

## **TITLE: Uninformative and misleading comparison of EuroSCORE and EuroSCORE II**

**Gary S. Collins**, *associate professor*

Centre for Statistics in Medicine, Botnar Research Centre,  
University of Oxford, Windmill Road, Oxford OX3 7LD, UK;

Email: [gary.collins@csm.ox.ac.uk](mailto: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](mailto:yannick.lemanach@phri.ca)

The author reports no conflict of interest.

In their recent paper, Kieser and colleagues compared the predictive performance of EuroSCORE against its successor EuroSCORE II in a consecutive series of isolated CABG patients with total arterial grafting by a single surgeon [1]. Whilst comparative validation studies such as these are extremely important, we have a number of concerns on the study design and analysis, for which we will highlight only a couple of issues, that question how anyone can meaningfully interpret their findings.

Validation studies are an important aspect of evaluating a risk score, and methodological rigor and transparent reporting are key to ensure the results are meaningful and interpretable. An important aspect often overlooked in validation studies is study design. Recommendations for sample size is that a minimum of 100 (and preferably 200) events (i.e., deaths) should be included in the study so that model performance and in particular calibration can be adequately assessed [2, 3]; a value much higher than observed 36 deaths in the Kieser study.

The authors correctly assert that the widely used Hosmer-Lemeshow test is problematic for assessing calibration and should be avoided; it neither assesses direction nor magnitude of calibration. The recent TRIPOD Statement for reporting risk scores cautions against its use with preference for calibration plots [4, 5]. However, the calibration plot of Kieser is also of limited usefulness (ignoring the annoyance that the two axes are not on the same scale; the y-axis is squashed), grouping by predicted risk also suffers from limitations including groups with no events and deciding how many groups. In the study by Kieser, we can observe four out of the 10 groups have no deaths, thereby making the interpretation of their calibration plot somewhat difficult. A calibration plot should indicate of those with a predicted risk of x% how many patients died (which for a well calibrated model should be close to x observed deaths); this information is not presented or inferable from their Figure. Recommendations are that loess smoothed calibration plots (preferably with confidence intervals) should be presented so that calibration can be examined across the range of predicted values [6]. The calibration plot

can then be supplemented with estimates of the calibration slope and intercept (as calculated by Kieser).

A final comment is related to the temporal analysis, as previously noted given the very small number of deaths (median of 4 per time-period between 2003 and 2014) and an analysis which doesn't actually investigate model performance, very little can be concluded whether the calibration *evolved* over time. Given these concerns, and others, including unclear handling of the large amount of missing data for ejection fraction, or whether a small single surgeon case series is at all interesting beyond the surgeon himself, the conclusions have limited utility and should be interpreted with a large *pinch of salt*.

## References

- [1] Kieser TM, Rose MS, Head SJ. *Comparison of logistic EuroSCORE and EuroSCORE II in predicting operative mortality of 1125 total arterial operations*. Eur J Cardiothorac Surg 2016.
- [2] Van Calster B, Nieboer D, Vergouwe Y, De Cock B, Pencina MJ, Steyerberg EW. *A calibration hierarchy for risk models was defined: from utopia to empirical data*. Journal of Clinical Epidemiology 2016.
- [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] Collins GS, Reitsma JB, Altman D, G., Moons KG. *Transparent Reporting of a multivariable prediction model for Individual Prognosis Or Diagnosis: The TRIPOD statement*. Ann Intern Med 2015;**162**:55-63.
- [5] Moons KGM, Altman DG, Reitsma JB, Ioannidis JPA, 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**:W1-W73.
- [6] Austin PC, Steyerberg EW. *Bootstrap confidence intervals for loess-based calibration curves*. Stat Med 2014;**33**:2699-700.