Author’s response to reviews

Title: Development of a Risk Score to Identify Patients with Type 2 Diabetes Mellitus and Multivessel Coronary Artery Disease Who Can Defer Bypass Surgery

Authors:
Andrew Perry (perrya@wustl.edu)
Matthew Chung (matthew.chung@duke.edu)
Eric Novak (enoavak@wustl.edu)
Ronald Krone (rkrone@wustl.edu)
David Brown (d.brown@wustl.edu)

Version: 1 Date: 14 Nov 2018

Author’s response to reviews:

November 14, 2018

Editorial Office
Diagnostic and Prognostic Research

Dear Editor,

Thank you for considering our manuscript "Development of a Risk Score to Identify Patients with Type 2 Diabetes Mellitus and Multivessel Coronary Artery Disease Who Can Defer Bypass Surgery" (DAPR-D-18-00021) for publication in Diagnostic and Prognostic Research. We have reviewed the comments made by the reviewers and have revised our manuscript accordingly. Please see our responses to the reviewers’ comments below.
Responses to Reviewer #1:

1. The paper could benefit from a discussion translating the risk score to NNT or other decision analytic measures and the use of the median as the only cut point seems a bit arbitrary. I would suggest this series of papers from a group in Utrecht as a starting point for a variety of potentially useful options:

https://www.ahajournals.org/doi/10.1161/CIRCULATIONAHA.112.000712

https://www.bmj.com/content/352/bmj.i1548/

We appreciate the reviewer’s comments but are influenced by the comments of the statistician Frank Harrell regarding the lack of value of NNT in estimating an individual patient’s benefit from an intervention. His issues are summarized below and can be found at https://discourse.datamethods.org/t/problems-with-nnt/195

Loses Frame of Reference: Suppose that the quantity of interest is the estimated risk of a stroke by 5y, and the estimated risks under treatments A and B are 0.02 and 0.01. Then ARR=0.01 and NNT=100. But if the two risks are 0.93 and 0.92 the ARR and NNT remain unchanged. Yet the patient will likely perceive the two settings to be quite different for the purpose of their decision making.

Computing NNT from Group Data and Using it on an Individual: The vast majority of NNTs are computed using group averages. For example, a physician may read a clinical trial report that provides the overall unadjusted cumulative incidence of stroke at 5y to be 0.25 and 0.20 for treatments A, B and compute NNT=1/(0.25 - 0.2) = 20. What is the interpretation of 20? It really has no interpretation. That is because RCTs with even the most restrictive inclusion criteria have a variety of subjects with much differing risks of the outcome. ARR varies strongly with baseline risk as discussed in more detail here 28. Thus ARR is subject-specific and so is NNT. NNT is impossible to be a constant, and if it is computed as if it were a constant, it may not apply to any real subject.

NNT Invites Physicians to Make Group Decisions: NNT as almost always computed on groups and thus apply to some sort of ‘average person’ who may not exist. An NNT of 50 is interpreted by many physicians as “I need to treat 50 patients to save one”. Who are the 50? Who is the one? Decision making in medicine is one patient at a time, and needs to use that patient’s absolute risk or life expectancy estimates.

NNT Has Great Uncertainty: How often have you seen a confidence interval for NNT? Probably not very often. And when you compute them they are often so wide as to render the NNT point estimate meaningless, even if the problems listed above didn’t bother you. For example, an ARR that comes from the two proportions 1/100 and 2/100 is 0.01 with 0.95 confidence interval of about [0, 0.044]. Confidence limits for NNT are the reciprocals of these two numbers or [23, ∞].
2. The terms OMT and IMT are both used in the paper to refer to the same treatment group. One should be selected and used throughout.

Our apologies for the interchange of terms. We have selected the term IMT and used it consistently throughout.

Responses to Reviewer #2:

Overall, the main problem I had was on the reporting of the methods that the authors used. The reporting was rather minimal, making it at times very hard to understand what exactly they did. However, based on what I could infer, they seem to have used valid methods. Therefore, my comments are mostly about clarifying these methods, and most of them are of minor importance.

We apologize for not making the methods clearer. As there were many specific questions raised by Reviewer #2, we will address each of these individually. We have reviewed again our methods description and believe the description gives adequate detail for reproduction of our findings.

1. Page 5, line 53. What is "each potential predictor"? How were these selected? Was there some clinical consideration behind the selection? In essence what the authors do here is multiple tests, which of course runs the risks associate with multiple testing. Later in this sentence, it writes "the hazard ratio (HR) …". They should clarify what the HR refers at, i.e. the effect of each covariate. Otherwise it is a bit ambiguous.

All of the “potential predictors” are baseline characteristics contained within the Biolincc dataset. These were selected based on a univariate model results (P value < 0.10) and clinical relevance. The hazard ratio for each variable is the risk in predicting the composite outcome. The relevant sections of the manuscript have been revised to make this more clear.

2. Page 6. What is a "sequential imputation algorithm"? Please provide a short description, as well as a citation of a paper that describes the actual algorithm. What is a "discriminant function"?

3. Page 6. What is a "regression predictive mean algorithm"? Please provide a description and a citation.
Our apologies for not making the imputation algorithms clearer in the original manuscript. The imputation algorithm was from SAS using the Multiple Imputation procedure. For quantitative and continuous traits, we used a regression predictive mean algorithm which imputes an observed value closest to the predicted value from the simulated regression model for each missing value. For classification variables, the discriminant function was used to impute missing variables. The following references further describe these techniques.


4. Page 6: "The performance of the risk score… " Although it becomes later evident what the risk score is, at this instance it is confusing. Please write here what this is and how it is calculated.

We apologize for the confusion. We have clarified the purpose of the risk score in this section.

5. Page 6: Please describe shortly the jackknife cross-validation method, and give a citation.

We have added the following description of the jackknife method to the section.

Under this method a subject is removed from the sample and the model is developed on the remaining sample. The prediction of the model is then tested on the removed subject. This is repeated so that all subjects serve once to test model performance. (Efron, B and Gong, G. A Leisurely Look at the Bootstrap, the Jackknife, and Cross-Validation. Am Stat 1983;37:36–48.)
6. Page 6, it writes "The predictors included in the multivariable Cox proportional hazards model were identified based on clinical relevance and univariate model results". This is very unclear. Based on the univariate model but in what way? How was a decision made?

We apologize for the confusion over how the predictors were selected for the risk score. Predictors were chosen based on univariate model results and were considered for inclusion if the univariate model P value was <0.10. Upon review of the variables, we added history of stroke or transient ischemic attack (TIA) and hemoglobin A1c to the score because they have particular clinical relevance to patients with diabetes and cardiac surgery risk.

7. Although I know what a jack-knife cross-validation is, I am not really sure what the authors mean by "jack-knife sample". Maybe they can clarify.

We apologize for the typographical error, there is no jack-knife sample, just an IMT sample and CABG sample.

8. Page 7. The authors say that based on their Cox model results, they developed a point scoring system, and they cite Sullivan et al. I found this description quite small, I would like to see at least the outline of how this score is developed.

The following description of how the point score was developed has been added to the manuscript:

A point scoring system was developed from the model to help facilitate ease of use, based on the methods of Sullivan et al. This method estimates the predicted risk from the Cox model by assigning integer points to each level of risk factor. Levels are designed to reflect clinically relevant states of the risk factor. For example, we chose three levels of risk for HbA1c: less than 7%, 7 to 9% and greater than 9%. The risk estimate is then obtained by comparing the sum of points to a reference table generated by the Cox model.

9. Authors should mention which software and which package/command they use for each part of their analysis, so that other researchers can replicate it, or do something similar in another dataset.

We have reviewed our methods section and added specific packages where appropriate.
10. Page 8. It writes that "...and yielded a calibration slope of 0.76 (see Figure 1a)." The way it is written it seems as if Figure 1a includes a calibration slope, but it actually does not. Including a calibration curve might be also useful and interesting.

We have deleted the text to which the reviewer refers.

11. Page 8, it writes "...patients in the IMT group had significantly reduced event-free survival (Figure 2b)." What is the HR between the groups?

We have changed the text to specifically refer to event-free survival as determined by comparing the survival curves.

Reviewer #2 minor comments:

1. Page 3, line 58, it writes "Within the CABG strata, however, a reduction in the composite outcome of death myocardial infarction (MI) and stroke was observed in the patients who underwent prompt CABG compared with those managed with IMT alone that was driven by a reduction in non-fatal MI in the CABG group." The last phrase, "that was driven by..." confused me. A reduction compared to what? CABG is the whole strata. Maybe the authors meant "in the prompt CABG group"?

The sentence has been re-worded to clarify that the improved outcomes of prompt CABG group reported in the original BARI 2D are primarily driven by reduction in non-fatal myocardial infarction.

2. Page 4, line 21, it writes that "The aim of this post-hoc analysis was to create a multivariable risk score to identify patients in the CABG strata at the time of randomization of BARI 2D for whom surgery may be safely deferred." I got confused initially, I thought this was referring to a previous analysis described before. Maybe the authors could rephrase to something like "The aim of this paper/this analysis", to make clear they are referring to this paper.

We have changed the sentence as recommended by the reviewer.

3. Page 5, line 51, and in many other places. What is "OMT"? Is this the same as "IMT". This abbreviation has not be introduced
We have used OMT and IMT interchangeably. We have selected IMT as the preferred abbreviation and eliminated all usage of OMT.

4. Page 8, line 46, it writes "…validation cohort for the point". What does this mean?

This sentence should read as follows: “Subjects randomized to prompt CABG served as the external validation cohort for the point score."

5. Table 2, first row of the table: Please correct the typo for the CI for age (10.17, 1.066)

This error has been corrected (1.017, 1.066).

Thank you for the opportunity to revise our manuscript in response to that thoughtful and constructive critique. We feel the changes have significantly improved the paper. If I can be of any additional assistance, please do not hesitate to contact me.

Sincerely,

David L. Brown, MD
Professor of Medicine
Cardiovascular Division
Washington University School of Medicine
660 S. Euclid Ave, Campus Box 8086
St. Louis, MO 63110
Phone: 314-362-1291
Fax: 314-996-3268
Email: d.brown@wustl.edu