EDITORIAL

Validation and Updating of Clinical Prediction Models: Why and How?

Validación y actualización de los modelos clínicos predictivos: ¿cómo y por qué?

EWOUT W. STEYERBERG

Are predictions from a previously developed prediction model valid for my patients? This is a difficult question that was addressed carefully in a recent paper that focused on the validity of the GRACE score to predict in-hospital mortality. (1)

Why is this high-quality study by Mangariello and Gitelman important? Prediction models are increasingly published in the medical literature. Newer methods are proposed, such as labeled machine learning, deep learning and artificial intelligence. Whatever the method of development, the key issue is whether the prediction model or algorithm provides valid predictions for physicians and their patients who rely on them as a source of information and decision making. Indeed, the authors rightly argue that many differences between settings may be present. These include differences in patient characteristics, healthcare systems, and socioeconomic environment. Moreover, treatment may change over time. All these differences may make a previously developed model not valid for the particular setting where the model is applied, for example the Dr. Juan A. Fernández Hospital in Buenos Aires. Therefore, a model may need updating for a specific setting. (2)

Validation: how and why?

How should a validation study be performed? The first issue is whether a previously developed model would be expected to be applicable to the validation setting. In the presented paper, patients from 2 centers in Argentina were included in the development of the GRACE model, in an international study. So, the validation setting was plausibly related to the development setting. Moreover, the model included a clinically reasonable set of predictors, and had been proposed by an international group of experts.

Further guidance for validation is provided by the TRIPOD guideline, especially in the detailed Explanation and Elaboration document. (3) Important elements for validation include adequate sample size and adequate methods.

a) Sample size: an adequate sample size implies at least 100 events at validation. (4) So, if the in-hospital mortality is 5%, a total sample size of at least 2,000 patients is needed for reliable results. Also, missing values may be imputed with advanced statistical methods to make full use of all available information, even if some patient records are incomplete. Both conditions were fulfilled in the current study for the total patient group (2,104 patients, 117 events) [1]. Sample size was, however, a limiting factor for the NSTE-ACS subgroup, with only 35 events. It is then impossible to separate apparently adequate performance from lack of power to detect inadequate performance.

b) Adequate methods: Key aspects are discriminative ability and calibration [2]. Discrimination is usually evaluated by the area under the ROC curve (AUC), also known as the concordance, or c, statistic. Calibration evaluates whether the estimated risks agree with the observed frequency of the event. The authors rightly emphasize graphical assessment over statistical testing. If we want to evaluate the potential of a model for guiding decision making, more modern measures are needed, including the “Net Benefit”. (5) Net Benefit counts the number of true positive classifications and penalizes for false positive classifications when we classify those at high versus low risk with a prediction model. The relative weight is defined by the clinical context, which is better than using a statistically defined weight. (6)

Interpretation and consequences of invalidity

How should we interpret the results? The GRACE score was developed well. The model had adequate internal validity, without risks for overfitting since the sample size was very large (Table 1). The model is obviously far from perfect in predicting who will die and who will not, and may be inadequate for specific groups of patients. The current external validation confirms that the discriminative ability of the model remains adequate when it is applied in another setting. The AUC was 0.83 at development and 0.87 at

Rev Argent Cardiol 2019;87:256-257. http://dx.doi.org/10.7775/rac.v87.i4.15891

SEE RELATED ARTICLE: Rev Argent Cardiol 2019;87:291-299. http://dx.doi.org/10.7775/rac.v87.i4.15346

Table. Overview of key issues with internal and external validity of clinical prediction models. (2)

Type of validity	Problems	Description	Potential solutions
Internal validity	Overfitting	Model describes the development setting, but is not valid for the setting it was derived from.	Large sample size, careful modeling. Quantify by cross-validation or bootstrapping.
	Underfitting	Model misses some important patterns in the data, making it invalid for some types of patients.	Careful modeling, can never be excluded. Realize any prediction is based on the specific definition of the model.
External validity	Predictor definitions and measurement	Model was developed with different definitions of predictors, invalidating performance in the other setting.	Try to stick to the definition ('common data elements'). Realize potential differences may impact on validity.
	Missed predictors	Differences in predictor characteristics not included in the model related to the specific setting and not explained by differences in the distribution of values of predictors that are in the model.	Compare development and validation settings. Realize any prediction model is limited; many prognostically relevant characteristics may not be included as predictors in a model because difficult to measure or as yet unknown.

external validation. (1) This is attributable to larger heterogeneity (variability in predictor values between patients: "case-mix"). (2)

Calibration of predictions is the most relevant aspect of validity for individual patients. It is the Achilles heel of prediction models. (7) In the presented validation study, outcomes were worse than expected, which may have two very different interpretations. One is that the care was suboptimal in the validation setting. The other is that the model was inadequate due to some predictors that were different in the validation setting compared to the development setting, but were not included in the model (Table 1). So, differences in basal risk may exist between the development and validation settings that were not captured by the model. (1,2)

Can we now apply the GRACE model in this specific hospital, in Argentina, in South America? The current validation results support the notion of a global model, with predictor effects that are widely valid. (8) Local application however requires updating to the local setting. A simple approach is to update the model with a local model intercept, such that predictions are on average correct. (8,9) New approaches for such updating need attention, in line with the availability of more and more routinely collected data and the desire for self-learning systems. (10) Such more up to date predictions will support medical decision making and contribute to improved outcomes for individual patients.

REFERENCES

1. Mangariello BN, Gitelman PC. Validation of the GRACE Score (Global Registry of Acute Coronary Events) as Predictor of In-hospital Mortality in Acute Coronary Syndromes in Buenos Aires. Rev Argent Cardiol 2019;87:291-9.

2. Steyerberg EW. New York: Springer; 2009. Clinical Prediction Models: A Practical Approach to Development, Validation, and Updating. New York: Springer; 2009. http://doi.org/dtwhd3.

3. 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;162:1-73. http://doi.org/gfrkkz.

4. Vergouwe Y1, Steyerberg EW, Eijkemans MJ, Habbema JD. Substantial effective sample sizes were required for external validation studies of predictive logistic regression models. J Clin Epidemiol 2005;58:475-83. http://doi.org/bj7bng

5. Steyerberg EW, Vickers AJ, Cook NR, Gerds T, Gonen M, Obuchowski N, et al. Assessing the performance of prediction models: a framework for traditional and novel measures. Epidemiology 2010;21:128-38. http://doi.org/bj7bng

6. Vickers AJ, Van Calster B, Steyerberg EW. Net benefit approaches to the evaluation of prediction models, molecular markers, and diagnostic tests. BMJ 2016;352:i6. http://doi.org/gcx6nq

7. Shah ND, Steyerberg EW, Kent DM. Big Data and Predictive Analytics: Recalibrating Expectations. JAMA 2018;320:27-8. http://doi.org/gd4rgm

8. Steyerberg EW, Nieboer D, Debray TP, van Houwelingen HC. Assessment of heterogeneity in an individual participant data meta-analysis of prediction models: An overview and illustration. Stat Med 2019, in press: 10.1002/sim.8296

9. Steyerberg EW, Borsboom GJ, van Houwelingen HC, Eijkemans MJ, Habbema JD. Validation and updating of predictive logistic regression models: a study on sample size and shrinkage. Stat Med 2004;23:2567-86. http://doi.org/bb7f9c

10. Chen JH, Asch SM. Machine Learning and Prediction in Medicine - Beyond the Peak of Inflated Expectations. N Engl J Med. 2017;376:2507-9. http://doi.org/gfsjxv