Reviewer's report

Title: Extreme Verification Bias in Paired Continuous Tests Can Cause Researchers to Choose the Wrong Screening Modality

Version: 1 Date: 30 July 2008

Reviewer: Mariska M. G. Leeflang

Reviewer's report:

Major Compulsory Revisions

1 (general)
The paper addresses a very relevant point, although the authors only investigate a small part of the problem and they do it in a very simplistic way (which is not necessarily bad, but addresses in this case the problem not to its full extent). They do for example not show how the situation is when clinicians would base their decisions on the optimal sensitivity and specificity for each test, in stead of the complete ROC curve. Furthermore, I would like to know what happens when the (true) ROC curves of both tests cross each other in stead of running parallel.

2
In the introduction, the authors do not explain what paired extreme verification bias is. They should do so in either the first or the second paragraph of the introduction. They expect the readers to know or read the Begg-1987-paper, but the paper should be generally understandable without having to check the cited papers. Not all readers will immediately know what is meant with this bias, so a description of the corresponding (biased) study design may help. The authors should at least explain that in the case of verification bias, only a (selective) part of the tested individuals will be verified.

3
In the Study design section, the authors state that the true disease status is determined by a confirmatory test or follow up. This is what I would call differential verification bias (Whiting, 2004), while the previous text implicates that the paper was about partial verification bias (see also Whiting, 2004). These two biases may have different effects and that is not addressed in this paper. Can the authors tell us something about the results when follow-up was not included in the analyses at all?

4
A big factor in these analyses is the correlation between the two index tests. Could the authors please explain how the correlation between these two tests influences their results? What happens when there is much stronger correlation (which may often be the reality)? Or when there is no correlation at all?
The “Simpson’s rule trapezoidal numerical integration method” may not be the appropriate method for calculating AUCs, because it does not give confidence intervals around the AUCs. There are better, more formal ways of calculating AUCs in for example SAS that do give confidence intervals.

These missing confidence intervals are another crucial limitation in the paper. When researchers formally want to test or investigate whether one AUC is bigger than the other one, they will need confidence intervals to say something about the precision of their decision. In the example in this paper, the ROC curves of the two tests swap. And the AUC that is in the observed situation the biggest, is in reality the smallest. But confidence intervals may have shown that there was no significant difference between the two AUC at all. However, that is a question the authors do not address.

The observed curve for test for test 1 shows a “point of infliction”. Do the authors have an explanation for this effect? And why do we see it in the curve for test 1 and not for test 2? And why not in the true curves?

Minor Essential Revisions

1. The abstract is a bit obscure for readers who are not familiar with the phenomenon of paired extreme verification bias. So please rewrite it. It may already help if the second sentence is put up front.

2. I do not agree that in the case of verification bias the sensitivity will always be inflated and the specificity will always be deflated (as is stated in the introduction). Could that perhaps be illustrated with an example? Or do the authors have a reference to this statement?

3. The authors state that “a paired comparison of ROC curves is the most common trial design used to compare screening modalities”. Do they have a citation for this statement?

4. Furthermore, do they have a citation for the “Simpson’s rule trapezoidal numerical integration method”?

5. The first sentence of the Results section is a conclusion, not a result.

6. On page 8, on top: “The slope of the ROC curve increases strongly at the cutoff point above which all test results are confirmed”. If you read the figure from left to right, the slope decreases in stead of increases. So please reword or explain.

7. If all patients undergo all tests, how is it then possible that “one test was verified more often than the other test” (page 8, bottom)?
8. Can the author explain what “differential inflation in sensitivity” is and why it is unique to paired extreme verification bias? Please add some citations to this statement.

9. Sometimes the term paired extreme verification bias is written in Italic and sometimes it is not. Please be consistent in this.

10. The authors state that in “many situations, complete verification … is impossible, as there is no accurate, [etc] test”. I think the problem arises only when the reference standard (confirmatory test) is too invasive, too costly or otherwise too big a burden. It has nothing to do with a reference test not being accurate. That gives other problems (imperfect reference standard bias, verification bias in general).

11. Could the authors elaborate a bit more on the problem of follow-up as second confirmatory ‘test’? Not only may new cases occur in time, but (especially in cancer), cases may also resolve or never develop to disease (depending on how well the tests measure the disease).

12. What is an ROC score? How is it calculated?

13. In the legend with figure 1: “The means of the ROC distributions for cases … 1.6 and 0.35.”. How have the authors calculated the means of the ROC distributions and what does it mean?

Discretionary Revisions

1. The authors use the term “observational trial”. This may be confusing, because the term trial often refers to an experimental (and thus not observational) design. Furthermore, the design that the authors describe (study participants undergo tests), may not be really observational. Alternatives may be diagnostic study, diagnostic accuracy study.

2. The same may be true for the term “screening trial”, which may implicate a trial to study the effect of screening programs on patient outcomes.

3. The authors may want to consider some more recent publications on diagnostic test accuracy studies:


Bossuyt et al, BMJ, 2006 (about test comparisons)

Book: Knottnerus (ed.) The evidence Base of Diagnosis.

**Level of interest:** An article of importance in its field
Quality of written English: Acceptable

Statistical review: Yes, but I do not feel adequately qualified to assess the statistics.

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