Author’s response to reviews

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

Authors:

Melissa Assel (asselm@mskcc.org)

Daniel Sjoberg (Sjobergd@mskcc.org)

Andrew Vickers (vickersa@MSKCC.Org)

Version: 1 Date: 07 Jul 2017

Author’s response to reviews:

Reviewer #1:

Comment: The authors … [are] 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 [claim] Brier score has been used directly to measure the quality of a diagnostic test … They give a reference … but this example is not convincing.

Response: We agree with the reviewer that the motivating example in the introduction was unclear and we have clarified the language in the introduction. We now say: “As a practical example, a study by la Cour Freiesleben et al. aimed to develop prognostic models for identification of patients’ risks of low and excessive response to conventional stimulation for in vitro fertilization/intracytoplasmic sperm injection in order to ascertain if a low or a high dosage level should be used….and and may assist clinicians in individual dosage [between two alternatives] of their patients”. In the paper cited, the authors are not predicting a continuous dose but investigate two dose levels and make recommendations as to which level high or low should be used. We would like to assure the reviewer that this scenario involves a model predicting a binary endpoint. To clarify, net benefit can be used for continuous predicted probabilities in addition to binary decision rules.

Comment: I think the point the authors would like to make is the following one: … The quality of the prediction model measured by the Brier score does not necessarily translate into the quality of the diagnostic test.

Response: The reviewer is correct in understanding our goal was to demonstrate that the quality of a prediction model or binary test as measured by the Brier score does not necessarily correspond to the clinical utility.
Comment: …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. …

Response: We carefully considered this recommendation however given many of our examples were related to miscalibration and the AUC is unaffected by calibration, we decided against focusing on AUC.

Comment: I do not feel that the idea of the authors to introduce versions of the Brier score directly applicable to a diagnostic test ...

Response: We have found several examples where this was used in the literature and have seen the recommendation in statistical textbooks and wish to caution researchers against the use of the Brier score in lieu of alternative assessments of clinical utility. For instance, we conclude “We conclude that Brier score is an unsound metric of the clinical value of diagnostic tests and prediction models and advocate, as an alternative, the use of decision analytic measures such as net benefit.”.

Specific comments:

Comment: 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.

Response: The notation has been updated throughout the manuscript.

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

Response: We have updated the manuscript per the reviewers comment to indicate where the Brier score simplification is equivalent to the misclassification rate.

Comment: 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.

Response: Thank you for the careful check, these formulas have been corrected in the manuscript. We now represent the Brier score as the expected value of the squared difference of the outcome and prediction using the joint distribution of the outcome and prediction. We have removed all references to a Beta distribution approximation.

Comment: 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…. 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.
Response: We have updated the methods used. Rather than a simulation approach, we have updated our methods extensively and directly solve for the joint distribution of the outcome and the prediction. After the derivation of the joint distribution, the exact Brier score was calculated by evaluating \( E(Y-X)^2 \).

Comment: 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.

Response: We accept the reviewers point and did indeed confirm these results using simulation. We have added the following to the Table 1 title: “All values given were calculated directly form the formulae in the text and independently verified using a simulation approach.”.

Reviewer #2:

Comment: This research makes some assumptions that are not correct. First, it assumes that researchers do not want measures to be incidence-dependent. Second, it assumes that researchers routinely compare prediction quality across unequal incidence datasets. Third, it ignores the decompositions of the Brier score that factor out overall prevalence. While researchers sometimes need to compare predictive performance across different incidence situations, the paper is written as if this is always the case. A good deal of work could make this paper useful to readers.

Response: We thank the reviewer for the useful points regarding how prevalence factors into our discussion. We agree that there is a debate on this point. However, it turns out that is tangential to our main conclusions as net benefit can be divided by prevalence to give what is termed “relative utility”. Relative utility is a monotonic transformation of net benefit and so rank ordering of tests and models is this same. We have added the following to the manuscript with the corresponding citation: “Note that our choice of a decision analytic comparator, net benefit, depends on prevalence. However, net benefit can be divided by prevalence to give relative utility, a prevalence independent measure. Because relative utility is a monotonic transformation of net benefit any conclusions about the value of decision-analytic versus Brier scores are unaffected as they result in a consistent rank ordering to that as relative utility.[1] Hence our conclusions are unaffected by our choice of net benefit as the decision analytic statistic.”.

With respect to the third point, we are focusing on forms of the Brier score that have been recommended in the literature for evaluation of tests.

Reviewer #3:

Major comment
Comment: Although I agree with the message of this paper, I believe revisions are need to the methods.

Response: We thank the reviewer for their agreement of the message of the paper and value their thoughtful and thorough review of the manuscript regarding how we can revise the methods section and have made every attempt to fully adhere to their comments.

Comment: The expression for $E[BS]$ at line 110 is, I believe, incorrect. I can derive this expression as $E[BS] = E[(D-f)^2] = E[D^2] - 2E[D \cdot f] + E[f^2]$. Then using $p = E[D]$ (see comment 2b below) this becomes $p - 2pE[f] + E[f^2]$. However, this last step only works by using $E[D \cdot f] = E[D] \cdot E[f] = pE[f]$, which only holds if $D$ and $f$ are independent. Presumably this independence does not hold, since we hope that $f$ predicts $D$. Therefore, I believe the line 110 is incorrect. (I also get different expressions compared to what appears at lines 115 and 120.

Response: We thank the reviewer for their careful review of our derivations. We are in agreement that we made an error in the assumption that $D$ and $f$ are independent. We now represent the Brier score as the expected value of the squared difference of the outcome and prediction using the joint distribution of the outcome and prediction. We have removed all references to a Beta distribution approximation. Despite this error our conclusions and the message of the paper were not altered as the rank-orderings of the Brier score were not changed despite the values of the Brier scores in the continuous setting being updated.

Comment: Note also that the paper says that for a binary test the Brier score can be defined in two ways but then gives neither definition. I think what they mean is there are two choices for $f$ - either the test result 0/1 or the PPV from the test.

Response: The reviewer was correct in assuming that we were intended to refer to the two methods outlined on line 109-117 for ascertaining the Brier score for a binary test using either the test result or the PPV/NPV of the test. We have removed the statement on line 107 which stated that the Brier score can be defined in two ways for clarity as it was more fully developed in subsequent lines.

Comment: The paper's methods use beta distributions for predicted risks and use the (incorrect, I believe) formula for $E[BS]$ on line 110 to say what the Brier score is for different distributions of predicted risks. However, I don't see anything in the methods that relates the beta distribution of $f$ to $D$. It appears again that the methods assume $D$ is independent of predicted risk $f$. If my interpretations here are correct, then much of the specific results (e.g. Table 1) need to be fixed.

Response: We have updated the methods used. Rather than a simulation approach, we have updated our methods and directly solve for the joint distribution of the outcome and the prediction. After the derivation of the joint distribution, the exact Brier score was calculated by evaluating $E(Y-X)^2$. 
Comment: I also don't understand how the authors produce Figure 3, since I cannot find anything in their methods that explains how they relate f to D. (This makes me wonder whether there is some basic aspect to their approach I failed to appreciate.)

Response: We thank the reviewer for pointing out areas where we could improve the clarity of our manuscript. We added the formula based on the logistic regression to show how the specified miscalibration relates to the predicted probabilities.

Other Major Comments

Comment: For a binary test, the authors' "Method 1" merits a note that the Brier score is equivalent to the misclassification rate (MCR) ….

Response: We thank the reviewer for this suggestion that will help to provide some meaningful context to the reader. We have added a sentence indicating where the Brier score is equivalent to the misclassification rate in the binary test setting: “This is equivalent to the misclassification rate in this binary test setting”.

Comment: The section starting on line 87 giving background on the Brier score should be revised.

Response: We hope that our responses to points 2.a. through 2.e. have adequately addressed this series of comments.

Comment: The paragraph that starts at line 94 belongs later in the paper.

Response: We chose to leave the paragraph starting at line 94 as is because we felt that it was useful to introduce the formula for the brier score early on but in response to another comment we have reordered other pieces of the methods section.

Comment: b. line 107: "true probability of disease" is p. I think the authors really mean something like the prevalence or population rate of disease, i.e. E[D]=p is what I assume is meant. The notion of "true probability" is fraught and should be avoided.

Response: We thank the reviewer for the comment. Reference to “true probability” and p have been removed.

Comment: The authors never actually write the definition of BS for binary tests. They only write E[BS].

Response: We are unsure of the reviewers’ point here. The formula given is applicable to both binary and continuous tests.

Comment: ….the authors are merely using beta distributions as a convenient way of describing distributions of predicted risks to make their points. First, make this clear. Second, move this out of the section giving background on the Brier Score. It is really about the paper's Methods.
Third, the mention of logistic regression here is a distraction; lots of other methods of constructing risk prediction models could provide "a range of predicted probabilities."

Response: We have restructured the paper in hopes to clarify and address this comment.

Comment: I suggest a more informative title for the paper. It is not be clear what it means for a metric to be "sound" or "unsound." Something like: "The Brier score does not evaluate the clinical utility of diagnostic tests or prediction models."

Response: We have updated the title accordingly. It is now: “The Brier score does not evaluate clinical utility of diagnostic tests or prediction models”.

Comment:

Minor Comments

4. Abstract. "determine" is a very, well, deterministic word. I suggest replacing with "assess" or "evaluate."

5. "favored the correct test or model". There is no "correct" model here. Consider revising to "favored the better test or model."

6. line 53. Consider citing the paper by Kerr et al (2014). While the request for this citation might seem self-serving, I think it is fair to say this is the most comprehensive evaluation of issues with NRI.

7. line 69. patients' risk → patients' risks

8. line 71. revise "with the models were selected"

9. lines 76-80. It would be helpful to give more details of the example. Also the phrase "prevalence and clinical consequences are discordant" does not make sense.

10. line 83 "would refute" → "refutes"

11. "thresholds probabilities" → "threshold probabilities"

12. line 156. "would be" → "is"

13. line 201. revise (a word is missing)

Response: We thank the reviewer for providing feedback on subtle wording changes that could provide clarity and have updated the manuscript to reflect all minor comments.

References