EuroSCORE II, illum qui est gravitates magni observe*

Paul Sergeant*, Bart Meuris and Matteo Pettinari

Department of Cardiac Surgery, Gasthuisberg University Hospital, Leuven, Belgium

* Corresponding author. Department of Cardiac Surgery, Gasthuisberg University Hospital, Herestreet 40, 3000 Leuven, Belgium. Tel: +32-16-344219; fax: +32-16-344616; e-mail: paul.sergeant@uzleuven.be (P. Sergeant).

Keywords: Risk assessment · EuroSCORE · Cardiac surgery · Mortality

“The science of risk prediction forces us to focus on three limitations of this approach and evaluate this manuscript versus those criteria. The sparsity in the reference class improves with the actuality, the size and the richness of the database, most of all in the density of the extremes of variability. In the presence of a sparsity in the reference class, statistical forecasting becomes unreliable. This unreliability is slightly improved through judgment adjustment.

Secondly, the inappropriateness of the reference class makes the statistical forecasting unreliable. An inappropriate reference does not include the essential variability or the appropriate outcome.

Thirdly, the inappropriate statistical model (e.g. oversimplified models, poor calibration etc.) destabilizes completely statistical forecasting. There is mixed evidence about judgement adjusting the model. The accuracy of a mathematical model evaluates calibration, discrimination and a combination of both. The calibration describes how well the predicted probabilities agree with the actual observed risk. The Hosmer–Lemeshow statistic compares proportions but is most certainly imperfect. A non-significant Hosmer–Lemeshow test means that there is no evidence of bad calibration, it does not mean that there is good calibration. The discrimination describes how well a model separates black from white. The methods used are the sensitivity (true positive), the specificity (true negative), the positive predictive value (PPV), the negative predictive value (NPV), the misclassified, the ROC [3] (receiving operating characteristics) and the C-statistic (concordance index). The combination of both calibration and discrimination is best described using the likelihood statistics, the \( R^2 \) [4] and Brier [5] scores.
Continuous variables could indeed easily be transformed in a search for an optimal relation between the outcome event and the available variability. In addition, risk is never residual in the average value of a variable but in the density at the outliers. As an example, to possibly allow body mass index (BMI) to enter into a final model with a correct transformation and coefficient, the reference dataset needs sufficient patients with a BMI <20, or <15 or with a BMI >35, >40, >45, >50. This density information is not transparent in this manuscript; therefore, any variable selection or elimination cannot be discussed and evaluated.

The manuscript does not give sufficient information about missing values, since the authors classify variables as ‘compulsory’ and ‘non-compulsory’ after the analysis. They identify missing values in compulsory variables as important and not-important in non-compulsory variables. They also give the impression of not having taken action to improve the missing data or impute the missing data using any of the available methods.

### THE EUROSCORE II REFERENCE DATABASE

The reference class used for EuroSCORE II is an extremely large database of 24 385 records (patients), originating from volunteering units. The project managers need to be applauded for the collection and connection of so many records. There has been no external validation of this dataset, not even of a random sample. The quality is therefore dependent on the repeated testimony of the units responsible. This testimony is devalued by the observation of double, triple and quadruple submissions of the same record.

The domain studied is adult cardiac surgery and the project managers have chosen a common predictive model covering the complete domain. This is a philosophical and pragmatic decision that has both benefits and limitations. The system is indeed applicable to a complete unit of adult surgery, but the variables’ list loses specificity. Indeed different variables play a role, possibly with a different coefficient, in different pathologies or surgical therapies. Echocardiographic data, as an example, play a dominant role in valve surgery but are possibly less important in coronary surgery. Patients in cardiogenic shock or cardiopulmonary resuscitation, as in aortic dissections, coronary bypass surgery or endocarditis, demand a completely different list of variables [6] never encountered in traditional scoring systems. The project managers could have responded to their philosophical decision by including, for all patients, a list of variables from different subdomains of adult cardiac surgery that would possibly only play a role in some of them. They decided not to respond and thereby reduced the quality of their global reference class versus the outcome event.

The selection and the format of the collected variability assures the richness of the reference class. The variable list has been minimally improved with some additional variables, but considerable variability remains excluded (quality of life, frailty, mental reserves etc.). This is most certainly a missed opportunity that could have revolutionised cardiac surgery.

One of the major limitations of EuroSCORE I was the parsimoniously dichotomous (yes/no) registration of variability, even with the availability of validated, possibly transformed, continuous presentations. Except for renal function, it is unclear from the manuscript if all continuous variability is registered in a continuous format since the final model repeats the use of dichotomous presentations for this variability. For example, by not presenting pulmonary function in the format of vital capacity (or % of normal) or one-second value (or % of normal) and ventricular function in the format of ejection fraction, end-diastolic pressure, end-systolic/diastolic volumes, a repeated opportunity has been lost to enrich the reference dataset.

### THE EUROSCORE OUTCOME EVENT

EuroSCORE II has the ambition of predicting early mortality, so the analysed outcome event is of the utmost importance for the quality of the reference class. The authors have correctly chosen the most discrete and serious event: mortality. Early mortality after adult cardiac surgery extends for coronary bypass patients [7] up to 3 months and even further for valve patients in follow-up. The authors, have therefore correctly chosen the 90-day observation interval as a primary outcome variable. The reference database has 55.4% missing information for this appropriate interval, and even for the secondary non-appropriate end-point of 30-day 43.4% of the information is missing. This information destabilizes the reader and raises a series of questions. The 160 participating units only needed to follow-up on average 140 patients each. To complete a follow-up of a patient, an average of three contacts per patient are needed and an average time of <1 h per patient [8]. The authors have consequently redefined their outcome interval into what they have as available information, namely the biased hospital stay in the primary hospital. This is incomplete, biased and inadmissible.

It is indicative of a failure of the volunteering units and their unit’s responsibilities in their participation of this project. A possible ethical problem in cardiac surgery is magnified to society through this project.

These and the previous observations classify the reference class as sparse and inappropriate for forecasting rare events. Further reading of this manuscript places the reader and society at risk of false interpretation. Therefore, this model should not be used, as such, for quality monitoring or comparison, for differential therapy or informed consent and most certainly not for the public reporting of medical performance.

### IMPROVEMENTS IN CALIBRATION AND DISCRIMINATION

Appropriate statistical methods were used on this sparse reference class; we therefore remain reluctant to discuss the inferences proposed. The authors have been able to recalibrate

---

**Table 1**: Methods for prediction of rare events

<table>
<thead>
<tr>
<th>Method</th>
</tr>
</thead>
<tbody>
<tr>
<td>Statistical forecasting</td>
</tr>
<tr>
<td>Expert judgement</td>
</tr>
<tr>
<td>Structured judgemental decomposition</td>
</tr>
<tr>
<td>Structured analogies</td>
</tr>
<tr>
<td>Statistical forecasting with judgemental intervention or adjustment</td>
</tr>
<tr>
<td>Delphi</td>
</tr>
<tr>
<td>Prediction markets</td>
</tr>
<tr>
<td>Scenario planning</td>
</tr>
</tbody>
</table>
EuroSCORE II, although the Hosmer–Lemeshow test indicates a possible model-fitting problem. Concerning discrimination, this new model reaches an ROC of 0.80, versus 0.789 (logistic) and 0.789 (additive) for the previous versions. So, no significant improvement was observed. In fact, an ROC of that range is insufficient to make any valid statement about differences in risk or performance of units or individual medical professionals. These scores are probably obtained through a rather accurate prediction of the survivors but an inaccurate prediction of the patients suffering the event. This would have been clarified, had the authors given the additional criteria for discrimination and calibration as the sensitivity, specificity, PPV, NPV, misclassification and the $R^2$ and Brier scores.

**CONCLUSION**

The EuroSCORE II project has applied appropriate statistical methods on a sparse reference database. The volunteering units have failed to complete the follow-up and, therefore, in their participation in this project and possibly in the ethics of their profession. Any patient undergoing a resource-expensive and risk-related cardiac surgical procedure is entitled to a complete follow-up. This would confirm the appropriateness of the differential therapy and of the therapeutic process.

**Conflict of interest:** none declared.

**REFERENCES**