© 2019 EDIZIONI MINERVA MEDICA
Online version at http://www.minervamedica.it

Minerva Anestesiologica 2019 July;85(7):701-4
DOI: 10.23736/S0375-9393.19.13520-1

EDITORIAL

Predictive models in clinical practice: useful tools to be used with caution

Bruno M. CESANA [1] *, Sabino SCOLLETTA [2]

[1]G.A. Maccacaro Unit of Medical Statistics, Biometry, and Bioinformatics, Department of Clinical Sciences and Community Health, Faculty of Medicine and Surgery, University of Milan, Milan, Italy; [2]Unit of Critical and Intensive Care Medicine, Department of Medicine, Surgery, and Neurosciences, University Hospital of Siena, Siena, Italy

*Corresponding author: Bruno M. Cesana, Giulio A. Maccacaro Unit of Medical Statistics, Biometry, and Bioinformatics, Department of Clinical Sciences and Community Health, Faculty of Medicine and Surgery, University of Milan, Milan, Italy.
E-mail: bruno.cesana@guest.unimi.it

Diagnostic and prognostic predictive models, aimed at calculating the probability of occurrence of a certain event (disease or its evolution), are frequent in biomedical literature[1] and in clinical guidelines for formal risk assessment. So, the necessity of their systematic reviews led to the formation of the Cochrane Collaboration Prognosis Reviews Methods Group,[2] which developed and validated search strategies for identifying prediction model studies.[3]

Then, a Checklist for Critical Appraisal and Data Extraction for Systematic Reviews of Prediction Modelling Studies (CHARMS)[1] has been designed, followed by the Transparent Reporting of a Multivariable Prediction Model for Individual Prognosis or Diagnosis (TRIPOD) guidelines.[4, 5]

A search for "predictive" or "prognosis" or "risk factors" or "diagnosis" in the title of the original articles published on *Minerva Anestesiologica* from 2013 to 2017 returned 38 papers on building and eight papers about the validation of a predictive model. Furthermore, a review on the design, statistics, interpretation, and validation of the predictive models for postoperative pulmonary complications has been recently published[6] with an editorial[7] emphasizing their clinical impact.

Indeed, the actual question is: how much can these predictive models be considered as useful tools for the clinical practice? For an overview of the statistical methodology the readers are referred to a Tutorial in Biostatistics by Harrell *et al.*[8] and to four British Medical Journal papers mainly tackling the general methodology of these studies.[9-12]

First of all, predictive models must come from properly planned *ad-hoc* studies (prospective "multivariable" cohort observational studies, being randomized trials affected by the presence of the treatment although not statistically significant). Then, patients have to be consecutively selected according to well defined inclusion/exclusion criteria for having a sample with the best possible representativeness of the target population, which must also be as wide as possible by including broad clinical scenarios. Furthermore, the recorded variables must be able to best characterize the phenomenon of interest and have to be used the best validated methods of measurement for which the properties of accuracy, precision, repeatability and reproducibility have to be fulfilled. Finally, "hard" (*i.e.*, objective) variables should be preferred to "soft" (*i.e.*, subjective) variables.

The requirement of an adequate sample size is a particularly relevant aspect taking into account that problems arising from data-dependent selection, goodness of fitting, validation, etc. can be exacerbated by small sample sizes. The frequently used criterion of at least ten events per variable (EPV) to be included into the predictive model has indeed to be considered as a very lowest threshold and, actually, not satisfactory.

Harrell *et al.*[13] concluded that for regression modelling the EPV should be at least ten times the number of potential prognostic variables that could be included in the model. Peduzzi *et al.*[14] showed that an EPV equal to 10 has to be considered a minimum. Finally, Feinstein[15] suggested that an EPV of 20 is safer. It must to be pointed out that most published studies do not meet the above criteria.

Many statistical procedures can be used to build a predictive model: from the traditional regression models (multiple linear, logistic, Cox's proportional hazard regression) to the recent methods of regression trees (CART), neural networks, machine learning techniques that, however, seem to not bring any consistent advantage.

It is not possible to consider here the pros and the cons of these statistical methods: each of them has its strengths and weaknesses, but, generally, all can be useful and usable if the prediction obtained is accurate for groups of patients or for individual patient. Indeed, according to Burstein,[16] "Usefulness is determined by how well a model works in practice, not by how many zeros there are in the associated P values."

It has to be pointed out that predictive models have to pass two steps. Firstly, the internal validity is assessed on the same dataset used to develop the model by obtaining the "apparent performance," which obviously tends to be overestimated, and consequently biased. Secondly, and more relevant, the external validity has to be assessed on different validation samples. To these aims, a number of predictive performance measures and statistics have to be evaluated graphically and/or by formal statistical tests: goodness of fitting, calibration ("how well the predicted risks compare to the observed outcomes"), discrimination ("how well the model differentiates between those with and without the outcome"), classification measures (notably, sensitivity and specificity), and reclassification measures (such as net reclassification improvement).

In this issue of *Minerva Anestesiologica*, Ranucci *et al.*[17] present a retrospective analysis of hemodynamic data to assess discrimination and calibration properties of the Hypotension Probability Indicator (HPI) for prediction of hypotensive events in 23 patients undergoing vascular and cardiac surgery. Cardiovascular patients are quite often expose to intraoperative hemodynamic derangements (*e.g.*, cardiac arrhythmias, impaired myocardial contractility, changes in preload and afterload due to blood loss or systemic inflammatory reaction), which occur with arterial hypotension and potential postoperative complications.

HPI is obtained by a machine-learning approach and its development and external validation has been recently published by Hatib *et al.*[18] These authors declared that HPI has a very satisfactorily performance, but they did not explain the statistical details for calculating the coefficients (unknown for patent reasons) of the variables. On the contrary, Ranucci *et al.*[17] in the validation part of their paper reported that the HPI algorithm had a poor calibration performance not even satisfying the first step of the goodness of fitting assessment of a prediction model.

A fundamental point is that without an external validation, a predictive model should be used very cautiously in clinical practice and, particularly, in different institutions. So validation studies are mandatory and, even if sample size rules are not well established, they have to be carried out with a minimum of 100 events and ideally 200 (or more) events,[19] or a minimum of 100 events and 100 non-events and 20 participants per predictor in the case of continuous outcomes.

More challenging than the usual statistical methods are the frequentistic and Bayesian procedures aimed to build a prognostic model from repeatedly measured independent variables as predictors and a fixed outcome.[20] Of course, repeated measurements would provide more information about the variable's trajectory over the time than just a single measurement. However, among other things, it has to be taken into account the correlation among the measures and the fixing of a time lag between the measurements and the event of interest.

Sophisticated statistical techniques together with the involvement of a professional statistician are also required for jointly considering, as in Ranucci *et al.*[17] paper, repeated measures of a predictor and the possible occurrence of multiple events per subject (hypotensive episode) in which both assumptions of independence of the predictor and of the events are not fulfilled.

Readers have to be well aware that the application of sophisticated statistical techniques is not sufficient to confer validity to a predictive model; indeed, we may, rather provocatively, state that the (mis)use of the statistical methods would allow one to demonstrate almost everything.

A rule of thumb to suggest to readers for disregarding a paper about a predictive model is when the above outlined criteria for assessing the internal/external validity are missing. An even more definitive negative judgement is when sensitivity or specificity confidence intervals have the lower limit less than 0.5 or when it is not reported, to avoid showing that a predictive model is equal to the toss of a coin.

Indeed, it is precisely in this area that the statistical and clinical aspects should play an integrate role, requiring that a predictive model is valid only if it has passed tests of statistical validation and of utility in clinical practice.

The limited sample size of patients enrolled in the study by Ranucci *et al*,[17] even if the number of events is similar to the one recommended for independent events,[19] allows to stress the fact that building and validating predictive models have to be carried out with large sample sizes, without being satisfied with the lower values reported in the statistical literature. Finally, it has to stress the obvious fact that 100 events occurring on the same subject cannot be a satisfactory basis for validating a predictive model at the same way as 100 events occurring on 100 subjects.

In conclusion, Ranucci *et al.*[17] are to be commended for having focused on the value of HPI in cardiovascular surgery patients, in whom intraoperative hypotension is frequently observed due to bleeding, reduced ventricular function, and systemic vasodilation. However, as the authors stated in their article, a homogenous approach to the statistical methodology is strongly recommended, especially in HPI validation series. We agree with that statement. Indeed, the question about HPI is still open and requires further validation studies.

References

1. Moons KG, de Groot JA, Bouwmeester W, Vergouwe Y, Mallett S, Altman DG, *et al.* Critical appraisal and data extraction for systematic reviews of prediction modelling studies: the CHARMS checklist. PLoS Med 2014;11:e1001744.

2. Riley RD, Ridley G, Williams K, Altman DG, Hayden J, de Vet HC. Prognosis research: toward evidence-based results and a Cochrane methods group. J Clin Epidemiol 2007;60:863–5.

3. Keogh C, Wallace E, O'Brien KK, Murphy PJ, Teljeur C, McGrath B, *et al.* Optimized retrieval of primary care clinical prediction rules from MEDLINE to establish a Web-based register. J Clin Epidemiol 2011;64:848–60.

4. Collins GS, Reitsma JB, Altman DG, Moons KG. Transparent Reporting of a multivariable prediction model for Individual Prognosis or Diagnosis (TRIPOD): the TRIPOD statement. Ann Intern Med 2015;162:55–63.

5. Moons KG, Altman DG, Reitsma JB, Ioannidis JP, Macaskill P, Steyerberg EW, *et al.* Transparent Reporting of a multivariable prediction model for Individual Prognosis or Diagnosis (TRIPOD): explanation and elaboration. Ann Intern Med 2015;162:W1-73.

6. Mazo V, Sabaté S, Canet J. How to optimize and use predictive models for postoperative pulmonary complications. Minerva Anestesiol 2016;82:332–42.

7. Ball L, Pelosi P. Automated mechanical ventilation modes in the intensive care unit: an obstacle course in building evidence. Minerva Anestesiol 2016;82:621–4.

8. Harrell FE Jr, Lee KL, Mark DB. Multivariable prognostic models: issues in developing models, evaluating assumptions and adequacy, and measuring and reducing errors. Stat Med 1996;15:361–87.

9. Moons KG, Royston P, Vergouwe Y, Grobbee DE, Altman DG. Prognosis and prognostic research: what, why, and how? BMJ 2009;338:b375.

10. Royston P, Moons KG, Altman DG, Vergouwe Y. Prognosis and prognostic research: developing a prognostic model. BMJ 2009;338:b604.

11. Altman DG, Vergouwe Y, Royston P, Moons KG. Prognosis and prognostic research: validating a prognostic model. BMJ 2009;338:b605.

12. Moons KG, Altman DG, Vergouwe Y, Royston P. Prognosis and prognostic research: application and impact of prognostic models in clinical practice. BMJ 2009;338:b606.

13. Harrell FE Jr, Lee KL, Califf RM, Pryor DB, Rosati RA. Regression modelling strategies for improved prognostic prediction. Stat Med 1984;3:143–52.

14. Peduzzi P, Concato J, Kemper E, Holford TR, Feinstein AR. A simulation study of the number of events per variable in logistic regression analysis. J Clin Epidemiol 1996;49:1373–9.

15. Feinstein AR. Multivariable Analysis: an Introduction. New Haven, CT: Yale University Press; 1996.

16. Burstein AH. Fracture classification systems: do they work and are they useful? J Bone Joint Surg Am 1993;75:1743–4.

17. Ranucci M, Barile L, Ambrogi F, Pistuddi V; Surgical and

Clinical Outcome Research (SCORE) Group. Discrimination and calibration properties of the hypotension probability indicator during cardiac and vascular surgery. Minerva Anestesiol 2019;85:724-30.

18. Hatib F, Jian Z, Buddi S, Lee C, Settels J, Sibert K, *et al.* Machine-learning Algorithm to Predict Hypotension Based on High-fidelity Arterial Pressure Waveform Analysis. Anesthesiology 2018;129:663–74.

19. Collins GS, Ogundimu EO, Altman DG. Sample size considerations for the external validation of a multivariable prognostic model: a resampling study. Stat Med 2016;35:214–26.

20. Rondeau V, Mathoulin-Pelissier S, Jacqmin-Gadda H, Brouste V, Soubeyran P. Joint frailty models for recurring events and death using maximum penalized likelihood estimation: application on cancer events. Biostatistics 2007;8:708–21.

Conflicts of interest.—The authors certify that there is no conflict of interest with any financial organization regarding the material discussed in the manuscript.

Comment on: Ranucci M, Barile L, Ambrogi F, Pistuddi V; Surgical and Clinical Outcome Research (SCORE) Group. Discrimination and calibration properties of the hypotension probability indicator during cardiac and vascular surgery. Minerva Anestesiol 2019;85:724-30. DOI: 10.23736/S0375-9393.18.12620-4.

Article first published online: March 12, 2019. - Manuscript accepted: February 20, 2019. - Manuscript revised: January 18, 2019. - Manuscript received: December 11, 2018.

(*Cite this article as:* Cesana BM, Scolletta S. Predictive models in clinical practice: useful tools to be used with caution. Minerva Anestesiol 2019;85:701-4. DOI: 10.23736/S0375-9393.19.13520-1)