Reviewer’s report

Title: Combined Performance of Screening and Variable Selection Methods in Ultra-High Dimensional Data in Predicting Time-To-Event Outcomes

Version: 1 Date: 01 May 2018

Reviewer: Ben Van Calster

Reviewer's report:

The paper has certainly improved by adding LASSO and ALASSO (this was a good point from the other reviewers that I had overlooked), by presenting the results differently, and by adding some further explanations. I still have a few comments on this interesting study.

The paper is now framed in terms of overfitting, but I think it is premature to state that these methods can lead to robust prediction models (cf last sentence of conclusion) when the events per variable is close to 0 (n=300 and p=100000). Wouldn't you agree that the aim is to select useful markers to pursue further, not so much to find a reliable prediction model? I think this suggests further attention. The calibration slopes of around 0.90-0.95 in some situations may be caused by the very clear difference between noise predictors (coefficient of 0) and informative predictors: even in the weak signal setting, hazard ratios of the 6 predictive SNPs were around exp(1)=2.7, which I do not find very weak. In this sense, it would certainly be useful to calculate the true c-statistic in order to compare this with obtained c-statistics in Table 4. Perhaps the fact that all predictors were non-correlated may have played a role as well. Then, in the case study, methods that only select the validated score had the highest c statistic, compared to methods that included the validated score plus additional SNPs. That may be a sign of overfitting?

I still find many parts of the text hard to follow.

All simulations assume uncorrelated predictors (as opposed to e.g. Fan (ref 9)): that is not very realistic, so it is a pity that the simulations did not include a setting with correlated predictors. I think you cannot write in the abstract that the simulation study was extensive.

Minor issues:

- What does R2 really add, and why is it computed on the training sets?

- Why 2000 validation samples in one condition and 4000 in the other?

- Case study: number of selected SNPs per method would be nice to see in Table 6.
- Ref 10: wrong journal?

- The discussion states that Fan (ref 9) demonstrated that SIS combined with LASSO is efficient rather than using LASSO. I did not see where Fan evaluated SIS+LASSO, it seems they only checked SIS/ISIS variants vs LASSO?

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:

Needs some language corrections before being published

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