Reviewer's report

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

Version: 3

Date: 4 February 2015

Reviewer: Nandini Dendukuri

Reviewer's report:

Review of Bayesian Meta-analysis of Comparative Diagnostic Test Accuracy Studies by Menten and Lesaffre

With the growing interest in measuring the comparative effectiveness of competing health technologies, the methods in this article are likely to be of much interest to researchers studying comparative diagnostic test accuracy. The article illustrates how to combine multivariate meta-analysis models and latent class models, and in order to compare sensitivity and specificity of different diagnostic tests using several possible statistics. The article is well written and I appreciate that the programs and data are provided. However, it is very long and almost like 3 articles in one, with Appendices 1 and 3 each possibly being articles in their own merit. So my suggestions for improvement below are not for adding more methods or results but rather as a request for better explanation to help the reader navigate what is already presented and cutting down text which is redundant or very basic in nature:

Discretionary Revisions

• Suggest editing title as “A general framework for comparative Bayesian meta-analysis of diagnostic studies” reflecting the fact that this article provides a very general description that will need to be further elaborated and tested in future articles.

• Sub-section heading “Meta-Analytic Models Based on a Reference Standard” could be “Meta-Analytic Models when a perfect Reference Standard is available”. Similarly, suggest replacing “Meta-Analytic Models Based on Latent Class Analysis” by “Meta-Analytic Models when no perfect Reference Standard is available”, or something to that effect.

Major Compulsory Revisions

• First para of page 7, suggest replacing last two sentences beginning “In its basic format, ...” with “LCA models have been described for a variety of situations ranging from when a single imperfect test is observed in each study to more complex designs involving multiple tests. When multiple tests are involved, they may be treated as independent conditional on the disease status or the conditional dependence between them may be modeled using a variety of approaches” or equivalent. Especially, since the authors acknowledge that prior information may be used I think the requirement for 3 tests is unnecessary. Indeed, in some of the simulation settings there are only 2 tests in a study. Also,
though traditionally LCAs are described beginning with the conditional independence model, I think this is better described as a candidate model among many possibilities.

• When multiple imperfect tests are used it is important that they are in fact measuring the same underlying latent variable (see, for example, Dendukuri et al. Stats in Med, 2009). Ignoring this may result in mislabelling of the latent classes identified and also invalidate the resulting test accuracy estimates. This problem will be further aggravated in a meta-analysis setting with each study using a different combination of tests. The authors should emphasize the importance of this assumption and clarify whether this is a reasonable assumption for their motivating problem.

• The problem of inconsistency of treatment effects is an important problem in the multiple treatment comparisons framework that the authors draw inspiration from. What is the equivalent in the comparative diagnostic accuracy setting? Are the proposed models susceptible to these problems? If so, how have the authors check for and address these problems?

• Please provide more detail on the structure of the variance-covariance matrices in Models 4 and 5. The authors mention these will be complex, but it would be helpful if they could go further and illustrate how within-study correlation between observations on the same patients, and between-study correlations on different contrasts could be defined. Importantly, how are the models used related to those previously used for conditional dependence? We are used to the concept of modeling the covariance between sensitivity and specificity across studies (as in Model 1) but what would be reasoning behind modeling covariance between the contrasts that form the main outcomes in the models proposed?

• Related to the above, the authors should mention the consequences of the simplifying assumptions they make.

• In the simulation study described in Appendix 1, it is unclear from which model the data were generated (from equation (5)?) and which model was fit to the data. For example, if a conditional independence model was fit to data generated from a conditional dependence model then it is well documented in the literature that the estimates of sensitivity and specificity will be biased. The authors should refer to this literature and comment whether their results are in agreement or not. More interesting, of course, are their observations on bias in the comparative statistics.

• In this same study, please explain why prevalence of 50% was chosen. And mention how results would be different, if at all, if the prevalence were higher or lower than 50%. Please also provide some insight as to the observed direction of bias, e.g. why does greater correlation between T1 and T3 than between T2 and T3 result in a certain type of bias?

• The simulation study described in Appendix 1 does not cover the case when a perfect reference test is available. What can we say then about the comparison statistics in this case? Does the second simulation in Appendix 3 show the relevant results?

• Please explain with reference to a real-life problem why investigators may be
unaware that a reference standard is different across studies.

• Across the various simulation results in Appendix 3, besides the cases where there was significant bias, it was not clear to me which is the most preferable model. Please explain is there any apparent advantage of Model 5 over 4 in Table 2 and Model 5a vs 4a in Table 5 because it uses all available data? Similarly, is Model 3 to be preferred over Model 2 and Model 1 in Table 2.

Minor Essential Revisions:
• Under Model 4. Again, like Model 1, this model ignores the correlation …

**Level of interest:** An article of importance in its field

**Quality of written English:** Acceptable

**Statistical review:** No, the manuscript does not need to be seen by a statistician.

**Declaration of competing interests:**

I declare that I have no competing interests