Reviewer's report

Title: The uncertainty with using risk prediction models for individual decision making: an exemplar cohort study examining the prediction of cardiovascular disease in English primary care

Version: 0 Date: 17 Mar 2019

Reviewer: Peter R. Rijnbeek

Reviewer's report:

The authors touch upon an important topic: to assess the performance of prediction models for patient-level decision making. They argue that performance measures such as discrimination and calibration only provide a population-level assessment but the impact for a specific patient is not well assessed. I agree this is an important topic and enjoyed reading this well-written paper.

The approach taken is to do a sensitivity analyses by comparing different design choices during model development and assess their impact on individual patient risk. I think that topic is relevant, and more research is needed on this topic. I do think a proper assessment is only possible if this is performed at a much larger scale, i.e. many target cohorts and outcomes to predict.

I appreciate the discussion that prediction models are not and should not be used in practice without expert guidance when applied to a specific patient. It is simply not possible to build a model that works 100% for all patients, they are intended to provide guidance, prioritisation etc. The same if true for the clinician himself, he will be wrong many times but will use his prior knowledge to optimize his posterior. This is discussed rather late in the paper (in the discussion), I would recommend to also put this in the introductory section.

Models as QRisk are developed in a certain population (often not well defined) and models may need to be recalibrated when applied to other population. This is not discussed in the paper and the original thresholds are applied (which is how they are used by many of course so still valuable to assess). Stating that the new developed model F is working better than the original QRisk model is not completely fair, I think. The message that QRisk should not be applied out-of-the-box to any population is valuable and we need to train the community much better on this.
Furthermore, I think a big problem with models as QRisk is that the Target population (index date) is not well defined. If you apply the model to a 25 year old patient or a 70 year old patient this makes a big difference. Yes Age is in the model, but the case mix is completely different and this results in the uncertainty on patient level in my view. I think we need to develop better methods/approaches that require a stronger anchor point for prediction and a lot of data.

The paper compares QRISK using CPRD data which is as the author mention very comparable to the original dataset used the develop QRisk3. Would this not actually be a good reason not to use CPRD if you want to assess uncertainty? This is recognized at the end of the paper, but I think this is a real weakness of the study and deserves more attention than is given now in the limitation section.

The conclusion of the paper that variations in model settings have impact on the individual prediction is not a surprise in my view, but there is value in making this explicit. I agree that clinical judgement is needed next to the prediction models, however the goal of these models is exactly that and not to overrule (we do need to teach people this better). I also support the fact that deep sensitivity analyses should be part of any development process, external validation at scale is the first requirement to tackle though.

Line 78. I understand the point to make is that the model is probably wrong. I suggest to make this more explicit since the considerable publicity may be about the fact many people had an higher 'Heart Age'.

Line 93. For many models a small set of predictors are selected based on 'prior-knowledge' and it remains to be seen if these models are optimal. I miss a discussion on the data driven approaches throughout the paper, especially since there is a focus on 'lack of knowledge'. Moreover, there is a general problem that there is lack of transparency on how the prediction model is defined, i.e. what is the target cohort, what is the outcome cohort. I think this point can also be given more attention in the paper since this is in my opinion a big source for uncertainty.
Line 108. I strongly doubt that all uncertainty in prediction can be removed by adding more data. First, human live is not that simple, the human body has built in chaotic mechanisms that enable us to survive as a species, outcome will therefore be uncertain by definition. What you are hoping for is a pool of patients that are alike with all their characteristics from which you can fit the best model for that group. However, this will require a very large datasets that is bigger than the world's population. Moreover, what will be the negative set?

Line 122. Can the authors clarify what they mean with very large populations? Provide some numbers. Are these indeed big enough? Could the variables in these models be explained and their rationale?

Line 129. Were the QRisk Definitions indeed clear enough to implement? I doubt there were no issues in for example selection of codes, algorithms used to defined a covariate etc. (this is discussed briefly in the limitations). Would suggest to briefly touch upon this point in the main paper since it is a big source of uncertainty. I appreciate the addition of Additional file 1!

Line 142. What is the rational for adding these other variables? Is this not just as arbitrarily as you mentioned earlier in the paper (line 93)? The fact that other papers thought they are predictive is one thing, but they applied a selection themselves (based on other papers..?). You therefore never answer the real question "how good can we predict CVD?" if you do not also add a full data driven approach containing the full EHR. This requires advanced "machine learning" approaches. I am not suggesting the authors do this in this paper, but a reflection on this I would welcome.

Line 166. Imputation was used. Can the authors provide some insight in which variables needed to be imputed often (I mean in the text not only provide the stats in Additional File 2)? Any remarks to make about the applicability of the QRisk model on CPRD? Would it not actually be necessary to build a good model on datasets like CPRD ? When applying the model, you impute using your own data correct, you do not apply the imputation characteristic's as done in the original study. How big is the bias created by this? At Line 216 the approach is explained a bit more, but I think this is not straightforward to replicate, access to the code would help. Imputation is done using age, sex, ethnicity, why not also include other variables?
Line 181. Many different measures are discussed, some will not be known by most readers. I understand the goal is to assess if some of them provide a better measure for personalized risk prediction performance. However, is there any reason why this would be the case based on their definitions?

Line 224. Here a model F is introduced that was not in the earlier table. I suggest adding this there as well.

Line 249. The effect of secular trends in prediction has been studied by others. I miss references to literature on this topic (in discussion).

Line 260. I would strongly recommend the authors made this code available for the public so everyone can reproduce the study.

Line 267. The fact that the baseline characteristics are 'comparable' and the incidence rates are similar is a good sign and I like this addition. However, this does not mean that therefore the datasets are by definition comparable in my view. There can be manner other factors that play a role. Can the authors comment on that?

Table 3: Overall the performance seems to improve with model complexity. Can the authors reflect on this. Does this actually mean that even better models could be developed if more variables would have been included, more sophisticated machine learning models would have been used etc? Moreover, the fact that models that are trained on CPRD (B-F) perform better is not surprising right? Why not also recalibrate model A, is that not a better comparison?
Are the methods appropriate and well described?
If not, please specify what is required in your comments to the authors.

Yes

Does the work include the necessary controls?
If not, please specify which controls are required in your comments to the authors.

Yes

Are the conclusions drawn adequately supported by the data shown?
If not, please explain in your comments to the authors.

Yes

Are you able to assess any statistics in the manuscript or would you recommend an additional statistical review?
If an additional statistical review is recommended, please specify what aspects require further assessment in your comments to the editors.

I am able to assess the statistics

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

**Statement on potential review bias**
Please complete a statement on potential review bias, considering the following questions:

1. Did you co-author any publication with an author of this manuscript in the last 5 years?

2. Are you currently or recently affiliated at the same institution as an author of this manuscript?

If you can answer no to all of the above, write 'I declare that I did not publish with these authors in the last 5 years and also meet the affiliation criteria'. If your reply is yes to any, please give details below.

I declare that I did not publish with these authors in the last 5 years and also meet the affiliation criteria

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/](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