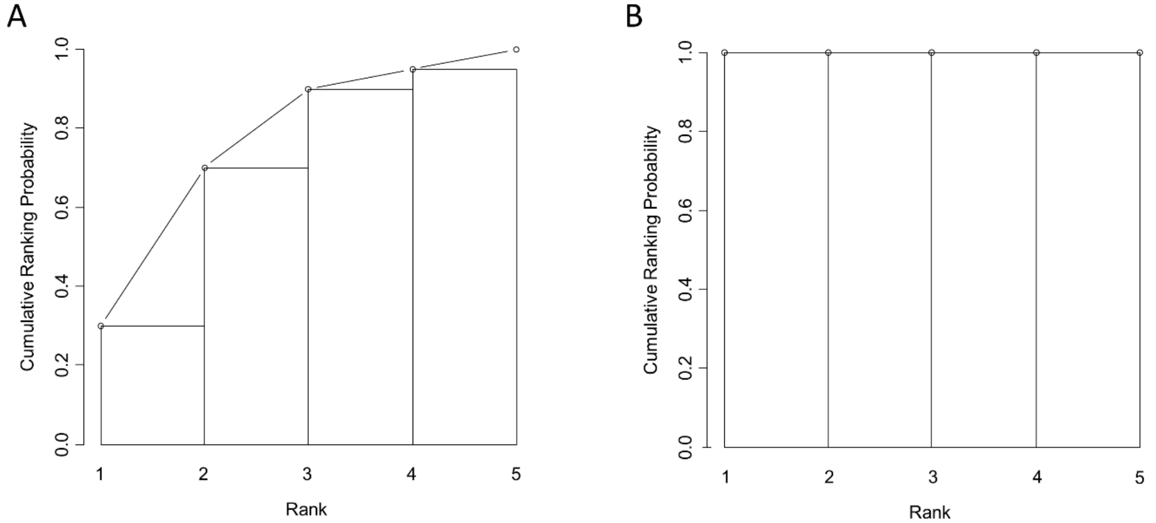


***Figure 2**: Examples of cumulative ranking curves for (A) a treatment that does not always rank best and (B) a treatment that always ranks best in Markov chain Monte Carlo simulations of the posterior distributions of the relative effects.*

To ensure SUCRA varies from 0 to 1 so that it may be interpreted as a probability of being the best, the maximum possible area under a SUCRA curve must be 1. This is possible if a treatment ranks best for 100% of Markov chain Monte Carlo simulations of the posterior distributions of the treatments' relative effects. The cumulative ranking curve for such a treatment in a network of 5 treatments is plotted in Figure 2B. Since the area of a rectangle is calculated as length × width, and length varies from 0 to 1, the width of each rectangle must be standardized so that the sum of the rectangle widths equals 1. Thus, for 5 treatments, and thus 4 rectangles, $w=\frac{1}{4}$. Therefore,

$$SUCRA(i)=A(i)=\frac{1}{4}\sum_{j=1}^{4}F\left(i,j\right)$$

in the case of 5 treatments. In general, for an NMA involving n treatments, $w=\frac{1}{n-1}$, and thus SUCRA for a particular treatment i is estimated as

$$SUCRA(i)=\frac{1}{n-1}\sum_{j=1}^{n-1}F\left(i,j\right).$$

Partial surface under the cumulative ranking curve (pSUCRA)

The closer a diagnostic test's ROC curve or a treatment's cumulative ranking curve is to the upper left corner of the plot, the better its performance. Thus, one can argue that the upper left quadrant of a plot is the key area of interest. The cumulative ranking curve should then be restricted to the more favorable ranks (ie 1st, 2nd, 3rd best, etc), denoted by $p=1,...,n-1$, or the larger cumulative ranking probabilities. Note that pSUCRA can be viewed as a special case of SUCRA, where if we were to consider the entire cumulative ranking curve, then SUCRA and pSUCRA at $p=n-1$ are identical.

Suppose a decision maker is interested in the top p-ranked treatments. If we solely truncate the cumulative ranking curve at the p^{th} rank, pSUCRA for the i^{th} treatment to the left of this cut-off value can be calculated as:

$$pSUCRA\left(i,p\right)=\frac{1}{n-1}\sum_{j=1}^{p}F\left(i,j\right).$$

However, note that based on this definition of pSUCRA, the maximum possible area, which is achieved when all cumulative ranking probabilities are equal to 1, is $\frac{p}{n-1}$, and this does not necessarily equal 1. In order to facilitate the interpretation of pSUCRA so it provides a measure of a treatment's probability of being the best among a restricted range of ranks, the widths of the rectangles must be standardized to add to 1. Since pSUCRA is summed over ranks 1 to p, the widths of the rectangles should be $\frac{1}{p}$. Thus, a more interpretable calculation of pSUCRA is:

$$pSUCRA\left(i,p\right)=\frac{1}{p}\sum_{j=1}^{p}F\left(i,j\right).$$

Interpretation of pSUCRA

The ranking probabilities for treatment i, denoted by $P\left(i,j\right), j=1,...,n$, are interpreted as the probability that treatment *i* ranks j^{th} best among all *n* treatments considered in the NMA.[3] Therefore, a cumulative ranking probability, $F\left(i,k\right)=P\left(i,j\le k\right)=\sum_{j=1}^{k}P\left(i,j\right)$, is then the probability that treatment *i* ranks at least k^{th} best among the available *n* treatments. As such, the area under this cumulative ranking curve, which is referred to as SUCRA, is an average of treatment i 's $(n-1)$ cumulative ranking probabilities. This observation is similar to an interpretation of the area under the ROC curve, which may be interpreted as, the average sensitivity across all possible values of specificity and vice versa.[10,15] SUCRA has also been shown to be an inversely scaled average rank and may also be described as "the average proportion of treatments worse than i ".[8] Regardless of this interpretation derived from first principle mathematical argument, SUCRA is often used for ranking the treatments, where the treatment with the highest SUCRA is considered to be the best (ranked number one) and the treatment with the lowest SUCRA is considered the least favorable treatment.

Similarly, $pSUCRA(i, p)$ is the average of treatment i 's top p cumulative ranking probabilities. We would like to highlight that this is also similar to an interpretation of the standardized pAUC, which is interpreted as the average sensitivity of a diagnostic test that is expected to be within the stated range of high specificities.[10,16] Another interpretation of the standardized $pSUCRA(i, p)$ may be, it is the probability that treatment i is one among the top p treatments, assuming that we are interested in only the top *p*-ranked treatments. This is particularly important in NMAs involving large number of treatments. For instance, NMAs of complex interventions may involve lots of different combinations of interacting components (eg up to 78 interventions in Tricco et al[17]).

We would like to highlight here that $pSUCRA(i, p)$ has an important feature, where it can be described as a weighted sum of the ranking probabilities, $P(i, j)$, up to rank p , where the higher ranks receive more weight. This allows an intuitive and more meaningful interpretation in practical applications. To illustrate this, let us consider $pSUCRA(i,3)$ for a given treatment i and assume the top 3 ranked treatments are relevant to a decision maker. $pSUCRA(i,3)$ can be rewritten as

$$\begin{aligned} pSUCRA(i,3) &= \frac{1}{3}\left[F(i,1)+F(i,2)+F(i,3)\right] \\ &= \frac{1}{3}\left[\left(P(i,1)\right)+\left(P(i,1)+P(i,2)\right)+\left(P(i,1)+P(i,2)+P(i,3)\right)\right] \\ &= \frac{1}{3}\left[3P(i,1)+2P(i,2)+P(i,3)\right] \\ &= P(i,1)+\frac{2}{3}P(i,2)+\frac{1}{3}P(i,3) \end{aligned}$$

By extension, SUCRA is also a weighted sum of the ranking probabilities. As such, pSUCRA maintains this intuitive and suitable interpretation of SUCRA.

Determining cut-off values

Determining the range of ranks to consider is a very challenging task and might be subjective in nature. This can be perceived as a limitation of pAUC as well, where a proposed solution is to specify this a priori, with clear justification.[12,13] Nevertheless, a decision maker (eg a clinician in the pAUC case and a policy maker in the pSUCRA case) often has an evidence informed argument for the acceptable ranges of sensitivities and specificities (in pAUC) and acceptable number of treatments (in pSUCRA) to consider. As such, to increase transparency and avoid a phenomenon similar to p-hacking, the cut-off values for the calculation of pSUCRA should be determined a priori and where possible, justified based on reasons such as policy and other relevant knowledge user factors.

As mentioned previously, a researcher might simply be interested in the top *p*-ranked treatments when the NMA involves a large number of treatments. In such situations, pSUCRA might be the most optimal measure since it quantifies the probability that a treatment is among the top *p* treatments. Another scenario where pSUCRA is the obvious candidate for an optimal measure is when the cumulative ranking curves cross. The intersection point, therefore, can be used as a cut-off value in such situations.

It is important to highlight that, one can also present pSUCRA values across all cut-off values, $p = 1, 2, 3, ..., n-1$, allowing researchers to see the distribution of pSUCRA across all possible values of p. This allows researchers and decision makers to see where the evidence lies in terms of the probabilities that inform pSUCRA and SUCRA. This may be of interest to a decision maker considering the results of a systematic review of n treatments, but only $m < n$ of those treatments are available in their region. If resources are limited, then they may wish to see how likely those m treatments are to be among the m best before conducting their own review. A systematic review may then wish to present $pSUCRA(i, p)$ across all $n-1$ ranks, like what is done in diagnostic decision making. This may appeal to researchers who aim to systematically review the literature and are not commissioned by an agency for the purpose of a health technology assessment. This may increase the uptake of the results of a systematic review and NMA by decision makers looking to interpret the literature in the context of their resources and a justifiable cut-off.

In this study, this is how we presented the results and we have provided possible interpretations by considering different choices of p.

Results

In this section we provide empirical evaluations of our method by considering a simulated dataset and a published NMA dataset to illustrate calculations, applications, and interpretations of pSUCRA. For each dataset, we highlight some of their characteristics prior to analyses. The NMAs, including the calculation of ranking probabilities, were performed in WinBUGS[18] and the figures were generated using basic plotting features in the R statistical software,[19] as well as the forestplot() command in the metafor package.[20]

Simulated data

Recall that the comparative location (measure of central tendency such as mean) and scale (variation) of the posterior distributions of the relative treatment effects are reflected by the magnitude of SUCRA. It is important to understand this fully as it might explain why two treatments with crossing cumulative ranking curves may have similar SUCRA values. To illustrate this, we simulated a dataset consisting of 19 studies, which provided estimates of the effects of 19 treatments relative to a common reference treatment (Treatment 20), resulting in a star network of evidence (Figure 3).[21] A fixed effect NMA model was fitted to the data in a Bayesian framework, where the relative effects were modelled with a normal likelihood and identity link.[22] Ranking probabilities and SUCRA values were calculated for all 20 treatments. A summary of the NMA results is presented in Figure 4.

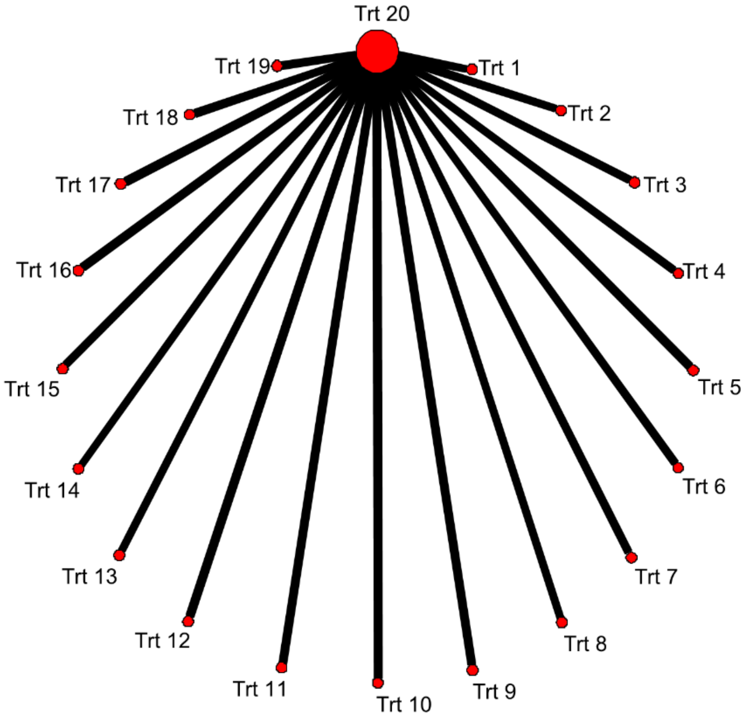


Figure 3. *Network diagram of simulated data. The sizes of the nodes are proportional to the number of patients randomized to the treatments, and the widths of the edges are proportional to the number of studies comparing two nodes. Abbreviations: Trt, treatment.*

The forest plot displays the mean of the posterior distributions of the treatment effects relative to Treatment 20, along with their 95% credible intervals (CrIs), defined by the 2.5th and 97.5th percentiles (Figure 4). The corresponding ranking probabilities and SUCRA values were estimated assuming that smaller values were preferred. Treatment 1 has the largest estimated treatment effect, however there is a lot of uncertainty around this estimate compared to the relative effects of other treatments. As the means of the posterior distributions decrease, the corresponding SUCRAs generally decrease. However, there are some exceptions observed for Treatments 2 and 5. The estimated effect of Treatment 2 is slightly smaller compare to Treatment 1, but the increased precision of the estimate results in a slightly larger value of SUCRA (Treatment 1: 0.809, Treatment 2: 0.811). A similar observation can be made for Treatment 5, where the estimated effect is the 5th strongest, but its greater precision gives rise to a SUCRA value similar to those of Treatments 1, 2, and 3.

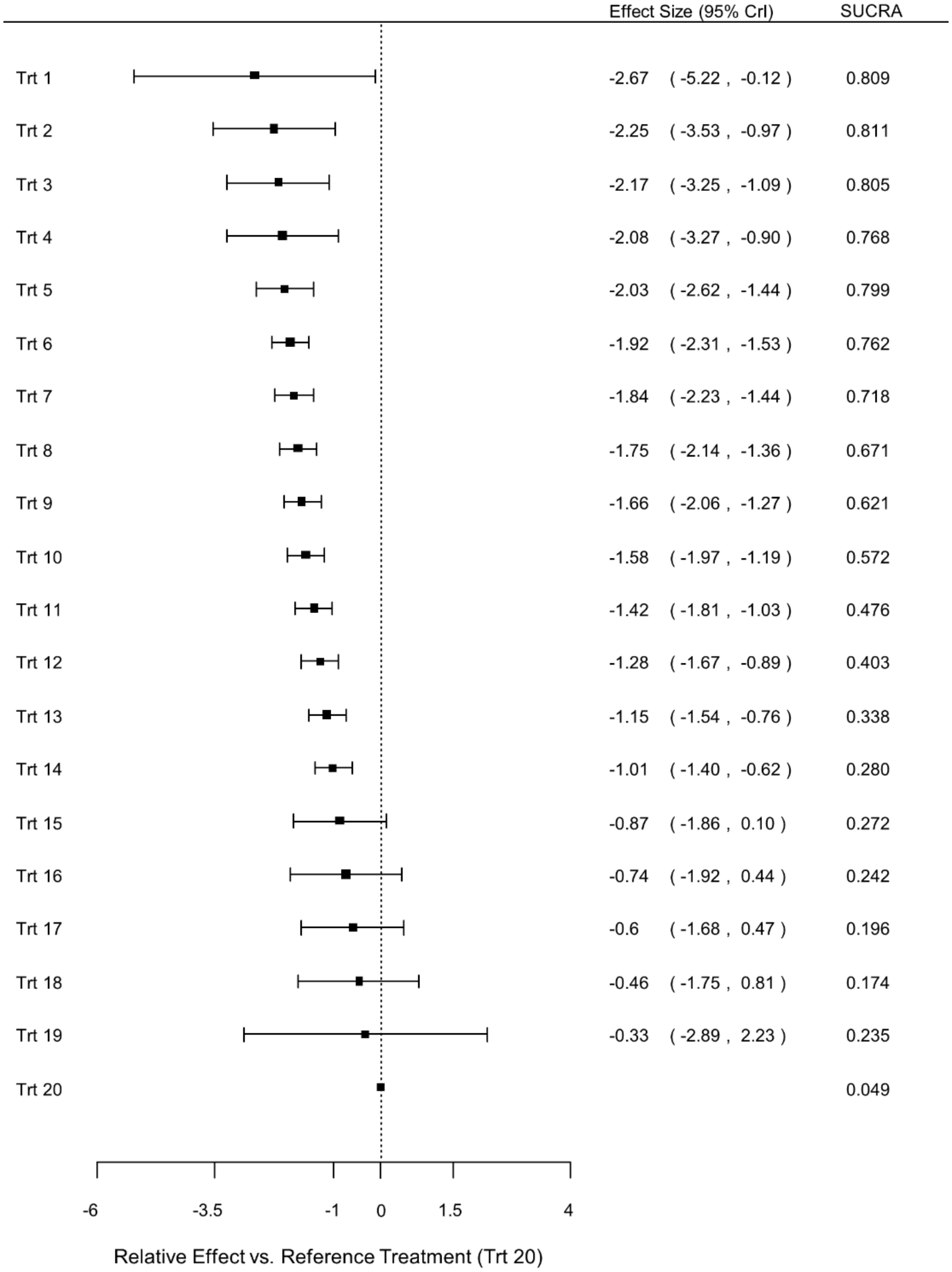


Figure 4: *Summary of results from a network meta-analysis of a simulated dataset of relative effects between Treatments (Trts) 1 through 20.*

To illustrate how the trade-off between the magnitude and uncertainty of the estimates is captured by the ranking probabilities, we focus on Treatments 1, 2, and 5. These treatments were selected because of the notable differences in the precision (scale) of their estimated effects, as well as the magnitude (location) (Figure 4). Consider the rank-o-grams for Treatments 1, 2, and 5 in Figure 5. Distributions of the ranking probabilities that are skewed towards the better ranks are preferable, and Treatment 1 looks promising (Figure 5A). Although it may not immediately obvious, the uncertainty in Treatment 1's effect is reflected by its small probability of being the best, which is less than 0.50, as well as its lower (worser) ranking probabilities, which are slighter larger than those for Treatments 2 and 5 for ranks 15 through to 20 (Figure 5A).

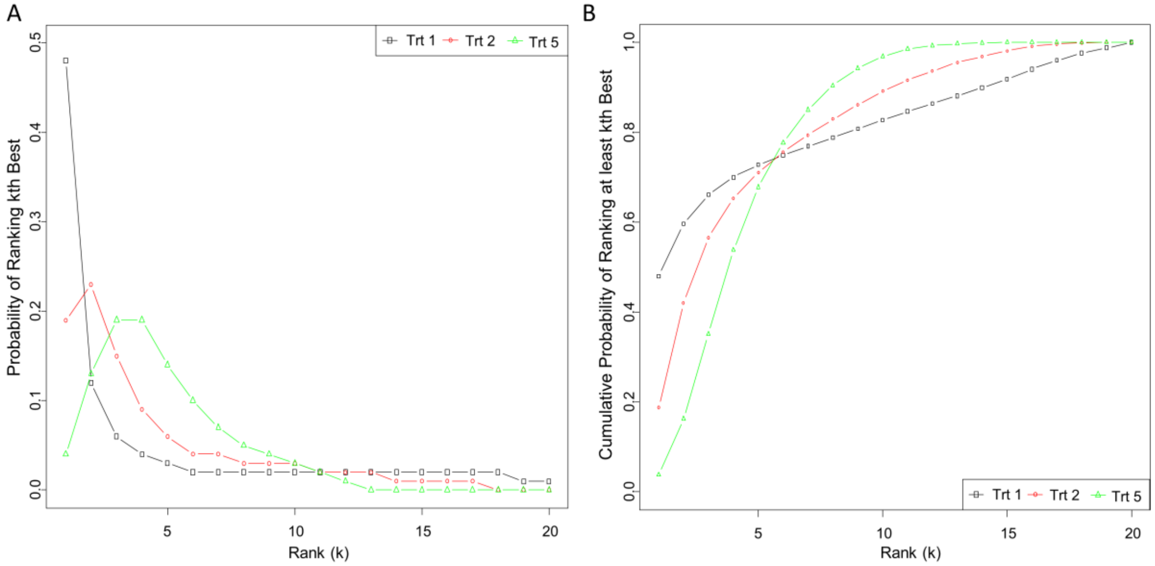


Figure 5: *Rank-o-gram (A) and cumulative ranking curves (B) for Treatments (Trts) 1, 2, and 5 in the simulated data.*

The trade-off in the location and scale of the posterior distributions of the treatments' effects is clearly demonstrated by the crossing of the cumulative ranking curves (Figure 5B). Between these three treatments, Treatment 1 has the highest cumulative ranking probabilities of being the best, at least 2nd best, at least 3rd best, at least 4th best, and at least 5th best. Hence, it has the largest area under its cumulative ranking curve between ranks 1 and 5. Treatment 2 has the second highest cumulative ranking probabilities of being the best, at least 2nd best, at least 3rd best, at least 4th best, and at least 5th best, while Treatment 5 has the third highest. However, Treatment 5 has the highest cumulative ranking probabilities of being at least 6th best through to at least 19th best. Hence, the area under Treatment 5's cumulative ranking curve is larger than the other two treatments between ranks 6 and 19. All cumulative ranking curves meet at $F\left(i,20\right)=1$. Since the cumulative ranking curves intersect between the 5th and 6th rank, it is possible that their SUCRA values, which are calculated across the entire curve, will mask the difference in these 3 treatments' top-ranking probabilities.

Indeed, we already noted this was the case in Figure 4, where the SUCRA values corresponding to these treatments are 0.809, 0.811, and 0.799, respectively. Not only are these values similar, but their SUCRA-based ranks would favor Treatment 2 over Treatment 1, despite Treatment 1 having the largest area under the curve for the top 5 ranks. We will show next that, in such situations, pSUCRA might be more optimal in distinguishing these treatments' performances compared to SUCRA. The pSUCRA values corresponding to the 20 treatments in the simulated data, for all possible cut-off values, are presented in Table 1.

***Table 1**: pSUCRA for treatments in the simulated data. The largest pSUCRAs values, at any cut-off value, are **bolded**. Note that pSUCRA(p=1) is equal to the probability of ranking best, while pSUCRA(p=19) is equal to SUCRA.*

Trt[a]	p = 1	p = 2	p = 3	p = 4	p = 5	p = 6	p = 7	p = 8	p = 9	p = 10	p = 11	p = 12	p = 13	p = 14	p = 15	p = 16	p = 17	p = 18	p = 19
1	**0.480**	**0.538**	**0.579**	**0.609**	**0.633**	**0.652**	**0.669**	**0.684**	**0.698**	**0.711**	**0.723**	**0.735**	**0.746**	**0.757**	**0.768**	**0.778**	**0.789**	0.799	0.809
2	0.188	0.304	0.391	0.457	0.508	0.549	0.584	0.615	0.642	0.667	0.690	0.710	0.729	0.746	0.762	0.776	**0.789**	**0.801**	**0.811**
3	0.135	0.241	0.333	0.406	0.464	0.512	0.552	0.588	0.619	0.648	0.673	0.696	0.717	0.736	0.753	0.768	0.782	0.794	0.805
4	0.119	0.212	0.293	0.359	0.413	0.458	0.497	0.531	0.563	0.592	0.619	0.644	0.667	0.688	0.707	0.724	0.740	0.755	0.768
5	0.038	0.100	0.184	0.273	0.354	0.424	0.485	0.537	0.582	0.621	0.654	0.682	0.706	0.727	0.746	0.761	0.775	0.788	0.799
6	0.009	0.032	0.078	0.144	0.223	0.303	0.377	0.443	0.500	0.549	0.589	0.623	0.652	0.677	0.699	0.717	0.734	0.749	0.762
7	0.004	0.016	0.043	0.086	0.145	0.212	0.282	0.350	0.412	0.467	0.514	0.554	0.588	0.617	0.643	0.665	0.685	0.702	0.718
8	0.002	0.008	0.022	0.048	0.087	0.137	0.195	0.258	0.321	0.380	0.433	0.479	0.519	0.553	0.583	0.609	0.632	0.652	0.671
9	0.001	0.003	0.010	0.024	0.046	0.079	0.121	0.174	0.232	0.292	0.349	0.401	0.446	0.485	0.519	0.549	0.576	0.599	0.620
10	0.000	0.001	0.004	0.011	0.023	0.042	0.071	0.109	0.157	0.213	0.271	0.326	0.375	0.419	0.458	0.492	0.522	0.548	0.572
11	0.000	0.000	0.001	0.002	0.004	0.009	0.018	0.032	0.055	0.089	0.134	0.186	0.240	0.290	0.337	0.378	0.414	0.447	0.476
12	0.000	0.000	0.000	0.000	0.001	0.002	0.004	0.009	0.018	0.034	0.060	0.098	0.145	0.196	0.246	0.292	0.333	0.370	0.403
13	0.000	0.000	0.000	0.000	0.000	0.000	0.001	0.002	0.004	0.010	0.021	0.042	0.074	0.117	0.165	0.214	0.260	0.301	0.338
14	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.001	0.002	0.006	0.014	0.031	0.059	0.100	0.147	0.195	0.240	0.280
15	0.000	0.001	0.003	0.006	0.009	0.012	0.017	0.022	0.029	0.037	0.047	0.060	0.076	0.096	0.121	0.154	0.192	0.232	0.272
16	0.001	0.002	0.005	0.008	0.011	0.016	0.020	0.025	0.032	0.039	0.048	0.059	0.072	0.087	0.107	0.133	0.166	0.203	0.242
17	0.000	0.001	0.001	0.002	0.004	0.005	0.007	0.010	0.013	0.017	0.022	0.029	0.038	0.050	0.065	0.087	0.118	0.156	0.196
18	0.000	0.001	0.002	0.004	0.006	0.008	0.010	0.013	0.016	0.020	0.025	0.032	0.039	0.049	0.062	0.079	0.104	0.136	0.174
19	0.024	0.038	0.051	0.062	0.071	0.080	0.088	0.096	0.104	0.112	0.121	0.130	0.140	0.150	0.162	0.175	0.192	0.212	0.235
20	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.002	0.015	0.049

[a] Abbreviation: Trt, treatment

Table 1 shows that Treatment 1 has the largest area under its cumulative ranking curve until a cut-off of $p = 17$ (ie pSUCRA), despite the cumulative ranking curves of Treatments 1, 2, and 5 crossing between $p = 5$ and $p = 6$ (Figure 5B). Moreover, there is a difference of at least 0.05 between Treatment 1's pSUCRA and that of all other treatments for $p = 1,...,9$ (Table 1). Both Figure 4 and Table 1 show that this treatment is clearly the best treatment in terms of the magnitude of its effect and its pSUCRA, respectively. However, this is not reflected by the magnitude of SUCRA, since Treatments 1, 2, 3, and 5's SUCRA values are similar, and Treatment 1 would rank second based on SUCRA nevertheless.

Consider a cut-off value of $p = 1$, that is the probability of being ranked best. Treatment 1 has a considerably larger probability (0.480) than treatments 2, 3, 4, or 5, with corresponding probabilities 0.188, 0.135, 0.119, and 0.038. Here, we would like to note the large difference in these probabilities between Treatment 1 and Treatment 5 (a difference of 0.442!). Now compare that with the difference in SUCRA for these treatments (0.809 for Treatment 1 and 0.799 for Treatment 5), a difference of only 0.01. Treatments 1 and 2 also have a considerable difference in terms of being ranked number one (0.480 and 0.188, respectively). Nevertheless, Treatment 2 is ranked best based on SUCRA, with a value of 0.811.

Suppose now a decision maker is interested in the top 2 treatments, that is a cut-off value of $p = 2$. Treatment 1 again has the highest probability (0.538) of being ranked in the top 2, compared to Treatments 2, 3, 4, and 5, with the corresponding probabilities of 0.304, 0.241, 0.212, and 0.100. Here, we would like to note again the large difference in probabilities between Treatments 1 and 5, where the probability of being in the top 2 for Treatment 1 is much larger (by 0.438) compared to Treatment 5. Nevertheless, these two treatments are indistinguishable with respect to SUCRA, with only 0.01 difference in SUCRA values.

Treatment 1's pSUCRA continues to be considerably larger than that of Treatments 2, 3, 4, and 5, with a difference of at least 0.10 maintained until a cut-off of $p=6$. The difference between the pSUCRAs of Treatments 2 and 5 is also at least 0.10 until a cut-off of $p=6$ and is maximized at a cut-off of $p=3$ (a difference of 0.207). Ranks based on the magnitude of the estimated effectiveness and pSUCRA agree that Treatments 1, 2, 3, 4, and 5 are first, second, third, fourth, and fifth best, respectively, until a cut-off of $p=8$. At this cut-off, Treatment 5's pSUCRA exceeds that of Treatment 4, hence achieving a rank of fourth best based on pSUCRA. Treatments 1, 2, and 3 remain first, second, and third best based on pSUCRA until a cut-off of $p=17$, although the differences between them becomes increasing negligible, with differences of less than 0.05 between all three treatments after $p=12$. Overall, presenting pSUCRA values across all cut-off points provides some insight into potentially important differences in treatments' relevant cumulative ranking probabilities that would have otherwise been masked by SUCRA values.

Pharmacological treatments for acute mania

As a second illustrative evaluation, we considered a published NMA comparing the effectiveness of 13 pharmacological treatments and placebo for acute mania in adults for which the data are publicly available.[23] We selected the proportion of patients who responded to treatment, which was a secondary outcome in the study. The network of evidence is given in Figure 6. A random effects NMA model was fitted to the data in a Bayesian framework, where a binomial likelihood and logit link were used to model the proportion of responders and pool the relative effects as log odds ratios.[22] The posterior mean odds ratios measuring relative response to treatment, along with 95% credible intervals, are presented in Figure 7 for each treatment vs. placebo.

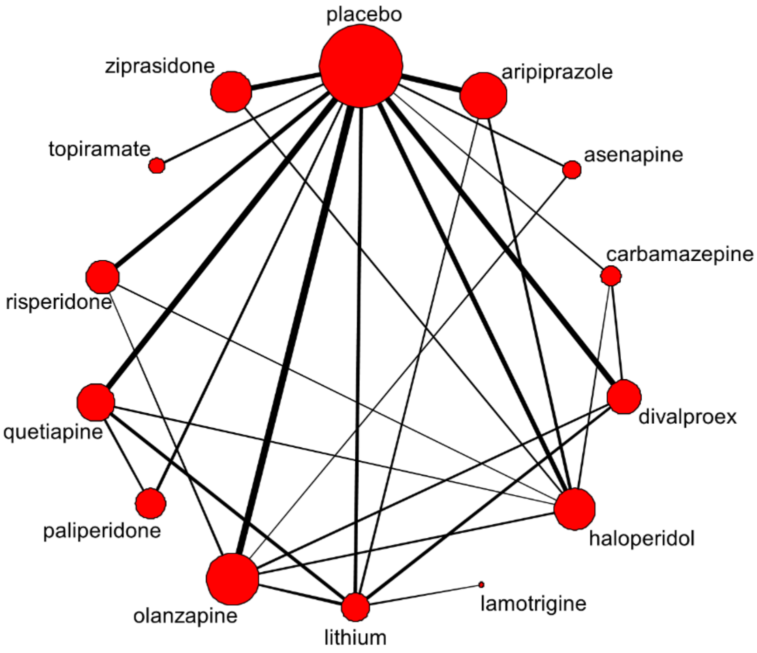


Figure 6. *Network diagram of acute mania dataset. The sizes of the nodes are proportional to the number of patients randomized to the treatments, and the widths of the edges are proportional to the number of studies comparing two nodes.*

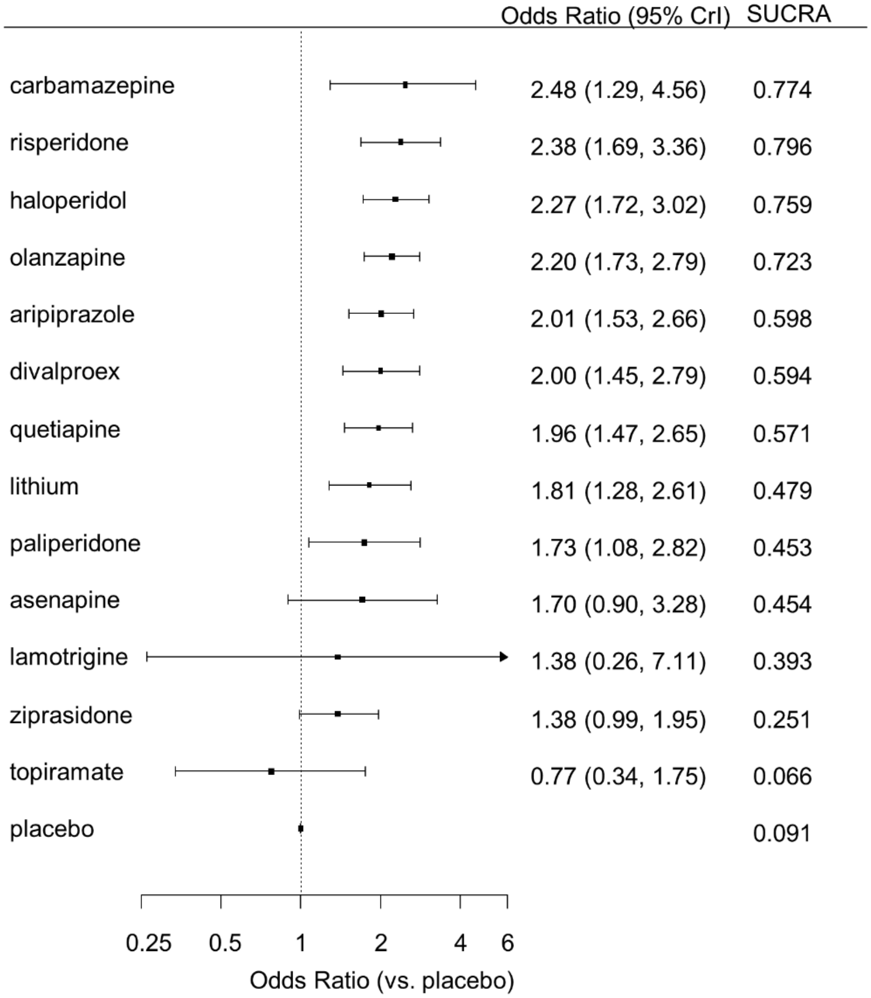


Figure 7: *Summary of the relative effects, presented as odds ratios vs. placebo, and SUCRA of all treatments in Cipriani et al.*[23]

Carbamazepine and risperidone have the first and second largest estimated effect, respectively (Figure 7). However, risperidone's effect estimate is more precise, leading to a slightly larger SUCRA value (0.796) compared to carbamazepine (0.774). Haloperidol has the third largest estimated effect, which is also more precise than that of carbamazepine's and risperidone's. Its SUCRA value (0.759) is within 0.4 of carbamazepine's and risperidone's. These observations are not obvious through the point estimate of SUCRA. Interestingly, the 95% credible interval of olanzapine's estimated effect (1.73, 2.79) is contained within the that of haloperidol (1.72, 3.02), which is contained within risperidone's (1.69, 3.36), which is contained within carbamazepine's (1.29, 4.56). This pattern of overlap does not continue to the treatments of lower estimated effects (eg aripiprazole), and there is a large difference between the SUCRA values of these four treatments and the remaining treatments (at least 0.125). As such, we focus on the rankings of these four treatments.

The ranking probabilities of carbamazepine, risperidone, haloperidol, and olanzapine are depicted by the rank-o-gram and cumulative ranking curves in Figure 8. Each treatment's distribution of ranking probabilities peaks at a rank that agrees with its ranked estimated effect (Figure 7 & 8A). Among these four treatments, carbamazepine, risperidone, haloperidol, and olanzapine have the highest probabilities of ranking best, 2nd best, 3rd best, and 4th best, respectively (Figure 8A). Since carbamazepine's estimated relative effect is slightly less precise compared to risperidone, haloperidol, and olanzapine, it also has slightly higher probabilities of ranking 9th to 12th. This is reflected by the crossing of the cumulative ranking curves (Figure 8B).

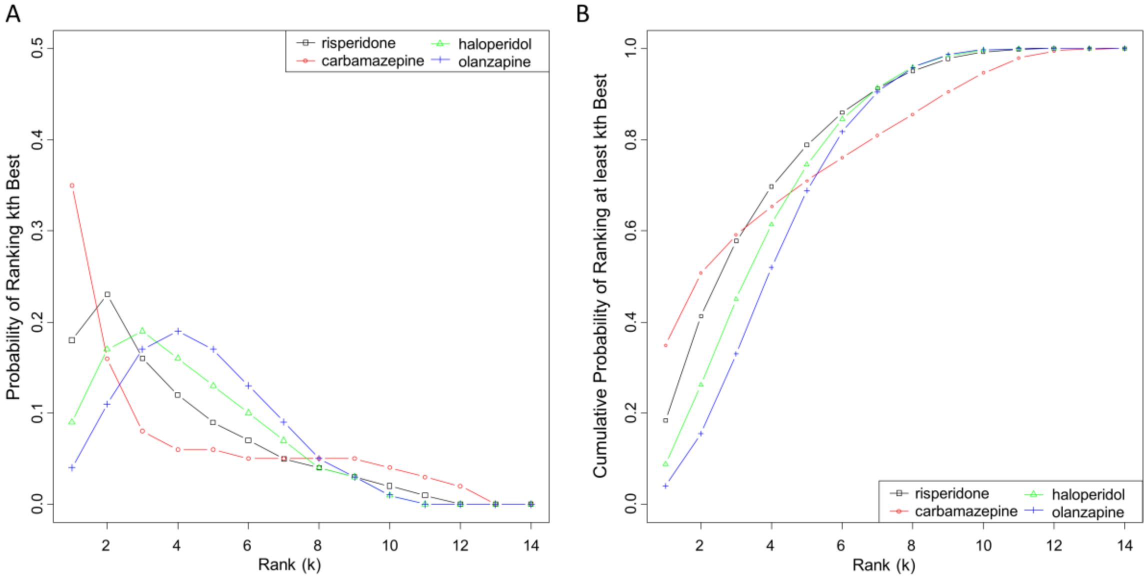


***Figure 8**: Rank-o-gram (A) and cumulative ranking curves (B) for the top four ranking treatments based on SUCRA in Cipriani et al.*[23]

Carbamazepine has the highest cumulative probabilities of ranking at least best, 2nd best, and 3rd best, and hence the largest area under the curve between these ranks (Figure 8B). The cumulative ranking curve of risperidone is the first to cross carbamazepine's curve, followed by haloperidol, then olanzapine. Risperidone has the highest cumulative probabilities of ranking at least 4th, 5th, and 6th best, and its curve is intersected by haloperidol's and olanzapine's around the 7th rank. Nevertheless, the difference between these three treatments' cumulative ranking probabilities are negligible after this rank. All treatments' cumulative ranking curves appear to converge around rank 13. Risperidone's slightly larger SUCRA compared to carbamazepine suggests the difference between the areas under their cumulative ranking curves between ranks 1 and 3 is slightly less than what is it between ranks 4 and 12. Next, we formally compare their partial areas under their cumulative ranking curves through pSUCRA, which is presented for all 14 treatments in Table 2.

***Table 2**: pSUCRA for treatments considered in an NMA presented in Cipriani 2011. The largest pSUCRAs, at any cut-off value, are **bolded**. Note that pSUCRA(p=1) is equal to the probability of ranking best, while pSUCRA(p=13) is equal to SUCRA.*

Treatment	p = 1	p = 2	p = 3	p = 4	p = 5	p = 6	p = 7	p = 8	p = 9	p = 10	p = 11	p = 12	p = 13
carbamazepine	**0.349**	**0.428**	**0.482**	**0.525**	**0.562**	**0.595**	0.626	0.655	0.682	0.709	0.733	0.755	0.774
risperidone	0.184	0.299	0.392	0.468	0.532	0.587	**0.633**	**0.673**	**0.707**	**0.735**	**0.759**	**0.779**	**0.796**
haloperidol	0.088	0.175	0.267	0.354	0.432	0.501	0.560	0.610	0.652	0.686	0.715	0.738	0.759
olanzapine	0.040	0.098	0.175	0.261	0.347	0.425	0.494	0.552	0.600	0.640	0.673	0.700	0.723
aripiprazole	0.019	0.043	0.078	0.122	0.174	0.234	0.297	0.361	0.422	0.477	0.524	0.564	0.598
divalproex	0.027	0.058	0.096	0.141	0.190	0.244	0.302	0.361	0.419	0.473	0.520	0.560	0.594
quetiapine	0.016	0.037	0.067	0.103	0.147	0.199	0.258	0.322	0.385	0.443	0.493	0.536	0.571
lithium	0.009	0.022	0.038	0.059	0.085	0.117	0.157	0.205	0.263	0.325	0.384	0.435	0.479
paliperidone	0.027	0.046	0.067	0.089	0.113	0.140	0.170	0.206	0.249	0.300	0.355	0.407	0.453
asenapine	0.057	0.085	0.109	0.133	0.157	0.182	0.209	0.238	0.272	0.313	0.361	0.410	0.454
lamotrigine	0.182	0.206	0.224	0.238	0.251	0.263	0.275	0.287	0.301	0.317	0.337	0.362	0.393
ziprasidone	0.000	0.001	0.002	0.003	0.005	0.008	0.012	0.021	0.036	0.065	0.120	0.189	0.251
placebo	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.002	0.028	0.091
topiramate	0.001	0.002	0.003	0.004	0.005	0.006	0.007	0.009	0.012	0.016	0.023	0.037	0.066

As expected, carbamazepine has the largest area under its cumulative ranking curve until a cut-off of $p = 6$ (Table 2). This treatment's pSUCRA is larger than all other treatments by at least 0.05 for $p = 1,...,4$. Recall risperidone's cumulative ranking curve was the first to cross carbamazepine's between ranks 3 and 4 (Figure 8B); its pSUCRA did not exceed that of carbamazepine until a cut-off of $p = 7$ (Table 2). After this point, risperidone's and carbamazepine's pSUCRAs were consistently the largest and second largest, respectively, although the difference between the two is less than 0.03 between $p = 7$ and $p = 13$.

Consider the probability of ranking best at $p = 1$. Carbamazepine has a considerably larger probability (0.349) than risperidone, haloperidol, and olanzapine, with corresponding probabilities 0.184, 0.088, and 0.040. The difference between carbamazepine's and haloperidol's probability of ranking best is more than 0.25, yet their difference in SUCRA is 0.015. In addition, despite a difference of 0.165 in carbamazepine's and risperidone's probabilities of ranking best, risperidone ranked best based on SUCRA.

Another treatment, lamotrigine, has the third highest probability of ranking the best (0.182). Lamotrigine had the 11th largest estimated effect vs. placebo, but this estimate was very imprecise, giving rise to a large probability of being the best compared to other treatments with larger estimated effect (eg haloperidol) (Figure 7, Table 2). However, as the cut-offs increase, lamotrigine's pSUCRA increased at a slower rate compared to carbamazepine and risperidone, which have much stronger and precise estimated effects. For instance, at a cut-off of $p = 2$ carbamazepine's and risperidone's pSUCRAs (0.428 and 0.299, respectively) are larger than lamotrigine's by a difference of at least 0.09 (Table

2). Lamotrigine (0.206) is closely followed by haloperidol (0.175), which exceeds lamotrigine when the cut-off is increased to $p = 3$, by an absolute difference of 0.043. Lamotrigine's pSUCRA rank does not agree with its ranked estimated effect (11th) until the cut-off value is at least $p = 11$ (Table 2, Figure 7). This provides some motivation for presenting pSUCRA at all possible cut-offs, which may have otherwise been misleading in terms of lamotrigine at the top cut-offs.

Among the treatments with stronger estimated effects, carbamazepine maintains a considerably larger pSUCRA compared to risperidone at cut-offs $p = 2$ and $p = 3$ (differences of 0.129 and 0.090, respectively) (Table 2). At a cut-off of $p = 4$, the point after which risperidone's cumulative ranking curve intersects carbamazepine's, there is a difference of 0.057 between the two treatments pSUCRA (carbamazepine: 0.525, risperidone: 0.468). This difference becomes increasingly negligible for subsequent cut-offs (less than 0.05). Finally, both carbamazepine's and risperidone's pSUCRAs are least 0.096 larger than haloperidol's for $p = 1,...,5$. These considerable differences among the higher ranks are not evident in their SUCRA values ($p = 13$) alone, where carbamazepine and risperidone's SUCRAs were larger than haloperidol by a difference of 0.015 and 0.037, respectively.

Discussion

In this study, we proposed the partial surface under the cumulative ranking (pSUCRA) curve as an alternative measure for ranking treatments in NMA and explored its properties and interpretations in light of SUCRA. This concept of considering partial curves is adapted from diagnostic medicine, where the partial area under the ROC curve is used to evaluate and compare the accuracy of diagnostic tests. pSUCRA is motivated by instances where 2 or more treatments have similar SUCRA values, despite notable differences in the magnitude and uncertainty of their estimated relative effects as well as their cumulative ranking probabilities of being at least p best. This is exemplified by the crossing of cumulative ranking curves for the treatments.

Through empirical illustrations using simulated data as well as data from a real NMA, we showed situations where pSUCRA is better than SUCRA in identifying treatments with larger probabilities of being among the preferred ranks. For instance, in the simulated example, the treatment with the strongest estimated effect (Treatment 1) had a SUCRA value of 0.809, while the treatment with the fifth strongest estimated effect (Treatment 5) had a SUCRA value of 0.799. Both Treatments 1 and 2 had notably higher probabilities of ranking best, at least 2nd best, at least 3rd best, and at least 4th best compared to Treatment 5, yet there was not much difference in their SUCRA values. This also occurred in the published NMA example on acute mania therapies. Although risperidone and haloperidol had smaller probabilities of ranking best, at least 2nd best, and at least 3rd best compared to carbamazepine, carbamazepine ranked second to risperidone based on SUCRA, and was not much better than haloperidol in terms of SUCRA (0.774 vs. 0.759). Differences among more important cumulative ranking probabilities became apparent when pSUCRA was considered at earlier cut-offs (eg $p = 1, 2, 3, \text{etc}$), thus distinguishing treatments (eg Treatment 1, carbamazepine) that were more likely to be among the p best treatments.

In addition to addressing such limitations of SUCRA, pSUCRA addresses limitations of solely considering the probability of a treatment ranking best. In the published NMA example, the two treatments with the strongest and second strongest estimated effects had the two largest probabilities of ranking the best (ie carbamazepine and risperidone). However, a treatment with the 11th strongest effect had the third largest probability of ranking the best (lamotrigine). This is likely due to the large uncertainty in its effect and the literature has warned about this possibility when using the probability of ranking the best.[24] In such situations, pSUCRA may provide an optimal measure in between the probability of ranking best ($p=1$) and SUCRA ($p=n-1$, where n is the number of treatments in the network). If a cut-off of $p=1$ was selected, then a treatment with a very imprecise effect estimate may appear to be among the best. However, if a cut-off of $p=n-1$ was selected, then differences between estimated effects may be misrepresented by the single value of SUCRA if the corresponding cumulative ranking curves crossed. A cut-off value in between $p=1$ and $p=n-1$ may then address both issues. For example, carbamazepine and risperidone were distinguished as better treatments than lamotrigine, while carbamazepine was notably better than risperidone for cut-off values $p=2,3,4$.

In both examples, the treatment with the second strongest effect estimate ranked better than the treatment with the strongest, but less precise, effect estimate according to SUCRA. Nevertheless, the differences in SUCRA were small in these examples. However, we wonder if there are situations where treatments with much smaller effect estimates have much larger SUCRA values than the intervention with the strongest effect estimate. Hypothetically, this may be more likely to occur in networks with very large numbers of treatments, which is often the case in NMAs involving complex interventions. For instance, a recent study evaluating the comparative effectiveness of falls interventions considered networks of up to 78 interventions involving components such as exercise, diet modification, and surgery.[17] With a larger spectrum of ranks to plot the cumulative ranking probabilities against (eg 1 to 78), there may be more opportunities for cumulative ranking curves to cross and mask important trade-offs. If a treatment's area under the curve among less important ranks (eg 30 to 78) exceeded that of another treatment among more important ranks (eg 1 to 30), then we may be concerned that the more effective treatment is not discerned from the less effective treatment, and may even rank worse.

Related to this, since ranking probabilities are driven by the degree of overlap between the posterior distributions of the relative effects, cumulative ranking curves are more likely to cross when there is overlapping uncertainty in the comparative effects. As such, we predict that this is more likely to happen in sparse networks with larger numbers of treatments. It would therefore be of interest to examine the uncertainty of pSUCRA across networks of varying size, intervention compositions and complexities, and sparseness which are likely to contribute to varying degrees of uncertainty of the relative effects. One advantage of pAUC in diagnostic medicine is increased precision.[13] It is, therefore, important to conduct extensive simulations exploring various optimality characteristics of pSUCRA as well as its sensitivities to different factors.

Here we note pSUCRA is a solution to distinguishing treatments based on the preferred cumulative ranking probabilities (eg of ranking best, at least 2nd best, at least 3rd best, etc). SUCRA provides an objective measure of the trade-off between the estimated magnitude and uncertainty.[8] The same is true for pSUCRA and this was particularly evident in the real NMA example. As noted above, lamotrigine, a treatment with the 11th strongest and very imprecise effect estimate had one of the top three probabilities of ranking best ($p = 1$). As the cut-off increased to $p = 11$, this treatment decreased in rank. In addition, in both examples, the treatments with the strongest effect estimates had varying degrees of precision. Upon examining the pSUCRA of the treatments with the stronger effects, it appeared as though more emphasis was placed on the magnitude or strength of the estimated effects, rather than the precision or uncertainty, as the left portion of the cumulative ranking curve, or p , decreased. As the cut-off grows to consider a larger portion of the curve, including the entire curve (ie SUCRA), more precise estimates are favored, even if the magnitude of the effect is smaller than others. Whether this trade-off reflects a decision maker's preferences is unclear and is worth future investigation.

Another option may be to consider the upper region of a plot of the cumulative ranking curves, instead of or in addition to the left region. This may be done by conditioning pSUCRA on the cumulative ranking probabilities being at least $\alpha,\ 0 \leq \alpha \leq 1$. Treatments with strong, but more precise, estimated effects will have cumulative ranking curves that reach 1 faster than those with less imprecise estimates. This is because the probabilities of ranking the $n^{th}, (n-1)^{th}, (n-2)^{th}, \ldots$ best (ie the less favorable ranking probabilities) are more likely to be greater than 0 for treatments with imprecise effect estimates. Thus, focusing on the area under the cumulative ranking curve once it exceeds α may place more emphasis on the precision of strong effect estimates. Such adaptations of pSUCRA should be explored in the future.

While some of the limitations of SUCRA may be revealed through the inspection of the cumulative ranking curves or the estimated relative effects, it is worth developing a statistic that captures a decision maker's preferences. This is because ranking statistics have been proposed as measures to compare interventions across multiple outcomes on the same scale,[25] and they may be integrated to develop an overall rank. As a result, it is important to make sure the ranking statistics convey the trade-offs a decision maker would have made based on the magnitude and uncertainty of the relative effects. We have also shown that pSUCRA and SUCRA are weighted sums of the ranking probabilities, where the probabilities of better ranks receive more weight. In the future, it is of interest to consider different weighting schemes to reflect decision makers' preferences.

Conclusion

The debate on how to best present NMA results is ongoing, especially in regard to the presentation of treatments ranks. Here, we have introduced the idea and highlighted the advantages of considering a portion of the cumulative ranking curve. We have the laid the groundwork for calculating the partial area under this curve among more favorable ranks and have discussed plausible adaptions that may capture various trade-offs in the magnitude and uncertainty of estimated effects. Further studies with a wide range of empirical and simulated data are required to establish the characteristics of pSUCRA as well as its comparative optimality with SUCRA and other ranking statistics.

References

1. Caldwell DM, Ades AE, Higgins JPT. Simultaneous comparison of multiple treatments: combining direct and indirect evidence. *BMJ.* 2005;331:897-900.

2. Lu G, Ades A. Combination of direct and indirect evidence in mixed treatment comparisons. *Stat Med.* 2004;23:3105-3124.

3. Salanti G, Ades A, Ioannidis J. Graphical methods and numerical summaries for presenting results from multiple-treatment meta-analysis: an overview and tutorial. *J Clin Epidemiol.* 2011;64(2):163-171.

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

5. Mbuagbaw L, Rochwerg B, Jaeschke R, et al. Approaches to interpreting and choosing the best treatments in network meta-analyses. *Syst Rev.* 2017;6(1):79-83.

6. Mavranezouli I, Megnin-Viggars O, Daly C, et al. Psychological treatments for post-traumatic stress disorder in adults: a network meta-analysis. *Psychol Med.* 2020;50(4):542-555.

7. Dias S, Ades AE, Welton NJ, Jansen JP, Sutton AJ. *Network Meta-analysis for Decision Making.* Hoboken, NJ: Wiley; 2018.

8. Rücker G, Schwarzer G. Ranking treatments in frequentist network meta-analysis works without resampling methods. *BMC Med Res Methodol.* 2015;15(1):28.

9. Mallett S, Halligan S, Thompson M, Collins G, Altman D. Interpreting diagnostic accuracy studies for patient care. *BMJ.* 2012;345:e3999.

10. Zhou X, Obuchowski N, McClish D. *Statistics in Diagnostic Medicine*. 2nd ed. Hoboken, NJ: Wiley; 2011.

11. Bamber D. The area above the ordinal dominance graph and the area below the receiver operating characteristic graph. *J Math Psychol.* 1975;12(4):387-415.

12. McClish D. Analyzing a portion of the ROC curve. *Med Decis Making.* 1989;9:190-195.

13. Ma H, Bandos A, Rockette H, Gur D. On use of partial area under the ROC curve for evaluation of diagnostic performance. *Stat Med.* 2013;32:3449-3458.

14. Stewart J. *Calculus: Early Transcendentals*. 6th ed. Belmont, CA: Thomson Higher Education; 2008.

15. Metz C. ROC methodology in radiologic imaging. *Invest Radiol.* 1986;21:720-733.

16. Jiang Y, Metz C, Nishikawa R. A receiver operating characteristic partial area index for highly sensitive diagnostic tests. *Radiology.* 1996;201:745-750.

17. Tricco AC, Thomas SM, Veroniki AA, et al. Comparisons of interventions for preventing falls in older adults: A systematic review and meta-analysis. *JAMA.* 2017;318(17):1687-1699.

18. Lunn DJ, Thomas A, Best N, Spiegelhalter D. WinBUGS -- a Bayesian modelling framework: concepts, structure, and extensibility. *Stat Comput.* 2000;10:325-337.

19. R Core Team: R: A language and environment for statistical computing. Vienna, Austria: R Foundation for Statistical Computing; 2020. https://www.R-project.org/.

20. Viechtbauer W. Conducting meta-analyses in R with the metafor package. *J Stat Softw.* 2010;36(3):48.

21. 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(2):80-97.

22. Dias S, Welton NJ, Sutton AJ, Ades AE. *NICE DSU* Technical Support Document 2: A generalised linear modelling framework for pair-wise and network meta-analysis of randomised controlled trials. 2011.

23. Cipriani A, Barbui C, Salanti G, et al. Comparative efficacy and acceptability of antimanic drugs in acute mania: a multiple-treatments meta-analysis. *Lancet.* 2011;378(9799):1306-1315.

24. Jansen J, Trikalinos T, Cappelleri J, 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 Health.* 2014;17(2):157-173.

25. Naci H. Communication of treatment rankings obtained from network meta-analysis using data visualization. *Circ Cardiovasc Qual Outcomes.* 2016;9(5):605-608.

Chapter 5

Spie charts for quantifying treatment effectiveness and safety in multiple outcome network meta-analysis: A proof-of-concept study

Caitlin H Daly[1,2], Lawrence Mbuagbaw[2,3], Lehana Thabane[2,3],
Sharon E Straus[4,5], Jemila S Hamid[1,6]

[1] Department of Health Research Methods, Evidence, and Impact, McMaster University

[2] Population Health Sciences, Bristol Medical School, University of Bristol

[3] Biostatistics Unit, Father Sean O'Sullivan Research Centre, St Joseph's Healthcare Hamilton

[4] Knowledge Translation Program, Li Ka Shing Knowledge Institute, St. Michael's Hospital

[5] Department of Medicine, Faculty of Medicine, University of Toronto

[6] Department of Mathematics and Statistics, University of Ottawa

Citation: Daly, C. H., Mbuagbaw, L., Thabane, L., Straus, S. E., & Hamid, J. S. (2020b). Spie charts for quantifying treatment effectiveness and safety in multiple outcome network meta-analysis: A proof-of-concept study. *Submitted to BMC Medical Research Methodology*. https://doi.org/10.21203/rs.3.rs-36139/v1

Abstract

Background: Network meta-analysis (NMA) simultaneously synthesises direct and indirect evidence on the relative efficacy and safety of at least three treatments. A decision maker may use the coherent results of an NMA to determine which treatment is best for a given outcome. However, this evidence must be balanced across multiple outcomes. This study aims to provide a framework that permits the objective integration of the comparative effectiveness and safety of treatments across multiple outcomes.

Methods: In the proposed framework, measures of each treatment's performance are plotted on its own pie chart, superimposed on another pie chart representing the performance of a hypothetical treatment that is the best across all outcomes. This creates a spie chart for each treatment, where the coverage area represents the probability a treatment ranks best overall. The angles of each sector may be adjusted to reflect the importance of each outcome to a decision maker. The framework is illustrated using two published NMA datasets comparing dietary oils and fats and psoriasis treatments. Outcome measures are plotted in terms of the surface under the cumulative ranking curve. The use of the spie chart was contrasted with that of the radar plot.

Results: In the NMA comparing the effects of dietary oils and fats on four lipid biomarkers, the ease of incorporating the lipids' relative importance on spie charts was demonstrated using coefficients from a published risk prediction model on coronary heart disease. Radar plots produced two sets of areas based on the ordering of the lipids on the axes, while the spie chart only produced one set. In the NMA comparing psoriasis treatments, the areas inside spie charts containing both efficacy and safety outcomes masked critical information on the treatments' comparative safety. Plotting the areas inside spie charts of the efficacy outcomes against measures of the safety outcome facilitated simultaneous comparisons of the treatments' benefits and harms.

Conclusions: The spie chart is more optimal than a radar plot for integrating the comparative effectiveness or safety of a treatment across multiple outcomes. Formal validation in the decision-making context, along with statistical comparisons with other recent approaches are required.

Background

Health technology assessments and clinical guidelines are increasingly supported by evidence synthesised through network meta-analysis (NMA) [1, 2]. The main output from an NMA is a coherent set of relative treatment effects, based on pooled direct and indirect evidence typically contributed by randomised controlled trials (RCTs) [3, 4]. The estimated treatment effects relative to a common comparator may then be used to inform a ranked list of treatments, from which knowledge users may be able to deduce which treatment is best for a given clinical problem.

Interpreting NMA results is challenging, particularly as the number of treatments and outcomes increase. Several pieces of literature have aimed to ease the interpretative burden of NMA. For example, three graphical tools were developed to display key features of an NMA (i.e. relative effects and their uncertainty, probabilities of ranking best, and between-study heterogeneity) for a single outcome [5]. The rank heat plot has been proposed as a visual tool for presenting NMA results across multiple outcomes [6]. However, knowledge users could also benefit from the quantification of the overall integrated results across multiple outcomes to facilitate interpretation in a more objective way. This is particularly important in situations where the comparative rankings of treatments on each outcome contradict each other.

Radar plots are often used as a visualisation tool to communicate multivariate data [7]. Recently, they have been used to visually compare the surface under the cumulative ranking curves (SUCRAs) in an NMA evaluating multiple interventions for relapsing multiple sclerosis [8]. Another NMA on dual bronchodilation therapy for chronic obstructive pulmonary disease has compared the area within radar charts of SUCRA values to deliver a single ranking of their efficacy-safety profile [9]. However, in this NMA, the quantification of the area weighed each outcome equally, which may not reflect a knowledge user's preferences. The use of radar plots for the purpose of comparing the overall performance of treatments is also limited by the fact that the area depends on the ordering of the outcomes on the plot. For this reason, the spie chart has been suggested as a better alternative [10].

A spie chart is a combination of two pie charts, where one is superimposed on another, allowing comparisons between two groups on multiple attributes [11]. For example, in the context of NMA, this could be the comparison of a treatment against a hypothetical treatment that is uniformly the best across multiple outcomes. The former's area will be a fraction of the latter's, thereby facilitating the comparison of multiple treatments in a manner similar to comparing areas on a radar chart.

To address the limitations of the aforementioned NMA summary tools, the objective of this paper is to lay the groundwork for conceptualising a treatment's likelihood of being the best overall in terms of its coverage area inside a spie chart. This circular plot may be divided into segments representing a treatment's level of efficacy or safety for each outcome. We provide a methodological framework and assess the feasibility of adapting the area on a spie chart to numerically integrate the efficacy and safety of treatments estimated by NMAs of multiple outcomes. Since radar plots have not been formally investigated and generalised for NMA, we also present the area on a radar plot and compare it to that of spie charts. We illustrate how the spie chart may be used to overcome the limitations of the radar plot, as well as its flexibility for further adaptations.

Methods

Measuring the coverage area inside a spie chart

Consider for example a situation where the performance of a treatment has been measured in terms of $J = 8$ outcomes valued between 0 and 1. Simulated values are plotted on a spie chart in Figure 1. In general, the resulting shape on any spie chart is a series of J sectors, each with their own radius equal to the value of the J outcome measures. The area covered by these sectors may be calculated as the sum of the areas of the individual sectors.

In Figure 1, the shaded area, A, is the sum of the area of the 8 sectors, $A_j, j = 1,\ldots,8$

$$A_j = \tfrac{1}{2}\theta_j y_j^2,$$

where y_j and θ_j are the radius and angle of sector j, respectively. In Figure 1, all angles are equal, i.e., $\theta_j = \frac{2\pi}{8} = \frac{\pi}{4}$ radians, and the shaded area on the spie chart is then:

$$A = \frac{\pi}{8}\sum_{j=1}^{8} y_j^2,$$

which is an average of the squared values of the 8 outcomes, scaled by a factor of π. In general, the shaded area within a spie chart informed by $J \geq 1$, outcomes for an intervention is

$$A = \frac{1}{2}\sum_{j=1}^{J}\theta_j y_j^2.$$

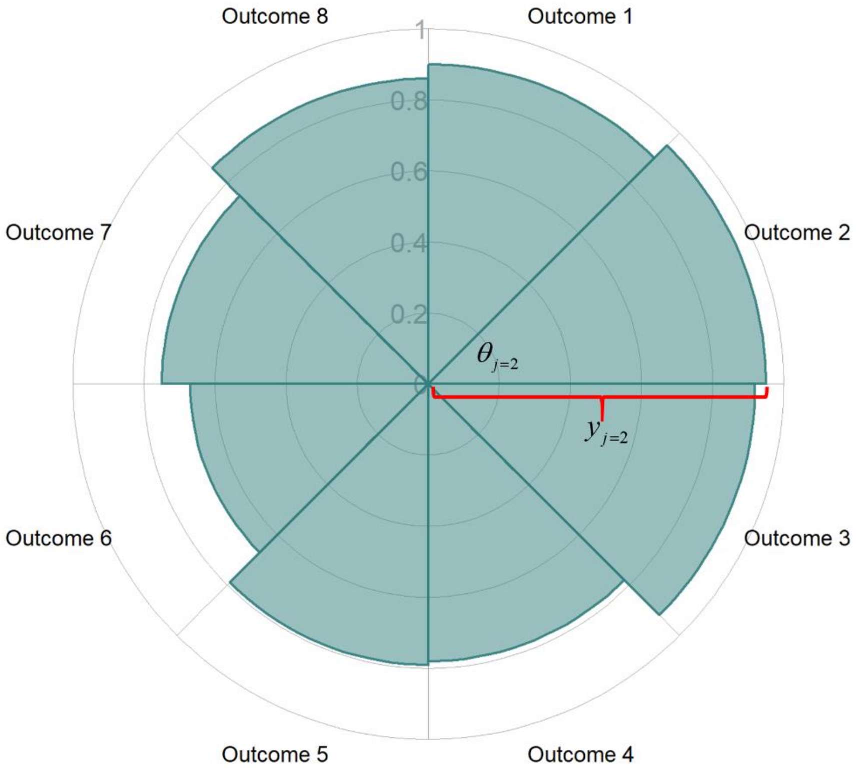


Figure 1: *Example spie chart informed by the values of 8 outcomes. To calculate the area of sector* $j = 2$*, the required parameters are denoted:* $\theta_{j=2}$ *is a known angle,* $y_{j=2}$ *is the radius of sector* $j = 2$*, equal to the value of outcome 2.*

Choice of outcome measure

To enable a fair comparison of the areas defined by the treatments' performances across multiple outcomes, the outcomes should be plotted on the same or comparative scales. This is not the case in most NMA studies involving multiple outcomes. As such, ranking probabilities and their summaries (e.g. the Surface Under the Cumulative RAnking curve (SUCRA) or P-score) may provide valid measures for this purpose [12, 13]. These measures transfer the comparative relative effects to a scale between 0 and 1. Alternatively, the posterior mean or median ranks may be used. However, note that the probability of a treatment ranking best should be avoided because treatments with high uncertainty around their estimated effects are likely to be ranked best [14], and this ranking probability has the potential to be biased [15]. SUCRA values, which are calculated in a Bayesian framework, provide a less sensitive and less biased alternative to rank treatments. The posterior mean rank is a scaled transformation of SUCRA, and the P-score is its frequentist equivalent [13]. These measures account for the uncertainty of a treatment's relative effect and are thus preferable [12].

Another option may be to use the absolute probabilities of response or risk for each treatment, as was done in a multicriteria decision analysis of statins [16]. Note that NMA pools relative effects such as log-odds ratios. To obtain estimates of the absolute probabilities for all treatments, an estimate of the absolute effect (e.g., log-odds) of a treatment in a contemporary population of interest must be assumed. This may be any treatment in the network [2]. The assumed absolute effect of this treatment would then be applied to the relative effects (e.g., log-odds ratio) vs. the chosen treatment, to obtain estimates of absolute effects (e.g., log-odds) of all other treatments, which may be subsequently converted into probabilities [2, 17]. If an NMA pools evidence on important outcomes measured on a continuous scale, response rates may be estimated [18] or standardised mean differences may be converted to log-odds ratios [19], provided that the underlying assumptions of these conversions are reasonable for the data. Note that plotting absolute probabilities of response or risk would limit the generalisability of the area to the population informing the assumed absolute effect of the chosen reference treatment.

In this paper, to simplify the presentation of our novel methodological framework, we use SUCRA values as a measure of the comparative effectiveness and safety of the treatments. We would like to highlight that this choice is made without loss of generality and the method is valid for any other measure.

Standardised area inside a spie chart

To facilitate interpretation of the coverage area inside a spie chart, we standardise it by the maximum possible area. Its interpretation would then be comparable to the interpretation of SUCRA [12]. As such, the minimum possible standardised area of 0 corresponds to a treatment that always ranked the worst (i.e., SUCRA = 0 across all outcomes). The maximum possible standardised area of 1 corresponds to a treatment that always ranked best for each outcome (i.e., SUCRA = 1 across all outcomes).

First, consider the maximum possible area of each sector in a spie chart defined by SUCRA,

$$A_j^{\max} = \frac{1}{2}\theta_j\left(1\right)^2 = \frac{\theta_j}{2}.$$

If there are J outcomes and the angles of each sector are equal, i.e., $\theta_j = \frac{2\pi}{J}, j = 1, ..., J$, then the maximum possible area on a spie chart is

$$A^{\max} = \frac{1}{2}\sum_{i=1}^{J}\left(\frac{2\pi}{J}\right) = \frac{J}{2}\left(\frac{2\pi}{J}\right) = \pi.$$

In fact, regardless of the size of the individual sector's angles θ_j , as long as $\max\left\{y_j\right\}=1$, the spie chart consists of a unit circle. Consequently, $A^{\max}=\pi$, for all $0 \leq \theta_j \leq 2\pi$. Therefore, in general, the standardised area on a spie chart is

$$A^{std}=\frac{1}{2\pi}\sum_{j=1}^{J}\theta_j y_j^2$$

where y_j and θ_j are the SUCRA value and angle of the sector corresponding to outcome j , respectively. Note that $0 \leq \theta_j \leq 2\pi$, where $\theta_j = 0$ implies outcome j does not contribute the area, and $\theta_j = 2\pi$ implies outcome j is the sole contributor to the area. In the case of equal angles, the standardised area on a spie chart for a given treatment is a weighted average of the squared SUCRA values:

$$A^{std}=\frac{1}{J}\sum_{j=1}^{J}y_j^2.$$

Incorporating stakeholder preferences

An advantage of the spie chart's circular design is the ability to incorporate preferences of the knowledge user. Some outcomes may be more important than others, and this can be incorporated in the plots by adjusting the contribution each outcome has to the overall area. By adjusting the angles of the sectors in a spie chart, we can adjust the proportion of the entire plot each sector covers. Noting that the sum of the angles in a spie chart must be 2π , given a set of weights, $w_j,\ j=1,...,J$ for a set of J outcomes, the corresponding angles may be calculated as

$$\theta_j = \frac{2\pi w_j}{\sum_{j=1}^{J} w_j}.$$

There are various ways to derive the contribution of the outcomes in terms of weights, which may be informed by preferences supported by evidence in the literature or through weights elicited from knowledge users themselves. For example, risk-prediction or prognostic models may be used to inform the weights of outcomes when the goal is to reduce the risk of an unfavourable event or disease such as cardiovascular disease (CVD). If all outcomes are included in a regression model, and measured in the same units, the magnitude of the unstandardised coefficients, $\beta_j, j = 1, ..., J,$ capture the influence each outcome has on the overall risk, adjusted for any additional factors included in the model:

$$w_j = \frac{|\beta_j|}{\sum_{j=1}^{J} |\beta_j|}.$$

If the outcomes are measured on different scales, then standardised coefficients may be considered. There are more optimal approaches to deriving the relative importance of predictors (e.g., outcomes) when individual patient data (IPD) are available to create multiple regression models [20]. In fact, the use of standardised coefficients for this purpose has been criticised because the dependencies between predictors are not fully taken into account [21]. Nevertheless, researchers undertaking NMA often have limited resources in terms of time and access to IPD, and thus have to make secondary use of aggregate or summary level data.

If there are important dependencies between the outcomes, this should be accounted for at the synthesis stage. There are several approaches available for this and guidance is provided by López-López and colleagues [22] and multi-parameter evidence synthesis methods should also be considered [2]. Nevertheless, if there is a need to avoid double counting the contribution of related outcomes, and we know the correlations between them, the contribution of each outcome to the overall area can be adjusted. The weights of each outcome may be informed by a $J \times J$ correlation matrix, denoted as

$$\begin{matrix} 1 & \rho_{12} & \rho_{13} & \cdots & \rho_{1J} \\ \rho_{12} & 1 & \rho_{23} & \cdots & \rho_{2J} \\ \rho_{13} & \rho_{23} & 1 & \cdots & \rho_{3J} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \rho_{1J} & \rho_{2J} & \rho_{3J} & \cdots & 1 \end{matrix} .$$

However, since correlation can be negative, the squares of the pairwise correlations, i.e., the coefficients of determination, i.e., $R_{ij}^2 = \rho_{ij}^2$, should be used instead. The weight of each outcome can then be proportional or equal to the marginal sums of the squared correlation matrix:

$$w_j = \sum_{i=1}^{J} \rho_{ij}^2, i = 1, ..., J .$$

Finally, there are several methods for eliciting preferences from decision makers, such as direct rating, where the decision makers rate outcomes on a scale from 1 to 100 and weights are derived by normalising these scores [23]. Regardless of the method to inform the weights, the application of the proposed framework remains the same.

Selecting outcomes to inform the area

The number of outcomes that may be plotted on a spie chart ranges from one to infinity. Nevertheless, a knowledge user would not benefit from either extreme. The purpose of the spie chart is to facilitate the combination of multiple outcomes, accounting for the desired contribution of the overall summary. As such, a minimum of two outcomes is sensible for this purpose. Plotting an overwhelming number of outcomes will not be visually appealing, although the area inside a spie chart is intended to overcome the visual interpretative burden. Increasing the number of outcomes will limit the contribution of important outcomes, to a degree that depends on the weights. Researchers presenting results of an NMA should be wary of this, though they do need not restrict themselves to a maximum number of outcomes.

Outcomes which are critical to the decision-making process should be plotted on the spie chart. For example, any outcomes for which lack of evidence would exclude a treatment from consideration should be plotted. Evidence for any plotted outcome should be available for every treatment under consideration. It is important that every treatment is compared based on the same set of outcomes. If evidence on a critical outcome is not available for a treatment within a decision set, then imputation methods may be considered at the NMA stage [24].

Efficacy and safety outcomes should be plotted on two separate spie charts for each treatment, as it is important for a knowledge user to recognise that a very effective outcome may not be safe. Plotting these on the same spie chart and summarising the area inside as a single numerical value may mask important information on harms. A knowledge user should be able to simultaneously compare both the benefits and the harms of a treatment. This is possible by plotting the area inside an efficacy spie chart against the area inside a safety spie chart on a scatter plot [25]. Points towards the upper right quadrant of a scatterplot (e.g., towards (1,1)) would represent the most efficacious and safe treatment.

Results

In this section, our proposed framework is illustrated using results from two published reviews [26, 27]. At the same time, we empirically compare the use of the spie chart and the radar plot for quantifying a treatment's overall performance. The formula for the standardised area inside a radar plot has been derived in Additional File 1.

The two reviews used in this section were selected as all interventions had complete outcome information, and they highlight conceptual issues that drive the development of this framework. The first example illustrates one way of weighting outcomes of unequal importance to reflect the preferences of decision makers. Since there are four outcomes in this example, there are different ways to order the outcomes on a radar plot, and we show how this impacts the area inside a radar chart. In the second example, there are three outcomes, and thus one unique ordering of the outcomes which allows us to fairly compare the areas inside the radar plot and spie chart. The second example also underlines the importance of considering efficacy and safety outcomes separately. We emphasize here that any observations made in these examples are for illustrative purposes only and should not impact clinical practice.

Lipids study

The effects of thirteen dietary oils and fats on total cholesterol (TC), low-density lipoprotein cholesterol (LDL-c), high-density lipoprotein cholesterol (HDL-c) cholesterol, and triglycerides (TG), were investigated by Schwingshackl and colleagues [26]. Blood tests measuring these lipoproteins are carried out to assess a person's risk for cardiovascular disease (CVD) [28]. The NMAs in this review pooled data from RCTs on thirteen treatments for the four outcomes of interest, and the SUCRA values are listed in Table 1. There is no treatment that clearly ranks the best across all outcomes.

Note that, lower values of TC, LDL-c, and triglycerides are preferred, while higher values of HDL-c are preferred. The SUCRA values were computed in this NMA so that higher values of SUCRA reflect the preferred direction (i.e., improvement) of the outcomes. This is important when plotting outcomes on a spie chart, so that larger areas reflect treatments that are better at improving outcomes.

Table 1: *SUCRA values and rankings produced based on all trials included in [26].*

	Outcome							
	Total Cholesterol[a]		LDL-c[a]		HDL-c[b]		Triglycerides[a]	
Treatment	SUCRA	Rank	SUCRA	Rank	SUCRA	Rank	SUCRA	Rank
Safflower oil	0.90	1	0.82	1	0.06	13	0.68	3
Rapeseed	0.85	2	0.76	2	0.53	7	0.58	7
Sunflower oil	0.72	4	0.71	4	0.57	5	0.61	6
Corn oil	0.72	4	0.66	6	0.29	11	0.66	4
Hempseed oil	0.61	5	0.69	5	0.59	4	0.63	5
Soybean oil	0.59	7	0.50	8	0.13	12	0.72	2
Flaxseed oil	0.59	7	0.71	4	0.47	9	0.56	8
Olive oil	0.43	8	0.37	9	0.52	8	0.32	10
Beef fat	0.41	9	0.50	8	0.74	3	0.06	13
Palm oil	0.34	10	0.33	11	0.80	2	0.74	1
Coconut oil	0.22	11	0.33	11	0.88	1	0.29	11
Lard	0.11	12	0.10	12	0.55	6	0.50	9
Butter	0.03	13	0.02	13	0.37	10	0.17	12

[a] Higher values of SUCRA reflect treatments that are better in terms of reducing levels of these lipids.
[b] Higher values of SUCRA reflect treatments that are better in terms of increasing levels of this lipid.

Spie chart

To compare the rankings of areas inside a spie chart, we first calculated the standardised areas, assuming equal weights i.e., equal angles (Table 2). The area corresponding to the spie chart for Safflower oil is displayed in Figure 2A. The percentages of the unit circle covered by the shaded areas for each treatment are small (Table 2), indicating that there is no treatment which is certainly the best across all outcomes. Knowing this, a stakeholder may then direct their attention to differences, if any, between treatments for more important outcomes.

Table 2: *Standardised areas inside spie charts of SUCRA values from multiple outcomes.*

	Outcomes weighted equally		Outcomes weighted for women 50 - < 65 years[a]	
Treatment	Standardised Area	Rank	Standardised Area	Rank
Safflower oil	0.150	1	0.597	1
Rapeseed	0.147	2	0.566	2
Sunflower oil	0.132	3	0.469	3
Corn oil	0.113	5	0.409	4
Hempseed oil	0.123	4	0.406	5
Soybean oil	0.087	8	0.271	7
Flaxseed oil	0.107	7	0.378	6
Olive oil	0.053	11	0.176	11
Beef fat	0.075	10	0.248	8
Palm oil	0.109	6	0.237	9
Coconut oil	0.078	9	0.196	10
Lard	0.044	12	0.080	12
Butter	0.013	13	0.026	13

[a] Outcomes are weighted differently according to a coronary heart disease risk prediction model for women aged 50 - < 65 years [29].

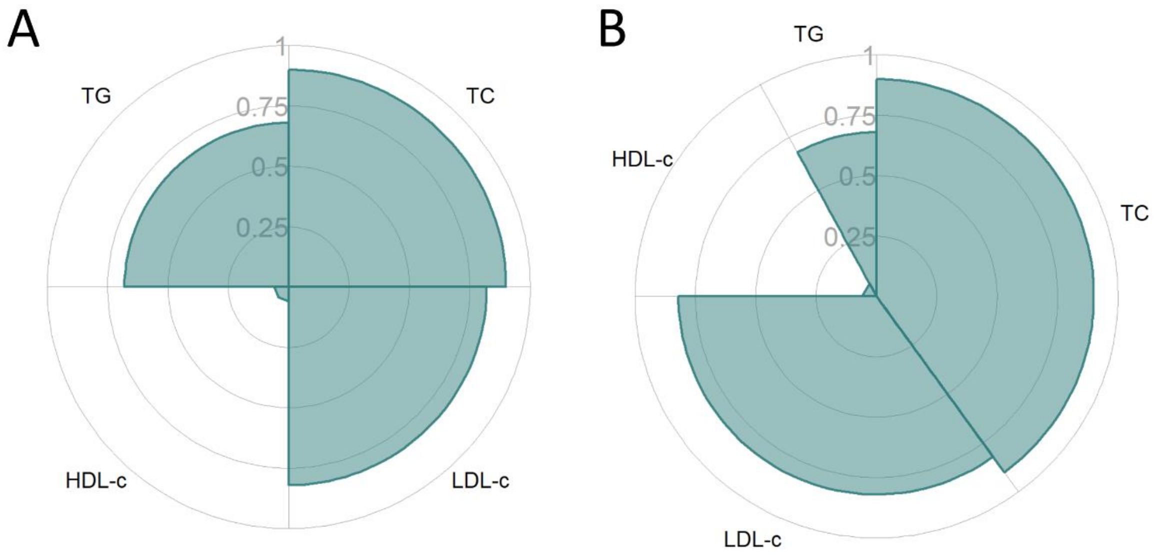


Figure 2: *Two possible spie plots of the SUCRA values corresponding to Safflower oil in [26]. The plot in (A) weighs each outcome equally, since they have the same angles. The plot in (B) weighs the outcomes based on a coronary heart disease risk prediction model for women aged 50 - < 65 years [29].*

For illustrative purposes, we make use of a multivariable regression model built by Castelli et al., where the outcome was coronary heart disease (CHD) [29]. This model was informed by data from the Framingham Study, and the reported regression coefficients for TC, LDL-c, HDL-c, and TG, for women aged 50 to less than 65 years old are 2.51, 2.19, -1.05, 0.48, respectively. These coefficients are adjusted for systolic blood pressure, glucose, and cigarette smoking status. We can calculate weights for each outcome based on the absolute values of these coefficients. For TC,

$$w_1 = \frac{|2.51|}{|2.51| + |2.19| + |-1.05| + |0.48|} = 0.40 .$$

The weights for LDL-c, HDL-c, and TG are 0.35, 0.17, and 0.08, respectively. The angle corresponding to TC may then be calculated as

$$\theta_1 = \frac{2\pi(0.40)}{1} = 0.8\pi .$$

The angles for LDL-c, HDL-c, and TG are 0.7π, 0.34π, and 0.16π, respectively. The resulting area corresponding to the spie chart for Safflower oil is displayed in Figure 2B. The standardised areas and ranks for each treatment, tailored to women aged 50 to less than 65 years are provided in Table 2. Based on this weighting scheme, the best treatment for reducing a 50 - < 65-year-old woman’s risk for CHD by improving lipid levels is Safflower oil.

Radar plot

In this example, there are four outcomes and thus four radii defining the radar plot. When using a radar plot, one must decide the order of the outcomes around the plot. The placement of the first outcome does not matter, but the ordering of the remaining $J-1$ outcomes will impact the area enclosed in the radar plot [10]. There are $\frac{1}{2}\left(J-1\right)!$ options to order outcomes around a circle. The $\frac{1}{2}\left(4-1\right)!=3$ possible orderings of outcomes in this example are displayed in Supplementary Figure 2 in Additional File 2 for a single intervention, Safflower oil.

Summary of Findings tables in Cochrane Reviews may provide some guidance on how to order the outcomes, since the outcomes are listed in decreasing order of importance [30]. Of course, this importance ordering will vary across different stakeholders. For example, one Cochrane Review examining the effectiveness of a Mediterranean-style diet in preventing CVD has listed the decreasing order of importance of the four lipids as TC, LDL-c, HDL-c, TG [31]. Another Cochrane Review examining the effectiveness of polyunsaturated fatty acids in preventing CVD orders the importance of the lipids as TC, TG, HDL-c, LDL-c [32]. Nevertheless, these separate orderings will produce the same area, assuming the angles between the outcomes are equal, $\theta_j = \frac{2\pi}{4} = \frac{\pi}{2}, j = 1,2,3,4$. For example, the areas enclosed in the radar plots for Safflower oil (Supplementary Figure 2A&B, Additional File 2), based on the formula provided in Additional File 1, are

$$
\begin{aligned}
A^{std}_{\text{Safflower, Rees}} &= \frac{1}{4}\left(y_{TC}y_{LDL-c} + y_{LDL-c}y_{HDL-c} + y_{HDL-c}y_{TG} + y_{TG}y_{TC} \right) \\
&= \frac{1}{4}\left(y_{TC}y_{TG} + y_{TG}y_{HDL-c} + y_{HDL-c}y_{LDL-c} + y_{LDL-c}y_{TC} \right). \\
&= A^{std}_{\text{Safflower, Abdelhamid}}
\end{aligned}
$$

There is only one other ordering that will produce a unique area: TC, HDL-c, LDL-c, TG. This is because of the triangles formed by TC & HDL-c and LDL-c & TG; these outcomes were not congruent in the plots generated by Rees' and Abdelhamid's orderings (Supplementary Figure 2C, Additional File 2). The standardised areas produced by these two ordered datasets, assuming equal angles, are provided in Table 3. The rankings of some of the treatments change, and although the differences between standardised areas may seem trivial in this example, this is an important feature of radar plots to highlight, as the differences could be exacerbated in other applications. For example, the areas for one treatment may be quite different if the outcomes are arranged in such a way that those reflecting higher scores alternate with those that have lower scores vs. an ordering where all outcomes with high scores are placed beside together.

Table 3: *Standardised areas inside radar plots of SUCRA values from multiple outcomes.*

	Ordering A[a]		Ordering B[b]	
Treatment	Standardised Area	Rank	Standardised Area	Rank
Safflower oil	0.360	4	0.318	6
Rapeseed	0.462	1	0.447	1
Sunflower oil	0.426	2	0.422	2
Corn oil	0.333	6	0.328	5
Hempseed oil	0.396	3	0.397	3
Soybean oil	0.220	8	0.232	8
Flaxseed oil	0.337	5	0.335	4
Olive oil	0.164	10	0.168	10
Beef fat	0.161	11	0.182	9
Palm oil	0.305	7	0.258	7
Coconut oil	0.171	9	0.161	11
Lard	0.099	12	0.055	12
Butter	0.019	13	0.007	13

[a] Ordering A: TC, LDL-c, HDL-c, TG
[b] Ordering B: TC, TG, HDL-c, LDL-c

Psoriasis study

The effectiveness and safety of seven biologic therapies plus placebo for treating psoriasis were investigated by Jabbar-Lopez and colleagues to support the development of a guideline [27]. Randomised trials informed the NMAs, which synthesised evidence on the following outcomes measuring efficacy: clear/nearly clear skin (defined as minimum residual activity, Psoriasis Area and Severity Index (PASI) > 90, or 0 or 1 on physician's global assessment), mean change in dermatology life quality index (DLQI), and PASI 75 (defined as PASI > 75). The first 2 outcomes were deemed "critical" outcomes by the guideline development group, while the latter outcome, PASI 75, was deemed "important". An additional outcome measuring safety, referred to in the review as tolerability or withdrawal due to adverse events, was also deemed an "important" outcome. For illustrative purposes, the published SUCRA values corresponding to each treatment under investigation are displayed in Table 4.

Table 4: *SUCRA values and rankings produced based on all trials included in [27].*

	Outcome							
	Clear/nearly clear		Mean change in DLQI		PASI 75		Withdrawal due to adverse events	
Treatment	SUCRA	Rank	SUCRA	Rank	SUCRA	Rank	SUCRA	Rank
Ixekizumab	0.99	1	0.70	3	0.96	1	0.14	7
Secukinumab	0.85	2	0.85	1	0.79	3	0.80	3
Infliximab	0.67	3	0.80	2	0.81	2	0.04	8
Ustekinumab	0.60	4	0.70	4	0.52	4	0.82	1
Adalimumab	0.46	5	0.51	5	0.49	5	0.81	2
Etanercept	0.28	6	0.31	6	0.28	6	0.46	6
Methotrexate	0.15	7	0.15	7	0.15	7	0.47	4
Placebo	0.00	8	0.00	8	0.00	8	0.47	5

As was the case in the lipids example, there is no treatment that is universally the best according to SUCRA across all outcomes. Ixekizumab has the largest SUCRA value in terms of the critical “Clear/nearly clear” outcome, but it is not the best in terms of the other critical outcome, mean change in DLQI. It also ranks the second worse in terms of tolerability, highlighting the importance of considering efficacy and safety outcomes separately.

Radar plot vs. spie chart

For illustrative purposes, we first combine the evidence on the three efficacy outcomes (clear/nearly clear, DLQI, PASI 75), considering them to be of equal importance (although the guideline committee suggested otherwise [27]) on both the spie chart and the radar plot.

Since there are only three outcomes, there is only one way to arrange the order of the outcomes, and thus one unique area. As such, this example provides an opportunity to fairly compare the area on the radar plot with that on the spie chart.

The standardised areas on the radar plot and spie chart are provided in Table 5. The standardised areas for each treatment on both plots are quite similar, and the corresponding ranks are the same. Nevertheless, the efficacy outcomes equally contributed to the standardised area, which is unlikely to reflect a knowledge user's preferences. There are some dependencies between the outcomes. For example, treatments that clear or nearly clear psoriasis for a large proportion of patients are also likely to have a higher proportion of patients that achieve a PASI score of at least 75. These dependencies should be accounted for using methods such as the ones suggested earlier in the Methods section.

Table 5: *Standardised areas inside radar plots and spie charts of SUCRA values in [27].*

	Radar Plot		Spie Chart	
Treatment	Standardised Area[a]	Rank	Standardised Area[a]	Rank
Ixekizumab	0.775	1	0.801	1
Secukinumab	0.686	2	0.687	2
Infliximab	0.572	3	0.578	3
Ustekinumab	0.362	4	0.370	4
Adalimumab	0.236	5	0.237	5
Etanercept	0.084	6	0.084	6
Methotrexate	0.022	7	0.022	7
Placebo	0.000	8	0.000	8

[a] These areas solely summarise the comparative ranking in terms of efficacy outcomes.

Scatter plot of efficacy vs. safety

The purpose of this illustration is to show the consequences of naively plotting all efficacy and safety outcomes on a spie chart and summarising them with a single numerical value. As such, the standardised areas on a spie chart containing all efficacy and safety outcomes were calculated, assuming they were of equal importance. Of course, in practice, this is unlikely to be true. A knowledge user may want the contribution of the safety outcome to be the same as the contribution of the collection of efficacy outcomes. This is possible by dividing the spie chart in half, where the safety outcome is plotted on one half of the chart, and the three efficacy outcomes contribute equally to the other half. Nevertheless, the numerical summary of the coverage area will not allow a knowledge user to simultaneously compare the benefits and harms of the treatments, and so a scatter plot comparing the two is more desirable.

The results show that Ixekizumab has the second highest SUCRA value (Figure 3A), agreeing with the ranks solely based on efficacy (Table 5), but the message that it is one of the least tolerable is lost in this result (Table 4). The standardised area on the spie chart containing the efficacy outcomes only is plotted against the reported SUCRA values for the safety outcome in Figure 3B. In this scatter plot, treatments in the top right corner are preferred. The benefit-risk trade-off is clearer for Ixekizumab, and Secukinumab seems to have the best benefit-risk ratio.

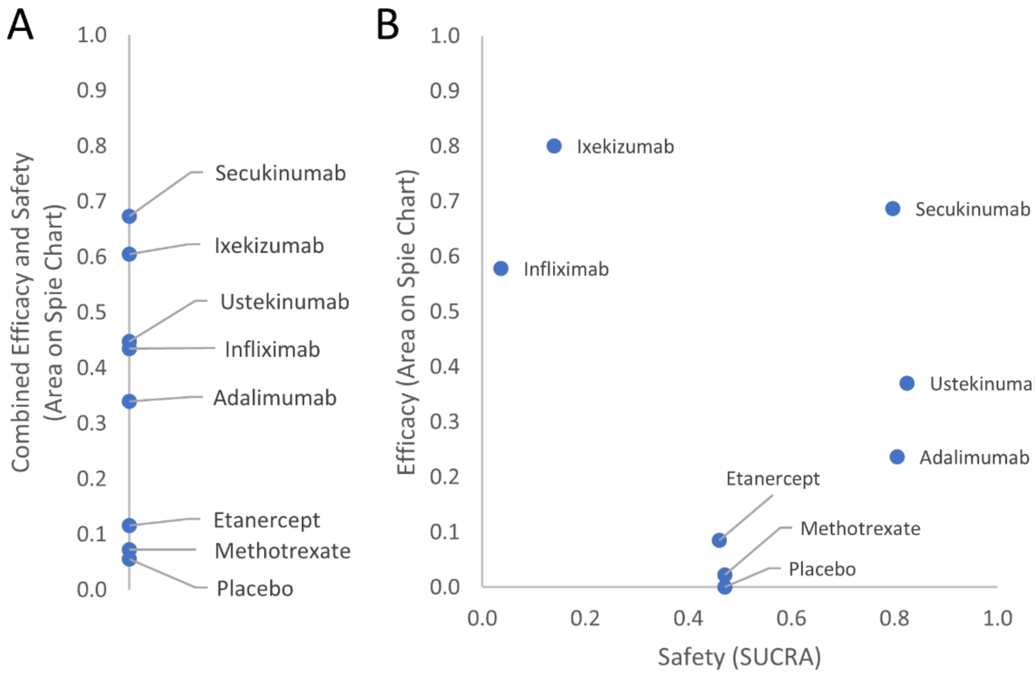


Figure 3: *Comparison of two approaches for balancing efficacy and safety outcomes in [27]. In (A), the areas inside a spie chart containing both efficacy and safety outcomes are plotted on a number line, where larger values (areas) are preferred. In (B) the areas inside a spie chart containing efficacy outcomes only are plotted against the SUCRA values for the single safety outcome on a scatter plot, where values in the top-right corner are preferred.*

Discussion

We have developed and presented a framework for obtaining the overall performance of treatments in NMA, summarised across all outcomes. Similar to SUCRA, the standardised area on a spie chart is a ratio of the maximum possible area, which a treatment could have if it always ranked best [12]. This paper lays the groundwork for integrating evidence across multiple outcomes, including some direction on how to incorporate key considerations for decision makers (e.g., outcome preferences).

Radar plots have been used in the past to compare outcomes in health research. More recently, they have been used to summarise the performance of treatments in an NMA context [8, 9]. Despite this, there are several limitations of radar plots that systematic reviewers should consider, and spie charts may be a more suitable alternative. A radar plot may be sufficient when evidence on three outcomes needs to be combined, and these three outcomes are of equal importance. If there are any additional outcomes, subjectivity can arise through the ordering of outcomes on the plot. Nevertheless, this may be mitigated by specifying outcome preferences a priori, which can be informed by preferences in Cochrane Summary of Findings tables or through a survey of stakeholders' preferences.

Spie charts, however, are a more generalisable option and they have nicer mathematical properties compared to a radar chart. For example, the area on a spie chart informed by a single outcome will output the same value that was inputted. In addition, adjusting the contribution of several outcomes on a spie chart is mathematically straightforward. Weighting schemes should be specified a priori to minimise subjectivity. This is also important when using coefficients from a risk-prediction model to inform the weights, as it is important to select a risk-prediction model that has been validated and covers the population of interest. Some risk-prediction models may even present coefficients tailored to subgroups, as shown in the lipids example, permitting subgroup-specific ranks.

Nevertheless, the practice of using coefficients to inform the relative importance of predictors has been criticised [21, 33]. More optimal methods require individual patient data [20], which NMA researchers may not have access to. Formally eliciting the relative importance of outcomes from decision makers may offer a better alternative in an NMA context [23]. In the future, it would be useful to design a weighting scheme that accounts for both the dependencies between the outcomes, as well as the preferences of knowledge users.

This framework was illustrated using SUCRA values; however, other outcome measures could be used. Nevertheless, the cited examples of systematic reviews presenting evidence across outcomes through radar plots have done so using SUCRA values [8, 9]. Another recent review averaged SUCRA values on LDL-c, HDL-c, and TG to give an overall indication of the effectiveness of diets on the lipid profile [34]. SUCRA is an attractive measure to compare treatments across multiple outcomes as it summarises both the strength and uncertainty of the relative treatment effects on the same scale [13]. The standardised area inside a spie chart informed by SUCRA values clearly conveys the degree of uncertainty in the evidence across outcomes. This is because the outcome values are squared in the calculation of the area, and smaller SUCRA values, which indicate less plausibility or certainty in a treatment ranking best, are penalised. The standardised area for a particular treatment will only be close to 1 if there is large certainty supporting a treatment being more effective than all other treatments across all outcomes.

While a treatment may be very effective, it could also be unsafe, and so it is important to consider efficacy and safety outcomes separately and not summarise them with one measure. Efficacy and safety outcomes should be combined separately, and they may be simultaneously compared in scatter plots such as the one plotted in Figure 3B in the psoriasis example. Nevertheless, we pause to reflect whether safety outcomes should be combined at all. A treatment's harmful effects in terms of one outcome could be diluted by the appearance of its safety in terms of several other outcomes. It might be better to pool evidence on efficacy outcomes together as a single measure and then compare it against critical safety outcomes one by one.

Additional aspects of the evidence also need to be considered, such as the internal and external biases of the RCTs informing the networks. This goes beyond assessing whether the evidence supporting a treatment ranking best is at high risk of bias. The decision maker must grasp how the biased trials effect the network estimates, and this depends on the geometry of the networks and size of the trials. Sensitivity analyses which remove the trials at high risk of bias, threshold analyses, or CINeMA may provide some insight into this [35-38].

There may be instances where there is no evidence on a treatment for a particular outcome. This treatment could still be included in the overall evaluation through spie charts, where a value of 0 is assumed. This would penalise the treatment's performance for missing outcome data. However, if a treatment cannot be considered without information on a critical outcome, then perhaps it should be excluded from the evaluation of the overall performances. Note that SUCRA depends on the number of treatments informing it. As such, the number of treatments should be equal across all outcomes to allow fair comparison. If a treatment is excluded from the decision set, then it should not be included in the calculation of the ranking probabilities, and thus SUCRA.

Conclusion

We have established the foundation of a framework that objectively summarises the comparative effectiveness of a treatment across multiple outcomes. This eliminates any subjectivity that may be introduced by a decision maker balancing contradictory rankings of treatments across different outcomes. The proposed framework is not meant to be a standalone presentation of the NMA results. Rather, it is intended to supplement the more detailed results that must be considered when evaluating the evidence. Forest plots or pairwise relative effect estimates should also be inspected to confirm whether there are any significant differences between treatments, a feature which may be masked by ranking statistics. Future research should investigate ways to adapt this framework when outcomes are missing for some treatments. The general approach should also be compared with existing approaches for integrating ranks across outcomes [39, 40]. Moving forward, we recommend the spie chart over the radar plot for summarising effectiveness across multiple outcomes.

References

1. Petropoulou M, Nikolakopoulou A, Veroniki A, Rios P, Vafaei A, Zarin W, et al. Bibliographic study showed improving statistical methodology of network meta-analyses published between 1999 and 2015. J Clin Epidemiol. 2017, 82:20-28.

2. Dias S, Ades AE, Welton NJ, Jansen JP, Sutton AJ. Network meta-analysis for decision making. Hoboken, NJ: Wiley; 2018.

3. Caldwell DM, Ades AE, Higgins JPT. Simultaneous comparison of multiple treatments: combining direct and indirect evidence. BMJ. 2005, 331:897-900.

4. Lu G, Ades A. Combination of direct and indirect evidence in mixed treatment comparisons. Stat Med. 2004, 23:3105-3124.

5. Tan SH, Cooper NJ, Bujkiewicz S, Welton NJ, Caldwell DM, Sutton AJ. Novel presentational approaches were developed for reporting network meta-analysis. J Clin Epidemiol. 2014, 67:672-680.

6. Veroniki AA, Straus SE, Fyraridis A, Tricco AC. The rank-heat plot is a novel way to present the results from a network meta-analysis including multiple outcomes. J Clin Epidemiol. 2016, 76:193-199.

7. Saary MJ. Radar plots: a useful way for presenting multivariate health care data. J Clin Epidemiol. 2008, 61:311-317.

8. McCool R, Wilson K, Arber M, Fleetwood K, Toupin S, Thom H, et al. Systematic review and network meta-analysis comparing ocrelizumab with other treatments for relapsing multiple sclerosis. Mult Scler Relat Disord. 2019, 29:55-61.

9. Rogliani P, Matera MG, Ritondo BL, De Guido I, Puxeddu E, Cazzola M, et al. Efficacy and cardiovascular safety profile of dual bronchodilation therapy in chronic obstructive pulmonary disease: A bidimensional comparative analysis across fixed-dose combinations. Pulm Pharmacol Ther. 2019, 59:101841.

10. Stafoggia M, Lallo A, Fusco D, Barone AP, D'Ovidio M, Sorge C, et al. Spie charts, target plots, and radar plots for displaying comparative outcomes of health care. J Clin Epidemiol. 2011, 64:770-778.

11. Feitelson D. Comparing partitions with spie charts. pp. 1-7. School of Computer Science and Engineering: The Hebrew University of Jerusalem; 2003:1-7.

12. Salanti G, Ades A, Ioannidis J. Graphical methods and numerical summaries for presenting results from multiple-treatment meta-analysis: an overview and tutorial. J Clin Epidemiol. 2011, 64:163-171.

13. Rücker G, Schwarzer G. Ranking treatments in frequentist network meta-analysis works without resampling methods. BMC Med Res Methodol. 2015, 15:58.

14. Jansen J, Trikalinos T, Cappelleri J, 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 Health. 2014, 17:157-173.

15. 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, 6:451-460.

16. Naci H, van Valkenhoef G, Higgins JPT, Fleurence R, Ades AE. Combining network meta-analysis with multicriteria decision analysis to choose among multiple drugs. Circ Cardiovasc Qual Outcomes. 2014, 7:787-792.

17. Dias S, Welton NJ, Sutton AJ, Ades AE. *NICE DSU* Technical Support Document 5: Evidence synthesis in the baseline natural history model. 2011.

18. Furukawa TA, Cipriani A, Barbui C, Brambilla P, Watanabe N. Imputing response rates from means and standard deviations in meta-analyses. Int Clin Psychopharmacol. 2005, 20:49-52.

19. Chinn S. A simple method for converting an odds ratio to effect size for use in meta-analysis. Stat Med. 2000, 19:3127-3131.

20. Lebreton JM, Ployhart RE, Ladd RT. A Monte Carlo comparison of relative importance methodologies. Organ Res Methods. 2004, 7:258-282.

21. Johnson JW. A heuristic method for estimating the relative weight of predictor variables in multiple regression. Multivariate Behav Res. 2000, 35:1-19.

22. López-López JA, Page MJ, Lipsey MW, Higgins JPT. Dealing with effect size multiplicity in systematic reviews and meta-analyses. Res Synth Methods. 2018.

23. Riabacke M, Danielson M, Ekenberg L. State-of-the-art prescriptive criteria weight elicitation. Adv Decis Sci. 2012, 2012:276584.

24. Riley RD, Jackson D, Salanti G, Burke DL, Price M, Kirkham J, et al. Multivariate and network meta-analysis of multiple outcomes and multiple treatments: rationale, concepts, and examples. BMJ. 2017, 358:j3932.

25. Bellanti F. From data to models: reducing uncertainty in benefit risk assessment: application to chronic iron overload in children. Leiden University, Faculty of Science; 2015.

26. Schwingshackl L, Bogensberger B, Bencic A, Knuppel S, Boeing H, Hoffmann G. Effects of oils and solid fats on blood lipids: a systematic review and network meta-analysis. J Lipid Res. 2018, 59:1771-1782.

27. Jabbar-Lopez ZK, Yiu ZZN, Ward V, Exton LS, Mohd Mustapa MF, Samarasekera E, et al. Quantitative evaluation of biologic therapy options for psoriasis: A systematic review and network meta-analysis. J Invest Dermatol. 2017, 137:1646-1654.

28. Pagana KD, Pagana TJ. Mosby's Canadian manual of diagnostic and laboratory tests. 1st ed. Toronto: Mosby; 2013.

29. Castelli WP, Anderson K, Wilson PW, Levy D. Lipids and risk of coronary heart disease. The Framingham Study. Ann Epidemiol. 1992, 2:23-28.

30. Higgins JPT, Thomas J, Chandler J, Cumpston M, Li T, Page MJ, Welch VA (editors). Cochrane Handbook for Systematic Reviews of Interventions version 6.0. Cochrane; 2019. Available from www.training.cochrane.org/handbook.

31. Rees K, Takeda A, Martin N, Ellis L, Wijesekara D, Vepa A, et al. Mediterranean-style diet for the primary and secondary prevention of cardiovascular disease. Cochrane Database Syst Rev. 2019.

32. Abdelhamid AS, Brown TJ, Brainard JS, Biswas P, Thorpe GC, Moore HJ, et al. Omega-3 fatty acids for the primary and secondary prevention of cardiovascular disease. Cochrane Database Syst Rev. 2018.

33. Bring J. How to Standardize Regression Coefficients. Am Stat. 1994, 48:209-213.

34. Neuenschwander M, Hoffmann G, Schwingshackl L, Schlesinger S. Impact of different dietary approaches on blood lipid control in patients with type 2 diabetes mellitus: A systematic review and network meta-analysis. Eur J Epidemiol, 2019, 34:837-852.

35. Phillippo D, Dias S, Ades A, Didelez V, Welton N. Sensitivity of treatment recommendations to bias in network meta-analysis. J R Stat Soc Ser A Stat Soc. 2017, 181.

36. Phillippo DM, Dias S, Welton NJ, Caldwell DM, Taske N, Ades AE. Threshold analysis as an alternative to GRADE for assessing confidence in guideline recommendations based on network meta-analyses. Ann Intern Med. 2019, 170:538-546.

37. Papakonstantinou T, Nikolakopoulou A, Higgins JPT, Egger M, Salanti G. CINeMA: Software for semiautomated assessment of the confidence in the results of network meta-analysis. Campbell Syst Rev. 2020, 16:e1080.

38. Salanti G, Giovane CD, Chaimani A, Caldwell DM, Higgins JPT. Evaluating the quality of evidence from a network meta-analysis. PLoS One. 2014, 9:ee99682.

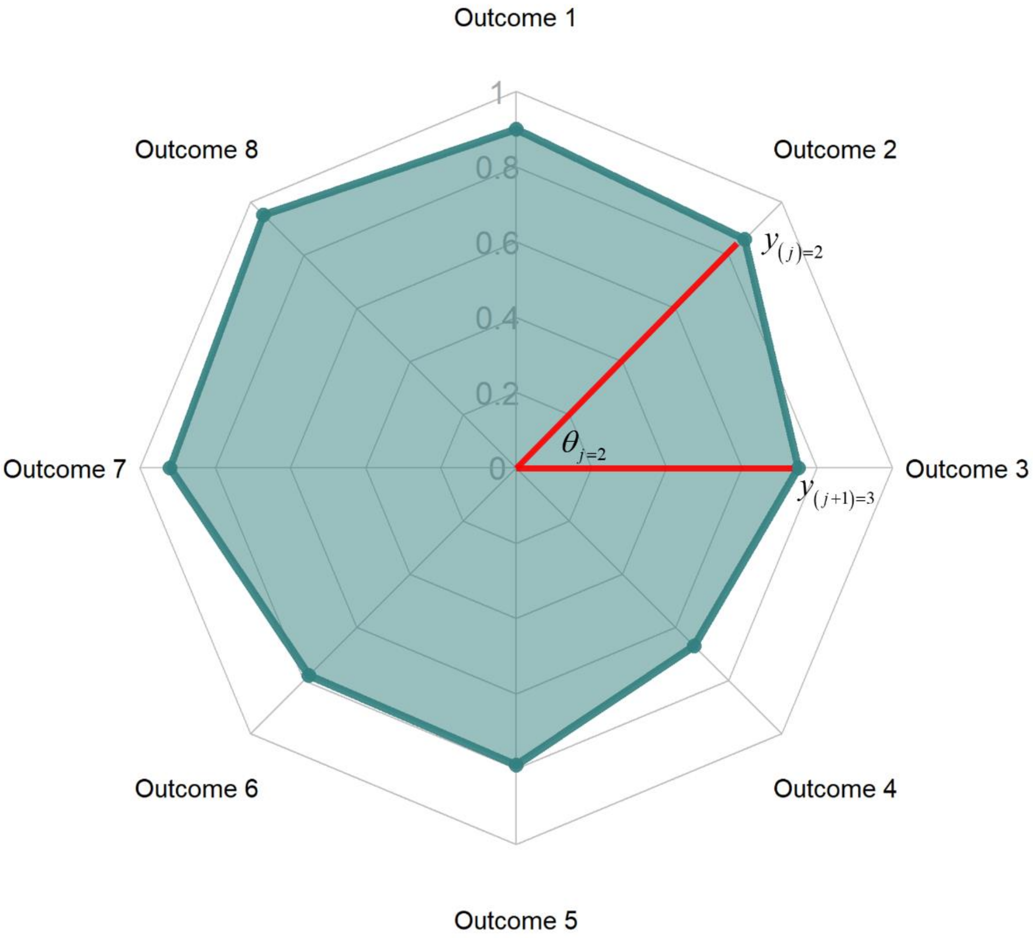
39. Rucker G, Schwarzer G. Resolve conflicting rankings of outcomes in network meta-analysis: Partial ordering of treatments. Res Synth Methods. 2017, 8:526-536.

40. Mavridis D, Porcher R, Nikolakopoulou A, Salanti G, Ravaud P. Extensions of the probabilistic ranking metrics of competing treatments in network meta-analysis to reflect clinically important relative differences on many outcomes. Biometrical J. 2020, 62:375-385.

Additional File 1

Measuring the area inside a radar plot

The traditional radar plot is equally divided by at least 3 axes or radii which measure an attribute of interest, e.g., health outcomes. To illustrate, Supplementary Figure 1 displays a radar plot that is informed by simulated values between 0 and 1 for $J = 8$ outcomes.



Supplementary Figure 1: *Example radar plot informed by the values of 8 outcomes. To calculate the area of triangle* $j = 2$*, the required parameters are denoted:* $\theta_{j=2}$ *is a known angle,* $y_{(j)=2}$ *and* $y_{(j+1)=3}$ *are the lengths of two sides of triangle* $j = 2$*, which are equal to the values of outcomes 2 and 3, respectively.*

The resulting shape on any radar plot is typically an irregular polygon, and the area it covers may be calculated as the sum of the areas of the triangles forming the shape. In Supplementary Figure 1, the area enclosed, A, is the sum of the area of 8 triangles, $A_{\Delta_j}, j = 1,...,8$,

$$A_{\Delta_j} = \frac{1}{2} y_{(j)} y_{(j+1)} \sin \theta_j ,$$

Where $y_{(j)}, y_{(j+1)}$ are the lengths of the sides congruent to the angle of each triangle j, equal to the values of the outcome measure on the corresponding radii, and θ_j are the angles between the radii. Since the 8 radii defining this radar plot are equidistant, the angle corresponding to each triangle's vertex at the centre are equal, i.e., $\theta_j = \frac{2\pi}{8} = \frac{\pi}{4}$ radians, and the area of inside this radar plot is then:

$$A = \frac{1}{2} \sin\left(\frac{\pi}{4}\right) \sum_{j=1}^{8} y_{(j)} y_{(j+1)} .$$

In general, the area within a radar plot informed by $J \geq 3$ outcomes for an intervention is

$$A = \frac{1}{2} \sum_{j=1}^{J} y_{(j)} y_{(j+1)} \sin \theta_j^*$$

where $j = 1,..., J$, $(j+1) = (1)$ when $j = J$, and $\theta_j^* = \theta_j$ if $0 < \theta_j < \frac{\pi}{2}$ radians, otherwise θ_j^* is the related acute angle.

Standardised area on a radar plot

The maximum possible area of each triangle in a radar chart is achieved when $y_{(j)} = y_{(j+1)}$ = maximum possible value. When the maximum possible value is 1, as is the case for SUCRA, then

$$A_{\Delta_j}{}^{\max} = \frac{1}{2}(1)(1)\sin\theta_j = \frac{1}{2}\sin\theta_j \,.$$

If there are J outcomes and the angles between the radii are equal, i.e., $\theta_j = \frac{2\pi}{J}, j = 1,..., J$, then

$$A^{\max} = \frac{1}{2}\sum_{i=1}^{J}\sin\left(\frac{2\pi}{J}\right) = \frac{J}{2}\sin\left(\frac{2\pi}{J}\right).$$

So, in the case of equal angles, the standardised area on a radar chart for a given treatment is then

$$\begin{aligned} A^{std} &= \frac{2}{J\sin\left(\frac{2\pi}{J}\right)}\frac{1}{2}\sin\left(\frac{2\pi}{J}\right)\sum_{j=1}^{J} y_{(j)}y_{(j+1)} \\ &= \frac{1}{J}\sum_{j=1}^{J} y_{(j)}y_{(j+1)} \end{aligned}$$

where $y_{(j)}, y_{(j+1)}$ are the SUCRA values congruent to the angle $\theta_j = \frac{2\pi}{J}$ of each triangle j.

In general, the standardised area on a radar chart is

$$\begin{aligned} A^{std} &= \frac{2}{\sum_{j=1}^{J} \sin\theta_j^*} \frac{1}{2} \sum_{j=1}^{J} y_{(j)} y_{(j+1)} \sin\theta_j^* \\ &= \frac{1}{\sum_{j=1}^{J} \sin\theta_j^*} \sum_{j=1}^{J} y_{(j)} y_{(j+1)} \sin\theta_j^* \end{aligned}$$

where

$$\theta_j^* = \begin{cases} \theta_j & \text{if } 0 < \theta_j \leq \dfrac{\pi}{2} \\ \pi - \theta_j & \text{if } \dfrac{\pi}{2} < \theta_j < \pi \end{cases}$$

and all other parameters are as defined above. Note that $0 < \theta_j < \pi$, since by definition, the sum of the angles inside a triangle is π radians.

Incorporating stakeholder preferences

In the spie chart, we were able to incorporate a stakeholder's preferences through the angles of the segments. However, this is not straightforward in a radar plot since the vertex angles are functions of 2 outcomes:

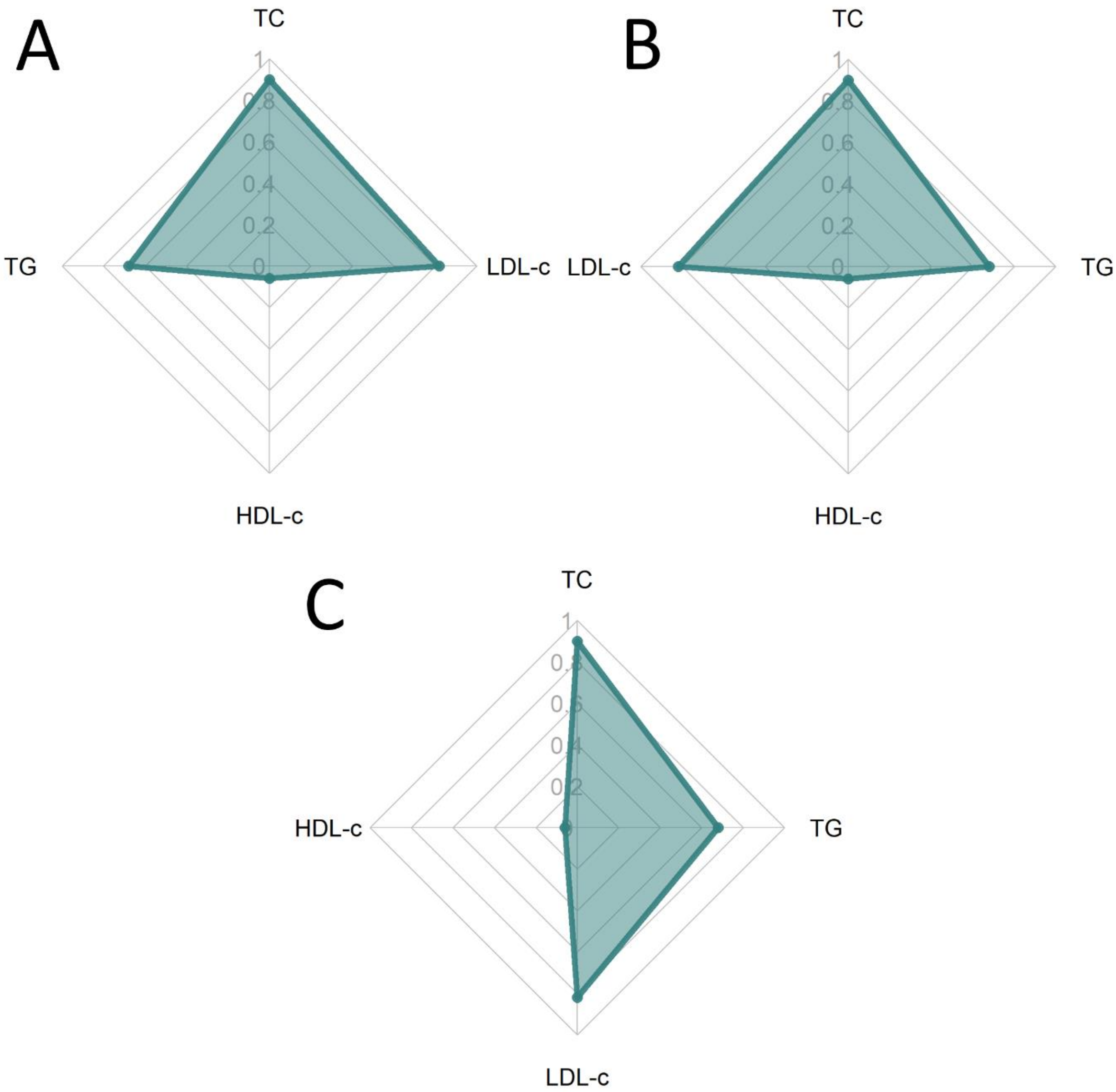
$$\theta_j = \sin^{-1}\left(\frac{2A_{\triangle_j}}{y_{(j)} y_{(j+1)}} \right).$$

The spie chart is much more favourable plot for this purpose since the angle of a sector can be adjusted to solely adjust the contribution of 1 outcome:

$$\theta_j = \frac{2A_{\bigcirc_j}}{y_j^2} \ .$$

As such, we recommend spie charts when there is a need to weight each outcome differently.

Additional File 2



Supplementary Figure 2: *Three possible radar plots of the SUCRA values corresponding to Safflower oil in [26]. The plots in panel A and B have the same area, since they are the same shape flipped at the vertical axes. The plot in panel C has a different area due to the different triangles formed by TC & HDL-c and TC & LDL-c.*

Chapter 6

Summary and future directions

The main output of a network meta-analysis (NMA) is a coherent set of synthesized relative effects from which a decision maker may determine which treatment is best (Higgins & Welton, 2015). The keyword is 'coherent': this implies if treatment A is better than treatment B (A > B), and treatment B is better than C (B > C), then treatment A must be better than treatment C (A > C). As such, the key NMA result provides an ordinal list of the treatments. Nevertheless, the use of ranking statistics to summarize NMA results has been criticized because of their sensitivity to evidence, uncertainty, and their perceived oversimplification of the results (Kibret et al., 2014; Mbuagbaw et al., 2017; Mills et al., 2013; Trinquart et al., 2016). This thesis sought to investigate and address some of these concerns and limitations, examine the robustness of existing methods, and explore situations where alternative ranking statistics might be useful. This was accomplished through the development of novel frameworks and measures for summarizing NMA results, which are presented in three separate papers comprising Chapters 3 – 5 of this sandwich thesis. These contributions centred around the use of ranking probabilities and the surface under their cumulative curve (i.e., SUCRA), since these measures account for the uncertainty of the posterior distributions of the relative effects (Salanti et al., 2011; Salanti et al., 2014), a key component in decision making (Sniazhko, 2019).

6.1 Summary of contributions

The first major contribution of this thesis is a novel conceptual and methodological framework for investigating the robustness of SUCRA-based treatment ranks to evidence contributing to the network, which constitutes Paper I of this thesis (Daly et al., 2019) and is presented in Chapter 3. Within this framework, Cohen's kappa is used to quantify the agreement between the ranks estimated in a base-case (original) NMA consisting of all eligible and relevant randomized controlled trials (RCTs), and an NMA of sub-networks constructed with a leave one (or more) RCT out approach (Cohen, 1960, 1968). This quantification enables an objective investigation and comparison of the impact studies (and the characteristics of their evidence) may have on the estimated hierarchy of treatments. For example, the Cochrane Risk of Bias Tool (version 2) evaluates the risk of bias in RCTs across five domains: the randomization process, deviations from the assigned treatment, missing outcome data, outcome measurement, and selective reporting of results (Sterne et al., 2019). RCTs may be at high or low risk of bias on each of the domains, or somewhere in between (i.e., some concerns of bias). A decision maker may then ask, are RCTs at high risk of bias more likely to have an impact on the treatment hierarchy compared to those with some or low concerns of bias? The distributions of rank agreement may then be visually compared among these three groups of RCTs. Similar empirical investigations may be conducted at a higher level with many NMAs to obtain an understanding of why SUCRA-based treatment ranks are less robust in certain NMAs compared to others. For example, are SUCRA-based treatment ranks less reliable in sparse networks (i.e., networks with many treatments with minimal head-to-head comparisons)? This may also answer questions such as, which bias domains have more impact on SUCRA-based treatment ranks, and is this specific to certain fields? For example, do RCTs comparing surgical interventions have more impact on ranks compared to RCTs comparing pharmacological treatments because of the difficulties in blinding treatment allocation in the former?

Empirical data from published NMA datasets were used to illustrate the proposed framework in Chapter 3, where RCTs were removed one at a time. These empirical evaluations show that the SUCRA-based treatment ranks were robust to most of the RCTs included in the networks, and most of the changes in rank were between treatments with smaller differences in SUCRA. Notably large changes in rank were observed when RCTs contributing a large degree of between-study heterogeneity to the network were removed. Treatment effect modifiers contribute to between-study heterogeneity, and this highlights the importance of controlling or adjusting for the effects of these modifiers, either through the inclusion and exclusion criteria of a systematic review or through meta-regression. Nevertheless, these observations are limited to the illustrative data explored, and further evaluations using diverse NMAs are required to establish the pattern of robustness, and what factors contribute to it (e.g., study sample size, contribution to between-study heterogeneity). In addition, although the framework for evaluating robustness was built around SUCRA-based ranks, it is also applicable to ranks based on any measure that already exists or may developed in the future.

Chapter 4 corresponds to Paper II (Daly et al., 2020a) of this thesis, in which a novel approach was developed to provide an alternative measure for ranking treatments in an NMA. This measure was derived as the partial area under cumulative ranking curves, and hence the name partial surface under the cumulative ranking curve (pSUCRA). This approach was motivated by and adopted from applications in diagnostic medicine, where the partial area under the receiver operating characteristic curve is used to compare diagnostic tests in relevant ranges of sensitivity and/or specificity (McClish, 1989). Similarly, the left region of the cumulative ranking curve in NMA is more relevant to decision makers, as it reflects the probabilities of a treatment ranking the best, second best, third best, etc. As the number of treatments increases in an NMA, the distribution of ranking probabilities for each treatment may be spread out across regions of irrelevant (poorer) ranks. Consequently, when the cumulative ranking curves of at least two treatments cross, a treatment with the highest probabilities of ranking the best, second best, third best, etc. may be associated with a smaller SUCRA value, and hence leading to much lower ranking probabilities. This was illustrated in the paper using simulated data and a published NMA, and the results showed that pSUCRA was able to highlight treatments with larger probabilities of being among the best and was an optimal alternative in such circumstances. As such, the cumulative ranking curves should always be visually inspected. If the curves of the top-performing treatments cross, pSUCRA should considered as an alternative summary to SUCRA.

Another important contribution arising from the derivation of pSUCRA is the finding that pSUCRA can be represented as a weighted sum of the ranking probabilities, where the better ranking probabilities receive higher weights. This lays out a structure from which different weighting schemes may be applied. Since ranking probabilities and pSUCRA are objective trade-offs of the magnitude and uncertainty of the estimated relative effects, weighting schemes may be developed to tailor the trade-off according to a decision maker's aversity to uncertainty. Potential adaptations are discussed in the next section as possible future directions.

Finally, a novel and flexible framework for quantitatively integrating NMA results across multiple outcomes was proposed in Paper III (Daly et al., 2020b) and presented in Chapter 5. This framework makes use of the spie chart (Stafoggia et al., 2011), which is divided into sectors representing the benefits or harms of a treatment. Comparing treatments based on the area inside their spie chart facilitates an objective means to balance their performance across multiple outcomes. An outcome's contribution to the summary measure (i.e., the area) may be adjusted through the angle of its corresponding sector, based on a suitable weighting scheme that reflects a decision maker's preferences. In this thesis, separate spie charts are recommended for efficacy and safety outcomes, as it is important for a decision maker to balance the trade-off between the two. The areas inside an efficacy spie chart may then be plotted against the area inside a safety spie chart through two-dimensional scatter plots, which are traditionally used to compare two outcomes in NMA (Chaimani et al., 2013). The spie chart offers the advantage of (separately) integrating multiple efficacy and safety outcomes so that more than two outcomes can be simultaneously considered.

In this thesis, the area inside a spie chart is developed and presented in terms of SUCRA, but the same principles apply for any outcome measure (e.g., absolute probabilities of risk or pSUCRA). pSUCRA should be considered over SUCRA when there are large numbers of treatments in the network, and the cumulative ranking curves cross.

Similarly, while published coefficients of prognostic models, matrices of the pairwise correlations between outcomes, or preferences elicited from decision makers are suggested as sources for evidence-based weights, any weighting scheme may be applied. Furthermore, the calculation of the area inside a spie chart is relatively simple, and thus researchers undertaking an NMA may easily compute this using standard software. The flexibility and simplicity of this framework increase its adaptability to a variety of scenarios. These features also make the area inside a spie chart a better alternative to the area inside a radar plot, a tool used recently in two NMAs (McCool et al., 2019; Rogliani et al., 2019).

6.2 Future directions

Most contributions made in this thesis are presented as proof of concept, and there are several opportunities to extend them and investigate their validity. For example, the framework in Chapter 3 was illustrated with unweighted and quadratic weighted kappa (Cohen, 1968), but other weighting schemes may be more useful in investigating the robustness of ranks. For instance, a decision maker may only be concerned if the ranks of the top three-ranked treatments change and this could be accounted for by discounting the weights of the remaining treatments' ranks. In addition, the weights should reflect the differences in the SUCRA values themselves, as changes in rank between two treatments with similar SUCRA will be less important than changes in rank between two treatments with a large difference in SUCRA. The framework in its current form also assumes that all treatments in the base-case network (the full network consisting of all studies) remain in the network of the subset of data. This condition is required in order to use Cohen's kappa to quantify the robustness of the treatment ranks. Nevertheless, when studies are removed, some of the treatments may also be removed. As such, when treatments are removed from the network, a rank agreement measure that accounts for missing treatments, perhaps an adaption of Cohen's kappa that includes a penalization factor, may offer a solution. The framework can, therefore, be extended to allow various weighting schemes or agreement measures to increase its generalizability. In addition, more extensive empirical evaluations across various characteristics (e.g., sparsity of networks, strength and quality of evidence) are required to further evaluate the framework and to identify potential characteristics associated with lack of robustness (e.g., risk of bias). This proposed framework may also be used to compare the robustness of pSUCRA and SUCRA-based ranks, either for a single outcome or across multiple outcomes based on the area inside a spie chart.

More guidance on the rationale for selecting a cut-off point on the cumulative ranking curve will help facilitate the uptake of pSUCRA. This is likely to be based on how many treatments are available, and so more empirical evaluations will provide insight into this. Sparse networks (i.e., networks with minimal direct comparisons) along with those with large numbers of treatments (e.g., several complex interventions) are hypothesized to be situations that will benefit the most from the application of pSUCRA. As such, future research on pSUCRA should make use of these networks.

In addition, the restricted region of the cumulative ranking curve relates to a decision maker's preferred trade-off between the magnitude and uncertainty of the relative effects. In Paper II (Chapter 4), the left region, that is the region of better ranks, was the primary focus in the development of pSUCRA. However, considering the upper region of the cumulative ranking curve may also be useful to decision makers averse to uncertainty. For example, a treatment with a very imprecise treatment effect estimate may have a larger probability of ranking best compared to treatments with more precise estimates (Jansen et al., 2014). Requiring a minimum probability of ranking best, which is equal to the first cumulative ranking probability, before calculating pSUCRA may penalize treatments with imprecise treatment effects.

Moreover, since pSUCRA is a weighted sum of the ranking probabilities, this may provide an opportunity to incorporate aversity to uncertainty through the weights. This is done in the inverse-variance approach to meta-analysis, where the contribution of an individual study is inversely proportional to the variance of its estimated effect (Borenstein et al., 2010). The variance of an option's estimated payoff may also be used to penalize its expected payoff in economics (Nagengast et al., 2011). As such, it seems reasonable to include the variance of an estimated relative effect in the weights of ranking probabilities, and this should be explored in future work. Regardless of how pSUCRA is defined, the robustness of its corresponding ranks may be assessed through the framework set out in Chapter 3.

A decision maker ultimately needs to balance all available evidence across multiple outcomes to select a treatment. It is thus important to make sure the inputs defining a spie chart reflect the information a decision maker needs to arrive at their decision. For example, concerns about bias may be incorporated in bias-adjustment models that synthesize the evidence (Dias et al., 2010b; Welton et al., 2009). The adjusted relative effects would then account for these concerns and so would SUCRA. The trade-off between the magnitude and uncertainty of the relative effects should also reflect a decision maker's preferences. If this can be captured through pSUCRA, then a treatment's performance may be better defined by this measure rather than SUCRA.

Weighting outcomes is another area that requires future work to optimize the framework proposed in Chapter 5. For one, it is important to properly account for the dependencies in the outcomes (Johnson, 2000). At the same time, the decision maker's preferences should also be incorporated into measures of the relative importance of the outcomes. The best approach for combining dependencies and decision maker's preferences is not clear. One promising lead is through the adaption of methodologies used to derive composite indicators in social research (Becker et al., 2017). Nevertheless, this requires individual patient data, which NMA authors may not have access to or the time to collect.

Another issue associated with integrating probabilities across multiple outcomes is that of missing data, as there may be no RCT evidence on some of the outcomes for one or more treatment. The framework presented in Chapter 5 relies on the assumption that there is evidence of each treatment's effect on all critical outcomes. Or alternatively, decision makers will only consider a treatment if there is evidence on all critical outcomes. However, if this assumption is not valid, then the use of this framework will require the imputation of missing outcomes. The use of surrogate outcomes may assist in this. If two outcomes are highly correlated, evidence on one of these outcomes may be used to impute information on the missing outcome through multivariate NMA (Riley et al., 2017). Alternatively, observational evidence may help impute this information, and this is an active area of research in NMA (Efthimiou et al., 2017). The framework in Chapter 3 may be used to assess the robustness of overall ranks to the inclusion of various sources of imputed evidence.

Finally, a common theme among these potential areas of future research is the opportunity to assess the robustness of the treatment ranks to the evidence informing the ranking measures, as well as the choice of the ranking measure itself. These robustness assessments are not only useful to methodological researchers, but also to decision makers themselves. For example, decision makers may wish to see how differences in their outcome preferences might impact the final treatment rankings. Therefore, user friendly software, such as a web-based R Shiny App, could be developed to aid decision makers in conducting these robustness assessments (R Studio Inc., 2020). Randomized comparisons of this tool informed by different input (e.g., pSUCRA or SUCRA, outcome weighting schemes) should be considered to measure preferences and understanding among end users. Such comparisons would also be useful in evaluating the established tool against other overall ranking measures and visual tools recently proposed in the literature.

6.3 Concluding remarks

While there are concerns about the use of ranks to summarize NMA results, it is not time to dismiss their usefulness in this field. This thesis provides three novel and objective measures and frameworks that may improve the use of ranks for summarizing NMA results. Ranking probabilities, pSUCRA, or SUCRA may be used to inform a framework that combines information across multiple outcomes. This will assist a decision maker in developing the single ordinal list of the treatments they need to make an informed decision. Treatment ranks have been noted to be unstable when there is large uncertainty in the network (Dias et al., 2018), and so it is important to select the most reliable ranking measure for such situations, and to be transparent about its robustness. More research is required to establish if and in what situations pSUCRA or SUCRA are suitable. The contributions in this thesis should be used to determine which ranking statistics (e.g., pSUCRA) are reliable enough to summarize NMA results for varying degrees of sparseness, and if they need to be adapted to improve their reliability.

References

Aromataris, E., & Munn, Z. (Eds.). (2020). *JBI Reviewer's Manual*. JBI. https://doi.org/10.46658/JBIRM-19-01

Becker, W., Saisana, M., Paruolo, P., & Vandecasteele, I. (2017). Weights and importance in composite indicators: Closing the gap. *Ecological Indicators, 80*, 12-22. https://doi.org/10.1016/j.ecolind.2017.03.056

Berkey, C. S., Hoaglin, D. C., Mosteller, F., & Colditz, G. A. (1995). A random-effects regression model for meta-analysis. *Statistics in Medicine, 14*(4), 395-411. https://doi.org/10.1002/sim.4780140406

Borenstein, M., Hedges, L. V., Higgins, J. P. T., & Rothstein, H. R. (2009a). Chapter 40: When does it make sense to perform a meta-analysis? In *Introduction to meta-analysis* (pp. 357-364). John Wiley & Sons, Ltd. https://doi.org/10.1002/9780470743386.ch40

Borenstein, M., Hedges, L. V., Higgins, J. P. T., & Rothstein, H. R. (2009b). Chapter 5: Effect sizes based on binary data (2 x 2 tables). In *Introduction to meta-analysis* (pp. 357-364). John Wiley & Sons, Ltd. https://doi.org/10.1002/9780470743386.ch5

Borenstein, M., Hedges, L. V., Higgins, J. P. T., & Rothstein, H. R. (2010). A basic introduction to fixed-effect and random-effects models for meta-analysis. *Research Synthesis Methods*, *1*(2), 97-111. https://doi.org/10.1002/jrsm.12

Borenstein, M., & Higgins, J. P. T. (2013). Meta-analysis and subgroups. *Prevention Science*, *14*(2), 134-143. https://doi.org/10.1007/s11121-013-0377-7

Borenstein, M., Higgins, J. P. T., Hedges, L. V., & Rothstein, H. R. (2017). Basics of meta-analysis: I^2 is not an absolute measure of heterogeneity. *Research Synthesis Methods, 8*(1), 5-18. https://doi.org/10.1002/jrsm.1230

Brignardello-Petersen, R., Bonner, A., Alexander, P. E., Siemieniuk, R. A., Furukawa, T. A., Rochwerg, B., Hazlewood, G. S., Alhazzani, W., Mustafa, R. A., Murad, M. H., Puhan, M. A., Schünemann, H. J., & Guyatt, G. H. (2018). Advances in the GRADE approach to rate the certainty in estimates from a network meta-analysis. *Journal of Clinical Epidemiology, 93*, 36-44. https://doi.org/10.1016/j.jclinepi.2017.10.005

Buchanan, J., & Kock, N. (2001). Information overload: A decision making perspective. In M. Köksalan & S. Zionts (Eds.), *Multiple criteria decision making in the new millennium* (pp. 49-58). Springer. https://doi.org/10.1007/978-3-642-56680-6_4

Bucher, H. C., Guyatt, G. H., Griffith, L. E., & Walter, S. D. (1997). The results of direct and indirect treatment comparisons in meta-analysis of randomized controlled trials. *Journal of Clinical Epidemiology, 50*(6), 683-691. https://doi.org/10.1016/S0895-4356(97)00049-8

Caldwell, D. M., Ades, A. E., & Higgins, J. P. T. (2005). Simultaneous comparison of multiple treatments: Combining direct and indirect evidence. *BMJ, 331*, 897-900. https://doi.org/10.1136/bmj.331.7521.897

Casella, G., & George, E. I. (1992). Explaining the Gibbs sampler. *The American Statistician, 46*(3), 167-174. https://doi.org/10.1080/00031305.1992.10475878

Casella, G., & Berger, R. L. (2002). *Statistical inference* (2nd ed.). Thomson Learning.

Chaimani, A., Higgins, J. P. T., Mavridis, D., Spyridonos, P., & Salanti, G. (2013). Graphical tools for network meta-analysis in STATA. *PLoS One*, *8*(10), e76654. https://doi.org/10.1371/journal.pone.0076654

Chalmers, I., Bracken, M. B., Djulbegovic, B., Garattini, S., Grant, J., Gülmezoglu, A. M., Howells, D. W., Ioannidis, J. P. A., & Oliver, S. (2014). How to increase value and reduce waste when research priorities are set. *The Lancet, 383*(9912), 156-165. https://doi.org/10.1016/S0140-6736(13)62229-1

Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement, 20*(1), 37-46. https://doi.org/10.1177/001316446002000104

Cohen, J. (1968). Weighted kappa: Nominal scale agreement provision for scaled disagreement or partial credit. *Psychological Bulletin, 70*(4), 213-220. https://doi.org/10.1037/h0026256

Cooper, H., & Hedges, L. V. (2009). Research synthesis as a scientific process. In H. Cooper, L. V. Hedges, & J. C. Valentine (Eds.), *The handbook of research synthesis and meta-analysis* (pp. 3-16). Russell Sage Foundation.

Cooper, N. J., Sutton, A. J., Morris, D., Ades, A. E., & Welton, N. J. (2009). Addressing between-study heterogeneity and inconsistency in mixed treatment comparisons: Application to stroke prevention treatments in individuals with non-rheumatic atrial fibrillation. *Statistics in Medicine, 28*(14), 1861-1881. https://doi.org/10.1002/sim.3594

da Costa, B. R., & Jüni, P. (2014). Systematic reviews and meta-analyses of randomized trials: Principles and pitfalls. *European Heart Journal, 35*(47), 3336-3345. https://doi.org/10.1093/eurheartj/ehu424

Daly, C. H., Neupane, B., Beyene, J., Thabane, L., Straus, S. E., & Hamid, J. S. (2019). Empirical evaluation of SUCRA-based treatment ranks in network meta-analysis: Quantifying robustness using Cohen's kappa. *BMJ Open, 9*(9), e024625. https://doi.org/10.1136/bmjopen-2018-024625

Daly, C. H., Mbuagbaw, L., Thabane, L., Straus, S. E., & Hamid, J. S. (2020a). Partial surface under the cumulative ranking curve (pSUCRA) as an alternative measure for ranking treatments in network meta-analysis. *Submitted to Clinical Epidemiology*.

Daly, C. H., Mbuagbaw, L., Thabane, L., Straus, S. E., & Hamid, J. S. (2020b). Spie charts for quantifying treatment effectiveness and safety in multiple outcome network meta-analysis: A proof-of-concept study. *Submitted to BMC Medical Research Methodology*. https://doi.org/10.21203/rs.3.rs-36139/v1

Decision. (2020). In *Oxford Advanced Learner's Dictionary*. https://www.oxfordlearnersdictionaries.com/definition/english/decision?q=decision

Deeks, J. J., Higgins, J. P. T., & Altman, D. G. (2019). Chapter 10: Analysing data and undertaking meta-analyses. In J. P. T. Higgins, J. Thomas, J. Chandler, M. Cumpston, T. Li, M. J. Page, & V. A. Welch (Eds.), *Cochrane handbook for systematic reviews of interventions* version 6.0 (updated July 2019). Cochrane. www.training.cochrane.org/handbook

DerSimonian, R., & Laird, N. (1986). Meta-analysis in clinical trials. *Controlled Clinical Trials, 7*(3), 177-188. https://doi.org/10.1016/0197-2456(86)90046-2

Devereaux, P. J., & Yusuf, S. (2003). The evolution of the randomized controlled trial and its role in evidence-based decision making. *Journal of Internal Medicine, 254*(2), 105-113. https://doi.org/10.1046/j.1365-2796.2003.01201.x

Dias, S., Welton, N. J., Caldwell, D. M., & Ades, A. E. (2010a). Checking consistency in mixed treatment comparison meta-analysis. *Statistics in Medicine, 29*(7-8), 932-944. https://doi.org/10.1002/sim.3767

Dias, S., Welton, N. J., Marinho, V. C. C., Salanti, G., Higgins, J. P. T., & Ades, A. E. (2010b). Estimation and adjustment of bias in randomized evidence by using mixed treatment comparison meta-analysis. *Journal of the Royal Statistical Society: Series A (Statistics in Society), 173*(3), 613-629. https://doi.org/doi:10.1111/j.1467-985X.2010.00639.x

Dias, S., Welton, N. J., Sutton, A. J., & Ades, A. E. (2011). *NICE DSU Technical Support Document 2: A generalised linear modelling framework for pair-wise and network meta-analysis of randomised controlled trials.* Decision Support Unit. http://nicedsu.org.uk/wp-content/uploads/2017/05/TSD2-General-meta-analysis-corrected-2Sep2016v2.pdf

Dias, S., Ades, A. E., Welton, N. J., Jansen, J. P., & Sutton, A. J. (2018). *Network meta-analysis for decision making*. Wiley. https://doi.org/10.1002/9781118951651

Djulbegovic, B., & Guyatt, G. H. (2017). Progress in evidence-based medicine: a quarter century on. *The Lancet, 390*(10092), 415-423. https://doi.org/10.1016/s0140-6736(16)31592-6

Dobson, A. J., & Barnett, A. G. (2008). *An introduction to generalized linear models* (3rd ed.). Chapman and Hall/CRC.

Donnelly, C. A., Boyd, I., Campbell, P., Craig, C., Vallance, P., Walport, M., Whitty, C. J. M., Woods, E., & Wormald, C. (2018). Four principles to make evidence synthesis more useful for policy. *Nature, 558*(7710), 361-364. https://doi.org/10.1038/d41586-018-05414-4

Edwards, P., Clarke, M., DiGuiseppi, C., Pratap, S., Roberts, I., & Wentz, R. (2002). Identification of randomized controlled trials in systematic reviews: accuracy and reliability of screening records. *Statistics in Medicine, 21*(11), 1635-1640. https://doi.org/10.1002/sim.1190

Efthimiou, O., Debray, T. P. A, van Valkenhoef, G., Trelle, S., Panayidou, K., Moons, K. G. M., Reitsma, J. B., Shang, A., Salanti, G. (2016). GetReal in network meta-analysis: a review of the methodology. *Research Synthesis Methods, 7*(3), 236-263. https://doi.org/10.1002/jrsm.1195

Efthimiou, O., Mavridis, D., Debray, T. P. A., Samara, M., Belger, M., Siontis, G. C. M., Leucht, S., Salanti, G. (2017). Combining randomized and non-randomized evidence in network meta-analysis. *Statistics in Medicine, 36*(8), 1210-1226. https://doi.org/10.1002/sim.7223

Efthimiou, O. (2018). Practical guide to the meta-analysis of rare events. *Evidence-Based Mental Health, 21*(2), 72-76. http://dx.doi.org/10.1136/eb-2018-102911

Eriksen, M. B., & Frandsen, T. F. (2018). The impact of patient, intervention, comparison, outcome (PICO) as a search strategy tool on literature search quality: a systematic review. *Journal of the Medical Library Association, 106*(4), 420-431. https://doi.org/10.5195/jmla.2018.345

Fu, R., Vandermeer, B. W., Shamliyan, T. A., O'Neil, M. E., Yazdi, F., Fox, S. H., & Morton, S. C. (2008). Handling continuous outcomes in quantitative synthesis. In *Methods guide for effectiveness and comparative effectiveness reviews*. Agency for Healthcare Research and Quality.

Grant, M. J., & Booth, A. (2009). A typology of reviews: An analysis of 14 review types and associated methodologies. *Health Information and Libraries Journal, 26*(2), 91-108. https://doi.org/10.1111/j.1471-1842.2009.00848.x

Greenland, S. (1990). Randomization, statistics, and causal inference. *Epidemiology, 1*(6), 421-429.

Higgins, J. P. T., Whitehead, A., Turner, R. M., Omar, R. Z., & Thompson, S. G. (2001). Meta-analysis of continuous outcome data from individual patients. *Statistics in Medicine, 20*(15), 2219-2241. https://doi.org/10.1002/sim.918

Higgins, J. P. T., & Welton, N. J. (2015). Network meta-analysis: A norm for comparative effectiveness? *The Lancet*, *386*(9994), 628-630. https://doi.org/10.1016/S0140-6736(15)61478-7

Higgins, J. P. T., Thomas, J., Chandler, J., Cumpston, M., Li, T., Page, M. J., & Welch, V. A. (Eds.). (2019a). *Cochrane handbook for systematic reviews of interventions* version 6.0. Cochrane. www.training.cochrane.org/handbook

Higgins, J. P. T., Li, T., & Deeks, J. J. (2019b). Chapter 6: Choosing effect measures and computing estimates of effect. In J. P. T. Higgins, J. Thomas, J. Chandler, M. Cumpston, T. Li, M. J. Page, & V. A. Welch (Eds.), *Cochrane handbook for systematic reviews of interventions* version 6.0. Cochrane. www.training.cochrane.org/handbook

Jansen, J. P., Fleurence, R., Devine, B., Itzler, R., Barrett, A., Hawkins, N., Lee, K., Boersma, C., Annemans, L., & Cappelleri, J. C. (2011). Interpreting indirect treatment comparisons and network meta-analysis for health-care decision making: Report of the ISPOR Task Force on Indirect Treatment Comparisons Good Research Practices: Part 1. *Value in Health, 14*(4), 417-428. https://doi.org/10.1016/j.jval.2011.04.002

Jansen, J. P., Trikalinos, T., Cappelleri, J. C., Daw, J., Andes, S., Eldessouki, R., & Salanti, G. (2014). 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*, *17*(2), 157-173. https://doi.org/10.1016/j.jval.2014.01.004

Johnson, J. W. (2000). A heuristic method for estimating the relative weight of predictor variables in multiple regression. *Multivariate Behavioural Research, 35*(1), 1-19. https://doi.org/10.1207/s15327906mbr3501_1

Kastner, M., Antony, J., Soobiah, C., Straus, S. E., & Tricco, A. C. (2016). Conceptual recommendations for selecting the most appropriate knowledge synthesis method to answer research questions related to complex evidence. *Journal of Clinical Epidemiology, 73*, 43-49. https://doi.org/10.1016/j.jclinepi.2015.11.022

Khangura, S., Konnyu, K., Cushman, R., Grimshaw, J., & Moher, D. (2012). Evidence summaries: the evolution of a rapid review approach. *Systematic Reviews, 1*, 10. https://doi.org/10.1186/2046-4053-1-10

Kibret, T., Richer, D., & Beyene, J. (2014). Bias in identification of the best treatment in a Bayesian network meta-analysis for binary outcome: A simulation study. *Clinical Epidemiology, 6*, 451-460. https://doi.org/10.2147/CLEP.S69660

Korteling, J. E., Brouwer, A-M., & Toet, A. (2018). A neural network framework for cognitive bias. *Frontiers in Psychology*. *9,* 1561. https://doi.org/10.3389/fpsyg.2018.01561

Lambert, P. C., Sutton, A. J., Burton, P. R., Abrams, K. R., & Jones, D. R. (2005). How vague is vague? Assessment of the use of vague prior distributions for variance components. *Statistics in Medicine, 24*(15), 2401-2428. https://doi.org/10.1002/sim.2112

Lasserson, T. J., Thomas, J., & Higgins, J. P. T. (2019). Chapter 1: Starting a review. In J. P. T. Higgins, J. Thomas, J. Chandler, M. Cumpston, T. Li, M. J. Page, & V. A. Welch (Eds.), *Cochrane handbook for systematic reviews of interventions* version 6.0. Cochrane. www.training.cochrane.org/handbook

Lefebvre, C., Glanville, J., Briscoe, S., Littlewood, A., Marshall, C., Metzendorf, M-I., Noel-Storr, A., Rader, T., Shokraneh, F., Thomas, J., & Wieland, L. S. (2019). Chapter 4: Searching for and selecting studies. In J. P. T. Higgins, J. Thomas, J. Chandler, M. Cumpston, T. Li, M. J. Page, & V. A. Welch (Eds.), *Cochrane handbook for systematic reviews of interventions* version 6.0. Cochrane. www.training.cochrane.org/handbook

Liberati, A., Altman, D. G., Tetzlaff, J., Mulrow, C., Gøtzsche, P. C., Ioannidis, J. P. A., Clarke, M., Devereaux, P. J., Kleijnen, J., & Moher, D. (2009). The PRISMA statement for reporting systematic reviews and meta-analyses of studies that evaluate healthcare interventions: explanation and elaboration. *BMJ, 339,* b2700. https://doi.org/10.1136/bmj.b2700

Littell, J. H. (2018). Conceptual and practical classification of research reviews and other evidence synthesis products. *Campbell Systematic Reviews, 14*(1), 1-21. https://doi.org/10.4073/cmdp.2018.1

Lu, G., & Ades, A. E. (2004). Combination of direct and indirect evidence in mixed treatment comparisons. *Statistics in Medicine*, *23*(20), 3105-3124. https://doi.org/10.1002/sim.1875

Lu, G., & Ades, A. E. (2006). Assessing evidence consistency in mixed treatment comparisons. *Journal of the American Statistical Association, 101*(474), 447-459. https://doi.org/10.1198/016214505000001302

Lu, G., & Ades, A. E. (2009). Modeling between-trial variance structure in mixed treatment comparisons. *Biostatistics, 10*(4), 792-805. https://doi.org/10.1093/biostatistics/kxp032

Lumley, T. (2002). Network meta-analysis for indirect treatment comparisons. *Statistics in Medicine, 21*(16), 2313-2324. https://doi.org/10.1002/sim.1201

Lunn, D., Jackson, C., Best, N., Thomas, A., & Spiegelhalter, D. (2013). *The BUGS book*. CRC Press.

Macleod, M. R., Michie, S., Roberts, I., Dirnagl, U., Chalmers, I., Ioannidis, J. P. A, Al-Shahi Salman, R., Chan, A-W., & Glasziou, P. (2014). Biomedical research: Increasing value, reducing waste. *The Lancet*, *383*(9912), 101-104. https://doi.org/10.1016/s0140-6736(13)62329-6

Mays, N., Pope, C., & Popay, J. (2005). Systematically reviewing qualitative and quantitative evidence to inform management and policy-making in the health field. *Journal of Health Services Research & Policy, 10*(Suppl 1), 6-20. https://doi.org/10.1258/1355819054308576

Mbuagbaw, L., Rochwerg, B., Jaeschke, R., Heels-Andsell, D., Alhazzani, W., Thabane, L., & Guyatt, G.H. (2017). Approaches to interpreting and choosing the best treatments in network meta-analyses. *Systematic Reviews, 6*(1), 79-83. https://doi.org/10.1186/s13643-017-0473-z

McClish, D. K. (1989). Analyzing a portion of the ROC curve. *Medical Decision Making, 9*(3), 190-195. https://doi.org/10.1177%2F0272989X8900900307

McCool, R., Wilson, K., Arber, M., Fleetwood, K., Toupin, S., Thom, H., Bennett, I., & Edwards, S. (2019). Systematic review and network meta-analysis comparing ocrelizumab with other treatments for relapsing multiple sclerosis. *Multiple Sclerosis and Related Disorders*, *29*, 55-61. https://doi.org/10.1016/j.msard.2018.12.040

McGough, J. J., & Faraone, S. V. (2009). Estimating the size of treatment effects: Moving beyond p values. *Psychiatry (Edgmont), 6*(10), 21-29.

McGowan, J., Sampson, M., Salzwedel, D. M., Cogo, E., Foerster, V., & Lefebvre, C. (2016). PRESS peer review of electronic search strategies: 2015 guideline statement. *Journal of Clinical Epidemiology, 75*, 40-46. https://doi.org/10.1016/j.jclinepi.2016.01.021

McKenzie, J. E., Brennan, S. E., Ryan, R. E., Thomson, H. J., & Johnston, R. V. (2019). Chapter 9: Summarizing study characteristics and preparing for synthesis. In J. P. T. Higgins, J. Thomas, J. Chandler, M. Cumpston, T. Li, M. J. Page, & V. A. Welch (Eds.), *Cochrane handbook for systematic reviews of interventions* version 6.0. Cochrane. www.training.cochrane.org/handbook

Metelli, S., & Chaimani, A. (2020). Challenges in meta-analyses with observational studies. *Evidence-Based Mental Health, 23*(2), 83-87. https://doi.org/10.1136/ebmental-2019-300129

Mills, E. J., Ioannidis, J. P. A., Thorlund, K., Schünemann, H. J., Puhan, M. A., & Guyatt, G.H. (2012). How to use an article reporting a multiple treatment comparison meta-analysis. *JAMA, 308*(12), 1246-1253. https://doi.org/10.1001/2012.jama.11228

Mills, E. J., Kanters, S., Thorlund, K., Chaimani, A., Veroniki, A-A., & Ioannidis, J. P. A. (2013). The effects of excluding treatments from network meta-analyses: survey. *BMJ, 347*, f5195. https://doi.org/10.1136/bmj.f5195

Moat, K. A., Lavis, J. N., & Abelson, J. (2013). How contexts and issues influence the use of policy-relevant research syntheses: A critical interpretive synthesis. *The Milbank Quarterly, 91*(3), 604-648. https://doi.org/10.1111/1468-0009.12026

Moher, D., Hopewell, S., Schulz, K. F., Montori, V., Gøtzsche, P. C., Devereaux, P. J., Elbourne, D., Egger, M., & Altman, D. G. (2010). CONSORT 2010 explanation and elaboration: Updated guidelines for reporting parallel group randomised trials. *BMJ, 340*, c869. https://doi.org/10.1136/bmj.c869

Moher, D., Booth, A., & Stewart, L. (2014). How to reduce unnecessary duplication: Use PROSPERO. *BJOG, 121*(7), 784-786. https://doi.org/10.1111/1471-0528.12657

Moher, D., Shamseer, L., Clarke, M., Ghersi, D., Liberati, A., Petticrew, M., Shekelle, P., Stewart, L. A., & PRISMA-P Group. (2015a). Preferred reporting items for systematic review and meta-analysis protocols (PRISMA-P) 2015 statement. *Systematic Reviews, 4*(1), 1. https://doi.org/10.1186/2046-4053-4-1

Moher, D., Stewart, L., & Shekelle, P. (2015b). All in the family: Systematic reviews, rapid reviews, scoping reviews, realist reviews, and more. *Systematic Reviews, 4*, 183. https://doi.org/10.1186/s13643-015-0163-7

Nagengast, A. J., Braun, D. A., & Wolpert, D. M. (2011). Risk-sensitivity and the mean-variance trade-off: Decision making in sensorimotor control. *Proceedings of the Royal Society B: Biological Sciences, 278*(1716), 2325-2332. https://doi.org/10.1098/rspb.2010.2518

Nikolakopoulou, A., Higgins, J. P. T., Papakonstantinou, T., Chaimani, A., Del Giovane, C., Egger, M., & Salanti, G. (2020). CINeMA: An approach for assessing confidence in the results of a network meta-analysis. *PLOS Medicine, 17*(4), e1003082. https://doi.org/10.1371/journal.pmed.1003082

Osborne, M. J. (2004). *An introduction to game theory*. Oxford University Press.

Oxman, A. D., & Guyatt, G. H. (1992). A consumer's guide to subgroup analyses. *Annals of Internal Medicine, 116*(1), 78-84. https://doi.org/10.7326/0003-4819-116-1-78

Perrier, L., Lightfoot, D., Kealey, M. R., Straus, S. E., & Tricco, A. C. (2016). Knowledge synthesis research: A bibliometric analysis. *Journal of Clinical Epidemiology, 73*, 50-57. https://doi.org/10.1016/j.jclinepi.2015.02.019

Petropoulou, M., Nikolakopoulou, A., Veroniki, A-A., Rios, P., Vafaei, A., Zarin, W., Giannatsi, M., Sullivan, S., Tricco, A. C., Chaimani, A., Egger, M., & Salanti, G. (2017). Bibliographic study showed improving statistical methodology of network meta-analyses published between 1999 and 2015. *Journal of Clinical Epidemiology, 82*, 20-28. https://doi.org/10.1016/j.jclinepi.2016.11.002

Phillippo, D. M., Dias, S., Ades, A. E., Didelez, V., & Welton, N. J. (2017). Sensitivity of treatment recommendations to bias in network meta‐analysis. *Journal of the Royal Statistical Society, Series A, Statistics in Society, 181*(3), 813-867. https://doi.org/10.1111/rssa.12341

Phillippo, D. M., Dias, S., Welton, N. J., Caldwell, D. M., Taske, N., & Ades, A. E. (2019). Threshold analysis as an alternative to GRADE for assessing confidence in guideline recommendations based on network meta-analyses. *Annals of Internal Medicine, 170*(8), 538-546. https://doi.org/10.7326/M18-3542

Puhan, M. A., Schünemann, H. J., Murad, M. H., Li, T., Brignardello-Petersen, R., Singh, J. A., Kessels, A. G., & Guyatt, G. H. (2014). A GRADE Working Group approach for rating the quality of treatment effect estimates from network meta-analysis. *BMJ, 349*, g5630. https://doi.org/10.1136/bmj.g5630

R Studio Inc. (2020). *Learn Shiny*. https://shiny.rstudio.com/tutorial/

Reeves, B. C., Deeks, J. J., Higgins, J. P. T., Shea, B., Tugwell, P., & Wells, G. A. (2019). Chapter 24: Including non-randomized studies on intervention effects. In J. P. T. Higgins, J. Thomas, J. Chandler, M. Cumpston, T. Li, M. J. Page, & V. A. Welch (Eds.), *Cochrane handbook for systematic reviews of interventions* version 6.0. Cochrane. www.training.cochrane.org/handbook

Richardson, W. S., Wilson, M. C., Nishikawa, J., & Hayward, R. S. A. (1995). The well-built clinical question: A key to evidence-based decisions. *ACP Journal Club, 123*(3), A12-3.

Riley, R. D., Higgins, J. P. T., & Deeks, J. J. (2011). Interpretation of random effects meta-analyses. *BMJ, 342*, d549. https://doi.org/10.1136/bmj.d549

Riley, R. D., Jackson, D., Salanti, G., Burke, D. L., Price, M., Kirkham, J., & White, I. R. (2017). Multivariate and network meta-analysis of multiple outcomes and multiple treatments: rationale, concepts, and examples. *BMJ, 358*, j3932. https://doi.org/10.1136/bmj.j3932

Rogliani, P., Matera, M. G., Ritondo, B. L., De Guido, I., Puxeddu, E., Cazzola, M., & Calzetta, L. (2019). Efficacy and cardiovascular safety profile of dual bronchodilation therapy in chronic obstructive pulmonary disease: A bidimensional comparative analysis across fixed-dose combinations. *Pulmonary Pharmacology & Therapeutics*, *59*, 101841. https://doi.org/10.1016/j.pupt.2019.101841

Rücker, G. (2012). Network meta-analysis, electrical networks and graph theory. *Research Synthesis Methods*, *3*(4), 312-324. https://doi.org/10.1002/jrsm.1058

Rücker, G., & Schwarzer, G. (2014). Reduce dimension or reduce weights? Comparing two approaches to multi-arm studies in network meta-analysis. *Statistics in Medicine*, *33*(25), 4353-4369. https://doi.org/10.1002/sim.6236

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

Salanti, G., Higgins, J. P. T., Ades, A. E., & Ioannidis, J. P. A. (2008). Evaluation of networks of randomised trials. *Statistical Methods in Medical Research*, *17*(3), 279-301. https://doi.org/10.1177/0962280207080643

Salanti, G., Ades, A. E., & Ioannidis, J. P. A. (2011). Graphical methods and numerical summaries for presenting results from multiple-treatment meta-analysis: An overview and tutorial. *Journal of Clinical Epidemiology*, *64*(2), 163-171. https://doi.org/10.1016/j.jclinepi.2010.03.016

Salanti, G., Giovane, C. D., Chaimani, A., Caldwell, D. M., & Higgins, J. P. T. (2014). Evaluating the quality of evidence from a network meta-analysis. *PLoS One, 9*(7), e99682. https://doi.org/10.1371/journal.pone.0099682

Sekhon, M., Cartwright, M., & Francis, J. J. (2017). Acceptability of healthcare interventions: An overview of reviews and development of a theoretical framework. *BMC Health Services Research, 17*(1), 88. https://doi.org/10.1186/s12913-017-2031-8

Shamseer, L., Moher, D., Clarke, M., Ghersi, D., Liberati, A., Petticrew, M., Shekelle, P., & Stewart, L. A. (2015). Preferred reporting items for systematic review and meta-analysis protocols (PRISMA-P) 2015: Elaboration and explanation. *BMJ, 349,* g7647. https://doi.org/10.1136/bmj.g7647

Shea, B. J., Reeves, B. C., Wells, G., Thuku, M., Hamel, C., Moran, J., Moher, D., Tugwell, P., Welch, V., Kristjansson, E., & Henry, D. A. (2017). AMSTAR 2: a critical appraisal tool for systematic reviews that include randomised or non-randomised studies of healthcare interventions, or both. *BMJ, 358,* j4008. https://doi.org/10.1136/bmj.j4008

Sniazhko, S. (2019). Uncertainty in decision-making: A review of the international business literature. *Cogent Business & Management, 6*(1), 1650692. https://doi.org/10.1080/23311975.2019.1650692

Royal Pharmaceutical Society. (2011). *Different types of evidence/literature reviews*. https://studylib.net/doc/8647248/series-two--article-four--different-types-of-evidence-lit

Sox, H. C. (2010). Defining comparative effectiveness research: The importance of getting it right. *Medical Care, 48*(6), S7-S8. https://doi.org/10.1097/MLR.0b013e3181da3709

Stafoggia, M., Lallo, A., Fusco, D., Barone, A. P., D'Ovidio, M., Sorge, C., & Perucci, C. A. (2011). Spie charts, target plots, and radar plots for displaying comparative outcomes of health care. *Journal of Clinical Epidemiology, 64*(7), 770-778. https://doi.org/10.1016/j.jclinepi.2010.10.009

Sterne, J. A. C., Savović, J., Page, M. J., Elbers, R. G., Blencowe, N. S., Boutron, I., Cates, C. J., Cheng, H.-Y., Corbett, M. S., Eldridge, S. M., Emberson, J. R., Hernán, M. A., Hopewell, S., Hróbjartsson, A., Junqueira, D. R., Jüni, P., Kirkham, J. J., Lasserson, T., Li, T., McAleenan, A., Reeves, B. C., Shepperd, S., Shrier, I., Stewart, L. A., Tilling, K., White, I. R., Whiting, P. F., & Higgins, J. P. T. (2019). RoB 2: a revised tool for assessing risk of bias in randomised trials. *BMJ, 366,* l4898. https://doi.org/10.1136/bmj.l4898

Stevens, S. S. (1946). On the theory of scales of measurement. *Science, 103*(2684), 677-680. https://doi.org/10.1126/science.103.2684.677

Stewart, L., Moher, D., & Shekelle, P. (2012). Why prospective registration of systematic reviews makes sense. *Systematic Reviews, 1*(1), 7. https://doi.org/10.1186/2046-4053-1-7

Sullivan, G. M., & Feinn, R. (2012). Using effect size-or why the p value is not enough. *Journal of Graduate Medical Education, 4*(3), 279-282. https://doi.org/10.4300/JGME-D-12-00156.1

Tierney, J. F., Stewart, L. A., Ghersi, D., Burdett, S., & Sydes, M. R. (2007). Practical methods for incorporating summary time-to-event data into meta-analysis. *Trials, 8*(1), 16. https://doi.org/10.1186/1745-6215-8-16

Tricco, A. C., Antony, J., Soobiah, C., Kastner, M., Cogo, E., MacDonald, H., D'Souza, J., Hui, W., & Straus, S. E. (2016a). Knowledge synthesis methods for generating or refining theory: A scoping review reveals that little guidance is available. *Journal of Clinical Epidemiology, 73*, 36-42. https://doi.org/10.1016/j.jclinepi.2015.11.021

Tricco, A. C., Soobiah, C., Antony, J., Cogo, E., MacDonald, H., Lillie, E., Tran, J., D'Souza, J., Hui, W., Perrier, L., Welch, V., Horsley, T., Straus, S. E., & Kastner, M. (2016b). A scoping review identifies multiple emerging knowledge synthesis methods, but few studies operationalize the method. *Journal of Clinical Epidemiology*, *73*, 19-28. https://doi.org/10.1016/j.jclinepi.2015.08.030

Tricco, A. C., Zarin, W., Ghassemi, M., Nincic, V., Lillie, E., Page, M. J., Shamseer, L., Antony, J., Rios, P., Hwee, J., Veroniki, A. A., Moher, D., Hartling, L., Pham, B., & Straus, S. E. (2018). Same family, different species: Methodological conduct and quality varies according to purpose for five types of knowledge synthesis. *Journal of Clinical Epidemiology, 96*, 133-142. https://doi.org/10.1016/j.jclinepi.2017.10.014

Trinquart, L., Attiche, N., Bafeta, A., Porcher, R., & Ravaud, P. (2016). Uncertainty in treatment rankings: reanalysis of network meta-analyses of randomized trials. *Annals of Internal Medicine, 164*(10), 666-673. https://doi.org/10.7326/M15-2521

Tversky, A., & Kahneman, D. (1974). Judgment under uncertainty: Heuristics and biases. *Science, 185*(4157), 1124-1131. https://doi.org/10.1126/science.185.4157.1124

Veroniki, A. A., Mavridis, D., Higgins, J. P. T., & Salanti, G. (2014). Characteristics of a loop of evidence that affect detection and estimation of inconsistency: A simulation study. *BMC Medical Research Methodology, 14*(1), 106. https://doi.org/10.1186/1471-2288-14-106

Veroniki, A. A., Straus, S. E., Rücker, G., & Tricco, A. C. (2018). Is providing uncertainty intervals in treatment ranking helpful in a network meta-analysis? *Journal of Clinical Epidemiology, 100*, 122-129. https://doi.org/10.1016/j.jclinepi.2018.02.009

Welton, N. J., Ades, A. E., Carlin, J. B., Altman, D. G., & Sterne, J. A. C. (2009). Models for potentially biased evidence in meta-analysis using empirically based priors. *Journal of the Royal Statistical Society. Series A, Statistics in Society, 172*(1), 119-136. https://doi.org/10.1111/j.1467-985X.2008.00548.x

White, I. R., Barrett, J. K., Jackson, D., & Higgins, J. P. T. (2012). Consistency and inconsistency in network meta-analysis: Model estimation using multivariate meta-regression. *Research Synthesis Methods, 3*(2), 111-125. https://doi.org/10.1002/jrsm.1045