Reviewer’s report

Title: Informative missingness in electronic health record systems: the curse of knowing.

Version: 0 Date: 19 Feb 2020

Reviewer: Matthew Sperrin

Reviewer's report:

This paper highlights the 'opportunity' of informative missingness in prediction - in that it can be highly predictive of outcome - but also the associated challenge that the missing mechanism is not likely to be stable across setting. I think it's important to highlight this underappreciated issue - with much of the prediction literature still seeing missingness as a 'nuisance' that needs to be dealt with. Some specific comments follow.

1. The notion of a missing mechanism that changes once a model is implemented fits in the more general problem of the data generating mechanism changing between model development and deployment. In particular, feedback loops, in which the presence of the model influences the setting it's deployed in. Relevant references are the paper of Lenert et al https://academic.oup.com/jamia/article/26/12/1645/5559573, and our letter concerning it https://academic.oup.com/jamia/article/26/12/1675/5625126. Your work on measurement error (e.g. https://www.ncbi.nlm.nih.gov/pubmed/31706963) is another example of the same phenomenon, as is calibration drift. Bringing this into the discussion would place this work into the more general context (Obviously, author's discretion concerning the reviewers' self-cites.)

2. Throughout the paper you give the example of a categorical variable with a "separate category" for the missing indicator approach. However, the simulation study considers a missing indicator in a continuous variable, where a value needs to be imputed for the continuous variable as well, so is somewhat different. Suggest consistency.

3. 'classification and regression trees can easily accommodate such features' - as I understand what these models actually do is equivalent to a pattern missing approach. Would be good to be explicit to avoid the impression that missing data is simply no problem (or magically solved) for the machine learning approaches.

4. Section 3: I question the value of this section. Clearly, no one will ever believe that there is a causal effect between a missing indicator and outcome. Also, this is much stronger than simply noting that missingness patterns may change in response to implementation of the model. Indeed, no predictor in a prediction model should ever be interpreted in a causal way such as this, as prediction models are not designed to allow such causal interpretation (even though this is of course incorrectly done in practice).

5. section 5.1: it should be made explicit at the outset that the first dataset is for development of the model and the second represents a dataset as used in deployment of the model.
6. Section 5.1: I assume that the missing mechanism’s lack of transportability will cause issues regardless of how missing data are handled. It would be helpful therefore to also show what would happen under complete case analysis and multiple imputation approaches, for example.

7. It was suggested in a recent paper by Fletcher-Mercaldo, section 1.4 https://doi.org/10.1093/biostatistics/kxy040 that a lot of the problems of missing indicator are not in introducing the indicator per se, but rather that this is usually accompanied by a bad way of imputing the corresponding continuous variable. This seems relevant here given that a default value of 0 is used in the simulation study, and indeed I'd suggest using at least mean imputation.

8. I'm uncomfortable with the notion of 'actual risk' used to report the simulation results. While I accept this can be defined in a simulation where the data generating mechanism is known, it does not have a real-world analogue. I'd prefer, therefore, that measures of RMSE are with respect to outcome rather than the 'actual risk' (i.e. a Brier score), and consider also reporting calibration.

9. In the conclusion, the question is what we do about this changing missingness mechanism. Can any recommendation be offered? One way forward would be to propose that such models are dynamically updated to reflect the changing missingness mechanism. Another might be to suggest that missing indicators/patterns should not be used at all in such scenarios - or more research needed to understand better which approaches to handling missing data might be most robust against the changing mechanism.

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 that 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