Reviewer’s report

Title: Untapped potential of multicenter studies: a review of cardiovascular risk prediction models revealed inappropriate analyses and wide variation in reporting.

Version: 0 Date: 09 Oct 2018

Reviewer: Long Nguyen

Reviewer's report:

Dear Editor, Dear Authors,

Manuscript ID: DAPR-D-18-00022,

Title: 'Multicenter studies: A systematic review of multicenter risk prediction models shows wide variation in reporting and inappropriate analyses',

I am grateful to both the Editor and the Authors for giving me the opportunity to review this interesting article.

Overall, I enjoyed reading the manuscript, which addresses an important and under-considered topic in clinical prediction modelling. It displays a short systematic review of multicentre clinical prediction models published in cardio-vascular diseases, which highlights some flaws in both statistical analysis and reporting; in addition, two short simulations are provided to illustrate common issues found in the literature. Though I suggest some revisions (in particular, regarding the illustrative cases), I think this article would be of interest to the readers of Diagnostic and Prognostic Research.

Please find below my review, hoping it will help improve this interesting work.

Yours sincerely,

T.L. Nguyen, Pharm.D. Ph.D.
*** Major comments:

Though I recognize the primary goal of this review is not to conduct comprehensive simulations, I would recommend to address some concerns about the illustrations.

Illustration 1:

- I find the description of the simulation methods quite confusing: There seems to be some difference between what is written in Methods and in Appendix. For instance, P.6 line 119, Methods: 'We generated data for 20 centers, with 1000 observations each [...]'; Appendix, Illustration 1: 'We generated a dataset of 500000 patients. [...] Random effects (ri) of 500 centers (1000 patients each) were generated [...]'. What did the authors really do? Also, what is the event rate: 0.35 (Methods) or 0.33 (Appendix)?

- I understand the simulation study is not meant to be the core of the article, but further clarification would be much appreciated (I find the explanation very brief, even in Appendix).

- The only scenario that is provided relies on: \( P(Y) = \expit(-1+ri+0.8X) \), where \( ri \) follows a distribution \( N(0,1) \) that is cluster-specific, and \( X \) follows a distribution \( N(0,1) \) that is not cluster-specific. From this, the authors show that mixed effects logistic regression outperforms standard logistic regression. Though I agree that this might be true - as shown in cited references - I find the results displayed by the authors quite 'tricky'. Indeed, following the equation, \( P(Y) \) highly depends on the intercept (i.e. \(-1+ri\), which thus follows \( N(-1,1)\)); in comparison, the coefficient 0.8 defines but a small effect of the predictor \( X \) on the outcome probability (i.e. 0.8\( X \) follows \( N(0,0.64) \)). Thus, it comes to no surprise that the standard logistic regression has some under-fitting issues: \( P(Y) \) depends much more on the random cluster-level intercept than on the predictor. Given the simulation model, it follows that some centres have a very high outcome prevalence (>0.9), whilst others have a very low one (<0.1), as depicted in lower panels of Figure 5. This seems, from a clinical and physiopathological rationale, quite 'extreme' and unrealistic to me: I would hardly imagine in clinical practice a disease which causes outcomes in less than 10% of patients in some settings, whilst causing them in more than 90% of patients in other facilities. I would suggest the authors define more realistic a scenario, in which the outcome risk depends more on patient characteristics (i.e. \( X \)) than on the random intercept.

- I am not sure that the 'true' probability (i.e. the theoretical value of \( P(Y) \), if I understand it well) should be plotted against the predicted probability in the calibration plots (Figure 5). Instead, I would plot the observed probability against the predicted probability. (Over the 500 000 observations [or 20 000?], this might result in no big changes; however, this
pedantic suggestion is to highlight that, in clinical practice, one is interested in the actual rather than in the theoretical probability.)

Illustration 2:

- Again, I find the description of this real data-based simulation quite brief.

- I am not sure to understand why the authors had to define 'true' outcome models, instead of merely sampling outcomes? This, I assume, would be more realistic.

- After stating earlier than mixed effects logistic regression outperforms standard logistic regression, it would be informative to see whether there is any difference in terms of performance between the two approaches in this realistic illustrative case. In fact, I would conjecture no substantial advantage of a mixed effects model, as the residual variance attributable to differences between centres is relatively small (as opposed to the first illustration scenario). But first and foremost, doing so - I suppose - would not address the issue of transportability posed by the differences across tertiary and secondary clinical settings.

In this regard, I would suggest clarifying the two separate issues highlighted by the illustrations:

1) When dealing with highly between-centre-heterogeneous data, mixed effect regression might outperform standard regression (more realistic illustration to be provided here);

2) When applying prediction models to clinical settings that differ from those used for model development (e.g. application of prediction models, previously derived in tertiary care, to secondary care), transportability might not hold - and this might be true for both standard regression and mixed effects regression.

As the manuscript currently stands, these two issues follow one another. This creates, at the first reading, the confusing idea that mixed effects regression would result in greater transportability. These two issues are, in fact, to be more clearly separated throughout the manuscript for they relate to two different concerns about generalizability of clinical prediction models: model fitting (statistical problem) and transportability (clinical problem). (About this distinction, see, for instance, Debray et al. 2015, J Clin Epidemiology; 68:279-289.)
*** Minor comments:

- Introduction: I think the manuscript would benefit from earlier clarification and more references about the issues posed by inappropriate statistical analysis (as displayed in Box 1), and also transportability across clinical settings.

- P.6, line 136 (Results): ‘After 2000, 65% of published models used multicenter data. In total, 52% (390/747) of all studies included in the Tufts PACE Clinical Prediction Model registry were multicenter studies published after 2000 (see flowchart in Figure 2).’ Where do the 65% come from (as compared with the 52%)?

- For the sake of readability and homogeneity, could the authors report absolute frequencies (with proportions) and medians (with interquartile ranges) for all results? I find it confusing to have some partly missing information (e.g. p.6 line 136: ‘After 2000, 65% [absolute frequency not reported] of published models […]’; p.7 line 153: ‘[…] with a median of 2 263 [IQR not reported]’).

- P.16 line 109 (Box 2): ‘The prediction model was less useful in secondary care […]’ I would suggest the authors rephrase this passage. Indeed, one cannot infer on usefulness from a calibration curve, as this depends also on the clinical usage. For instance, if a prediction model is used to prioritise care (e.g. to sort patients by decreasing risk, then offer care to the N patients with highest risk), then I presume a mis-calibrated model would not necessarily lose clinical usefulness as long as the calibration function remains monotonous (i.e. as long as the risk order holds properly).

- Figure 2: Could the authors provide further information on the first exclusion step (5 348 excluded records)?

*** Cosmetic comments:

- P.13 line 27: The text of sub-section ‘Clustered data is commonly ignored during analysis' is completely redundant with regard to Figure 4. (One of these two could be removed to save space.)

- Overall, the manuscript would benefit from proof-reading to avoid some typos and grammar rule violation (e.g. no numeral at the beginning of a sentence).
Level of interest
Please indicate how interesting you found the manuscript:

An article whose findings are important to those with closely related research interests

Quality of written English
Please indicate the quality of language in the manuscript:

Needs some language corrections before being published

Declaration of competing interests
Please complete a declaration of competing interests, considering the following questions:

1. Have you in the past five years received reimbursements, fees, funding, or salary from an organisation that may in any way gain or lose financially from the publication of this manuscript, either now or in the future?

2. Do you hold any stocks or shares in an organisation that may in any way gain or lose financially from the publication of this manuscript, either now or in the future?

3. Do you hold or are you currently applying for any patents relating to the content of the manuscript?

4. Have you received reimbursements, fees, funding, or salary from an organization that holds or has applied for patents relating to the content of the manuscript?

5. Do you have any other financial competing interests?

6. Do you have any non-financial competing interests in relation to this paper?

If you can answer no to all of the above, write 'I declare that I have no competing interests' below. If your reply is yes to any, please give details below.

I declare that I have no competing interests.

I agree to the open peer review policy of the journal. I understand that my name will be included on my report to the authors and, if the manuscript is accepted for publication, my named report including any attachments I upload will be posted on the website along with the authors' responses. I agree for my report to be made available under an Open Access Creative Commons CC-BY license (http://creativecommons.org/licenses/by/4.0/). I understand that any comments which I do not wish to be included in my named report can be included as confidential comments to the editors, which will not be published.

I agree to the open peer review policy of the journal