Reviewer’s report

Title: Chronic Disease Population Risk Tool (CDPoRT): a study protocol for a prediction model that assesses population-based chronic disease incidence

Version: 0 Date: 22 May 2018

Reviewer: Kym Snell

Reviewer's report:

The authors present a protocol for the development and validation of a prediction model for chronic disease incidence in a Canadian population. In general, the protocol is fairly well structured and written, however I have the following comments/questions for the authors:

1) If the intended use of the model is to make predictions for a population, it is not immediately clear to me how the model will be used as it will provide predictions for individuals based on their individual predictor values. Will the model be applied to existing cohorts to estimate the expected incidence or is interest really in the associations between predictors and outcome? Please clarify.

2) Pg 5, lines 14-22: Can the authors further justify the need for a model with a composite outcome rather than predicting the individual conditions? I'm just thinking that the association between a predictor (e.g. smoking) and each of the outcomes may differ, and therefore becomes some average when modelling the composite outcome. For that reason, I would think that the predictive performance of the model may suffer due to the heterogeneity in predictor effects.

3) Pg 7: Is there any clustering in the data, e.g. different centres? If so, how is this being accounted for?

4) Pg 7, line 11: A random split-sample will be used for the internal validation. Would it not be better to use a cross-validation approach, especially if the combined data will be used for the final model (pg 17, line 9)?
5) Pg 9, line 22: Do the authors mean that the cohorts have not yet been used for prediction modelling of this chronic disease outcome? Other prediction models have been published by the researchers using the Ontario and Manitoba cohorts.

6) Pg 11-12: Many continuous predictors will be categorised. Would it not be better to model these predictors continuously and perhaps look at non-linear functions to avoid this loss of information and potentially also predictive performance?

7) Pg 13, line 4: Please say how income will be imputed. Will this be done using multiple imputation? Also, why impute income but not the other missing predictor values?

8) Pg 14, line 31: Shrinkage will only be applied if less than 0.9. In my experience, even when there is considerable overfitting in small datasets, the shrinkage estimate is greater than 0.9. Considering the size of the data (no. of events), I would expect this estimate to be very close to 1, but would suggest shrinking the coefficients even if above 0.9. If the expected optimism is 10%, this can make quite a difference to predicted probabilities if high.

9) Pg 15, line 24: I understand wanting to fit the initial model first and then assessing the other sets of predictors. It says forward and backward stepwise selection will be used. However it is not quite clear how this will be applied with the sets of predictors. Based on lines 39-44, it sounds like the whole set will be included and then backward elimination applied. Please just clarify in the text whether stepwise selection or backward elimination is being used (backward is usually preferred). Also, what criteria will be used to select or remove variables e.g. p-value, AIC?

10) Pg 16, line 26: For the calibration plots, will this be assessed at certain time points?

11) Pg 16, lines 36-41: "...where the clinically relevant standard of calibration - 20% difference... will be used'. I don't understand what is meant by this. Please provide more information here and references if possible. 20% seems rather large if this is an 'acceptable' amount of mis-calibration.

12) Pg 17, line 14-19: Can you please clarify which cohorts are being referred to for 'cohort-specific intercepts'. Both Ontario, or Ontario and Manitoba? Also, is it likely that the model would be applied elsewhere? If so, how would the model be used?

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

An article of importance in its field

**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