

TITLE: Nomograms need to be presented in full

Gary S. Collins, *professor of medical statistics*

Centre for Statistics in Medicine, Botnar Research Centre,
University of Oxford, Windmill Road, Oxford OX3 7LD, United Kingdom
Email: gary.collins@csm.ox.ac.uk

Yannick Le Manach, *assistant professor*

Departments of Anesthesia & Clinical Epidemiology and Biostatistics, Michael DeGroote
School of Medicine, Faculty of Health Sciences, McMaster University and the Perioperative
Research Group, Population Health Research Institute, Hamilton, Canada
Email: yannick.lemanach@phri.ca

750 words (max 750 words)

The authors report no conflict of interest.

Letter Re: Prognostic nomogram for refining the prognostication of the proposed 8th edition of the AJCC/UICC staging system for nasopharyngeal cancer in the era of intensity-modulated radiotherapy. By Pan et al, Cancer 2016; DOI: 10.1002/cncr.30198

We read with great interest, the recent paper by Pan and colleagues describing the development of a nomogram to predict the 5 and 8-year survival probability for patients with nondisseminated nasopharyngeal cancer¹. 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 two main characteristics on a particular dataset are discrimination and calibration as noted in the recent TRIPOD reporting guideline for prediction model studies². Discrimination is defined as the proportion of all patient pairs in which the predictions and outcomes are concordant, commonly measured by the *c*-index, whilst calibration reflects how accurate the predictions from the model reflect the survival in the observed data. Whilst the authors have 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. Whilst bootstrapping is the preferred and recommended approach for internal validation, it is important that all model building steps (including the flawed univariate analyses³ and backwards elimination). If these steps are omitted and only the ‘final model’ is included, then the resulting bias-corrected *c*-index will itself be biased and overoptimistic. It is unclear, what exactly the authors have done, as the reporting is somewhat brief, but we suspect that only the final model has been bootstrapped.

The authors have presented the traditional and commonly seen calibration plot, of predictions against observed outcomes by fourths of predicted risk (i.e., four equal sized groups), though usually this is done by tenths of predicted risk (i.e., 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. The authors then report 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’), whilst the value of the slope should be around 1 (values less than 1 reflect overfitting)⁴. Unfortunately, the estimates of the calibration intercept and slope reported by Pan et al are incorrect. The authors have incorrectly taken the four x and y coordinates (for each calibration plot) and fit linear regression lines and taken the values from this 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 $\log(\text{hazard}(y = 1)) = h_0 + b \times \text{linear predictor}$ (where h_0 is the baseline hazard at a single time point, e.g., 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)^{5, 6}. To improve the calibration plot, the authors could have overlaid the plot with a smoothed regression line using flexible adaptive hazard regression^{2, 78}. This enables readers to judge agreement across the spectrum of predictions (i.e. 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 have repeated the steps taken during the development of the model on the validation data concluded that c-index is similar and consistent with the proposed nomogram. This is flawed and does not constitute validation, it merely creates a new model which 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. The authors have presented two nomograms, presumably to aid in the uptake of the model. Unfortunately, whilst a nomogram can be used on individual patients, for other who wish the 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 have reported the hazard ratios for the predictors, but have not reported the baseline hazard 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.

REFERENCES

1. Pan JJ, Ng WT, Zong JF, et al. Prognostic nomogram for refining the prognostication of the proposed 8th edition of the AJCC/UICC staging system for nasopharyngeal cancer in the era of intensity-modulated radiotherapy. Cancer. 2016.

2. Collins GS, Reitsma JB, Altman DG, Moons KGM. Transparent Reporting of a multivariable prediction model for Individual Prognosis Or Diagnosis (TRIPOD): The TRIPOD Statement. *Ann Intern Med.* 2015;162: 55-63.
3. Sun GW, Shook TL, Kay GL. Inappropriate use of bivariable analysis to screen risk factors for use in multivariable analysis. *J Clin Epidemiol.* 1996;49: 907-916.
4. Steyerberg EW, Vickers AJ, Cook NR, et al. Assessing the performance of prediction models: a framework for traditional and novel measures. *Epidemiology.* 2010;21: 128-138.
5. Steyerberg EW. Clinical prediction models: a practical approach to development, validation, and updating. New York: Springer, 2009.
6. Crowson CS, Atkinson EJ, Therneau TM. Assessing calibration of prognostic risk scores. *Stat Methods Med Res.* 2013.
7. Moons KGM, Altman DG, Reitsma JB, et al. Transparent Reporting of a multivariable prediction model for Individual Prognosis Or Diagnosis (TRIPOD): Explanation and Elaboration. *Ann Intern Med.* 2015;162: W1-W73.
8. 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: 214-226.