Reviewer’s report

Title: Methodological standards for the development and evaluation of clinical prediction rules: A review of the literature

Version: 0 Date: 05 Nov 2018

Reviewer: Ben Van Calster

Reviewer's report:

This paper aims to provide a comprehensive overview of the literature on standards for developing, validating, and implementing prediction models. I commend the authors for their efforts. The challenge of such manuscripts is to decide the optimal depth of addressing each topic. One can either write a short piece, with referral to relevant literature and perhaps supplementary material with more in depth discussions, or write a more extensive paper. The authors have chosen the latter option, the choice of which also depends on the target audience and journal. This resulted in a very long paper, I guess around 10,000 words? I am not sure whether this is acceptable, and it does not seem to fall under the category 'commentary'. Because of this, my review is also quite lengthy, but I hope it is useful to the authors.

Given that the authors aimed to write a comprehensive manuscript (which it is), I here summarize topics about which I believe there is little or no information:

- Timing of predictors is a logical issue, but apparently it is not always respected that you should only consider predictors that are available at the time you want to make your prediction (Whittle et al, JCE 2018)

- Penalization of models

- Assessment of nonlinear effects and interaction terms (Collins et al, Stat Med 2016; Royston & Sauerbrei book 2009)

In addition, the authors aimed to provide a review of the literature. Although the authors identified a large body of important literature, I give a chronological list of publications that I think are highly relevant but not discussed. All of these are mentioned somewhere in my review comments.

- Vergouwe et al, Am J Epidemiol 2010

- Steyerberg et al, J Clin Epidemiol 2011
Further key comments

- Section on missing data: this is quite elaborate and general. I would consider shortening and focusing more on prediction modeling. E.g. the fact that outcome should be included in imputation. Also, some people state that MI is not imperative in prediction model research (Masconi et al, PLoS One 2015). Hotdeck imputation appears a bit old-fashioned. Is this similar to predictive mean matching (which is used in mice)? And what about missing outcomes? How to use a model in practice when something cannot/is not measured? Is Figure 2 needed?

- P14, "although it is more appropriate to consider the total number of regression coefficients estimated, including those considered prior to predictor selection": please be more clear and decisive. EPV should not be judged based on variables in the final model, but on all considered coefficients (i.e. a nominal predictor will have >1 coefficient) prior to eventual data-driven variable selection.

- P14, "sample size could be determined by the level of precision in the CI around the performance metric": This is a quick statement to close this section... how would that be done? Does this focus on CI around performance metrics in the development data, although you know that development performance is optimistic? This statement mainly raises questions, I think. I'd say that you need a sample size that ensures stable models. In that respect, the findings by van Smeden are important, but I do not think they considered settings in which data-driven variable selection was performed. This must be another influential aspect (Steyerberg et al, J Clin Epidemiol 2011;64:1464-5; and your ref 119).

- P18, 'giving an average estimate of the optimism': how this is done is not clearly explained. Either elaborate or refer the readers to the relevant literature. (Also, there are bootstrap methods like .632 and .632+ that are not mentioned.)
- P19, regarding calibration, a comprehensive overview is given in Van Calster et al (J Clin Epidemiol 2016). Apologies for advertising my own work!

- P20, 'an intercept of 0 and a slope of 1 represents perfect calibration': this is hard to understand because the logistic recalibration framework has not been explained.

- P21, 'a CPR with good discrimination, but poor calibration, may still be appropriate and useful depending on its intended use': the tone of this statement is not appropriate, as it downplays the importance of calibration. For risk models, accurate predictions are essential, although I agree that calibration is something that can be fixed, as you write. In Van Calster and Vickers (Med Decis Making 2015), we illustrate that poor calibration can make a model useless or harmful for clinical decision making. I think this is a very relevant finding that should be mentioned. Again, apologies for advertising my own work!

- P22, 'some researchers simply present the final regression formula': it has to be clear that this is imperative, see TRIPOD. This also relates to the next section of the paper (reporting the derivation of a clinical prediction rule).

- P26, 'in principle, temporal validation is the same as the split-sampling approach described above, and carries the same limitations': this is not true. It is correct that differences between temporal validation and derivation populations can be small (cf p25), but this statement undervalues temporal trends (e.g. see Davis et al, JAMIA 2017). You cannot state that temporal validation is the same as split-sample internal validation.

- P28: based on the issues with NRI and IDI, mainly that they are non-proper metrics, these statistics cannot be recommended. Please be clear on that. See also Antolini et al (Biom J, in press).

- P32, 'measures of impact include safety and efficiency': but not just that, right? One can look at mortality, unnecessary procedures, etc. This part comes across too much as paraphrasing the Reilly & Evans paper.

- P34, 'if the CPR does not add anything … then the use of the CPR … would not be warranted': why? A CPR may make the judgment objective, and may help less experienced clinicians. This part about CPR vs clinical judgment foregoes the role of experience. Later on, the authors state 'differences between the two methods of judgment are likely due to different diagnostic thresholds'. Then you can argue that a CPR allows you to set the threshold to the desired level, something that is not so easy with clinical judgment?

- P36, 'incorporation of CPRs into clinical guidelines may also be of benefit': this should be more prominent, in my view. Uptake in national/international guidelines (e.g. NICE) may be essential to help the uptake of models in practice. This touches upon medicolegal aspects: if a clinician manages a patient based (partially) on a model result, and the management turns out to be
problematic, I can imagine that it makes a difference as to whether the model was recommended by national guidelines.

Other comments, including details

- Background, 'small number of highly valid indicators': this is not always the case, it depends on how the considered predictors were chosen (i.e. strong and informative a priori selection or not).

- Ref 2 is heavily used in the background section, but does not seem to be a key reference.

- P4, 'guidelines for the reporting of…': explicitly mention TRIPOD here already.

- P4, assistive vs directive rules and models vs rules: the difference between these two distinctions is not fully clear, but they by and large seem to boil down to the same thing: there is the original risk prediction model, and there are classifications based on such models after specifying one or more risk thresholds.

- P5, 'if a model has been externally validated in multiple settings or populations': here the authors do not simply refer to the existence of multiple external validation studies, but to the existence of multiple successful external validation studies, right? This raises the question as to when a model has successful broad external validation? What if 1 study reports poor performance, but many others report good performance? What constitutes poor performance? The c-statistic depends on the population studies (case-mix, homogeneity), so lower c-statistics may not indicate the model does not work well. Would calibration be more central to the definition of successful validation? Even then, it is overoptimistic to state that 'clinicians can use its predictions confidently in future patients'. We know that performance is heterogeneous (Riley et al, BMJ 2016), so in fact one may take the position that validation/impact/implementation is something that is very setting-dependent?

- There is a lot of literature on how to assess the added value of new markers. Where does this fit in? Stage 1: if a good CPR exists, it may still be of value to evaluate the added value of a novel marker?

- P9: perhaps state why summing up individual risk factors are 'much less accurate'.

- P10, 'multiplied by every 5 years': this is not fully clear. Does this relate to continuous age (in years) divided by 5, or to age categories?

- P10, 'an individual with no predictors present': I suggest to use a different formulation for 'no predictors present'. 
- P10: I thought it was well known that recursive partitioning has inferior performance to multivariable regression, or at least is more greedy and variable. If you state that neither is superior, please reference the appropriate literature.

- Missing data section, minor comments
  o MNAR: depends on the unobserved values/predictors after conditioning on known values/predictors, right?
  o 'missing values are normal': perhaps consider a different formulation
  o P13: 'MI typically assumes that data are MAR'
  o P13, 'work in this area is ongoing': but ref 99 is from 2003!

- P13, causality vs prediction: Shmueli (Stat Sci 2010) is a common reference here, where the common distinction is between explanation and prediction; I guess because causality is a very strong term. Also, I do not see why one should not report odds ratios (or hazard ratios) in prediction studies. It can be of interest to judge the logic of the model, which is related to the issue of acceptability.

- P15, univariable significance testing: up to then, explanations related to a priori selection, but univariable testing is data-driven which is commented on later in the section 'selection of predictors during multivariable modelling'. In line with my previous comment, I suggest to make a clear distinction between a priori selection and data-driven selection (i.e 'during modeling', as you label it), and suggest a prior selection where possible.

- P18, 'if the derivation dataset is split randomly, the training and test samples will differ only by chance and thus produce optimistic results': cross-validation and bootstrapping have the same results, they are all random splits/subsets of the same dataset.

- P18-19, 'the advantages of bootstrapping over other internal validation methods are that it can account for the influences of the predictors selection method used': cross-validation, or repeated train-test splitting can do this too? Nevertheless I prefer bootstrapping, e.g. because CV and (repeated) data splitting train and validate on (multiple) splits of the data and hence training is done on smaller samples than in bootstrapping.

- P22, presentation of a CPR: I think presentations can be interesting to understand a model (odds ratios, but nomograms are preferable), or can focus on using a model (web-based calculator or app). Perhaps that distinction can be made more clearly. Regarding understanding: a colleague of mine worked on color-based alternatives for nomograms (Van Belle & Van Calster, PLoS One 2015;10:e0132614).
- P24, 'performance should be assessed using calibration and discrimination': and measures for clinical usefulness such as decision curve analysis.

- P25: it is interesting to recommend the assessment of inter-rater reliability, but I think this is rarely done. Have you got an idea about the frequency of such efforts? You may perhaps comment on this.

- P25, 'ideally 200 events': in Van Calster et al (J Clin Epidemiol 2016), we also argue for at least 200 events (i.e. cases in the smallest outcome group) in order to make useful calibration plots.

- P25 about the interpretation of external validation studies: ref 34 is indeed important. But Vergouwe et al (Am J Epidemiol 2010) and van Klaveren et al (Stat Med 2016) are key references as well in this regard.

- P25 at the bottom: first explain the 3 types of external validation before comparing temporal and domain validation.

- P26: I disagree that 'fully independent validation' is a type of validation much like temporal-geographical-domain-methodologic-spectrum, because this relates only to who is doing the validation. Interestingly, Siontis et al (J Clin Epidemiol 2015) do not report different results for independent external validation vs non-independent external validation.

- P27, 'there are no strict criteria to determine an acceptable level of accuracy of a CPR': I do not think this is possible. Regarding c-statistic, this may depend a lot on case-mix. E.g. if the validation sample is more homogeneous, this will naturally tend to reduce the c-statistic. I'd say that calibration is key?

- P28, 'inappropriate in the case of nested models': this deals with testing for the difference (Demler et al, Stat Med 2012), and I think it only holds when testing on the derivation data (not sure though).

- P28: 'found that … appropriate risk reclassification measures were rarely reported': which ones should be reported then?

- P30, 'found that patient outcomes are rarely reported in impact studies': which outcomes were reported, then?

- P30-31, study design for impact analysis: can you give any information about cluster size and sample size calculations? I would think you need quite a lot of clusters for an impact cluster RCT?

- P31, 'practical, logistic and economic challenges': what are you alluding to here? Is the required sample size often a problem?

- P31-32, 'if possible … enhance performance': shouldn't this point be made upfront?
- P33, 'qualitative research can be invaluable': agree, but do you have any idea how common this is?

- P34, 'understanding the performance of a CPR in contrast to clinician judgment may aid clinicians' acceptability of the CPR': in what way?

- P35, phase 4 of impact analysis: this is the same as stage 6 in this paper, right? This might be mentioned.

- P35, stage 5 (cost-effectiveness): isn't this a part of impact assessment?

- P37, barriers: complex models (e.g. based on machine learning algorithms) may have additional barriers.

- P38, concluding remarks: this is more a summary than a conclusion.

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:

Acceptable

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