Reviewers report

Title: The Brier score does not evaluate the clinical utility of diagnostic tests or prediction models

Version: 0 Date: 14 Mar 2017

Reviewer: Werner Vach

Reviewers report:

The authors address the relevant question about the best way to describe the quality of a diagnostic test. However, the paper suffers from the basic drawback that they focusing on comparing two methods which are hard to compare, as they address very different questions: The Brier score is a measure to evaluate a prediction model generating probabilities for each patient, whereas net benefit is a measure for a binary decision rule. The authors try to give the impression that the Brier score has been used directly to measure the quality of a diagnostic test, but I do not think this is true. They give a reference for what they feel is an example for such a misuse, but this example is not convincing. In the example the authors refer to a prediction model as a useful tool to discuss individual dosage, but individual dosage is a type of a clinical decision making which is beyond the simple application of a binary decision rule.

I think the point the authors would like to make is the following one: If a diagnostic test is based on dichotomizing a prediction model, then it can be misleading to interpret measures of the prognostic ability of the model (like the Brier score) as measures of the quality of the test. This is the point the authors try to make in the lower part of table 1 with some success: The quality of the prediction model measured by the Brier score does not necessarily translate into the quality of the diagnostic test.

So my first suggestion would be to focus on this aspect. And my second suggestion would be to consider not only two measures, but several ones to make the advantage of the net benefit approach more clear. My suggestion would be to include also the AUC (as a more common measure for diagnosis related use of a continuous predictor), and the Youden-index, overall missclassification or the diagnostic OR as simple overall measures for diagnostic tests. The latter would allow to demonstrate that net benefit is not only useful because it is a measure for a diagnostic test, but that it is useful because it allows to take into account the clinical knowledge about the importance of FP vs TP decision.

Together with an extension of the models considered (allowing also variation in AUC and prevalence), the paper may make a useful pedagogical contribution.

I do not feel that the idea of the authors to introduce versions of the Brier score directly applicable to a diagnostic test (as done for the upper part of Table 1) is very convincing, as long as this is an invention by the authors. If none has done it this way, there is little need to do this. In the first approach presented by the authors the Brier score reduces (in my opinion) just to the
missclassification rate. Probably, also in the second approach the resulting statistics is related to existing measures.

Specific comments:

1) The authors introduce in lines 109 and 110 expressions for E[BS] involving the symbol \( f \), without introducing the symbol \( f \). The formulae do not seem to hold in general, but only in specific situations.

2) In line 115 the authors claim that E[BS] is equal to \((1\text{-specificity}) + (1\text{-sensitivity})\). However, to my understanding the Brier score reduces to the missclassification rate here. This corresponds also to the results later shown in Table 1.

3) I am also in doubt about whether the formula given in line 120 is correct. Also a derivation of the formula given in lines 124/125 may be useful.

4) In their computations, the authors seem to assume that the probabilities they generate are both the true individual probabilities as well as those generated from some model. It is fair to make such an assumption, as the stochastic imprecision may be negligible here. But the authors should be more clear about, that they make such an assumption. Later, it becomes even more confusing, when they consider also misspecified models. It remains a little bit unclear, what the authors then actually compute in Table 1.

5) Even if neglecting stochastic uncertainty may be justifiable here, I think that many readers will be more convinced when presenting results based on simulation studies allowing also stochastic imprecision. In the moment, the reader may have the impression that the results only hold for the case of no stochastic imprecision.

In summary, I think the paper missed in the present form the point. However, with a more precise notation and the extensions mentioned above the paper has the potential to make a useful pedagogical contribution.

Level of interest
Please indicate how interesting you found the manuscript:

An article whose findings are important to those with closely related research interests

Quality of written English
Please indicate the quality of language in the manuscript:

Acceptable

Declaration of competing interests
Please complete a declaration of competing interests, considering the following questions:
1. Have you in the past five years received reimbursements, fees, funding, or salary from an organisation that may in any way gain or lose financially from the publication of this manuscript, either now or in the future?

2. Do you hold any stocks or shares in an organisation that may in any way gain or lose financially from the publication of this manuscript, either now or in the future?

3. Do you hold or are you currently applying for any patents relating to the content of the manuscript?

4. Have you received reimbursements, fees, funding, or salary from an organization that holds or has applied for patents relating to the content of the manuscript?

5. Do you have any other financial competing interests?

6. Do you have any non-financial competing interests in relation to this paper?

If you can answer no to all of the above, write 'I declare that I have no competing interests' below. If your reply is yes to any, please give details below.

I declare I have no competing interests

I agree to the open peer review policy of the journal. I understand that my name will be included on my report to the authors and, if the manuscript is accepted for publication, my named report including any attachments I upload will be posted on the website along with the authors' responses. I agree for my report to be made available under an Open Access Creative Commons CC-BY license (http://creativecommons.org/licenses/by/4.0/). I understand that any comments which I do not wish to be included in my named report can be included as confidential comments to the editors, which will not be published.

I agree to the open peer review policy of the journal