Reviewer's report

Title: Bayesian Meta-analysis of Comparative Diagnostic Test Accuracy Studies

Version: 3

Date: 1 February 2015

Reviewer: Yemisi Takwoingi

Reviewer's report:

This is a well written paper addressing the important issue of meta-analyses of comparative test accuracy. The authors propose a model that allows for direct and indirect test comparisons, and imperfect reference standard. This model is an important contribution to meta-analysis of diagnostic accuracy studies. However, I have some comments as follows.

Minor Essential Revisions

1. Background, page 1 lines 44-46: the authors state that the methods suggested tend to ignore trial effects by handling comparisons as subgroup analyses. The reference cited is Chapter 10 of the Cochrane Handbook for Systematic Reviews of DTA. The handbook recommends a regression modelling framework. For example in the case of two tests, the handbook states that “the data for the two tests must be analysed within study at level one of the analysis, and a binary covariate for test type included to identify which 2×2 table corresponds to each test”. This regression modelling approach is not mentioned at all in the paper or considered as one of the possible models.

2. I'm not sure scenario 2 where different reference standards were used depending on the test is realistic. I think it is more often the case that where a study uses more than one reference standard, the choice of reference standard was determined by test result or local availability and not the test per se.

3. For the simulation studies, what is the justification for 250 datasets?

4. For the simulation study about selection of the most appropriate statistic, what is the justification for 10,000 subjects and why was performance assessed only in terms of bias?

5. There is no mention of how you assessed convergence.

6. The logit was the link function used and so the natural output of the models is the log OR. How did you estimate absolute differences in S and C between tests? The same also applies to the relative S and C.

7. Since the model produces average estimates of the difference in diagnostic accuracy between two tests, please make it clear how you obtained the summary S and C for each test with 95% CrI.

8. The authors focus solely on bias and do not discuss power and coverage results at all in the main manuscript even though this is how they assessed the models. Power to detect a difference in S and C was generally poor. Any explanation for this?
9. A limitation of the proposed approach is that it does not account for the pairing of test results if all tests were applied to all patients in a study. This should be mentioned with reference to the approach by Trikalinos et al. [Trikalinos T. A., Hoaglin D. C., Small K. M., Terrin N. and Schmid C. H. (2014), Methods for the joint meta-analysis of multiple tests, Research Synthesis Methods, 5, pages 294–312, doi: 10.1002/jrsm.1115]

10. Page 10 lines 4-5 – I disagree with the statement that relative S and C is seldom used. Measures of relative accuracy are not often reported and when they are reported it is typically the rDOR or relative S and C. Absolute differences may be easier to interpret and accessible to readers but are rarely, if at all reported because this measure cannot be readily derived from the bivariate or HSROC models. The rDOR is a natural output of the HSROC model, and the relative S and C can be derived using functions of the HSROC or bivariate model parameters (e.g. by using ESTIMATE statements in SAS or post estimation commands in Stata to estimate log relative S and C and standard errors are obtained using the delta method).

11. Page 4 line 3 – Comparisons between tests can be summarised using the difference S-C or the relative risks S/C. This is an unusual measure that I have not come across before and don’t see why such a measure would be used. Is this an error?

12. Page 5 line 1 – do you mean compare the estimates... across meta-analyses and not across studies?

Discretionary Revisions
13. Maybe better to use the term ‘study’ instead of ‘trial’ when referring to studies of diagnostic accuracy.

Level of interest: An article of importance in its field

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.