Author’s response to reviews

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

Authors:
Ryan Ng (ry.ng@mail.utoronto.ca)
Rinku Sutradhar (rinku.sutradhar@ices.on.ca)
Walter Wodchis (walter.wodchis@utoronto.ca)
Laura Rosella (laura.rosella@utoronto.ca)

Version: 1 Date: 21 Jun 2018

Author’s response to reviews:

June 21, 2018

Dear Dr. Snell,

Re: DAPR-D-18-00010

We would like to thank the reviewers for their thoughtful review of our manuscript, “Chronic Disease Population Risk Tool (CDPoRT): a study protocol for a prediction model that assesses population-based chronic disease incidence”.

We have provided a response to each of the items identified by the reviewers, and have indicated corresponding changes in the manuscript in the highlighted text. There were some additional changes made to the manuscript regarding the reported sample size. Some of the numbers were misreported in the abstract and text, which we have revised appropriately and highlighted.
On behalf of my co-authors, I would like to thank you for considering our manuscript.

Sincerely,

Ryan Ng

Reviewer reports:

Reviewer #1: This paper documents a study protocol for a chronic disease prediction model to assess chronic disease incidence. The approaches that will be taken are fairly clearly described and in the large part well justified.

I have a few comments detailed below:

1. I agree that death free of a chronic disease is a competing risk for the first occurrence of any; but is it not also interesting to understand the breakdown of chronic diseases and also if multiple diseases are developed in certain individuals? If so, death from a lethal chronic disease will also be a competing risk and a more advanced modelling approach may be needed.

-You make a great point about understanding the breakdown of chronic diseases. We were considering examining chronic diseases separately, but decided not to add it initially because of the considerable amount of extra work required to model each disease separately (12 models total based on 6 six chronic diseases for each sex). However, after further consideration, we have decided to explore it as a sensitivity analysis as examining chronic diseases separately may improve model predictive performance (pg. 18, lines 8 to 21). In regards to examining multiple diseases, we agree that is an important area of research, and we will be exploring that in a separate study.
2. Following on from that, some are much worse outcomes than others and a lot have interlinked risk factors. I'm not 100% convinced that this composite outcome is the most informative outcome of choice? You do refer to multimorbidity as a potential sensitivity analysis - I cannot imagine that the same model will work well for diseases with highly correlated risk factors. Can you further justify the outcome choice and how you choose which are included?

-We agree that we were not explicit in our choice of chronic diseases for the outcome. We have added a sentence describing our reasons for choosing the six chronic diseases of choice based on prevalence, associations with behavioral risk factors and effects on morbidity and mortality. (pg. 8, lines 13 to 17). The chronic diseases were also chosen because they have validated algorithms in Ontario. We agree that the same model will likely not work well with multimorbidity.

3. Choice of validation approach - is it not better to use all of the data to build the model, and then use a repeated cross-validation technique?

-Thanks for the suggestion. We felt that split-sample validation was appropriate given the sample size of the data and that we are performing external validation in Manitoba. However, we agree that a cross-validation technique is better. We have added in bootstrap validation of the Ontario derivation cohort as another way to validate the model in the Ontario data (pg. 16, lines 19 to 23; pg. 17, line 1) in addition to the split-sample validation.

4. For people with multiple records in the survey - will you update their covariate values if they differ from survey to survey?
No, we will not update their covariate values from survey to survey. The CCHS was initially designed as a cross-sectional study and it was never intended to be used in a longitudinal manner. There are relatively few individuals with multiple records (1.2%).

5. Do you need to categorise BMI and other continuous covariates? Can you not fit continuously and non-linearly, and then report in groups later if needed for reporting purposes?

-Thanks for the suggestion. Based on our previous work with the knowledge users, they prefer the categories of BMI and other continuous covariates. Also, based on our previous work with other population risk tools, we have found the categories of BMI and other continuous covariates tend to perform well and did not result in a significant decline in model discrimination and calibration. We do not expect categorization of continuous covariates to affect the model predictive performance greatly, but we acknowledge that it may change the model more, which is why we are modelling these covariates continuously and nonlinear as well.

6. PH checking - you say you need a more interpretable model, and so will sometimes use PH - why? Some of the interpretability will be lost in other ways such as using splines for continuous covariates - transformations, graphical presentations and predictions of the model parameters will be needed then anyway. Best to fit the better model, no?

-Thanks for the comment. You are right. It is best to fit the better model. We have modified the protocol to reflect that any violations will be incorporated into the model (pg. 13, lines 7 to 9).

7. Could a full imputation approach not be incorporated to deal with the missing data?

-We agree that an imputation technique would be ideal for handling missing data, but imputation is not feasible with how the prediction tool will be used. Non-response to a question is an option
available to the CCHS interviewee, and so the knowledge user will have missing data in their population. Unfortunately, the knowledge users are not all trained in statistics, and will probably not have the tools or expertise to impute the data.

Reviewer #2: 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.

-Yes, the model will be applied to existing cohorts, in this case to Canadian Community Health Survey (CCHS) respondents, to estimate the expected incidence of a population. The CCHS is representative of 98% of the Canadian population, and each respondent is accompanied by a sampling weight. Using respondent information in conjunction with the survey weights produces the expected incidence in the population. The sampling weights are incorporated while developing and validating the model.

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.

-Thanks for the insightful question. This is something we grappled with during the design of the protocol. The composite measure is what the knowledge user is interested in, and there are tools that exist that look at some of the chronic diseases (e.g. diabetes) separately. Modeling each disease
separately would also mean the creation of 12 models total (6 models for each chronic disease for each sex). We do realize that the covariate coefficients will be an averaged estimator based on the composition of the outcome, which might impact predictive performance. We have modified the protocol to include a sensitivity analyses in which we examine the predictive performance by creating separate chronic disease outcomes (pg. 18, lines 8 to 21).

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

- There is no clustering of the data. However, the CCHS is a survey with a complex design, and to accommodate the survey effects we incorporate survey weights and bootstrap weights into model building (pg. 14, lines 13 to 16)

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)?

- Thanks for your suggestion. We have decided to include bootstrap validation of the Ontario derivation cohort as another validation approach in addition to the split-sample approach for robustness (pg. 16, lines 19 to 23; pg. 17, line 1). We don’t believe there will be much difference in the validation approaches as our sample size is large for split-sample. We expect to spend the majority of time with the external validation of the Manitoba cohort.

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.
We apologize for the confusion. In Ontario, the CCHS health survey data has been linked to health administrative for this study. However, in Manitoba, the survey data has not been linked to health administrative data for this study. The sentence has been updated to make this clear (pg. 10, lines 1 to 2). The other prediction models were created using different linked datasets that do not include newer cycles of the CCHS and/or up-to-date health administrative data.

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?

-Thanks for the suggestion. The two main continuous variables are age and BMI. Based on our previous work with the knowledge users, they prefer the categories of these variables. Also, based on our previous work with other population risk tools, we have found the categories of BMI and other continuous covariates tend to perform well and did not result in a significant decline in model discrimination and calibration. We do not expect categorization of continuous covariates to affect the model predictive performance greatly, but we acknowledge that it may change the model more, which is why we are modelling these covariates continuously and nonlinear as well.

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?

-Thanks for the question. Respondents did not like reporting their income in early cycles of the CCHS, and Statistics Canada decided to start imputing income values from the 2009/10 cycle onwards to deal with the high non-response. The imputation technique Statistics Canada used was a nearest neighbour imputation method based on a modeled household income. This processing is done before the CCHS is made available for public use. We decided not to use any imputation ourselves for the other variables because non-response to a question is an option available to the
CCHS interviewee, and so the knowledge user will have missing data in their population. Unfortunately, the knowledge users are not all trained in statistics, and will probably not have the tools or expertise to impute the data. We have made a note that imputation was done by Statistics Canada (pg. 13, line 5).

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.

-Thanks for your insight. We will explore data reduction techniques if the shrinkage estimator is large, and there is poor predictive performance (pg. 14, lines 12 to 13).

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?

-Sorry for the confusion. We will be using a stepwise selection process (pg. 15, line 10). We will be assessing overall model fit based on AIC and BIC as well as examining the discrimination and calibration of the models (pg. 15, line 17 and 18).

10) Pg 16, line 26: For the calibration plots, will this be assessed at certain time points?
-Yes, the calibration plots will be ascertained at certain time points. The time points have been added (pg. 16, line 10).

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.

-Sorry for the confusion. We have re-phrased the sentence to make this clearer (pg. 16, lines 14 to 16). We are interested in a less than 20% relative difference for adequate calibration. This is just a number we have used in the past, and is not based on any reference.

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?

-We have clarified that a cohort-specific intercept will be added for Manitoba (pg. 17, line 8). The model has drawn interest in other Canadian provinces, and it is our hope that the model will be used Canada-wide. The model will be validated against the Ontario CDPoRT model and updated accordingly. The model could potentially be used in other countries as long as there is a survey that is representative of the population that is available with similar predictors. See response to question 1 on more specifics on how the model is used to predict incidence.