Reviewer’s report

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

Version: 0 Date: 01 Mar 2020

Reviewer: David Muller

Reviewer's report:

This is a well written commentary that clearly describes an important issue in the development and application of prognostic models in the presence of missing data. I have a few major comments and some minor suggestions.

Major comments:

1. The motivating example of cholesterol measurement is well described and clearly illustrates the concept of informative missingness. It is not clear how this motivating example is linked to the simulated numerical example. I think the paper would benefit from a tighter integration between the theoretical discussion and the numerical example. This would be a useful aid for non-expert readers. I would like to see the parameters of the data generating models justified and tied to a (hypothetical) practical example.

2. I am concerned about the presentation of the "missing indicator" method. It is well known that this is an inappropriate method for handling missing data. In practice, and especially in the context of models being built using data from EHR, there is likely to be information available that can be used to better predict/model the missingness mechanism. Whilst the paper as it stands does a good job of describing the problems that can arise in the context of informative missingness, I am a little worried that, as it currently stands, it might be taken as an endorsement of using missing value indicators (with appropriate caution regarding transportability of the missingness mechanism). The author touches on this in the conclusion, but also suggests that other methods of dealing with missing data may not be useful... a claim that I consider to be stepping well beyond the scope of the manuscript, especially since no other methods are investigated or even discussed. There is a substantial amount of prior and ongoing work in this area that should not be dismissed out of hand.

3. In a few places the wording would benefit from more precise language around "risk". For example, on page 4 line 14: "After all, the model gives the lowest risk if the cholesterol is not measured. Once a measurement is being made, the risk increases already, regardless of what the result of that measurement is." This implies that the act of taking the measurement itself increases the risk (the author notes in the next paragraph that the measurement cannot affect the risk per-se). I suggest being careful to use "estimated risk" or "predicted risk" in these situations so as to clearly differentiate it from any true underlying risk.

4. I am not sure that a single numerical example (especially divorced of context, see comment 1) is of great use here. What are we meant to take away from the results? If this is included I suggest at least running a full simulation and reporting simulation-based expected values.
5. It is stated that the data generating model for the outcome was set up so as ~50% of observations had the outcome, but when X1 and X2 are distributed U(0,1) the expected value of the linear predictor is -1, not 0, so ~1/3 of observations will have the outcome.

Minor suggestions:

6. In the R code I suggest a couple of minor edits to aid the reader/user. Firstly to spell out "missing indicator" in the comments, as MI is commonly used to denote multiple imputation. Secondly, to rename the function "logit" to "invlogit" or "expit", as the logit function is typically defined as \( \log(x/(1-x)) \).

**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