Letters to the Editor

Assessing calibration in an external validation study

To the Editor

Evaluating a prediction model using a separate data set from which the model was developed is a crucial step in assessing its predictive performance, often referred to as external validation [1]. The recent study by Tetreault et al. [2] modified their previous prediction model by omitting one of the predictors and then refitting the model on the original development data from 12 sites from North America [3]. The modified prediction model was subsequently evaluated on a larger international cohort from the AOSpine CSM-I trial [4]. Although it is encouraging to see authors carrying out such external validation studies, there are concerns in the analysis that need highlighting.

It is widely accepted that the two main characteristics to report when evaluating the performance of a prediction model are discrimination and calibration [5]. In the study by Tetreault et al. [1], calibration has been incorrectly evaluated. The authors have presented the traditional and commonly seen plot of predictions and observed outcomes, by ranking and grouping individuals (47 groups of size 10), and calculating the mean observed outcome against the mean predicted probability. However, the authors have then incorrectly fit a linear regression line to the 47 points and examined whether the resulting intercept and slope are noticeably different from 0 and 1, respectively. The arbitrary creation of groups of size 10 will affect these estimates of the slope and intercept, and different values will result if groups of 5 or 20 were made. The correct approach would be to calculate the slope and intercept fitting the linear predictor (LP; \( LP = 1.59 - 0.81P + 0.19mJOA + 0.91G - 0.69S - 0.27DS \)) calculated for all individuals in the validation data set as the only predictor in a logistic regression model: log odds (outcome) = \( \alpha + \beta \times LP \) [6].

A second concern is that the authors refit the model on the validation data set and then made a judgment on the similarity of the regression coefficients. This does not constitute an assessment of validation and provides no useable information on the performance of the prediction model [6]. Validation concerns only the performance evaluation of the prediction model in new data.

The Transparent Reporting of a multivariable prediction model for Individual Prognosis or Diagnosis (TRIPOD) Initiative (www.tripod-statement.org) recently published the TRIPOD reporting guideline for clinical prediction models, where key issues in the development and validation of a prediction model are discussed. The TRIPOD guideline is similar to other well-known reporting guidelines (eg, CONSORT, STROBE, PRISMA) designed to help authors, peer reviewers, and journal editors in ensuring that the essential items describing the development or validation of a clinical prediction model are clearly reported [5]. Accompanying the reporting guideline is an extensive Explanation & Elaboration article describing the rationale for the checklist item but also highlighting many methodologic considerations when developing or validating a clinical prediction model [6].

References


Gary S. Collins, PhD
Emmanuel O. Ogundimu, PhD
Nuffield Department of Orthopaedics, Rheumatology and Musculoskeletal Sciences, Centre for Statistics in Medicine, Botnar Research Centre, University of Oxford, Windmill Rd, Oxford OX3 7LD, United Kingdom

We thank Dr Gary Collins, Dr Emmanuel O. Ogundimu, and Dr Yannick Le Manach for their interest in our work. It is a significant undertaking to develop and validate clinical prediction models. To date, such models have been lacking for degenerative cervical myelopathy, the most common cause of spinal cord impairment in the world [1]. Our work represents an initial effort to determine key predictors of functional outcome after surgery, and these results will help guide clinical decision making, manage patients’ expectations, and enable shared decision discussions with patients and their families. We recognize that validation of clinical prediction models is a dynamic process that requires the application of a number of complementary statistical approaches and is influenced by changes in clinical care and patient presentation.

The major concern outlined in the letter was the statistical approach used to evaluate calibration. Our approach resembles the logic of the goodness-of-fit test of Hosmer-Lemeshow, which is the most commonly used approach to calibration validation. We agree, however, that the calibration may be better evaluated by fitting the linear predictor. When we fit the linear predictor to our external validation dataset, we obtained an intercept (b0) of −0.4951 and a slope (b1) of 0.9410, both of which deviate slightly from the ideal model of b0 = 0 and b1 = 1 [2]. Compared with our estimates published in the article, the intercept is farther from 0, but the slope is closer to 1. Overall, both analyses reflect that our predictions are somewhat high and indicate overfitting, a phenomenon commonly observed in predictive modeling [3]. Our model, therefore, overestimated the outcome: only 46.7% achieved a score 16 or more on the modified Japanese Orthopedic Association in the external validation dataset compared with a model-predicted proportion of 57.6%. The literature does not offer guidance as to what calibration performance parameters are acceptable for clinical application. This area of clinical epidemiological research continues to evolve.

The authors of this letter were also concerned that we compared regression coefficients between the development and validation datasets. We did not use this analysis for the purpose of external validation. Instead, we ran this analysis for exploratory purposes to identify potential limitations in our model.

Finally, we agree that the development of the Transparent Reporting of a multivariable Prediction model for individual prognosis or diagnosis (TRIPOD) checklist is an important step in improving the quality of reporting of prediction models.

References


Michael G. Fehlings, MD, PhD, FRCSC
Toronto Western Hospital
University Health Network
399 Bathurst St, Rm 4W-449, Toronto Ontario, Canada M5T 2S8

Lindsay Tetreault, BSc
Toronto Western Hospital
University of Toronto
University Health Network
Rm 4F-412, 399 Bathurst St, Toronto Ontario, Canada M5T2S8

Pierre Côté, DC, PhD
Faculty of Health Sciences
University of Ontario Institute of Technology (UOIT)
UOIT-CMCC Centre for the Study of Disability Prevention and Rehabilitation
Science Building, Room 2031, 2000 Simcoe Street North
Oshawa, Ontario, Canada, L1H 7K4

Paul Arnold, MD, PhD
Department of Neurosurgery
University of Kansas