BMJ Open

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]

To cite: Daly CH, Neupane B, Beyene J, *et al*. Empirical evaluation of SUCRA-based treatment ranks in network meta-analysis: quantifying robustness using Cohen's kappa. *BMJ Open* 2019;**9**:e024625. doi:10.1136/bmjopen-2018-024625

► Prepublication history and additional material for this paper are available online. To view these files, please visit the journal online (http://dx.doi.org/10.1136/bmjopen-2018-024625).

Received 11 December 2018
Revised 02 July 2019
Accepted 24 July 2019

For numbered affiliations see end of article.

Correspondence to
Dr Jemila S Hamid;
jehamid@cheo.on.ca

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.

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

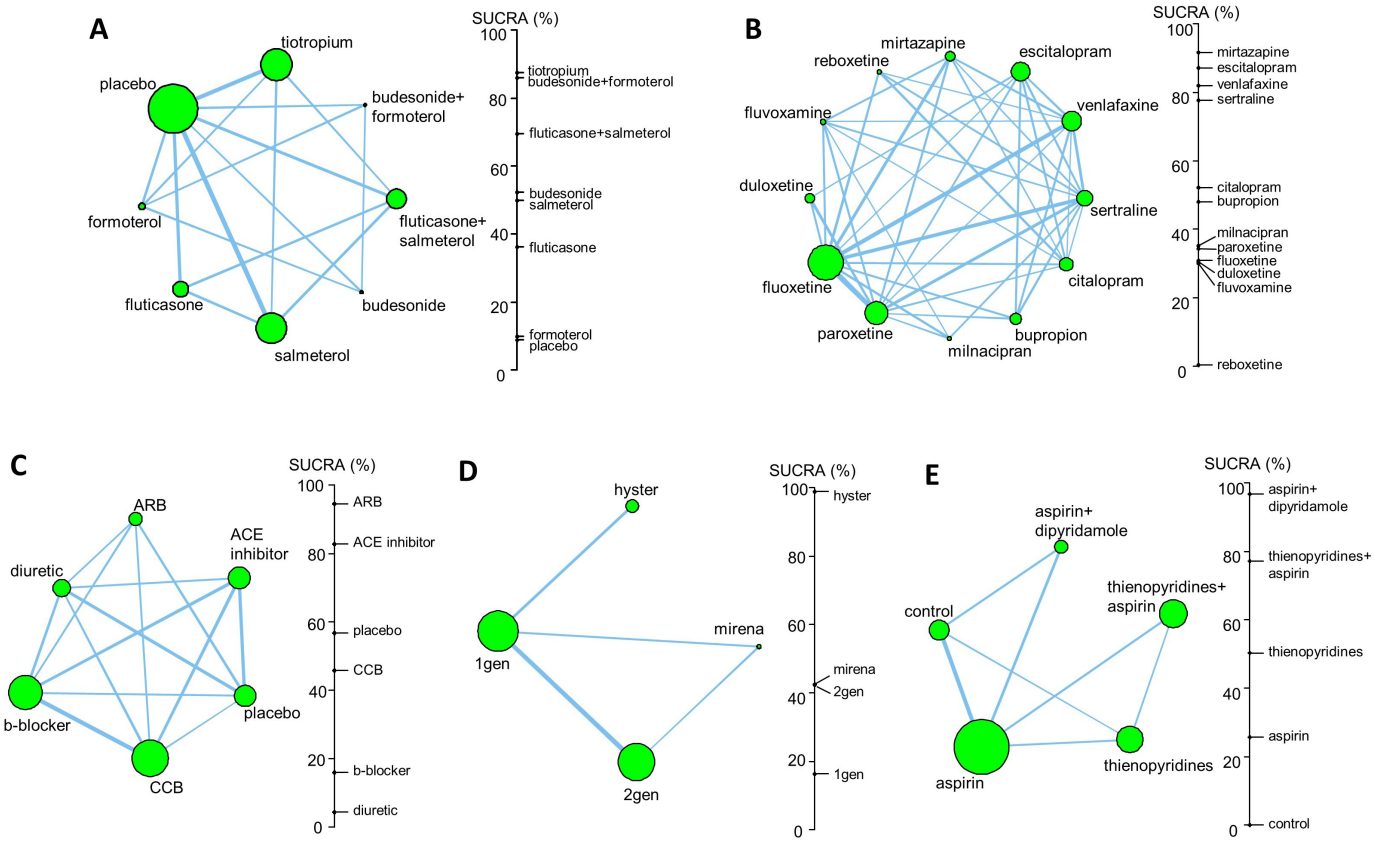


Figure 1 Network diagrams (left) and SUCRA (right) for (A) chronic obstructive pulmonary disease network,[21] (B) depression network,[22] (C) diabetes network,[23] (D) heavy menstrual bleeding network and[24] (E) stroke network.[25]The sizes of the nodes are proportional to the number of patients randomised 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.

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 at 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 100 000 samples with a thinning rate of 10 after an adaption phase of 20 000 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 $n_s = \frac{k_s(k_s-1)}{2}$ 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 (LL_S, UL_S) into CrIs for the expected rank (LL_r, UL_r) using this relation:

$$(LL_r, UL_r) = (n - (n-1)\,LL_S, n - (n-1)\,UL_S)$$

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 ($\kappa_W = 1$)). The smallest beyond chance agreement was observed in the heavy menstrual bleeding network ($\kappa_{UW} = 0\%$). 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 $\kappa_W = 40\%$.

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 $\kappa_{UW} = 82\%$. 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 $\kappa_W = 89\%$, while in the latter situation, the weighted agreement was $\kappa_W = 99\%$.

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 ($\kappa_W = 83\%$) and study 18 ($\kappa_W = 81\%$) (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).

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

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

Network	Number of trials that did not change any treatment rank (%)	Among studies that impacted treatment's ranks… Median of observed kappas (κ) (minimum, maximum)		Treatments*	Number (%) of studies displacing treatment's rank by…			
		Unweighted	Weighted		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%)	Angiotensin-receptor blockers	0 (0%)	0 (0%)	0 (0%)	0 (0%)
				Angiotensin-converting-enzyme inhibitors	0 (0%)	0 (0%)	0 (0%)	0 (0%)
				Placebo	1 (100%)	0 (0%)	0 (0%)	0 (0%)
				Calcium-channel blockers	1 (100%)	0 (0%)	0 (0%)	0 (0%)
				β blocker	0 (0%)	0 (0%)	0 (0%)	0 (0%)
				Diuretic	0 (0%)	0 (0%)	0 (0%)	0 (0%)

Continued

Table 1 Continued

Network	Number of trials that did not change any treatment rank (%)	Among studies that impacted treatment's ranks… Median of observed kappas (κ) (minimum, maximum) Unweighted	Weighted	Treatments*	Number (%) of studies displacing treatment's rank by… 1 rank	2 ranks	3 ranks	4 ranks
Heavy menstrual bleeding[24]	16 (80%)	33% (0%, 33%)	80% (40%, 80%)	Hysterectomy	0 (0%)	0 (0%)	0 (0%)	0 (0%)
				Mirena	3 (75%)	1 (0%)	0 (0%)	0 (0%)
				2nd generation endometrial destruction	4 (100%)	0 (0%)	0 (0%)	0 (0%)
				1st generation endometrial destruction	1 (25%)	0 (0%)	0 (0%)	0 (0%)
Stroke[25]	26 (100%)	N/A	N/A	Aspirin+dipyridamole	N/A			
				Thienopyridines+aspirin				
				Thienopyridines				
				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.

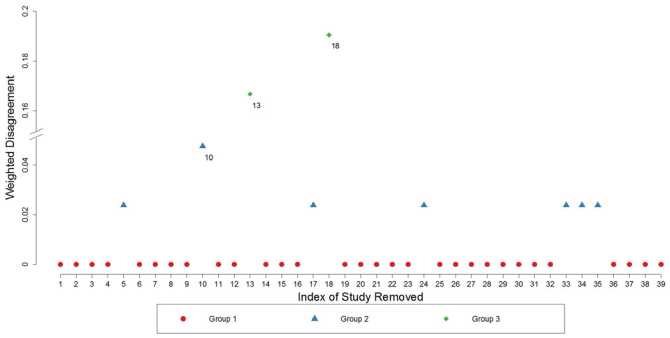


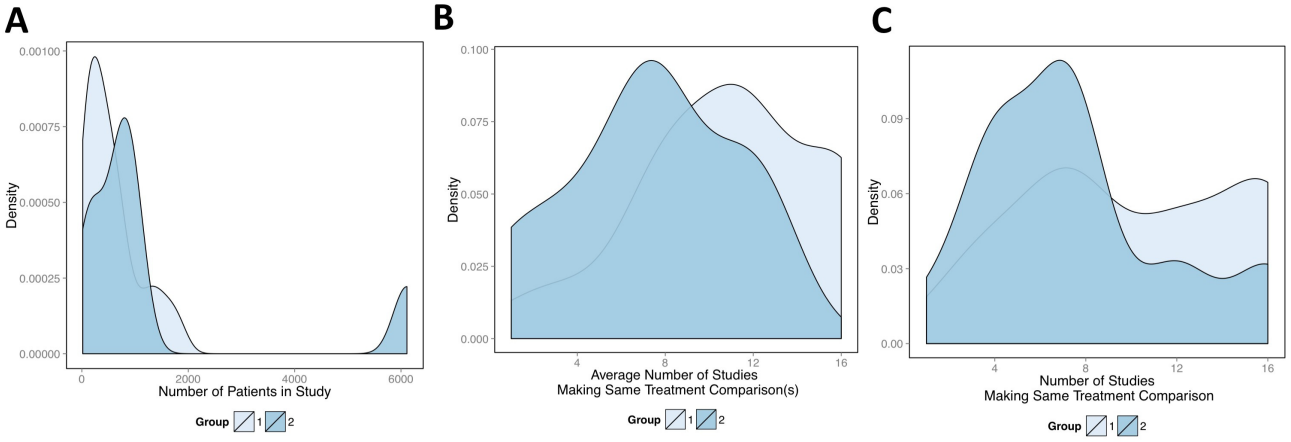
Figure 2 Observed 1−quadratic weighted kappa (κ_W) in the chronic obstructive pulmonary disease network,[21] which quantifies the weighted disagreement between the treatment ranks produced from the complete data set and the ranks produced from a sub-data set where one RCT (indexed on the x-axis) was 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.

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).

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



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 (K_W=1), and group 2, where the individual removal of these RCTs had a small impact on treatment rankings (0.95<K_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.

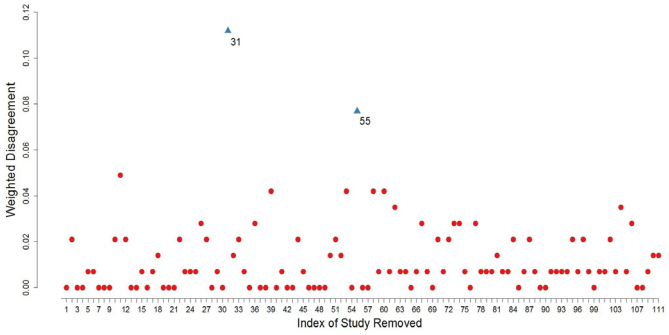


Figure 4 Observed 1−quadratic weighted kappa (κ_W) in the depression network,[22] which quantifies the weighted disagreement between the treatment ranks produced from the complete data set and the ranks produced from a sub-data set where one RCT (indexed on the x-axis) was removed. RCT, randomised controlled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

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.

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).

Author affiliations
[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

Contributors CHD conceptualised and designed the study, analysed the data, interpreted the results of the empirical evaluation, and drafted and revised the manuscript. BN and JB provided input into the study design, acquired the data and revised the manuscript. LT and SES provided input into the study design and revised the manuscript. JSH conceptualised and designed the study and revised the manuscript. All the authors approved the final version of the submitted manuscript.

Funding This work was supported in part by an Ontario Graduate Scholarship granted to CHD.

Competing interests None declared.

Patient consent for publication Not required.

Provenance and peer review Not commissioned; externally peer reviewed.

Data availability statement Data may be obtained from a third party and are not publicly available.

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 meta-analysis: 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 meta-analyses 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 multiple-treatments 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 meta-analyses. *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.