Author’s response to reviews

Title: Tufts PACE Clinical Predictive Model Registry: Update 1990 through 2015

Authors:

Benjamin Wessler (bwessler@tuftsmedicalcenter.org)
Jessica Paulus (jpaulus@tuftsmedicalcenter.org)
Christine Lundquist (clundquist@tuftsmedicalcenter.org)
Muhammad Ajlan (ajlan.muhammad@gmail.com)
Zuhair Natto (z_world@hotmail.com)
William Janes (wjanes@tuftsmedicalcenter.org)
Nitin Jethmalani (Nitin.Jethmalani@tufts.edu)
Gowri Raman (graman@tuftsmedicalcenter.org)
Jennifer Lutz (jlutz@tuftsmedicalcenter.org)
David Kent (dkent1@tuftsmedicalcenter.org)

Version: 2 Date: 2017-11-20

Author’s response to reviews:

CPM Update Manuscript
Response to Reviewer Comments

November 13, 2017

Response to Reviewers: The reviewer comments are noted followed by our responses. We reference the location where edits have been made to the original text and the new text appears in quotes.
Reviewer #2:

I thank that authors for their response to my revisions, which I think show that there is much potential for the manuscript. I think there are lot of interesting details to be unveiled from this registry, and hope that the further work planned by the authors can offer us much greater insight into the current state of the CPM field (at least in CVD outcomes). I think there are a few areas where further questions have been raised by the responses and have questioned these below.

RESPONSE: Thank you for this comment. We look forward to learning significantly more about this field through our Registry.

Minor comments

1. I think perhaps that my first original comment has been slightly misinterpreted. I appreciate the adjustment to the text, but the proportion given for those reporting calibration performance has remained the same, at 39%. This is still the proportion of all CPMs in the registry which reported calibration performance (e.g. 425/1083). But the real proportion of interest is excluding those articles which only performed model development. The question of interest is really - what proportion of studies which should report calibration have reported calibration? So the figure reported here should be the number of studies reporting internal/external validations that report calibration performance. I expect that this number should be higher than 39%, and it will be interesting to see if it is. The current statistic given in the paper, is like saying 'how many men take up paternity leave', as a proportion of all men, not just new fathers.

RESPONSE: Thank you for this comment. We apologize for misinterpreting your original comment and agree that the more interesting measure of reporting is as you suggest. We have changed the text with the Model Performance paragraph (page 8) to the following: “Of the CPMs included in this Registry reporting some type of validation (n=577), only 333 (57%) report some measure of model calibration. For the reports presenting an external validation, only 50% report calibration.”
We have also changed the Results of the Abstract to read: “Of the CPMs reporting validations, only 333 (57%) report some measure of model calibration. Reporting of discrimination but not calibration is improving over time (p for trend <0.0001 and 0.39 respectively).”

The Conclusion of the Abstract to read: “While the number of CPMs continues to increase, model performance is often inadequately reported and calibration is infrequently assessed.”

The Conclusion of the manuscript (page 12): “Model performance is often inadequately reported, though discrimination (but not calibration) reporting appears to be improving over time.”

2. Corresponding to the above, I suggest making sure the percentages in table 2 for calibration represent the appropriate proportions. And then please check the trend again, because as above the current percentages I believe correspond to the proportion of calibration reporting for the whole set of CPMs not just those that include some validation (internal/external).

RESPONSE: Thank you again. We agree. We have adjusted these trends to reflect the more relevant proportions and have added the text to the Time Trends paragraph (page 9). There is no change over time in the frequency of reporting CPM calibration (p for trend =0.39).

3. I appreciate your response to my 3rd point, it is good to see that statements on the reporting of particular statistics are possible. However I think this has been slightly misinterpreted, I should have been clearer in that the Hosmer-Lemeshow statistic is not something recommend to be reported (as it does not quantify the extent of calibration), more relevant calibration statistics include the E/O ratio, calibration slope and calibration-in-the-large. If the 56% of new models reporting the HL statistic is the only measure of calibration reported in these studies then the reporting quality is potentially still substandard. Perhaps rephrase the current statement to, "Of the CPMs reporting calibration that were published after May 2012, 93 (56%) only reported a Hosmer-Lemeshow statistic for calibration." If this is true of course.

RESPONSE: Thank you for this comment. We agree with your comments about the limitations of the HL statistic. We have ongoing work focused on assessing the measures of calibration that are reported. Unfortunately more work is needed before we can make a full assessment of how these measures are reported across the database. If our manuscript is acceptable for publication
we are certainly happy to work with the Editors to adjust or eliminate the comment regarding HL statistic. We do not advocate its use given the identified limitations though report it for readers interest.

4. Thank you for your response to point 9 in my revisions. I now understand the discrepancy in numbers of logistic and survival models, poor reporting of the baseline survival or hazard is a major issue in the reporting of Cox models. This is perhaps something worth highlighting as part of the manuscript as it would encourage others to report this information in the future. Indeed it could be valuable to include a table of reasons for exclusion, as this can inform new model development studies what is needed for good reporting (reinforcing the TRIPOD guidelines).

RESPONSE: Thank you again. This comment is incredibly helpful. We agree and have added the following sentence to the Discussion (page 11): “In our systematic review, many more Cox models than logistic models were excluded from the Registry since, in accordance with our original inclusion criteria; many Cox models did not describe the baseline hazard or provide an alternative way to generate individual patient predictions and thus were excluded.”

While more work is needed to understand the granular reasons for screen failure, we have added the following sentence to address this reporting concern in the Discussion (page 11): “Since 2012, 46% of full text articles that are screened for inclusion are excluded from our Registry because they do report a usable model.”