Author’s response to reviews

Title: A general approach to risk modeling using partial surrogate markers with application to perioperative acute kidney injury

Authors:
Derek Smith (derek.k.smith@vanderbilt.edu)
Loren Smith (loren.e.smith@vanderbilt.edu)
Frederic Billings (frederic.billings@vanderbilt.edu)
Jeffrey Blume (j.blume@vanderbilt.edu)

Version: 1 Date: 16 Nov 2017

Author’s response to reviews:

Reviewer #1:

We would like to thank the first reviewer for her/his insightful comments. These critiques have helped us improve the manuscript and make it more accessible and meaningful to readers, especially those who are more clinically inclined.

1. Abstract. Methods section needs to clearly state which outcome is being predicted from change in serum creatinine. In general, the abstract could be clearer in showing how the conclusions follow from the results (I appreciate this is difficult with a restricted word count)

We have made substantial changes to the abstract to improve clarity.

2. P19. `l22 onwards. The argument that a harm state is detected by an egfr90 needs to account for a broader range of patients with transient AKI. Recovery after AKI to an egfr90 within 10% of baseline may still confer risk of an accelerated decline in renal function. My point here is that using a dichotomous outcome at 90 days may fail to identify patients who will have a poorer longer term clinical outcome even though their egfr90 is back to baseline - reduced renal function at 90 days is not the only manifestation of harm in AKI survivors.

The reviewer brings up an excellent point here regarding transient AKI sufferers. Patients who have suffered from AKI and later have complete normalization of their baseline serum creatinine level or eGFR, sometimes referred to as full renal recovery, have an increased long term mortality risk and risk of chronic kidney disease development or progression compared to similar
patients who did not suffer from AKI. We acknowledge that eGFR90 decline is not the only adverse long term consequence of AKI and that this outcome alone may fail to identify the increased long term risk of kidney decline observed in patients who suffer transient AKI. If these additional, important long-term AKI outcomes were to be incorporated into our mixture model, the model would require validation with respect to the added long-term outcomes. We agree that adding these outcomes to our model could further improve our mixture model-based kidney injury risk score. Our current patient cohort has not been followed for long-term mortality, chronic kidney disease diagnosis, or development of dialysis dependence. However, our inability to incorporate these additional long-term outcomes does not diminish the importance of our demonstration of the partial surrogate behavior of serum creatinine, or our demonstrated of improvement in risk modeling with the application of a two-component mixture model to account for renal functional reserve. We have added a discussion to the manuscript of the limitations of eGFR90 to make it clearer to the reader that we are not advocating eGFR90 decline as the only adverse long term consequence of AKI, but rather as one possible target for an AKI risk prediction model which is of interest to clinicians and researchers. The manuscript now states:

“Despite being a clinically important marker of kidney function, eGFR90 has limitations. Patients who experience transient serum creatinine elevations that resolve by 90 days may still be at increased risk for adverse sequelae (Welten, Schouten et al. 2007, van Kuijk, Flu et al. 2010). The analysis we performed here does not fully capture these patients’ increased risk. This analysis demonstrates that the proposed technique produced a superior model for the prediction of eGFR90. Although it is plausible that similar models will improve the prediction of other AKI related endpoints, these models will require validation using data not available in this dataset.” in the discussion section.

We have also altered the language in the manuscript to prevent further confusion on this point. Each occurrence of “true outcome” has been replaced with “target outcome” to emphasize that we are referring to the variable which is the target of the prediction problem, which does not need to hold a gold standard status.

3. L51. There is a strong rationale to expect a confounding by indication for patients who have a renal function test at 90 days after surgery compared with survivors who do not have this test requested by their treating clinicians. I disagree that Table 1 shows equivalence of the two populations. Patients who do not have an eGFR90 taken have a higher mean eGFI at baseline (66.7 vs 74.5 in groups with >1000 patients in each). Patients with an eGFR90 result look like they are more likely to have diabetes. I suspect that missingness is not random. Some further exploration of this is required, as otherwise the analysis is built on a biased population - or at least a population that generalises to 'those in whom clinicians are worried about renal function 90 days after cardiac surgery'.
We have added a sensitivity analysis based on propensity scores to provide additional protection against any confounding by indication that may have resulted from differences between the group that had eGFR estimated at 90 days and those that did not. Our propensity score adjusted analysis (now added to the results) yielded very similar results as our non-adjusted analysis in terms of the models’ ability to discriminate, suggesting that the differences in patients that did and did not have eGFR estimated at 90 days do not substantially confound our analysis, and supporting generalization of our results to cardiac surgery patients. The results of this propensity score analysis has been added to the clinical results section of the paper and in a new Figure 3.

4. P20 L50 - Choice of egfr90 decline of 20ml/min. I think a stronger justification needs to be presented for why absolute decline was chosen instead of relative decline in eGFR e.g. 20% reduction. For lower baseline eGFR, greater levels of creatinine increase are required to generate an absolute reduction in eGFR of 20. From table 1, the range of baseline eGFR runs from 47 to 85 and epidemiologically there is huge difference in mortality between patients who have an egfr of 27 and an egfr of 65.

In order to address concerns about our chosen cutoff, we have constructed propensity score-adjusted, eGFR90-dependent ROC curves for each model which demonstrate that the statistical mixture model approach we suggest maintains superiority over the traditional linear model approach over a broad range of potential eGFR90 cutoffs. The difference between these curves along with its confidence interval are also given in the new Figure 3.

Although ratios may be very useful quantities in clinical practice, they can cause a great number of problems that compromise statistical validity. For example, the ratio of two centered, normally-distributed quantities is distributed via the Cauchy distribution. The Cauchy distribution is known to have poor statistical properties due to the fact that the expectation, the value we attempt to model in linear regression models, does not exist for a Cauchy distribution. Additionally, ratios can induce spurious correlation which produces spurious partial regression coefficients in multiple regression analyses, and the results of multiple regression models based on ratios are not comparable across studies where the distribution of the normalizing variable is not the same. The included reference (Kronmal 1993) addresses this issue in some detail. If we were to incorporate a relative decline in eGFR into our mixture model, these issues would be exacerbated because mixture models are fit by the expectation-maximization algorithm and successful separation of the subpopulations is intimately related to efficient estimation of the partial regression coefficient estimates. A sentence has been added to emphasize this point along with the supporting reference.

“An absolute change in eGFR90 was chosen over a relative change because of the many statistical issues that can arise from the inclusion of ratios of random variables such as improper error distributions and spurious associations (Kronmal 1993).”
5. P22 L3. Predicting AKI. The absolute increase of 0.3mg/dl in creatinine is not a universally agreed definition of AKI, from an international perspective. The NHS England AKI algorithm uses % increase in creatinine to define different stages of AKI - these have different mortality risks and require different levels of intensity of management. Some justification about this choice of definition (and dichotomising this outcome) would be appropriate.

We agree with the reviewer that dichotomization is non-ideal with continuous outcome measures. However, we felt it was important to include a dichotomized analysis for the benefit of clinician readers familiar with ROC curves. There is no universally agreed upon definition for AKI to guide our choice of serum creatinine cutoff value. Our chosen cutoff, 0.3 mg/dL, is an AKI cutoff suggested by both the Acute Kidney Injury Network and the Kidney Disease Improving Global Outcomes criteria. Both these criteria are commonly utilized in clinical research in many parts of the world. Further, we chose 0.3 mg/dL because it is often a more sensitive AKI criterion than fold changes which require a 50% or 100% increase or more in serum creatinine to diagnose AKI, equating to 0.5 or 1 mg/dL for patients with normal baseline serum creatinine. Using the more sensitive cutoff provides us with more events (greater effective sample size) with which to estimate the ROC curve whose illustration is the primary reason for the inclusion of a cutoff. For the reader who recognizes the limitations and potential pitfalls of dichotomization or is not interested particularly in a 0.3 mg/dL AKI cutoff, we included the Spearman’s rank correlation, which is a continuous measure of discrimination.

The following sentence has been added to the methods to justify the 0.3mg/dL cutoff to the reader.

“For comparison we have included the ROC curves that would result for the models predicting whether 2-day postoperative serum creatinine change exceeded 0.3 mg/dL, a relatively sensitive cutoff for AKI suggested by the Acute Kidney Injury Network and the Kidney Disease Improving Global Outcomes guidelines (Bellomo, Ronco et al. 2004, Kidney Disease: Improving Global Outcomes (KDIGO) Acute Kidney Injury Work Group 2012).”

6. In addition, is this assessment of AKI from early post operative blood tests in all patients, or just those who have an egfr90 taken? Are these post op tests within the first 48 hours, 72 or even the first week? Clarification here would be helpful. Presumably there is a larger dataset with post op renal function results for the whole patient cohort, and looking at AKI in this group would avoid the confounding by indication criticism of restricting all analyses to those in whom there is a blood test at 90 days.
Additional information regarding the timing of the perioperative tests has been added to the description of the clinical example in the manuscript. These tests are performed on the vast majority of cardiac surgery patients for the duration of their hospital stay. We therefore have very complete 48h data after which completeness begins to decline. There are a few reasons we chose eGFR90 as opposed to a shorter-term metric. These reasons are as follows: 1) shorter-term creatinine levels will be more highly correlated with the 2-day serum creatinine measurements making the surrogate argument less compelling, 2) 90days had good completeness in our dataset relative to other longer-term follow up times, 3) it provides a great example of when this statistical technique may be very useful (i.e. in observational studies surrogates can be used to fit risk prediction models when the desired outcome has a degree of missingness).

We acknowledge that our data is limited by missingness, variable length of stay, and other issues related to electronic health record data. We hope you will agree that the propensity score adjustment, which was performed as part of the eGFR-dependent ROC curve (Fig 3) analysis adequately addresses any concern of potential confounding by indication.

Reviewer #2:

We would like to take this opportunity to thank reviewer 2 for their insightful comments. We have made multiple changes to the presentation of the manuscript and completely rewritten the simulation section to improve clarity based on her/his comments. We have also added a number of details to the manuscript to clarify the estimation of eGFR90. Below we have summarized our manuscript alterations to address each concern.

1. The authors do not seem to have access to the true outcome (T) but only another surrogate marker (eGFR90). I think that paper which aims to resolve methods for dealing with partial surrogates should start with an example in which the true outcome is known.

We would like to thank the reviewer for raising this point for clarification. There is no universally recognized true outcome for kidney injury as renal function decline after an acute kidney injury is multidimensional and therefore difficult to quantify. Although this point is ancillary to the statistical method, it is worthwhile to note that when we referred to the true outcome in the manuscript, we actually mean the target of the predictive model. We have altered the language in the manuscript to prevent further confusion on this point. Each occurrence of “true outcome” has been replaced with “target outcome” to emphasize that we are referring to the desired variable of the prediction problem, and that this variable does not need to hold a gold standard status. We chose eGFR90 as our target outcome because we believe that it is well within the scope of plausibility that clinical researchers would want to fit a predictive model for patients’ post-operative renal function based on acute postoperative kidney injury.
2. Another problem here is eGFR90 itself. It is not clear from the manuscript how GFR has been estimated. They seem to be skirting around the issue of GFR being estimated at all. eGFR90 (and once or twice just eGFR) is used in the text many times but in the list of abbreviations it is written as just GFR?, there is no mention of estimated GFR. The problem with this is, is that GFR is often estimated using serum creatinine which is the partial surrogate in the example. Even if GFR is not estimated but measured using something like inulin clearance then I am still not sure that this is as good as a clinical diagnosis of AKI

We appreciate this point being raised and the opportunity to clarify. We have updated the manuscript to make it clear that eGFR90 is estimated using the CKD-EPI creatinine equation which has also been added to the references (Levey, Stevens et al. 2009). Clinically, direct measurement of GFR with substances like inulin is rare, while estimation of GFR is the most common clinical method for monitoring kidney function. Therefore, we selected eGFR90 as our target outcome in our clinical example to emphasize the real-world utility of our statistical methods.

The partial surrogate under study in our work is maximum serum creatinine change in the first 2 postoperative days, whereas, eGFR90 is a derivation which utilizes the serum creatinine measurement of the patient 90 days after surgery. In the section detailing the result of the clinical analysis we compare the performance of the linear model, mixture model, and the surrogate itself (maximum two-day postoperative serum creatinine change from baseline) for prediction of eGFR90. The surrogate itself displayed the worst discrimination of all three metrics emphasizing that using a creatinine-based surrogate and a creatinine-based target outcome is not problematic. Clinically, CKD-EPI eGFR90 is a more meaningful target outcome than 90 day serum creatinine change because it is an estimate of kidney function adjusted for sex, age, and race-related baseline serum creatinine differences, which allows clinicians to more accurately compare renal function between patients.

3. Finally, the suggested change of 20 is not standard as far as I am aware and probably too stringent given the reference change value of eGFR using the MDRD equation is 15%. Sensitivity analyses would be required to assess the importance of setting change at this level.

To address concerns about our chosen cutoff of eGFR90 decrease of 20 mL/min/1.73 m2, we have constructed propensity score adjusted, eGFR-dependent ROC curves for each model which demonstrate that our statistical mixture model approach maintains superiority over the traditional linear model approach over a broad range of potential cutoffs. The difference in AUC for the range of potential cutoffs from 5 to 25 mL/min/1.73 m2 is displayed in Figure 3.
4. The simulation method as per the introduction difficult to follow. The analysis was also incomplete. I want to know when using this approach might go wrong (fitting a two-mixtures when there are three) and I would want to know about coverage and bias. This is what I would expect in a methods paper. I am also surprised to see that there are no tables showing detailed results of the simulations.

The entire simulations section has been updated to improve clarity and presentation. With respect to the extent of the analysis, our work focuses on prediction of the target outcome (i.e. producing a model that results in good discrimination and is ‘clinically useful’). We are not trying to draw inference about model coefficients. Coverage is a result which primarily speaks to inference and is beyond the scope of the rest of the manuscript. Both bias and variance in the coefficients can affect a predictive model’s calibration and discrimination, therefore we include mean square error as a metric in our simulation studies. The second simulation demonstrates that despite having poor MSE, the linear model actually suffers only a moderate reduction in discrimination. Ultimately, predictive models fit using surrogate measures cannot be expected to be well calibrated as the relationship between the surrogate and the target outcome will generally be curvilinear and not an identity. Therefore, we feel that any bias/MSE result reported on the prediction would be misleading to the reader and overstate the method’s performance relative to what should be expected in real-world performance. As such, adequate discrimination is the best that can be hoped for in this scenario.

We have added a table summarizing the simulation results (Table 3).

5. In addition, it would be nice to have as an appendix, the code used to do the simulations as this is often much easier to follow than text.

The two data based simulations require the full dataset to be run and therefore the simulation code would not be useful for the reader. We have included code for two examples in an appendix that uses entirely simulated data that the reader can run on their machine if desired. The code for the first model will allow the reader to modify the relationship between the covariates and surrogate measure and visualize the consequence on the model’s prediction. The code for the second model simulates a situation where the models in the two subpopulations are distinct but too similar for the mixture model to resolve, which is an example of one case where the mixture model can fail and collapses back to something resembling the linear model. We have also added additional discussion of and references to various sources discussing the selection of the number of components and mixture model failures in Section 3 of the manuscript.

“In order to apply the mixture modeling approach to a clinical problem, it is necessary to decide how many components the model should have. In the case of partial surrogates this is the number
of subpopulations that are present. In many cases, the number of subpopulations may be strongly suspected based on clinical considerations, but in cases where the number is less certain there is a large literature that describes various methods for identifying the proper number of components and the consequences of selecting the wrong number (Henson, Reise et al. 2007, Nylund, Asparouhov et al. 2007, Tofghi and Enders 2007, Grimm, Mazza et al. 2017).

6. Finally, the title of the paper suggest that this will be a methods paper but the abstract reads as if the paper is setting out to apply this method to a particular clinical setting (AKI). I wonder whether this has been submitted previously to a clinical journal? I would suggest that if this were to be accepted then this abstract should be amended.

We have made substantial changes to the abstract to improve clarity.

Full Citations


