EuroSCORE II and the art and science of risk modelling

Samer A.M. Nashef a,*, Linda D. Sharples b, François Roques c and Ulf Lockowandt d

a Papworth Hospital, Cambridge, UK
b Medical Research Council Biostatistics Unit, Cambridge, UK
c CHU Fort de France, Martinique, France
d Karolinska Hospital, Stockholm, Sweden

* Corresponding author. Papworth Hospital, Cambridge CB23 3RE, UK. Tel: +44-1480-364299; fax: +44-1480-364744;
e-mail: sam.nashef@papworth.nhs.uk (S.A.M. Nashef).

Keywords: Risk assessment · EuroSCORE · Cardiac surgery · Mortality

EuroSCORE II [1] appeared in this journal in April 2012 alongside editorials by Sergeant et al. [2] and Kappetein and Head [3]. This month, Chalmers et al. [4] report an early validation of the model. We thank these authors for constructive and sometimes kind comments about our model, and appreciate the opportunity to respond.

APPLICABILITY

Chalmers et al. [4] applied the model to a 5500-patient cohort, concluded that EuroSCORE II is globally better calibrated and found better overall discrimination with a C-statistic of 0.79 (old model 0.77), with best performance in mitral (0.87) and coronary (0.79) surgery and weakest in isolated aortic valve replacement (0.69), marginally better than the old model (0.67).

Pooling contemporaneous multi-institutional data provides optimal model validation. A single institution performing exceptionally in one type of surgery may perceive a lack of fit in the entire cohort, as Chalmers indeed found (Hosmer-Lemeshow P-value <0.05) for EuroSCORE II overall, but not in any subdivision. Considering this and other limitations of a single-institution study, the model has achieved its objective of better calibration and discrimination in global cardiac surgery. EuroSCORE II risk-stratifies using factors including operation type. Lower discrimination when these are neutralized is therefore unsurprising. We advise caution in applying the model to narrow patient subsets.

In the editorials [2, 3], many comments reiterate issues already addressed in the discussion section of the original paper [1]. As Sergeant states, EuroSCORE is widely used and has surgeons’ confidence. There has been misuse, particularly in evaluating patients for transcatheter aortic valve implantation (TAVI), and in predicting non-mortality events. The former is avoidable by our recommendation for using risk-adjusted mortality ratios (RAMR) [1], but that is not enough. Conventional surgery risk factors differ from those of TAVI, and we shall explore our TAVI data to illuminate this. Predicting non-mortality events using EuroSCORE has often succeeded, but its purpose remains to evaluate expeditiously the risk of death.

SELECTING AND HANDLING RISK VARIABLES

We agree with Kappetein and Head [3] in desiring to restrict parameters numbers, where selection is a compromise between the ideal and the sensible. Many parameters satisfy the criteria of being common, available, objective, falsification-resistant and user-credible. More parameters improve risk prediction, but model usability suffers. Fewer parameters simplify the model, but user-credibility diminishes if common factors are ignored. Judgment guides decision, and EuroSCORE’s success indicates that we got that about right. We believe to have done so again in EuroSCORE II.

Including data from disparate centres increases generalizability but introduces variation. During development, we explored inter-centre variation using random effects models, finding little impact on either the score for the ‘average’ centre, or overall prediction, with scores differing by 0.2%. Here, the random effects model did not improve prediction. We believe EuroSCORE II can be widely used, but recommend RAMR adjustment, especially if mortality differs substantially from the average (3.77%) or if the RAMR mandates it.

Both editorials [2, 3] address risk factors and procedure-specific interactions. We conducted many more analyses than we published, but included few interactions. The power to detect interactions is much lower than the power to detect main effects and imprecise interaction estimates can decrease predictive accuracy.

Choosing between general and procedure-specific models is difficult. Seeking a usable score, we compromised between generalizability and specificity. Current practice contains many combined and complicated procedures: a general score is more likely to be usable and robust than a specific score based on diminishing numbers of isolated procedures. More variables undoubtedly affect outcome, but the price is more missing data and less usability, as recognized by Kappetein: ‘inclusion of a greater number of variables increases the risk of errors that can be caused by differences in the interpretation of definitions, typing errors or conflicting chart information’. There is an inverse relationship between the amount of data demanded and their quality.
We intended to retain continuity for creatinine clearance and pulmonary hypertension, but skewed creatinine clearance distribution was further complicated by dialysis patients, with lower risk than in renal dysfunction not on dialysis yet, hence change-point models. In pulmonary artery pressure, only elevated readings are routinely recorded and we compromised to achieve inclusion while retaining continuity.

**DATASET INTEGRITY**

Only 418 cases were excluded for missing ‘compulsory’ risk factor and hospital mortality data. Of these, 384 lacked a single measurement in centres where presumably it was not routinely collected. We confidently assume that these were missing at random. Only 34 cases had sporadic missing risk factor data, where imputation in over 22,000 cases had no discernible benefit in either bias reduction or precision. The 295 sporadic cases with missing outcome data neither occurred in a specific period, nor followed covariate patterns. Mortality in these seven centres (3.57%) resembled the overall rate (3.77%). Excluding them yielded similar overall predictions and a C-statistic of 0.8095, identical to imputing missing outcomes under a random assumption, with minuscule standard error changes [5], rendering multiple imputation unhelpful.

Sergeant states that data accuracy is ‘devalued by the observation of double, triple and quadruple submissions of the same record’, but 147 duplicate entries out of 23,451 records is 0.62%, far below expected in such large, multicentre data. We absolutely disagree that data quality is undermined by this observation.

**THE OUTCOME MEASURE**

Both editorials addressed hospital mortality as the outcome, a limitation indeed. Attempting to capture 30- and 90-day survivals succeeded in only 56 and 43% of patients. Centres contributed to EuroSCORE II voluntarily against competing clinical, financial and administrative pressures. Participation is a testimony to better data systems than most. If only half of these centres record 30- and 90-day survival, the fraction will be even smaller in non-participating centres. Ideally, follow-up of all patients ad infinitum would provide rich information about long-term cardiac surgery outcomes, but currently this is not available. Encouraging such data collection is a task for specialist and professional bodies but we are in no position to mandate it. Sergeant harshly criticizes participating centres, stating that they ‘failed to complete the follow-up and therefore, in their participation in this project and the ethics of their profession’. We disagree. Participants deserve praise, not opprobrium, for their hard work. We take this opportunity to thank them for their sterling effort and benevolence in making this project possible.

Sergeant et al. [2] questions our analytic process as cardiac surgical mortality is rare. We agree that the asymptotic properties of statistical methods are less reliable for rare events, but in our large sample size such properties should hold. Multiple analyses confirmed good model fit (Hosmer–Lemeshow), good prediction (cross-validation), good stratification (receiver operating characteristic curves, C-statistic) and good calibration (Hosmer–Lemeshow, overall risk). Finally, the strongest test of EuroSCORE II was application to external data created by randomly splitting the database into developmental and validation subsets, with a most satisfactory result.

**THE NAME**

The project group and most participating centres are European and the model is a development of the original EuroSCORE, so the name is appropriate.

**THE MODEL**

EuroSCORE was a good model, bringing widespread use of monitoring to this specialty. Cardiac surgical performance improved partly as a result of this, leaving the model uncalibrated. EuroSCORE II is superior, being better calibrated for the current era, with better discrimination. It was built on sound data and analysis. Other models could also be constructed from the same data. We selected this model on the basis of performance, credibility, usability and our own clinical and statistical judgement. It is not perfect. We agree with Chalmers et al. [4] that there is ‘room for improvement in risk modelling’, and welcome developments and improvements from our critics and interested parties.

**REFERENCES**