

TITLE: When data are scarce, model validation should be efficient

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. Tel: +44 (0)1865 223460

The author reports no conflict of interest.

Letter Re: Dólera-Moreno C, Palazón-Bru A, Colomina-Climent F, Gil-Guillén VF. Construction and internal validation of a new mortality risk score for patients admitted to the intensive care unit. Int J Clin Pract 2016; 10.1111/ijcp.12851

In their recent paper, Dólera-Moreno and colleagues describe the development of a model to predict mortality of patients admitted to ICU [1]. However, there are design aspects in the development and validation of their model which are of concern.

Evaluating the predictive accuracy of a model is an important, if not the most important, aspect of a prediction model. When developing a new model, investigators should attempt to demonstrate its predictive ability in the most robust and fair manner. When data are limited, deciding how best to use the available data to develop and evaluate (internally validate) the model is crucial. Dólera-Moreno and colleagues randomly split their data into separate development and validation cohorts. This approach, whilst common, has been shown to be inefficient, a waste of precious data and a weak test of model performance [2]. Randomly splitting data, reduces the sample size, so that fewer data are used to develop the model and the validation (or test) dataset is then also (too) small (26 events as in this study is too far for model evaluation; 100 events are recommended [3]). Furthermore, randomly splitting data merely creates two very similar datasets (as illustrated by the non-statistically significant p-values) and hardly a meaningful test of the model. More efficient approaches would be to use the entire dataset to develop the model and then use methods such as bootstrapping to evaluate the model; adjust for optimism and overfitting. Problems of the small validation dataset is also observed from the calibration plot, which is problematic in itself as smoothed calibration plots are usually preferred and recommended [4, 5], but due to the small sample size, the first two groups (of five) contain no events, with the majority of the events in the upper group, and thus assessing calibration is difficult.

My next point concerns sample size. Sample size considerations for developing a prediction model are based on the concept of events-per-variable (EPV), with an EPV of at least 10 usually accepted as the criteria to minimize model overfitting. EPV was mentioned in the study, but the authors incorrectly asserted that it is the number of predictors in the final model (six in their model) that the EPV is based on rather than the number of predictors considered (12 in the current study). As such, the sample size should comprise at least 120 (12 predictors * 10) events to build the model (91 deaths were observed in the current study). Therefore, the author conclusions that the statistical power (which is irrelevant in prediction model studies) of their study is close to 100% and that the prognostic value is *great* (sic) is nonsensical. Additional (independent) external validation studies assessing discrimination, calibration and clinical utility would be required to conclude the predictive ability and usefulness of the model [6]. The recent TRIPOD Statement describes all the information needed to report in a published

prediction model study [5], and the accompanying Explanation & Elaboration paper describes key methodological considerations [7].

REFERENCES

1. Dolera-Moreno C, Palazon-Bru A, Colomina-Climent F, Gil-Guillen VF. Construction and internal validation of a new mortality risk score for patients admitted to the intensive care unit. *International journal of clinical practice* 2016.
2. Steyerberg EW, Harrell Jr FE, Borsboom GJJM et al. Internal validation of predictive models: efficiency of some procedures for logistic regression analysis. *J Clin Epidemiol* 2001; **54**: 774-81.
3. 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.
4. Van Calster B, Nieboer D, Vergouwe Y et al. A calibration hierarchy for risk models was defined: from utopia to empirical data. *Journal of Clinical Epidemiology* 2016; **74**: 167-76.
5. 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.
6. Steyerberg EW, Vickers AJ, Cook NR et al. Assessing the performance of prediction models: a framework for traditional and novel measures. *Epidemiology* 2010; **21**: 128-38.
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.