Reviewer’s report

Title: The effects of misclassification in routine health care databases on the accuracy of prognostic prediction models: A case study of the CHA2DS2-VASc score in atrial fibrillation

Version: 0 Date: 04 Oct 2017

Reviewer: Gareth Ambler

Reviewer's report:

This paper considers the sensitivity of the CHA2DS2-VASc score to errors in the predictor variables. Specifically, the paper performs a validation of this score using both gold standard (reference) and potentially misclassified (index) predictors.

I am not convinced that the CHA2DS2-VASc score is the best example to use for this case study. For one, the score does not seem to actually make predictions (Authors: if it does, please add this to the Methods section and Table 1). Instead, it assigns patients to an integer score between 0 and 9 which, presumably, can be used to rank patients in terms of severity / expected mortality. Second, it is a relatively simple score with only 7 (binary/ordinal) predictors with simple integer weightings. Of these, misclassification is only considered in 5 of the predictors. Since the weightings are so simple (either 1 or 2), misclassification in a couple of the predictors may actually cancel each other out.

I have a number of concerns regarding the analysis section and the motivation behind each analysis. I will discuss these by table and figure.

Table 3: I am not convinced that the various measures presented in this table (i.e. kappa to NPV) are actually that helpful for summarizing the misclassification. I think it would be far clearer if the actual data (e.g. the information from the 2x2 tables) were presented in this table. This should be straightforward as there are only 5 predictors. If the diagnostic measures (sensitivity to NPV) are to be retained, it would be helpful if more interpretation was provided.

Table 4: It is not clear whether this analysis and table are necessary. The results are not discussed in the text.

Table 5: I do not understand why new Cox models are fitted to the reference and index predictors since this is (I presume) a validation exercise rather than an exercise in model development / updating. The analyses reported in this Table certainly need to be motivated better. Additionally, I also don't really understand why age and sex are intentionally left out of the re-fitted models.
Figure 1: It might be better to provide both numbers and row percentages (please state that these are row percentages in the caption). Also, more insight could be provided in the text, e.g. how many misclassifications cancel out?

Figure 2: Please see my comments for Table 5. This perhaps shows that the CHA2DS2-VASc score is not suitable for this case study as it does not produce mortality predictions and hence cannot be assessed with respect to calibration.

There is also, I think, an omission in the Results section. I couldn't find estimates of the c-index for the CHA2DS2-VASc score based on the index and reference scores respectively. To be clear, since this is a validation exercise, models do not need to be fitted to produce these values.

Some other more minor points:

Abstract | Aim: As mentioned earlier, the CHA2DS2-VASc score does not seem to predict (absolute) mortality.

Abstract | Conclusion: This conclusion is far too strong and almost certainly does not generalize.

Background (l7): Please add references.

Background (l44): Please add a reference for the model.

**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.

No.

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