Prediction models

Journal of Clinical Epidemiology 142 (2022) 330–331

Journal of Clinical Epidemiology

LETTER TO THE EDITOR

Prediction models: stepwise development and simultaneous validation is a step back

We have read the article 'Using a stepwise approach to simultaneously develop and validate Machine Learning based prediction models' as recently published by the Journal of Clinical Epidemiology [1]. While we agree with the paper's premise that 'clinical prediction models based on machine learning techniques are often not properly validated', the authors' stepwise approach is not a step forward but a step back in the proper validation of prediction models. We noted several inconsistencies and questionable claims in the paper, and highlight some of the most important issues for this commentary.

First, in step 7 of their stepwise development scheme the authors propose to "externally validate" the final model with all the available data, which includes both the training cohort and the test cohort. In this way, no unbiased assessment of model performance can be obtained as the model has been optimized on the training cohort. Such a procedure is misleading and does not deserve to be called "validation," let alone "external validation" [2]. External validation examines the generalisability of a model using data collected in different but plausibly related settings than development [3-4], and is ideally performed by different investigators [5]. Hence, any split-sample validation, including the variant proposed in [1], is no external validation. While the authors admit this fact in the discussion, the paper still uses the notion "external validation" for split-sample validation throughout the manuscript. More importantly, the authors state that their faulty external validation procedure is interesting because real external validation is time-consuming and can be done after implementing the model in clinical practice. We stress that truly independent, and preferably local, validation remains a prerequisite before introduction of a model into clinical practice. The suggested approach favours speed over patient safety and clinical utility, and cannot be recommended.

Second, the proposed repeated validation approach may easily lead to selection of an overfitted model and overoptimistic estimates of this model's performance. This point becomes most obvious when when the authors state: 'if after already a few steps a very high AUC is achieved, larger than a prespecified threshold, one can decide to reduce the number of steps, resulting in a larger test set to validate the final model on.' Clearly this approach is set up to select models that are on a random high in their performance in the test set. Model development will stop prematurely with overoptimistic estimates of performance. Moreover, the test set is used twice: once to decide to stop, and once in a final validation. Hence the final validation is no longer unbiased. In addition, we point to the extensive literature on the necessity of separating model validation from model selection [3-6]. This principle is violated by the set-up of the suggested stepwise approach. Note also that the authors are inconsistent on a fundamental aspect: whereas the text states that every step involves the evaluation of many models on the test set, Table 1 states that every step involves the evaluation of the best model (using cross-validation on the training cohort) on the test set.

Third, the suggested approach lacks theoretical underpinning and is not supported by empirical evidence from the literature. It was also not evaluated by means of a simulation study and/or case study to illustrate that overoptimism in the assessment of model performance is avoided or at least minimized. Such evaluations might have shown that the suggested sample size numbers are too low when many candidate predictors are available relative to sample size. Recent research provided recommendations on the necessary sample size for prediction model development and validation [7-9]. There is evidence that the more flexible a prediction algorithm is, the more data it needs [10].

Fourth, the suggested approach promises to "evaluate the stability of the final 'best' model over increasing sample size by predicting the subjects in the training sets and determine the AUC of the final model for each training set" (step 8). This proposal only evaluates the cumulative apparent model performance, not the model's stability. It is quite predictable that AUC will converge with the growing training set. Random sampling variability in the case-mix may explain some variation in the AUC [11]. Other performance measures such as calibration were ignored while such measures may take longer to stabilize [11].

Fifth, the claim is not substantiated that "if the amount of available data per patient increases drastically, modern and more flexible modelling techniques such as machine

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

https://doi.org/10.1016/j.jclinepi.2021.07.019

learning techniques might be preferred.” Machine learning methods use cross-validation or bootstrapping to tune regularization parameters. These should guard against overfit, but recent research suggested that the estimation of the regularization parameter can easily fail in small data sets [12-14]. This failure with only few candidate predictor variables will become even more problematic “if the amount of available data per patient increases drastically.” However, the authors do not discuss tuning of regularization parameters at all.

Considering our objections, we strongly advise against the use of the proposed stepwise validation. Instead, we encourage authors to invest more time in thorough validation of existing prediction models in various locations, and search for updating of the model to provide better predictions for future patients. Performance should include aspects such as discrimination, calibration and utility, assessed in appropriately sized datasets. We therefore encourage initiatives to improve thorough and timely validation of prediction models without lowering the bar of thoroughness.

Georg Heinze*
Section for Clinical Biometrics, Center for Medical Statistics, Informatics and Intelligent Systems, Medical University of Vienna, Spitalgasse 23, 1090 Vienna, Austria

Maarten van Smeden
Julius Center for Health Sciences and Primary Care, University Medical Center Utrecht, Utrecht University, Utrecht, the Netherlands

Laure Wynants
Department of Epidemiology, CAPHRI Care and Public Health Research Institute, Maastricht University, Netherlands
Department of Development and Regenaration, KU Leuven, Leuven, Belgium
EPI-center, KU Leuven, Belgium

Ewout Steyerberg
Department of Biomedical Data Sciences, Leiden University Medical Center, Leiden, the Netherlands

Ben van Calster
Department of Development and Regenaration, KU Leuven, Leuven, Belgium
EPI-center, KU Leuven, Belgium
Department of Biomedical Data Sciences, Leiden University Medical Center, Leiden, the Netherlands

*Corresponding author: Tel: +43-140-4006-6890, fax: +43-140-4006-6870.
E-mail address: georg.heinze@meduniwien.ac.at (G. Heinze)

References

[1] Haalboom M, Kort S, van der Palen J. Using a stepwise approach to simultaneously develop and validate Machine Learning based prediction models. J Clin Epidemiol 2021. doi:10.1016/j.jclinepi.2021.06.008.

[2] Steyerberg EW, Harrell Jr FE. Prediction models need appropriate internal, internal-external, and external validation. J Clin Epidemiol 2016;69:245–7. doi:10.1016/j.jclinepi.2015.04.005.

[3] Altman DG, Vergouwe Y, Royston P, Moons KGM. Prognosis and prognostic research: validating a prognostic model. BMJ 2009;338:b605. doi:10.1136/bmj.b605.

[4] Justice AC, Covinsky KE, Berlin JA. Assessing the Generalizability of Prognostic Information. Ann Internal Med 1999;130:515–24. doi:10.7326/0003-4819-130-6-199903160-00016.

[5] Altman DG, Royston P. What do we mean by validating a prognostic model? Statistics in Medicine 2000;19:453–73 10.1002/(SICI)1097-0258(20000229)19:4%3C453::AID-SIM350%3E3.0.CO;2-5.

[6] Moons KGM, Kengne AP, Grobbee DE DE, Royston P, Vergouwe Y, Altman DG, Woodward M. Risk prediction models: II. External validation, model updating, and impact assessment. Heart 2012;98:691–8. doi:10.1136/heartjnl-2011-301247.

[7] Riley RD, Snell KIE, Ensor J, Snell KI, Harrell Jr FE, Martin GP, Reitsma JB, Moons KGM, Collins GS, van Semeden M. Calculating the sample size required for developing a clinical prediction model. BMJ 2020;368:m441. doi:10.1136/bmj.m441.

[8] Riley RD, Debray TPA, Collins GS, Archer L, Ensor J, van Smeden M, Snell KIE. Minimum sample size for external validation of a clinical prediction model with a binary outcome. Statistics in Medicine 2021. doi:10.1002/sim.9025.

[9] van Smeden M, Moons KGM, de Groot JAH, Collins GS, Altman MJC Eijkemans DG, Reitsma JB. Sample size for binary logistic prediction models: Beyond events per variable criteria. Statistical Methods in Medical Research 2019;28:2455–74 10.1177%2F0962280218784726.

[10] van der Ploeg T, Austin PC, Steyerberg EW. Modern modelling techniques are data hungry: a simulation study for predicting dichotomous endpoints. BMC Medical Research Methodology 2014;14:137. doi:10.1186/1471-2288-14-137.

[11] Christodoulou E, van Smeden M, Edlinger M, Timmerman D, Wanitschek M, Steyerberg EW, van Calster B. Adaptive sample size determination for the development of clinical prediction models. In: Diagnostic and Prognostic Research, 5; 2021. p. 6. doi:10.1186/s41512-021-00096-5.

[12] B van Calster, M van Smeden, B De Cock, EW Steyerberg. Regression shrinkage methods for clinical prediction models do not guarantee improved performance: Simulation study. Statistical Methods in Medical Research 29:3166-3178, doi: https://doi.org/10.1177%2F0962280220921415

[13] Riley RD, Snell KIE, Martin GP, Whittle R, Archer L, Sperrin M, Collins GS. Penalization and shrinkage methods produced unreliable clinical prediction models especially when sample size was small. Journal of Clinical Epidemiology 2021;132:88–96. doi:10.1016/j.jclinepi.2020.12.005.

[14] Sinkovec H, Heinze G, Blagus R, Geroldinger A. To tune or not to tune, a case study of ridge logistic regression in small or sparse datasets. arXiv e-prints [Preprint]:210111230v1 [statME]; 2020.