Reviewer’s report

Title: Ranking treatments in frequentist network meta-analysis works without resampling methods

Version: 2 Date: 19 May 2015

Reviewer: Dan Jackson

Reviewer’s report:

This is an interesting paper that looks at ranking in classical network meta-analysis. The suggestion of P scores is a useful classical analogue of methods that are used in Bayesian network analyses (SUCRA). I am therefore broadly positive about the paper. I do however have a series of minor essential revisions. The authors can be trusted to make these, so I do not need to see the paper again, but the editor should ensure that these points are addressed; once they have been addressed the paper should, in my opinion, be accepted. I think that this journal is a suitable home for this work.

Minor Essential Revisions

1. On page 3 of 14, we have arm based effects. Thus I think the approach only makes sense under an arm based analysis? If we use contrasts (treatment effects) in a contrast based analysis then all the estimates of the average effects are correlated and things become more complicated? The subtle issue is that a contrast based analysis is of reduced dimension to the arm based analysis (it is of dimension one less of course) so it is clear what the arm based analysis implies in terms of the contrasts (as the authors use to their advantage on page 3 of 14) but not necessarily vice versa. Perhaps the authors should say that they are assuming an arm based analysis? If not they should explain how all of this translates to a contrast based analysis of reduced dimension.

2. The authors should be aware that White et al in "Consistency and inconsistency in network meta-analysis: model estimation using multivariate meta-regression" in RSM also performed ranking whilst using classical methods, using a bootstrap procedure (see their section "Ranking in the consistency model"). This alternative idea should be compared and contrasted to the proposed method.

3. The authors’ arguments seem to be mostly limited to comparing two treatments in the network. See equation (1) which continues to examine pairwise comparisons. Rather than being interested in the pairwise probabilities to perform ranking as in equation (1) in a Bayesian framework we might be interested in the probability that a particular treatment is best. If so we want P(Y_i > Y_j, all j not equal to i). See Table 1 of A "design#by#treatment interaction model for network meta-analysis with random inconsistency effects" in SiM and look at treatments D and U. Both have similar estimates (-0.78) but U has a markedly different
probability of being the best. Sure $P(Y_i > Y_j)$ is close to 0.5 but to be the best we have to beat all the other treatments as well. Rather perversely, U benefits from its large uncertainty (0.58 SD) rather than (0.16 SD) and so has some chance of being best - because we know rather little about it! The authors motivate a method of ranking that seems to be based on pairwise comparisons (the P score is based on the $P_{ij}$, which are pairwise comparisons) but this might be a poor indicator of $P$ (best), which might also be important. Issues like this should be discussed.

4. Continuing with point 3, the authors seem to be sold on SUCRA and want a classical analogue. But the SUCRA really only is one way of ranking treatments and the authors should question whether this is really always appropriate. SUCRA is not without its issues. Ranking is always likely to be sensitive for the criterion used you see! Take the Olympics medal table for eg. The conventional method for ranking is by number of golds. In the event of a tie we then rank by number of silvers. And then in the event of a tie we then rank by number of bronzes.

http://www.bbc.co.uk/sport/winter-olympics/2014/medals/countries

But we might instead let countries score 3 for a gold, 2 for a silver and 1 for a bronze and rank according to the total score. This will change the ranking! And so will other scoring systems! Thus, in a network, there will always be different methods for ranking treatments that will result in different rankings and the SUCRA is just one way to do this (and is not without its limitations as I say). I was hoping for much more intelligent discussion of these issues but I did not really find any.

5. The authors use $Y_i$ and $Y_j$, later we have $\hat{\mu}_i$ and $\hat{\mu}_j$. Why not just use $\hat{\mu}_i$ and $\hat{\mu}_j$ throughout, its the parameter estimates that we are interested in so I found the Y notation unnecessary and confusing. Similarly in (1) we have $P(Y_i > Y_j)$ and later we have $P_{ij} = \Phi(\ldots)$. Since $Y_i$ and $Y_j$ are the estimates $\hat{\mu}_i$ and $\hat{\mu}_j$, again I found the $Y_i$ and $Y_j$ annoying! Or is the point that $Y_i$ and $Y_j$ are the Bayesian quantities and $\hat{\mu}_i$ and $\hat{\mu}_j$ are the classical ones? I wonder if the use of both T and $\hat{\mu}$ is for some such subtle reason (which should be explained if so) or just unclear thinking and presentation of the methods?

Apologies for being rather picky in the above points - I do think the paper has many positive points and I am positive about it as I say. Hopefully the above points will help the authors make the paper even better -- and also make them question SUCRA more, which the authors seem to take as set in stone, whereas in reality it is anything but!

**Level of interest:** An article whose findings are important to those with closely related research interests

**Quality of written English:** Acceptable
Statistical review: Yes, and I have assessed the statistics in my report.

Declaration of competing interests:
I declare that I have no competing interests’