Author’s response to reviews

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

Authors:

Laure Wynants (laure.wynants@kuleuven.be)
D. Kent (dkent1@tuftsmedicalcenter.org)
D. Timmerman (dirk.timmerman@uzleuven.be)
C. Lundquist (CLundquist@tuftsmedicalcenter.org)
B. Van Calster (ben.vancalster@kuleuven.be)

Version: 1 Date: 11 Dec 2018

Author’s response to reviews:

Dear Dr. Debray,

Dear reviewers,

We wish to thank the editor and reviewers for their time invested in reading our manuscript, their useful suggestions, and the chance to revise and resubmit our work. We have replied to each of the reviewers' comments below, and attach a version of the manuscript with tracked changes.

Reviewer #1:

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.

REPLY: We thank the reviewer for the comments and constructive feedback. We have addressed each comment point by point, below.

*** 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)?

REPLY: Thank you for noticing us of these errors. Initially, we performed the simulation with only 20 centers, for convenience in plotting the center-specific curves. Later, we ran the simulation with 500 centers to better illustrate the asymptotic properties of the method, and plotted only a random selection of center-specific curves. The event rate changed slightly, from 0.35 to 0.33 We forgot to update the methods section. We apologize and have corrected this.
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).

REPLY: The explanation is very brief, but the simulation is equally brief and simple. It is not meant as a simulation study, but merely an illustration of some important concepts on simulated data. The description follows the simulation code, nearly line by line. To further improve transparency of what was done, we provided the R script 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. \( 0.8X \) 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.

REPLY: We fully acknowledge that the chosen example is extreme. It was chosen for didactical reasons. We intended to illustrate a theoretical concept (marginal regression coefficients are closer to zero, in theory resulting in underfitting in an average center), and the variance of the random intercept was set this large to visualize the effect on calibration. The published study on the topic used more realistic simulation scenarios (and real clinical data), but is rather technical. We wanted to take this opportunity to explain the effect on calibration with an easy, visual example.

Following to the reviewer’s justified comment, we now also provide a realistic illustration in appendix, in which the random intercept follows \( N(-1,0.5) \). The same effects can be seen, but the effect on under-fitting is of course much smaller. In our experience, this degree of between-
center differences is what may be found in a study using data from various similar centers (e.g., all tertiary referral centers).

- 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.)

REPLY: The reviewer is correct to point out that in applied prediction research, calibration plots have the observed probability on the vertical axis. The goal of calibration is to assess how close predicted probabilities are to the true probability. In empirical studies, we do not know the true probability, we can only estimate it (using some flexible smoother or grouping observations according to e.g. deciles of predicted risk), i.e., observed probability.

Because we know the true underlying probability in our simulation, we do not need to estimate it from the data. It is a subtle difference without implications for the conclusions. More extensive published simulation studies have used a large hold-out sample approach and estimated the observed probabilities in the validation set, as implicitly suggested by the reviewer, and came to the same conclusion.

Illustration 2:

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

REPLY: We have expanded the description in the Appendix.

- 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.

REPLY: We used resampling to augment the dataset in each setting (secondary and tertiary). We used augmentation to reduce sampling variability to a minimum. We have regenerated the outcome according to the “true” model (estimated from the real data) in each setting to avoid having exact duplicates. We have elaborated on this in the Appendix.
- 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.

REPLY: We have simulated a second example for Box 1 with smaller differences between centers. This real database has been used with random effect models in previous publications (e.g. Wynants, Stat Meth Med Res 2018, Timmerman, Am J Obstet Gynecol 2016, Van Calster, BMJ 2014). The reviewer is right that mixed effect models do not solve the issue of non-transportability. We have clarified this at multiple places in the manuscript now.

We’d like to note that if the development database had been composed of centers of both types, a mixed effect model would have allowed to estimate an effect of center type (whereas a standard logistic regression model would yield misleading standard errors for a center type effect). This has been addressed in the discussion section.

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.)

REPLY: We completely agree with the reviewer that model fitting and model transportability are two separate issues and explicitly addressed this in box 2 and in the introduction.
*** 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.

REPLY: We have followed the reviewers suggestion and included these concepts and additional references in the introduction.

- 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%)?

REPLY: 65% is the proportion of models built on multicenter data since 2000. The number of models published after 2000 is the denominator. This is included to illustrate that prediction models based on multicenter data are abundant in the literature.

52% is the proportion of studies that are multicenter and published after 2000 (i.e., eligible for this review), with all studies in the registry as the denominator. This is the relevant percentage in the study flowchart. Sometimes, one study proposes multiple models.

We have clarified this in the manuscript.

- 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]’).

REPLY: We have added absolute numbers and interquartile ranges where they were missing.

- 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).
REPLY: We do not fully agree, as clinical usefulness depends on both discrimination (based on the risk order) and calibration (see e.g. Van Calster & Vickers 2015 Medical Decision Making on the impact of miscalibration on decision-analytic performance). However, as this issue is not crucial for the current paper, we have rephrased it in the manuscript.

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

REPLY: For the current paper, the Tufts Registry is our point of reference, i.e. the “sampling frame”. The inclusion and exclusion criteria that were applied for papers to be included in the Tufts registry have been detailed elsewhere (Wessler, 2017; Wessler, 2015). The 5348 excluded records were identified in a pubmed search but after abstract screening did not meet the inclusion criteria for the Tufts registry: (1) the primary aim was to develop a CPM as indicated in an objective statement; (2) the model predicts binary clinical endpoints; (3) the model contains at least two independent (i.e. predictor) variables, and (4) the model provides a way of calculating a probability for an individual patient. We have added this in the flowchart.

*** 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.)

REPLY: We have shortened the text of the sub-section to avoid repetition.

- 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).

REPLY: We have checked the manuscript again and corrected the errors we detected.

Reviewer #2:

REPLY: We are grateful for the constructive comments and address them pointwise below.

1. Title doesn't accurately reflect that this is a review of one specific area - suggest rewording to include 'cardio-vascular'
2. The study is described as a systematic review. However, I don't believe that this study satisfies the criteria of a systematic review as it is based on a random sample of papers from a registry and so does not provide a complete, exhaustive summary of current literature. The title and methods should be adapted to reflect this.

REPLY: This is indeed not a systematic review in that we do not provide a complete, exhaustive summary of current literature to answer a particular research question. We have followed the reviewer’s suggestion and removed the word “systematic” from title and text. The study is systematic in that we used systematic methods to collect studies for our review (instead of a selection of case studies), and systematically extracted data from the studies and synthesized the findings. We felt the PRISMA guidelines were appropriate for this reason.

3. Whilst the aims of the descriptive study of characteristics of the 50 predictive models were clear to me, the aims of the two additional illustrative examples were not entirely. Illustrative example 1 describes the difference between marginal and predictive probabilities and sets out to illustrate the consequences of ignoring clustering yet is just based on one 'simulated' dataset. To most usefully describe the differences between marginal and conditional probabilities then a range of scenarios and multiple simulated datasets should be examined.

REPLY: Large simulations have been carried out in previous published studies, which we cite in the manuscript. We included the simulated examples for didactical reasons, because the published papers on the topic are very technical. Rather than only pointing out what published studies did, we wanted to illustrate why we the common approach is not optimal. We have now also added a second simulation for example 1 (which is perhaps more realistic).

In example 2 the concept of transportability is examined but this concept hasn't really been described in detail. As in example 1, the second illustrative example focuses on just one example and so is potentially misleading in terms of describing the properties and concept of transportability.

REPLY: Thank you for pointing this out. The concept of transportability has now been described in the introduction. It is well recognized in the literature, including the TRIPOD guidelines, that models are not simply transportable from one clinical setting to another, and that they should undergo extensive external validation in multiple settings before being introduced to clinical
practice. We have added references to our introduction section to highlight this. The example has been included for didactical reasons.

4. The paper is almost exclusively focussed on prediction models in cardio-vascular disease and so it's unusual that the data in illustrative example 2 is then taken from the cancer field.

REPLY: Because validating models in another clinical setting than the model development setting is rare in cardiovascular research, we did not have access to a real database to illustrate the problem with data from this field. However, we did find a published example, and discuss it in box 2. We wanted to keep the real data illustration in for didactical reasons, even though the data come from another field.

5. Referral bias and spectrum bias are mentioned on page 16 but not described.

REPLY: Thank you for pointing this out. The concepts of referral and spectrum bias have now been described in the introduction.

Reviewer #3:

The manuscript by Wynants et al details a systematic review of clinical prediction models in the field of cardiology. The aim of the study was to determine how authors commonly deal with clustering by center in the development of prediction models in multi-center studies. The authors found that of the included studies published after the year 2000, 65% were multi-center studies. From 390 eligible studies, the authors randomly selected 50 for full text review and data extraction.

They showed that out of these 50 studies, clustering was ignored in 39 (78%). Moreover, sample size imbalances between centers were common. However, in only 22 of the 50 studies sample size imbalance was reported and 9 out of 50 studies reported sample sizes by center. The author go on to highlight the possible consequences of ignoring clustering and center level differences during the development of a clinical prediction model. Importantly generalizability may be lacking due to mis-calibration and non-transportability of the models.

Finally, the author make a series of recommendations on study design, model development, validation, and reporting.
I found this an interesting and education manuscript on an important but often overlooked topic. Standard textbooks (such a Steyerberg Clinical Prediction Models) make little mention of the impact of clustered data. Hence, the manuscript may be of interest to the readership of BMC-DAPR. The manuscript was generally well written in appropriate English, barring a few minor mistakes (e.g. the use of a contraction in line 189).

REPLY: Thank you for this positive feedback. We have corrected the contraction and reviewed the manuscript for spelling and grammar mistakes. We are grateful for the constructive comments and address them pointwise below.

I only have several minor comments/questions.

* How was the sample size of 50 arrived at?

REPLY: A random sample of 50 studies was a pragmatic choice to provide an idea of what study characteristics are common or typical. The review was initially done by the first author alone to inform realistic values for simulation settings in a planned study on prediction models in multicenter data. We later realized our information might be valuable to a wider audience because multicenter data was so common and issues with multicenter data proved to be overlooked so often. We have addressed the limited sample size in the discussion section.

* The models were selected from cardiology, where there is a relatively long standing tradition of prediction modeling and research. For the purpose of their discussion and the generalizability of their findings, have the authors considered the state-of-the-art in other fields. For example, by checking a few reviews?

REPLY: Thank you for the suggestion. We have checked published reviews in other fields and commented on them in the discussion section.

* The figures were vague. May be the dpi was set too low. Exporting from R in a vector format (pdf or eps) should avoid this issue.

REPLY: Thank you for noticing. We have provided pdf versions of each figure.