Author’s response to reviews

Title: Clinical Risk Prediction with Random Forests for Survival, Longitudinal, and Multivariate (RF-SLAM) Data Analysis

Authors:
Shannon Wongvibulsin (swongvi1@jhmi.edu)
Katherine Wu (kwu@jhmi.edu)
Scott Zeger (sz@jhu.edu)

Version: 2 Date: 07 Oct 2019

Author’s response to reviews:

September 18th, 2019

Dear Dr. Maria Hodges and the BMC Medical Research Methodology Editorial Board:

Thank you for the review of our manuscript and for the opportunity to revise our submission. We appreciate the thoughtful review comments and have submitted our revisions for Clinical Risk Prediction with Random Forests for Survival, Longitudinal, and Multivariate (RF-SLAM) Data Analysis. The thorough feedback from the reviewers has allowed us to improve the clarity of the presentation of the material in the manuscript. We have included the specifics of our revisions in the red text in response to the review comments detailed at the end of this letter.

We believe that this paper will be a timely and informative article for BMC Medical Research Methodology readers. We confirm that the authors approve of the manuscript and agree to its submission to BMC Medical Research Methodology. The material is original, has not been previously published in another peer-reviewed journal, and has not been submitted for publication elsewhere.

Thank you for reviewing our revised manuscript on Clinical Risk Prediction with Random Forests for Survival, Longitudinal, and Multivariate (RF-SLAM) Data Analysis for publication in BMC Medical Research Methodology.

Sincerely,

Shannon Wongvibulsin, PhD
MD Candidate, Class of 2021
Reviewer reports:
Eugene Blackstone (Reviewer 1): I found your article very exciting! As you can imagine, the work on randomForestSRC carried out for a number of years at Cleveland Clinic and subsequently at the University of Miami by Dr. Ishwaran and his graduate students continues to evolve, starting as a proportional hazards approach and then morphing into its present non-proportional hazards version. Like you, the realities of real life data are far more complex than "baseline and outcome."

Your piecewise approach, as is necessary for analysis of time-varying covariates and repeated events (including weighted events) is ingenious (I think first suggested in Kalbfleisch and Prentice). I predict it will better manage time-varying hazards after procedures, which will of necessity require more closely spaced CPIUs. Indeed, I might argue that the length of CPIUs should in some sense mimic the underlying hazard function (we have previously compared, for example a piecewise Cox Model with a temporal decomposition non-proportional hazard model whose pieces were selected based on the shape of the underlying hazard [Myers WO, Blackstone EH, Davis K, Foster ED, Kaiser GC. CASS Registry long term surgical survival. Coronary Artery Surgery Study. J Am Coll Cardiol. 1999;33:488-98]. As you have done, the temporal decomposition model is proportional within a given phase of hazard.

I do know that like you, Dr. Ishwaran has been working on a hazard-based approach within random forests as proposed in a funded NIH grant. It will be interesting to watch how these methods might converge or diverge, but for sure, both share the same motivations. Thus I congratulate you on this remarkable work.

Thank you for your review of our manuscript. We appreciate your positive comments and your insights about the future directions. We will keep you and Dr. Ishwaran updated on future progress of this work.

Yuedong Wang (Reviewer 2): The manuscript proposes a new statistical method for clinical risk prediction using survival, longitudinal, and multivariate data. The research is timely, the approach is cutting edge, and the application is impressive. The paper is well-written. I think the paper can be accepted with minor revision.

Thank you for your review of our manuscript and your detailed comments to help improve the clarity of our equations and explanations.

Detailed Comments:
1. Some notations in equation (4) is only defined in the Supplemental Materials.

We added the out-of-bag notation to the main text for equation (4) as suggested.

The following comments are pertinent to the Supplemental Materials.
2. Page 1, line 8, subscripts of d and Y are inconsistent with the rest.

We have modified the subscripts for consistency. Also, to improve the clarity of the notation, we have replaced the Y notation with E to represent the number of individuals at risk.

3. \(x_i\) in equation 3 should be italic.
We have now italicized $x_i$.

4. Page 2, line 10, the subscript of $t$ should depend on $i$. Otherwise $t_{ij} = t_j - t_{j-1}$ does not depend on $i$.

We have added the $i$ subscript to $t$.

5. Page 2, it is stated that "Within each CPIU for the same subject, the covariates are constant but between CPIUs for the same subject, the covariate values can vary for time-dependent covariates." It is possible to have none or multiple observations of a covariate in the same interval. Please discuss how to deal with these situations.

Thank you for suggesting that we include additional details regarding these situations. We have added the following: “It is also possible to have no or multiple observations of a covariate in the same interval. In the case of no observations for the covariate, the covariate value is set to missing. In the case of multiple observations in the same interval, a shorter CPIU length can be selected so there is only one observation per interval. Alternatively, a summary value of the multiple observations (e.g. the mean value) can be used.”

6. Page 2, line 16, $z_{ij}$ is used to denote the covariates while the symbol $x$ was used in the main text.

We have modified the $z$ to $x$ for consistency.

7. Page 2, the statement "$Y_{ij}$ is the event indicator for subject $i$ during time interval $j$ that is 1 if subject $i$ experiences an event during time interval $j$ and is 0 otherwise" seems incorrect. I guess $Y_{ij}$ represents the number of events.

We have modified the statement to the following: “The number of events for individual $i$ in interval $j$ is denoted by $y_{ij}$.”

8. J$_i$ should be j$_i$ in equations (7) and (8).

We have modified the notation for the total number of CPIUs for subject $i$ to be J$_i$ for consistency with equations (7) and (8).

9. The maximum j$_i$, J, is used in equation (9). However, $y_{ij}$ and $\mu_{ij}$ are not defined for all J intervals.

We edited $J$ to J$_i$ in equation (9).

10. It is stated that equation (11) is an "estimate of the event rate for individual $i$ in time interval $j$". It seems to be an estimate of $\lambda_j$ assuming the same rate for all individuals in interval $j$. This estimate may be removed to avoid confusion since it is not used in the following arguments.
We agree there was potential for confusion. We have removed the equation as it was not essential to the presentation.

11. Define $t^L$ and $t^R$ in equation (12).

We have replaced the $t^L$ and $t^R$ with notation to define the sets of individuals at risk in each of the nodes in interval j for clarity.

12. Provide some details about the calculation of the best split.

We have added details about selection of the best split at the end of the section on the RF-SLAM Splitting Criteria.

13. Some notations in Figure 1 are inconsistent. For example, "pHF" and "iHF" are used in the figure and "phf" and "ihf" are used in the caption.

Thank you for making a note of this inconsistency. We have modified the caption accordingly for consistency.