Nomograms Need to Be Presented in Full

We read with great interest the recent article by Pan et al describing the development of a nomogram to predict the 5-year and 8-year survival probability for patients with nondisseminated nasopharyngeal cancer. However, there are some aspects surrounding the evaluation of the model and reporting of the study that are of concern.

When evaluating the performance of a clinical prediction model, it is widely accepted that the 2 main characteristics of a particular data set are discrimination and calibration, as noted in the recent Transparent Reporting of a multivariable prediction model for Individual Prognosis Or Diagnosis (TRIPOD) reporting guideline for prediction model studies. Discrimination is defined as the percentage of all patient pairs in which the predictions and outcomes are concordant, commonly measured by the c-index, whereas calibration reflects how accurately the predictions from the model reflect the survival in the observed data. Although Pan et al reported an assessment of both discrimination and calibration, their approaches are flawed.

Our first concern is the internal validation assessment using bootstrapping to produce a bias-corrected estimate of the c-index. Although bootstrapping is the preferred and recommended approach for internal validation, it is important that all model-building steps (including the flawed univariate analyses and backwards elimination) are replayed in each bootstrap sample. If these steps are omitted and only the “final model” is fit in each bootstrap sample, then the resulting bias-corrected c-index will itself be biased and overoptimistic. It is unclear what exactly the authors have done because the reporting is somewhat brief, but we suspect that only the final model was bootstrapped.

Pan et al presented the traditional and commonly observed calibration plot of predictions against observed outcomes by fourths of predicted risk (ie, 4 equal-sized groups), although usually this is done by tenths of predicted risk (ie, 10 equal-sized groups). It is unclear whether the presented calibration plots are for the 5-year overall survival endpoint or the 8-year overall survival endpoint. Pan et al then reported the calibration intercept and slope. The intercept indicates whether the predictions are systematically too low or too high (often referred to as “calibration-in-the-large”), whereas the value of the slope should be approximately 1 (values <1 reflect overfitting). Unfortunately, the estimates of the calibration intercept and slope reported by Pan et al are incorrect. The authors have incorrectly taken the 4 x and y coordinates (from each calibration plot) and fit linear regression lines and taken the values from these regression lines as estimates of the calibration intercept and slope. Using a different number of groups will clearly change the linear regression model and, therefore, the estimates of the calibration intercept and slope.

Unlike models based on logistic regression, calculating the calibration slope and intercept for models based on Cox regression is not straightforward, and in fact only the slope can be calculated by fitting a single-term regression model of the form log(hazard\(y = 1\)) = \(h_0 + b \times \text{linear predictor}\) (in which \(h_0\) is the baseline hazard at a single time point [eg, 5 years], the linear predictor is the sum of the regression coefficients multiplied by the individual patient values, and \(b\) is the estimate of the calibration slope). To improve the calibration plot, the authors could have overlaid the plot with a smoothed regression line using flexible adaptive hazard regression. This enables readers to judge agreement across the spectrum of predictions (ie, for every 100 patients given a prediction of x%, the observed number of patients with the outcome is close to x).

Our next point refers to the validation analyses. The authors repeated the steps taken during the development of the model on the validation data and concluded that the c-index was similar and consistent with the proposed nomogram. This is flawed and does not constitute validation; rather, it merely creates a new model that is then incorrectly (as identified earlier) validated.

Our final comment, and arguably the most important aspect, is the presentation of the model so that other clinicians and investigators can use or validate the model. Pan et al presented 2 nomograms, presumably to aid in the uptake of the model. Unfortunately, although a nomogram can be used on individual patients, for others who wish to evaluate the prediction model (for which the nomogram is a graphical presentation of the underlying Cox regression model), the full prediction model needs to be reported. The authors reported the hazard ratios for the predictors, but did not report the baseline hazard ratios at 5 and 8 years to enable predictions to be made. Without this information, investigators are unable to validate the prediction model in their own data.

FUNDING SUPPORT
No specific funding was disclosed.
CONFLICT OF INTEREST DISCLOSURES
The authors made no disclosures.

REFERENCES

Gary S. Collins, PhD
Centre for Statistics in Medicine
Botnar Research Centre
University of Oxford
Oxford, United Kingdom

Yannick Le Manach, MD, PhD
Department of Anesthesia
Department of Clinical Epidemiology and Biostatistics
Michael DeGroote School of Medicine
Faculty of Health Sciences
McMaster University;
Perioperative Research Group
Population Health Research Institute
Hamilton, Ontario, Canada

DOI: 10.1002/cncr.30347, Published online Month 00, 2016
in Wiley Online Library (wileyonlinelibrary.com)