Uninformative and misleading comparison of EuroSCORE and EuroSCORE II

Gary S. Collins* and Yannick Le Manach†

* Centre for Statistics in Medicine, Botnar Research Centre, University of Oxford, Oxford, UK
† Department of 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

Received 4 May 2016; accepted 4 July 2016

Keywords: Perioperative issues and risk analysis • Coronary disease

In their recent paper, Kieser et al. [1] compared the predictive performance of EuroSCORE against its successor EuroSCORE II in a consecutive series of isolated coronary artery bypass graft patients with total arterial grafting by a single surgeon. Although 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. Recommendation 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 the 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 assesses neither 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 et al. 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 et al., we can observe that 4 out of the 10 groups have no deaths, thereby making the interpretation of their calibration plot somewhat difficult. A calibration plot should indicate 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 in 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 et al.).

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 that does not 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


* Corresponding author. Centre for Statistics in Medicine, Botnar Research Centre, University of Oxford, Windmill Road, Oxford OX3 7LD, UK. Tel: +44-1865-223460; e-mail: gary.collins@csm.ox.ac.uk (G.S. Collins).

doi:10.1093/ejcts/ezw262