Author’s response to reviews

Title: A comparison of landmark methods and time-dependent ROC methods to evaluate the time-varying performance of prognostic markers for survival outcomes

Authors:

Aasthaa Bansal (abansal@uw.edu)

Patrick Heagerty (heagerty@uw.edu)

Version: 1 Date: 26 Feb 2019

Author’s response to reviews:

We thank the reviewers for their helpful comments on our manuscript. We address these comments below. We note that revisions to the text are in blue color.

Reviewer 1 Comments

The manuscript mainly introduced how to estimate time-varying hazard ratio in the Cox model by means of changing the length of follow-up duration and making coefficients vary with time, and how to evaluate accuracy in terms of time-dependent AUC by comparing incident or cumulative cases with dynamic controls.

1. If I understand you correctly, the lines 152 - 155 mean that the regression coefficients are interpreted as the (average) HR which becomes the exponential of coefficients, as shown in line 214. Which one of \beta and exp(\beta M) is the hazard ratio in line 104?

   - In line 104, the hazard ratio is exp(\beta), not \beta as we had previously stated. Thank you for pointing out this error. We have corrected it in line 104.

2. You say 'Hazard ratios are also sensitive to the scale on which the marker is measured' in line 518, that is, the model parameter coefficients are sensitive to the scale of variables? Is this an issue if you normalize every continuous variable, e.g., in the range of [0,1]?

   - No, this is not an issue if you normalize every continuous variable, as we did in the illustration. The point that we are trying to convey in line 518 is that in order to be able to compare hazard ratios, normalization is a necessary added step, hence a disadvantage compared to using the AUC, which does not require such normalization.
3. In page 16, the first paragraph involves the ties issue. Did you dichotomize the eight continuous markers in the process of AUC estimation?

- We did not dichotomize the eight continuous markers in our illustration. Although ties are not an issue in our analysis, they can generally be a concern with using ROC curves with categorical markers, which is what we are conveying on page 16.

4. In line 345, the precise representation could be like "the predictions for a comparable pair of subjects", as two controls (censored cases) are not comparable. The C-index is actually a generalization of AUC(t), is it possible to append the results in terms of C-index for each AUC curve in Figure 4(right)?

- Thank you, we agree that it would helpful to see the C-index value for each curve in Figure 4 (right panel). We have now added C-index values and corresponding 95% confidence intervals to the figure.

5. For ROC(p) and specificity=1−p in pages 9 and 10, what probability does p, in particular, refer to? Eq.1 has used such notation as the dimension of data. Similarly, k is reduplicate for a certain subject and a landmark time point (t^L_k) on page 9.

- p refers to the false positive rate, hence specificity = 1−p. We have clarified this in the text. Also, we thank the reviewer for pointing out that we earlier refer to p in equation 1, in reference to the number of variables in the Cox model. We have updated Equation 1, so that we no longer use p.

Thank you also for pointing out the duplication of k. We have updated the manuscript so that we no longer use k to represent an individual.

6. The ROC / AUC is sensitive to class/group imbalance. The landmark analysis would lead to short follow-up periods (e.g., 1-year study) and yield more Controls and fewer Cases. Did such an imbalance between cases and controls arise and affect the AUC performances?

- An imbalance between the number of cases versus controls is, in fact, not an issue for the estimation of AUC, except in terms of ultimate precision of the estimate (determined by both case and control sample sizes). However, we agree that doing landmark analysis in a short study with a short follow-up period could result in small numbers of cases and/or controls and impact estimation. Since landmark analysis is done to assess time-varying association/performance, it might not typically make much scientific or practical sense to do a landmark analysis in a short
study because one would not expect association/performance to change meaningfully over a short time period.

Reviewer 2 Comments

1. While the HRs is a popular measure of association, and it is very commonly reported even in the prognostic context, it is hard to argue it is a measure of prognostic performance. Moreover, in the new world where machine learning models of different type and usability emerge, one might want to use metrics that do not depend on one particular model formulation. While I find the discussion about landmark vs. time-dependent HRs of interest, it either needs to be placed in a more general context or relegated to problem exposition (exposing the nice connection with incident AUC) rather than a "prognostic methods" as the title suggests.

- Thank you. We agree that the discussion of HRs might not fit in a section on methods for assessing prognostic performance. We have updated the structure of the manuscript, so that hazard ratios are no longer discussed in the section titled ‘Summaries that Characterize Time-varying Performance’, but instead in a separate section titled ‘Time-varying Hazard Ratios’.

2. Along the similar lines, I would like to see one or more truly prognostic metrics that are applicable in whatever form the model was developed. This could be include one of the risk-based measures from Liang and Heagerty plus something that the ML community relies on.

- We appreciate the opportunity to clarify the important distinction between model formulation/development and model evaluation. We emphasize that the ROC curve (and AUC) do not depend on the particular model formulation. One may use standard Cox regression or more flexible, modern machine-learning approaches for model development with training data. Regardless of the chosen initial modeling approach, the ultimate prognostic model is then fixed and used in validation data to provide patient predictions of the disease outcome, i.e. a risk score, which can then be summarized using ROC or AUC methods. In this manuscript, we are agnostic to model development/selection. We focus on methods for evaluating any single “biomarker”, which may be a novel predictive measurement, such as a specific serum protein level measured in the laboratory, or more commonly may be the risk score derived from a model that includes multiple factors, i.e. a derived biomarker or classifier. The ROC curve approaches we discuss for evaluating a risk score in the validation data are independent of those used for model selection in the training data, in that they do not rely on the assumptions that may be necessary for the development of the risk score. We have addressed this point in the Discussion section of the manuscript.
Additionally, in the discussion, we do comment on alternative metrics such as the risk-based metric developed by Liang and Heagerty.

Finally, we note that sensitivity, specificity and ROC curves are standard metrics of prognostic/classification performance that the machine learning community also relies on, as discussed in these papers, for example:


3. The commentary and example presented read well and are informative. However, the paper does not give the reader a sense of what to expect from the metrics of performance in terms of magnitude. In the context of the example, we can rely on ranking of markers but how do we decide how many to ultimately select? This a more practical question than the ranking exercise.

A set of simulations aimed at establishing some intuitions about the expected magnitude under common, plausible scenarios would greatly enhance the paper.

- Thank you. We agree that the question of marker selection/combination and incremental value are important. The methods we present allow one to evaluate and compare the time-varying performance of a baseline model versus an expanded model in order to assess incremental value of additional predictors. We agree that simulations aimed at establishing intuition about the expected incremental value under common, biologically motivated scenarios are important. We have, in fact, previously published a paper implementing such simulations in the diagnostic setting with binary outcomes:


We expect similar results to hold in the setting of prognostic markers with survival outcomes. We have added the above to our discussion of multiple markers and incremental value in the Discussion section of the current manuscript.

4. Similarly, it is not very common that we are looking for just one marker. A section on multiple markers and incremental value would be helpful.
Please see our response to the above comment. We thank the reviewer for bringing up this point. We agree that a discussion of multiple markers and incremental value is important and we have expanded our discussion on this point in the Discussion section of the paper.