Reviewer's report

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

Version: 2 Date: 4 December 2013

Reviewer: Gerta Rücker

Reviewer's report:

General

The paper reports a simulation study comparing a number of confidence intervals for the area under the curve (AUC) for case-control diagnostic test accuracy (DTA) studies. AUC is a measure of suitability of a biomarker for discriminating persons with and without disease at all, not depending on the choice of a special cut-off. As such, its role is more important in early phase DTA studies. The restriction to the case-control design is due to the approximately balanced sample size of cases and controls that cannot be assumed in cohort studies that would represent the patient mix for which the index test would likely be used in clinical practice.

Many of the confidence intervals considered here have been proposed in the literature for probabilities in general. AUC can be interpreted as a probability: Given that larger values indicate disease, AUC is the probability that for a randomly drawn pair of observations (X,Y) where X is a case and Y is a control, the value of X is greater than the value of Y.

In the introduction, a number of confidence intervals for proportions is listed, referring, e.g., to the article by Newcombe (ref [11]) and other references. In the methods part the investigated intervals are explained and defined, and a comprehensive simulation design is presented. Criteria for evaluation are description of the interval length, coverage probability, power, and robustness with respect to unbalanced sample size up to 1:2, violation of normality assumption, discretisation leading to ordinal data, and heteroscedasticity. The intervals are applied to an example from the literature. The authors conclude that a modified Wald interval has the most adequate performance and is easily implemented.

Major Compulsory Revisions

None.

Minor Essential Revisions

page 4, middle: What is monotone posterior? Please explain.
page 4, middle: What does "stochastically comparable" mean?
page 5, top: What does "the AUC under the null hypothesis" mean here? As I
understand this, given the true variances, the means of cases and controls (truly) differ by (say) delta such that \( \text{AUC} = \text{AUC}_0 \). I would not call \( \text{AUC}_0 \) "the null hypothesis", but the true value of AUC, as you primarily want to estimate AUC, not to test a hypothesis.

**English language**

page 1, last line: Delete "And" and start sentence with "For".

page 2, line 2: Put a comma after "example".

page 2, line 5: Delete comma after "selected".

page 2, bottom: Replace "AUC is nothing more than a probability" with "AUC is nothing but a probability" or simply say "AUC is interpreted as a probability"

page 3, Methods, 3rd sentence. Better say: "Each individual observation \( X_{is} (s = 1, ..., n_i) \) in group \( i=0,1 \) follows the normalized version of ... (etc.)".

**Typo**

page 4, line 1: Replace "Cnfidence" with "Confidence".

**Table 1**
The formulas, particularly the various point estimates of AUC (\( \hat{\text{AUC}}, \tilde{\text{AUC}}, \text{AUC}^* \)) and standard deviation (\( s_{\{\text{AUC}\}} \)) and especially the standard error (\( s_{\{\hat{\text{AUC}}\}} \)) are printed very small, difficult to distinguish and easily confused. I propose to separate the table into two tables, one for the "ingredients" (point estimates, standard deviation and standard error), giving them short names, where possible without subscripts (such as \( a, sd, se, ... \)), and another table for the confidence intervals, using this slimmed-down notation.

**Discretionary Revisions**

I am not a specialist for confidence intervals of proportions. Thus I was a bit confused when reading about all the various intervals and their variants. For instance, (page 2) Qin and Hotilovac recommended 9 intervals, one (or two?) of them called the logit-transformation-based Mann Whitney interval(s) with (or without?) transformation. On page 3, we read about Qin and Hotilovac's nonparametric logit-transformation-based interval (LT), but this seems to be another one, as the Mann Whitney intervals (M-W) are mentioned some lines below, likewise many others. These parts of the paper (Confidence intervals for the AUC, Confidence intervals for a single proportion, Modified Wald intervals) are boring to read, confusing and not very informative for readers not familiar with the literature and notation. I propose (1) to use exactly one wording for each interval, (2) to associate acronyms (such as LT or M-W) for *all* intervals, and (3) to repeat these acronyms in table 1.

page 6, "Statistical power": "The Wald-cc interval has slightly higher power as a result of its liberal behaviour". You might consider to account for this by using adjusted power such as explained here. Let alpha be the significance level, F the observed cdf of the p-value under the null hypothesis (that is, F is not necessarily
uniform), and \( G \) the cdf of the p-value under some alternative hypothesis. The adjusted power is defined as the probability that a p-value under the alternative hypothesis passes the *actual* alpha-quantile of the observed p-value distribution under the null hypothesis:

\[
G(F^{-1}(alpha))
\]

If the test is liberal, then \( F^{-1}(alpha) < alpha \) and thus \( G(F^{-1}(alpha)) < G(alpha) \) which means a decrease of power. (See, e.g., Rücker G, Schwarzer G, Carpenter J. Arcsine test for publication bias in meta-analyses with binary outcomes. Stat. Med. 27(5):746-763, 2008.)

**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.