Author’s response to reviews

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

Authors:

Lira Pi (Lira.pi@duke.edu)

Susan Halabi (susan.halabi@duke.edu)

Version: 1 Date: 02 Apr 2018

Author’s response to reviews:

Dear Dr. Debray,

We would like to thank you for the thoughtful reviews of our manuscript and for granting us the opportunity to resubmit the manuscript. By means of this letter we would like to address each of the concerns raised by the reviewers. Their constructive input has resulted in considerable changes and, we believe, notable improvement in the manuscript, as outlined below. In addition, we have uploaded a letter responding to the reviewers critique in the supplementary materials.

In summary, we have addressed each of the concerns raised, and trust that the manuscript is suitable for publication in the Diagnostic and Prognostic Research. Thank you once again for your consideration.

Sincerely,

Susan Halabi
Reviewer #1: The authors present an interesting simulation and application study about the ability of nine combinations of pre-screening and prediction models to identify informative variables in a high dimensional setting and a time-to-event problem. I'd like to address several issues in the following listing:

- General: Throughout the text (e.g. page 2, line 32) it is mentioned that the purpose of this work was " … to compare pre-screening approaches with three popular variable selection methods …." This statement leads to confusion for several reasons:

  o The pre-screening approaches are actually not compared with the variable selection methods but combinations of pre-screening approaches and variable selection methods are compared against each other. Please make sure that, throughout the manuscript, it is clear that the different combinations of methods are compared.

  o Variable selection can be done by filter, wrapper and embedded methods. Pre-screening is a filter method and therefore a variable selection method itself. Please change the wording such that it is clear that both, pre-screening and the prediction models are variable selection methods and that the combination of different variable selection methods is in focus of investigation.

Response: We appreciate the reviewer’s concerns and have addressed to the two questions at the same time. It is certain that SIS and PSIS can select variables by themselves, that is, SIS and PSIS can play a role of selecting variables same as LASSO, ALASSO, and RSF. However, SIS or PSIS should be done prior to running LASSO, ALASSO, or RSF when SIS or PSIS are mutually combined with LASSO, ALASSO, or RSF. To emphasize the processing order among them in this paper, we divide the variable selection methods into two groups: SIS and PSIS in screening approach and LASSO, ALASSO, and RSF in variable selection method. Per your recommendation, we added the explanation on page 4, lines 19-24.

  o The term "ultra-high" needs some clarification. What is the difference to high dimensional (n<<p) data?

Response: We thank the reviewer for raising this point. The definition of ultra-high dimensionality was added on page 4, lines 14-16. While the high dimensionality refers to simply n<<p, the ultra-high dimensionality refers to the exponential growth of p relative to n.

  o Nine combinations of pre-screening approaches and variable importance measures are investigated and their performance is compared. It is not clear whether such a combination is necessary as a corresponding benchmark is missing. How does pre-screening perform, i.e. does it already identify all of the informative variables? How do LASSO or RF perform
without pre-screening? How does the ranking of variable importances of a RF without variable selection look like, i.e. what are the ranks of the informative variables?

Response: Thank you so much for your suggestion. In addition to the nine combinations, we have included only SIS, only PSIS, only LASSO, and only ALASSO in the simulation study. It is impossible to run only ISIS because, by the definition of ISIS, ISIS iteratively runs SIS and a variable selection method. For example, ISIS-LASSO means the running SIS->LASSO->SIS->LASSO->… For only RSF, it was computationally intensive. Linux system gave up running only RSF to analyze 100,000 covariates simultaneously.

- Abstract, Background: The authors refer to personalized medicine and a phase III clinical trial which gives the impression that treatment and treatment x feature interactions are investigated. This is not the case in the present manuscript. Please be clear on the research objective of the manuscript.

Response: We rewrote this part on page 2, lines 8-10.

- Abstract, Results: The acronyms "PSIS", "ISIS" and "ALASSO" have not been introduced.

Response: This has been spelled on page 2, lines 6-7 and page 4, lines 8-9.

- Abstract, Results: It is not clear under which curve the time-dependent area is computed. Please add the word "ROC".

Response: Since Harrell’s c-index provides the same information with the integrated area under the ROC curve proposed by Heagerty and Zheng (2005), we decided to calculate only the c-index so that the statement was removed.

- Abstract, Results, last sentence: A verb is missing.

Response: We added “achieved” as the verb on page 3, line 3.

- Page 4, lines 14-19: Random Forests is not a variable selection method. It is not until page 9 that the reader is acknowledged that an extension of this method is used to perform variable selection.
Response: Per your suggestion, we changed the words on page 4, lines 8-9.

- Page 4, line 21: What are "high dimensional covariates"?
Response: We appreciate the reviewer raising the question. It meant covariates in high dimensionality. It was clarified on page 4, line 12.

- Page 4, lines 19-26: The authors state that LASSO and RF are "poor in terms of both computational efficiency and estimation accuracy even when the number of covariates is close to the sample size." This statement about the predictive accuracy in a non-high dimensional setting is rather surprising to me. Please add a reference.
Response: We added explanation on page 4, lines 11-15 and page 5, lines 3-5 as we did not find an adequate reference.

- Page 4, line 36: Why are quotation marks used?
Response: It was only to emphasize that ALL important covariates are survived by screening approaches. The quotation marks have been removed.

- Page 4, lines 37-41: Please provide a reference for this very strong statement.
Response: We rewrote this sentence and included references on page 5, lines 7-9.

- Page 6, line 13: "m" or "M"?
Response: We apologize for the confusion. It should be m. m and M refer to the number of covariates contained in a reduced subset and a final model set, respectively.

- Page 6, lines 49-56: When s is a set of indices, what does "|.|" mean then? Also in the following lines it is often not clear when it is referred to an integer or a set of integers.
Response: s is the number of the selected covariates in each partition so that |.| indicates the number of covariates in a set. To make notations simpler in the following lines, capital letter
means a set of indices while small letter indicates the size of them in the section. This has been corrected on page 7.

- Page 7, lines 10-13: "... is most likely to keep unimportant covariates that are highly correlated with the important covariates." Please define "important". The authors are actually talking about marginal and conditional effects. Also variables with a marginal but no conditional effect can be useful in Random Forests, for example.

Response: We are in agreement, and have included the definition of the importance on page 4, lines 19-20.

- Page 7, lines 15-33: Please be clear on whether "O" is a set of covariates, a set of indices or a model. Line 28 for example says "... the second model O_2 ..."

Response: O is defined by a set of indices so that the “model” was changed to “subset” on page 8, line 10.

- Page 8, line 3: Please introduce "I_k".

Response: We apologize for omitting the definition. I_k is defined by the information matrix and the explanation was added on page 9, line 4.

- Page 10, line 15: What is the reason to set aside 1/3 of the simulated data for testing? In a simulation study one can compute the performance of a prediction model as it is applied to a large dataset generated from the underlying joint distribution (see T. Hothorn, F. Leisch, A. Zeileis, and K. Hornik. The design and analysis of benchmark experiments. Journal of Computational and Graphical Statistics, 14(3):675-699, 2005. doi:10.1198/106186005X59630. URL https://urldefense.proofpoint.com/v2/url?u=http-3A__dx.doi.org_10.1198_106186005X59630&d=DwIGaQ&c=imBPVzF25OnBgmVOlcsiEgHoGl6YHLR0Sj_gZ4adc&r=jVU6V2ymkbL- 3a__AaHfIfie7xASUQw82HJGRBQ&m=NN9csev2gMpBLLjLoO_BhXdidP2FpZdubpJVXN3DXQ&s=fgYpuPy6ZbazQjYUlZ69tzuSEoi8j1mSIH3MtJ4SBs&e=.)

Response: We thank the reviewer for the suggestion. We have used the reference and we generated large data sets for testing. The detailed explanation is provided on page 11, lines 9-13.
- Page 10, line 48: Has "MAF" been introduced?

Response: The abbreviation MAF was defined on page 12, lines 1-4.

- Page 11, lines 5-11: Please give a short explanation why it is a good idea to use a similar setting as in Fan et. al.

Response: We answer this question and the next one below.

- Page 11, lines 45-50: "By imposing a weak signal strength with a small sample size ..., we wanted to explore if the low coefficients had an adverse impact on the performance ..." Unfortunately both, the change in the sample size and the change in the coefficients will affect the results simultaneously. Therefore it is not possible to conclude on an adverse impact of low coefficients from this simulation setting. A fixed sample size and varied coefficients would be needed.

Response: These two are very good questions. We first thought that it was obvious that strong signal strength would make better performance than weak signal strength at fixed sample size. Then we expected the same design setting in accordance with Fan et al.’s paper would make more interesting results. However, we admit the difficulty of interpreting the results as the effects of different sample size and different signal strengths simultaneously make the results difficult to interpret. Thus we changed the design setting as indicated on page 11, lines 23-24. Both strong and weak signal strengths are done at fixed sample size.

- Page 12, line 52: What is a "random split point"?

Response: In contrast of deterministic splitting for an x-variable, split points are randomly chosen among the possible split points for x-variables in a maximum of number of splits. The random splitting can significantly increase speed over deterministic splitting. The explanation is provided on page 13, lines 4-5.

- Page 13, Metrics of Performance: The categorization into "undersize", "correct size" and "oversize" is simple and disguises information. The authors intend to show whether a model is of correct size and whether the correct variables are among the selected ones. To do so they should use the precision and recall of the models (next to the size and c-index). These
measures can also be plotted against each other for each method to see which methods are dominated by others.

Response: We appreciate the suggestion and we provided the mean and percentage of informative and uninformative features as well as final model size and removed the previous simple size measures in the results section.

- Page 14, line 1: The highest proportion of CorrectSize-All is "observed" for ISIS-LASSO but not "estimated".

Response: This has been corrected.

- Page 14, lines 14-17: "… a small sample size … was highly likely to be empty …" How can a sample size be "empty"?

Response: We apologize for the confusion. The sentence was wrong. It originally meant that at a small sample size such as 150 and with weak signal strength, it is highly likely for SIS and ISIS to select no covariate or few informative features. This has been corrected.

- Page 14, line 22: "Unlike SIS, ISIS, PSIS selected covariates …" Should it read "Unlike SIS and ISIS, PSIS …" or "Unlike SIS, ISIS and PSIS …"?

Response: This has been corrected to read that it was “Unlike SIS and ISIS, PSIS…”.

- Page 14, line 42: "… we observed that the concordance indices increased with larger sample sizes and weaker signal." The effect of a weaker signal cannot be separated from the effect of larger sample sizes based on the simulation settings used. A conclusion on the effect of a weaker signal is therefore not possible.

Response: We recognize the limitation and the simulation setting was changed per the previous suggestion.

- Page 15, line 39: What is "IDD-AR(1)"?

Response: It is typo. It should have been AR(1)-Exp. We have simulated now exponential distribution in new simulation setting. The notation does not exist on the manuscript.
- Page 17, line 21: "Among the 623 Caucasian men, 94% were dead." Do you mean that events have been observed in 94% of the 623 Caucasian men?

Response: This has been corrected on page 15, lines 4-5.

- Page 17, line 42: Heagerty and Zheng (2005) show that the integral over weighted AUC values, that are derived from the dynamic specificity and incident sensitivity, equals the c-index. Why do you use both? Does it add any extra information?

Response: Harrell’s c-index may be thought of as a generalization of finding the area under an ROC curve. Nevertheless, the exact numeric values of c-index and integrated AUC can be different so that both measures were given together.

- Page 18, lines 25-28: "Using the aggressive variant when sample size is small …" What is a "small" sample size? How can the reader make use of such a recommendation?

Response: Thank you for the suggestion. We clarified the term as n=150 and rewrote the paragraph as simulation setting was changed on page 18, line 19-20.

- Page 18, line 31: What is meant by "problematic"?

Response: we thank the reviewer for pointing it out. It meant that the performance from PSIS is lower value of c-index bad because of final model including too many uninformative covariates.

- Page 18, line 38: What is meant by "well-structured"?

Response: We meant to state that this refers to when a sample size is large enough and/or relationships between important covariates and survival outcomes are strong. This has been excluded.

- Table 1 is not needed.

Response: We removed Table 1.
- Tables 2-6: Results should be presented uniformly.

Response: We thank the reviewer for the suggestion. All results are now presented uniformly.

Reviewer #2:

General comments:

* The authors compare various versions of the lasso and the screening procedure. I would like to see a comparison that includes reference values. Firstly, performance values of the data-generating-mechanism could be included to show whether the methods' approach these or further improvement is possible. Secondly, the performance values of performing only screening and only the lasso could be added, to identify the additional value of applying the lasso after a screening procedure and of applying a screening procedure before the lasso. Any comparison between methods (such as the three screening methods) is meaningless when no comparison to the default (such as no screening) is made. Including these comparisons would make the manuscript significantly more interesting.

Response: We thank the reviewer for the suggestion. As explained above for reviewer 1, we included only SIS, only PSIS, only LASSO, and only ALASSO in the simulation study. It is impossible to run only ISIS because, by the definition of ISIS, ISIS iteratively runs SIS and a variable selection method. For example, ISIS-LASSO means the running SIS->LASSO->SIS->LASSO->… For only RSF, it was computationally expensive. Linux system gave up running only RSF to analyze 100,000 covariates simultaneously.

* The authors have only compared the methods on discriminatory performance and on variable selection, and not on calibration and overall predictive performance. Therefore any claims about predictive accuracy are a stretch.

Response: We are in agreement with the reviewer and included calibration and Graf et al’s R-square as measure of predictive accuracy.

* Why were specifically these screening and estimation methods chosen?

Response: This is a very good question in terms of showing our motivation in the paper. The detailed explanation was included on page 5, lines 6-14. The estimation methods such as LASSO, ALASSO, and RSF were included because of their wide application to analyze high
dimensional and bio-medical data in clinical studies. Moreover, we wanted to introduce SIS and
PSIS to those who are not familiar with them since prediction accuracy from the combinations
between the screening approach and the estimation method was shown as improved rather than
using only the estimation method.

Abstract, page 2,
Line 8 - 13. Is ultra-high dimensional data really restricted to molecular data? This sentence
seems to imply this.
Response: We removed “molecular” on page 2.

Line 13- 18. What is 'pre-screening'? Isn't what is being done here just 'screening'?
Response: We replaced “pre-screening” with “screening” throughout the manuscript following
Fan et al.’s paper.

Line 47-50. The abbreviations PSIS-LASSO, ISIS-LASSO and ISIS-ALASSO have not been
introduced. I personally think the way of introducing these as has been done in the Background
section is easier to read, e.g. after LASSO has been introduced, 'adaptive LASSO (ALASSO)' is
immediately clear. Though, I don't know whether the journal has any guidelines thereon.
Response: We have spelled out all acronyms on pages 2 and page 4.

Line 52-54. I don't understand this sentence.
Response: We apologize for the confusion. We removed the sentence and rewrote the paragraph
on pages 2-3, lines 22-4.

Page 3, conclusion. I think the conclusion of the abstract should reflect that the simulation was
limited in scope. Firstly, the performance of the methods was only investigated in the context of
SNP data. E.g. on page 10-11, the authors say the simulation was designed to reflect such data.
Secondly, only discriminatory performance and variable selection were investigated.
Response We thank the reviewer for the suggestion. We corrected that on page 3, lines 5-7 per
your recommendation and we added measures of overall performance and calibration.
Abstract, page 3

Line 1 seems to be missing a verb.

Response: We added the verb “achieved” as on page 3, line 3.

Background, page 4

Line 3 - 10. I would not call this a problem of the PH model, but rather the Maximum Likelihood estimation of the PH model.

Response: We rewrote the sentences on page 4, lines 2-6 to make them clearer.

Line 37-42. Is this well-known? I recall that variable selection methods only work well when there is sufficient data. Maybe the authors can provide a reference here.

Response: We thank the reviewer for the comment. To clarify, we meant that combinations between screening and variable selection methods work well to remove uninformative features while retaining all informative features in final models. To carry out the meaning, we changed the sentence and included references on page 4, line 7-9.

Page 5

Line 2, and later. I recall that the unit commonly used for thousand is k, whereas K is used for kelvin.

https://urldefense.proofpoint.com/v2/url?u=https-3A__en.wikipedia.org_wiki_SI-5Fbase-5Funit&d=DwIGaQ&c=imBPVzF25OnBgGmVOlcsiEgHoGl1i6YHLR0Sj_gZ4adc&r=jVU6V2ymkbL-RBgPqJnLyLnOEir7xASUQw82HJGRBQ&m=NN9csev2gMpB-ILjLoO_BhXdidP2FpZdubpJVXN3DXQ&s=KWC37bckYB3GqVDxlMDQ8FBOr5_yfNHIlDgzzyxgoY&e=

https://urldefense.proofpoint.com/v2/url?u=https-3A__physics.nist.gov_cuu_Units_current.html&d=DwIGaQ&c=imBPVzF25OnBgGmVOlcsiEgHoGl1i6YHLR0Sj_gZ4adc&r=jVU6V2ymkbL-RBgPqJnLyLnOEir7xASUQw82HJGRBQ&m=NN9csev2gMpB-ILjLoO_BhXdidP2FpZdubpJVXN3DXQ&s=hDGLHzrD9AAqPacXkcsOhskOpNGWG5wZ93bHNZBQqko&e=
Response: We replaced the unit K, with k on page 5, line 23, page 12, line 10 and page 13, line 19.

Methods, page 8
Line 58 -1. This sentence seems to suggest all covariates are non-zero. Surely the lasso can also set some coefficients to zero in this algorithm, similar to the regular lasso?

Response: This was only the explanation of regular LASSO and we rewrote the sentence as indicated on pages 10, lines 5-6.

Page 9
Line 43-46. "baseline hazard function is 1". Is it equal to the value 1, or equal to equation (1)?

Response: It is the value of 1 so we corrected the sentence on page 11, lines 1-2.

Simulation Studies, page 9
Line 57 - 2. Why is the lasso without screening procedure not considered? Similarly, I would like to see the number of variables selected by the screening procedures, before applying the lasso.

Response: We thank the reviewer for the suggestion. We now include results from only SIS and only PSIS without variable selection methods in the results section.

Page 10
Line 7-10. Why are calibration and/or overall performance not compared?

Response: Please refer to the answer below.

Line 7-22. I think a (reference to the) definition of the concordance index would be in order. The integrated time-dependent area under the curve is not mentioned in the methods section at all, and I think a (reference to the) definition would also be in order in this part of the manuscript. Further, I think the text would be clearer if the authors highlight the difference between these indices.
Response: We appreciate these two recommendations. Both performance measures were added and each reference for each measure were also included in order as indicated on page 11, lines 10-22.

Line 14. Concerning the simulation of the data sets, why was the test set smaller than the training set? In practice this makes sense: to allow for a large training set. But in a simulation study one can simply generate a larger test set.

Response: We thank the reviewer for the suggestion. We generated large data sets for testing. The detailed explanation was provided on page 12, lines 1-4.

Line 48. The abbreviation MAF is not introduced.

Response: We defined MAF on page 12, line 16.

Page 11

Line 5-32. If I understand correctly, the coefficients were fixed for the entire simulation and chosen at random. Why was this procedure used? And if they truly are random, how are they generated? I observe that strong > moderate > weak signal, but this seems rather coincidental if they were generated randomly.

Response: Please refer to the answer below.

Line 35-37. I don't fully understand this sentence: what average is being referred to?

Response: Please refer to the answer below.

Line 45-50. What is the purpose of altering sample size and coefficient strength simultaneously? Is there some other parameter that is fixed?

Response: We apologize for the confusion. Please allow us to explain the above three questions at the same time. We first thought that it was obvious that strong signal strength would have better performance than weak signal strength at fixed sample size. Then we expected the same design setting in accordance with Fan et al.’s paper would make more interesting results. However, we admit the difficulty of interpreting the results as the effects of different sample size
and different signal strengths simultaneously make. Thus we changed the design setting as on page 11, lines 23-24. Both strong and weak signal strengths are done at fixed sample size.

Page 13

Line 1-5. I don't see where the number of 520 simulated data sets comes from. Perhaps this could be explained, or a reference to the section of the text where this is explained could be added.

Response: We rewrote this sentence, as indicated on page 13, lines 19-21.

Line 16-26. I understand the notations "CorrectSize", "UnderSize" and "OverSize", but all situations are also labeled "all". If all conditions receive the label "all", then what does the label mean?

Response: We thank the reviewer for the suggestion. We admit that the measures are inefficient and too simple to correctly carry out performance of each combination or approach. We now provide the mean and percentage of informative and uninformative features as well as final model size and removed the previous simple size measures in the results section.

Line 46. The link appears to be dead.

Response: This has been fixed.

Results, page 15

Line 42-46. I don't see a causal relation here. The two observations go hand in hand.

Response: We agree with the suggestion and we removed it.

Discussion, page 18

Line 51 - 56. Perhaps I don't fully understand this, but it sounds like the authors knew on beforehand that the (A)LASSO would outperform RSF due to the linearity assumption, and that in cases where this assumption does not hold the results would turn around. If so, why was RSF included in the simulation? If not, could the authors clarify this paragraph?
Response: This is a very helpful comment. RSF is a non-parametric machine learning method and known to produce models that do rarely cause overfitting. We included the performance of RSF due to its wide application in clinical data analysis. Since the statement could be incorrect if it is not depending on linearity assumption, we have removed it.

Page 19

Line 15-28. In what proportion of the simulations did these discrepancies occur? Did the authors investigate these discrepancies and were any mistakes in either of the codes found?

Response: We took the R codes to run SIS and ISIS from the R library SIS version 0.6 which was available at the time of the simulation study. Thus, results from the aggressive variant of ISIS-LASSO using the latest R library SIS can produce slightly different results when the sample size is small and signal strength is weak.

Figures

A combination of 3 x 3 methods is being investigated. Why aren't there 9 lines in the plots?

Response: As we found it was inefficient way to visualize the results, it was removed for new simulations.

Tables

What are the numbers in brackets in tables 3 and 4? SE?

Response: The numbers is standard deviation and we have added a footnote to the tables.

It appears that in contrast with the other tables, no SE are given in table 6.

Response: We now provide the 95% confidence interval from 200 bootstrapping samples for the testing set in Table 6.
Reviewer #3:

1. I miss a more broad framing of the context and results: I think that ultra-HD data are a context in which you mainly want to select the most promising variables for further investigation, rather than develop a prediction model. Overfitting will no doubt be an issue, but this is not discussed. E.g. why is the c-statistic low in small datasets despite stronger coefficients for the six true predictors? In that sense, you can actually compare the observed c-statistics with the true c-statistic.

Response: We provide the original Harrell’s c-index and its corrected one from training set of real prostate cancer data in Table 6. As shown in the table, the corrected c-index proposed by Harrell et al. (1996) is smaller than the original c-index.

2. SIS and related methods: how often is this used in practice? I think many people are not very familiar with this, so I would devote sufficient attention to a clear explanation (perhaps with an appendix). It was not always easy to follow for me. (On p5 you write that you want to give applied statisticians a basic understanding.) Some specific comments:

Response: We thank the reviewer for the suggestion. It can explain the motivation of this paper. Because LASSO has a tendency of selecting too many uninformative features, SIS was introduced to obtain higher predictive accuracy as removing the uninformative features. We provide detailed explanation on page 5, lines 6-14.

- P6: how is s chosen?

Response: This was chosen as n/log(n) for aggressive variant of SIS, referred by Fan et al. This has been clarified on page on page 7, line 18.

- P7: SIS is asymptotically 'sure'. If asymptotically refers to sample size going to infinity, how relevant is that for the typical ultra HD situation?

Response: This asymptotic certainty implies that the reduced subset is highly likely to include all the important covariates in the HD problem. This has been clarified on page 8, lines 3-4.

- 'the subset filtered by SIS or PSIS includes not only the important p* covariates', p8: that sounds as if that is a certainty.
Response: This statement cannot be 100% certain, so we re-wrote the sentence on page 9, lines 7-8.

- P8 at the bottom: can you give more information regarding the details of the method to select lambda? Which grid, which kind of cv, which error?

Response: We thank the reviewer for asking. We used a function, cv.glmnet in the R library of glmnet, to calculate the lambda. According to the library, 10-fold cross-validation was used as a default and the lambda was selected with minimum of the partial likelihood error from survival model. We have provided the explanation on page 9, lines 3-5.

- P12, 'we set the maximum number of iterations to five': I am not sure what this refers to.

Response: We thank the reviewer for raising this question. ISIS runs iteratively until either a condition is satisfied, or a pre-specified number of iterations is reached. We chose the maximum number of iterations to be five as it was the default in the SIS library version 0.6 when running ISIS to prevent iterations from running extensively.

3. There are a lot of results, and then it inevitably becomes difficult to keep track of them. Given that you want to provide recommendations on the use of the tested approaches (p5), a clear overview of findings together with recommendations should be added. Related to this, the conclusions on p19 were rather uninformative.

Response: We apologize for the confusion. As we have performed simulations in different simulation settings and additional approaches, we entirely changed the results section with two additional measures. We expect that the current results section will help readers understand the results and the conclusions.

4. At the same time, several interesting parameters have been fixed in the simulations, raising the question of generalizability. E.g. p* is fixed at six because that is what is used in most other studies, q was fixed at 0.15, the combination of sample size and strength of the six variables was fixed. Do you plan to check the impact of this on the results, or do you assume that the impact is limited?

Response: We thank the reviewer for the suggestion. These are important assumptions that will affect simulation results. We wanted to make the simulation setting similar to our real SNP data so that, for example, the covariates in the simulation study were generated as 0, 1, and 2. In the
circumstance, our primary interest was which selection approach performs the best depending on different sample sizes and the signal strength. Thus, we admit that general use of the simulation results to the real data could be limited in some cases. The limitation was explained on pages 19-20, lines 22-8.)

5. P18, 'relaxing the linearity assumption would have led to different conclusions': how do you know? In your simulation, you analysed the SNPs as continuous variables to get 1 coefficient per SNP, right? If you would analyse the SNPs as categorical variables, and hence get 2 coefficients per SNP, you may increase overfitting? Perhaps that may make the assumption of linearity in the ultra HD setting a defendable one.

Response: We thank the reviewer for the suggestion. We regarded the SNPs as continuous variables. We only wanted to point out that there could be a penalty of using RSF in our simulation setting because only RSF is non-parametric method. We removed the paragraph per your suggestion.

6. P13, '520 simulation runs': why 520??

Response: We have now changed the simulation scenarios runs to 300. To generate 100,000 covariates and run SIS or ISIS, it is computationally intensive. This is higher than what Fan et al.’s reported as they simulated 200 data sets.

7. Selection results are summarized using the CorrectSize-All, UnderSize-All, OverSize-All categorization, and using the mean number of selected variables. Why not show the mean/percentage of important and non-important variables that are selected?

Response: Per the suggestion, we present the mean of important variables out of six and proportion of unimportant variables in each model are shown in Table 3. We computed the mean of the number of important variables finally selected over the total simulation data sets for each simulation scenario. Also the average of percentage of unimportant variables finally selected were driven. The values have been presented on Table 3.

8. P19: the fact that you obtained different results depending on which R code you use is very important. You cannot state this without further information. Can the code be trusted?

Response: We have addressed this issue in our response to reviewer 2. We have included the code on a link that will be publically available.
Minor comments

1. Background, first paragraph, 'this model is unable to select the important covariates': unclear whether I understand what you are referring to. Do you refer to stepwise selection methods?

Response: We rewrote the statement on page 4, lines 3-6.

2. Background, second and third paragraphs, 'current implementation of these methods shows that their performance is poor…' and 'it is well-known that the variable selection methods work almost perfectly…: you need to add references here to support these claims.

Response: We have addressed this point per reviewer 1.

3. About metrics of performance: which concordance statistic was used? There are several c statistics for survival data.

Response: For the new simulations, we used Harrell’s concordance statistic and we stated that on page 11, line 18,

4. P15, lines 8-25: some sentences in the paragraph can be interpreted wrongly (in my view), please consider rewording.

Response: We thank the reviewer for the suggestion. We rewrote the section based on the new simulations.

5. P17: missing SNP values: how common were missing SNP values?

Response: Distribution of missing SNP values is as following:

<table>
<thead>
<tr>
<th># of missing values for each SNP</th>
<th>0</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
</tr>
</thead>
<tbody>
<tr>
<td># of SNPs</td>
<td>424,792</td>
<td>59,508</td>
<td>9,713</td>
<td>2,317</td>
<td>972</td>
<td>500</td>
<td>279</td>
</tr>
<tr>
<td>Proportion (%)</td>
<td>85.29</td>
<td>11.95</td>
<td>1.95</td>
<td>0.47</td>
<td>0.20</td>
<td>0.10</td>
<td>0.06</td>
</tr>
</tbody>
</table>

We have indicated that in the manuscript.
6. Table 2 (and other tables): Percentages and means are averaged over the 520 simulation results?

Response: Percentages and means were averaged over the 520 simulation results. We have included an explanation on page 15, lines 11-12 for the new simulations.

7. Figure 3A-B: please state over what parameters you averaged the concordance index.

Response: It was averaged over 520 simulation data sets. These figures have been deleted as we did new simulation.

8. Figures are of poor quality

Response: We have changed the figures so that we have better resolutions.