Reviewer's report

Title: A modified Wald interval for the area under the ROC curve (AUC) in diagnostic case-control studies

Version: 2
Date: 11 November 2013
Reviewer: Philipp Doebler

Reviewer's report:

Major Compulsory Revisions

1.1) Motivation for work on confidence intervals: You name two potential applications of your methodology, neither of which I find completely convincing. In the introduction you mention the identification of biomarkers as an area where CIs for the AUC are needed. You mention the situation where an AUC_0 of 0.8 is used as selection criterion. I believe you need to address a different (multiple) testing problem in this situation, namely the one-sided null-hypothesis AUC < c, for some c. I agree that a well-working two-sided interval can be helpful here, but I guess it is really one-sided intervals that are of interest. Two-sided intervals would be better suited for comparisons of markers.

In the conclusion part you mention diagnostic meta-analysis. I am not convinced that anyone would reduce a diagnostic meta-analysis to pooling AUCs, especially as point estimates for AUCs are typically side-products from bivariate models (see for example the paper of Gatsonis and Paliwal (2006) on this point or the guidelines of Deeks et al (2001) in BMJ). Strong support for your work could stem from examples of biomarker research that uses two-sided CIs for AUCs. You might also want to consider providing an example in that direction (see comment below).

1.2) I understand, that the ratios 1:1 or 1:2 stem from the Cochrane and Ebmeier paper on Parkinson (you also base your sample size considerations on this paper). As AUC methodology is used in many other areas, I doubt that this meta-analysis is the best starting point for the design of your simulation study. In Figure 3 one can see, that the LT seems less affected by imbalanced designs than the MW; is this problematic for more extreme ratios? Did you look into ratios that are realistic in studies screening for rare diseases? say 1:10? Especially if short screens are evaluated in a community sample, the prevalence can become very small.

1.3) The majority of your simulations are based on homogeneous variances. A common (and rarely addressed) point is that variances are often larger in the cases than in the controls. You discuss this in a very small paragraph for a small subset of the intervals you consider. Clearly the (change in coverages) you report, tells us nothing about the variability of the point estimate and also we learn nothing about the lower bound of the interval (which might be essential for inference). I am curious to know whether the mean length of the confidence
intervals is affected in any way. I hence think it is worth studying the heteroscedastic case in more detail; this could also add value in terms of AUC point-estimation.

Minor Essential Revisions

2.1) I believe the cases and controls are mixed up in the example: The following is taken from the variable key supplied on the website you mention: "non-cancer patients (d==0) are pancreatitis controls" (so, there are more cases than controls here).

2.2) I am confused that you were unable to calculate the binormal interval in the example. I think this is well possible as all the quantities you state in Table 1 (mean, sd for each group) can be calculated from the data set.

2.3) Table 1:
- \sqrt(n) should be replaced by \sqrt{n} or n^{1/2} (can omit the \frac if using the latter)
- Use (quantiles of) the Beta distribution for CP interval, that explains what is going on.
- Some of the "AUC"s are really \widehat{AUC}, aren't they?
- You may want to replace \widehat{AUC} by AUC^\hat and the same goes for the wide tilde
- definition of s1, s2 is missing (binormal AUC)

2.4) page 3: "..., and it ." this sentence has no end

2.5) page 2: While it is clear to me, that the AUC is a probability, you should supply a reference to this fact or just mention the appropriate integral.

2.6) The example needs some improvement, also to motivate your work. Here you could consider to analyse both biomarkers from the original Wieand data set and compare their performance based on your AUC intervals.

2.7) Design of your simulation study: I would recommend to use fewer values of the sample size. There is virtually no difference between 80 and 100 and one can see a clear trend in the sample size figure, so I would probably choose 20 and 250 (and maybe 100 if I suspected something interesting in between). In general I would use three values for a single factor at most (following the logic of optimal design of experiments where one uses values in between the extremal values only if one wants to test for interaction effects or quadratic terms).

Discretionary Revisions

3.1) page 2: Omit reference to PROC LOGISTIC, also on page 7, the reason should be stated, not the error message. I would recommend to omit the references to SAS on page 2, as this is not a software tutorial. Your findings are independent of the software used.
3.2) You might want to make the point estimator for the AUC explicit. I feel the paper could be more self-contained here.

3.3) It would be nice to summarize the design of the simulation study, say in a table, so that it is instantly clear which factors your varied at what levels.

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

**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