Single centre, single domain validation of the EuroSCORE on a consecutive sample of primary and repeat CABG

Paul Sergeant*, Erik de Worm, Bart Meyns

Cardiac Surgery Department, Gasthuisberg University Hospital, 3000 Leuven, Belgium

Received 25 September 2000; received in revised form 11 June 2001; accepted 24 September 2001

Abstract

Objectives: Intra- and interdepartmental benchmarking require scoring systems with excellent performance on several properties: discrimination (resolution), reliability (calibration) and stability over the complete spectrum of peri-procedural risk. This single centre, single domain study validates the European system for cardiac operative risk evaluation (EuroSCORE) on an independent sample of primary and repeat coronary artery bypass grafting (CABG) patients and will evaluate these different properties. Methods: The study population is a consecutive series of 2051 isolated primary and repeat CABG patients, inclusive of patients in cardiogenic shock or resuscitation, operated on in a single institution from January 1997 to July 2000. The age of the patients was 66 ± 9 years, 77% were males and 7% were repeat procedures. The EuroSCORE was 5.0 ± 3%, with a range from 0 to 22. The studied event was in-hospital death, defined as mortality during hospital stay, which was unlimited in time and included a stay in a secondary hospital without discharge home. Results: The EuroSCORE predicted 102 deaths versus 81 deaths observed (P = 0.14, Fisher exact test). The EuroSCORE described only 20% of the variance of in-hospital mortality. The EuroSCORE created an area under the receiver operating characteristic curve of 0.83 ± 0.03. The highest discriminative accuracy was obtained with 8% EuroSCORE risk (only 64% sensitivity and 87% specificity). Further exploration identified an over score in the EuroSCORE range 0–8 (57%, P = 0.0001). There was an equal score (22%, P = 1) in the range 9–11, but an under score in the range 12–22 (−133%, P = 0.003). Conclusions: On the condition that these single centre results could be extended to any European cardiac surgery centre, it can be concluded that the overall acceptable performance of the EuroSCORE is the result of an over score in the lower risk and insufficient correction in the higher risk spectrum. The EuroSCORE is probably refined enough for improved informed consent versus aggregated results but should only be used for inter-institutional benchmarking with great caution, preferably below the 12% risk pivot.

© 2001 Elsevier Science B.V. All rights reserved.

Keywords: Coronary artery bypass grafting; EuroSCORE; Scoring system; Validation; Quality control

1. Introduction

Health authorities, and health service purchasers and providers demand the highest quality of care. They use this to optimize their limited resources. Quality of care is an amalgam of elements, but in-hospital mortality is certainly one of major importance and is very apparent. The analysis and probable comparison of this in-hospital mortality mandates the availability of tools to correct for variability in the patient spectrum.

Patients also demand the highest quality of care but in addition they have received legal guarantees to be informed of their individual risk. This knowledge of the individual risk is also essential to the attending physician in evaluating correctly the appropriateness of the therapy. These partners in health care trust the scientific community to provide simple prediction systems which are stable for larger samples of patients but also stable for the individual patient.

We have studied earlier the overall behaviour [1] and limitations of complex domain-specific prediction models in coronary surgery in early and late follow-up. The purpose of this manuscript is to study the overall and spectral behaviour of a simple but recent non-domain-specific prediction model of in-hospital mortality, the European system for cardiac operative risk evaluation (EuroSCORE) [2,3], in a large independent single centre population of coronary surgery patients with a wide spectrum of risk.

2. Materials and methods

2.1. Study population

A consecutive series of 2051 primary and repeat isolated
coronary bypass patients were operated on at the K.U. Leuven University Hospital between January 1997 and July 2000 and form the study sample of this analysis. This sample includes rescue procedures and patients in cardiopulmonary resuscitation. Their in-hospital mortality (observed value) was compared with the EuroSCORE (Table 1) prediction (predicted value). In-hospital mortality has been defined as mortality during hospital stay, which was unlimited in time and included a stay in a secondary hospital without discharge home. One patient was readmitted after a few hours (same calendar day as discharge) and died at readmission; he was considered to be an in-hospital death. The mean age within the study sample was 65.5 ± 9 years with 15% of the patients older than 75 years and 77% of the male gender. There were 151 repeat procedures in the study sample. Orally-treated diabetes was observed in 9% and insulin-treated in 6% of the patients. Single-vessel disease was active in 4%, two-vessel disease in 24% and triple-vessel disease in 72% of the patients.

2.2. Scoring system

The mean EuroSCORE value was 5.0 ± 4% (5.1% for extra-corporeal and 4.5% for off-pump procedures). See Table 2 for the number of patients at risk for each EuroSCORE value. Off-pump coronary surgery was performed in 450 patients. The mean duration of the extra-corporeal circulation (ECC) and of the aortic cross-clamp in the other patients was 64.2 ± 30 and 19 ± 11 min, respectively. The mean number of grafts was 2.6 ± 0.7 (arterial grafts 1.4 ± 0.6) and the mean number of anastomoses was 2.9 ± 0.9 per patient. The mean hospital stay of the hospital survivors was 13 ± 13 days (range 4–184 days).

2.3. Surgical technique

The surgical procedure was performed with or without cardiopulmonary bypass, but always under normothermia and with Lidoflazine pre-treatment. As many anastomoses as possible were constructed on the beating supported heart, thereby shortening the intermittent aortic cross-clamp times. Procedures on rotary support systems were considered as ECC procedures. Off-pump surgery was limited to single
and double grafts to the anterior and proximal inferior regions of the heart until October 1999, when it became the technique of choice for all procedures.

2.4. Statistical analysis

The % difference between the predicted and observed hospital mortality was calculated using the formula: 

\[(\text{predicted deaths} - \text{observed deaths}) \times 100/\text{predicted deaths}\]

The Fisher exact test was used for the contingency tables and logistic regression for the evaluation of the EuroSCORE value as an incremental predictor. The contingency tables were formed comparing the number of predicted events and N-predicted events versus observed events and N-observed events.

A receiver operating characteristic curve (ROC) [4,5] evaluated the predictive performance (discriminatory power) of the EuroSCORE. This was created by using each EuroSCORE (0–22) as a theoretical cut-off point to predict in-hospital mortality. The sensitivity and the specificity of the prediction were calculated for each EuroSCORE. The sensitivity was then plotted (Fig. 1) versus the 100-specificity and the points were interconnected. The area under this curve is a measure of the discriminatory power of the test.

The maximum efficiency of the EuroSCORE was identified as the point at which the rate of change of the tangent to the curve is maximal, the place where the least false negative and false positive results are observed. When one predicts a lower value, then the true positive fraction and sensitivity will increase. On the other hand the false positive fraction will also increase and therefore the true negative fraction and specificity will decrease. When one predicts a higher value, then the false positive fraction will decrease with increased specificity but on the other hand the true positive fraction and sensitivity will decrease. This was done for the complete group and for the patients operated on using extra-corporeal circulation. Insufficient events were present in the off-pump group (minimum 50 events needed [6]) for a subgroup analysis.

After the first step in the spectral analysis with the evaluation for every single risk category (Table 2 and Fig. 2), the patients were grouped (Table 3) in a second step of data exploration. In the third step a cumulative approach is tested. A gradually increasing sample size is created, starting with 0% predicted risk and in a stepwise manner adding one additional risk category. Each of these samples is considered a separate sample and for each sample the difference (Fig. 3) between predicted and observed mortality is calculated.

3. Results

The overall observed in-hospital mortality was 81 patients (3.9%). The interval between surgery and death ranged from 1 to 139 days, and 21 deaths were observed beyond 30 days of hospital stay.

The EuroSCORE predicted 101.8 deaths (5.0%). The Fisher test \(P\) value of the total number of observed versus expected hospital deaths was 0.14. With the extra-corporeal procedures 69 deaths were observed versus 81.6 predicted \((P = 0.32)\) and with the off-pump procedures 12 deaths were observed versus 20 predicted \((P = 0.21)\).

Logistic regression identified the EuroSCORE risk prediction as an incremental predictor \((P < 0.0001)\) but describing only 20% of the variability.

The EuroSCORE created an area under the ROC curve of \(0.83 \pm 0.03\) for the complete dataset \((0.81 \pm 0.03\) for the extra-corporeal procedures). The highest discriminative accuracy in predicting in-hospital death for the complete dataset was obtained with 8% EuroSCORE risk \((64\%\,\text{specificity and } 87\%\,\text{specificity})\).

The spectral analysis is performed in a stepwise manner: individual, grouped and then cumulative.

Table 2 and Fig. 2 are the first steps in the spectral analysis and present the number of patients at risk, the number of predicted deaths, the number of observed deaths and the % observed mortality for every EuroSCORE risk value.

The reduced number of events in the low-risk groups and the small sample sizes in the large risk groups induce statistical limitations. These are avoided by grouping the patients in Table 3 in three larger samples. This exploration identified an over score in the EuroSCORE range 0–8 \((57\%, P < 0.0001)\). This over score pivots in the range 9–11 with a correct prediction \((-2\%, P = 1)\), but an under score pivots in the range 12–22 \((-133\%, P = 0.003)\).
The cumulative approach is the third step of the spectral analysis. It simulates the predictive accuracy of the EuroSCORE in samples increasing in risk and sample size in a stepwise manner (Fig. 3). The plot demonstrates that the considerable over score in the lower risk domains is gradually reduced by bringing additional risk into the analysis.

4. Discussion

4.1. Validation

The word valid has Middle French (valide) or Medieval Latin (validus) origins according to the Webster Dictionary. Its definitions also include descriptions about ‘well-grounded or justifiable’ as ‘appropriate to the end in view’. In this manuscript we have focused on ‘the appropriateness to the end in view’, and in particular on the correct estimation of risk for informed consent, therapeutic indication and quality control.

4.2. Strengths and weaknesses of the study sample

A validation dataset can have limitations in several domains: sample size, origins (single- or multi-institutional), sample population, sample timeframe and finally sample spectrum.

The size of the study dataset is adequate for overall validation and for ROC analysis (50 events needed). The study sample size exceeds the original validation sample \( n = 1497 \) of the EuroSCORE system. The size of the study set, on the contrary, is inadequate for analysis of every single calculated value due to the limited number of events in the low-risk patients and an insufficient number of patients at risk in the high-risk domain. This becomes apparent in Fig. 2 with the 95% confidence limits of the observations. This limitation has been partly solved by grouping the risk domains and by the final stepwise approach. The total sample certainly exceeds the annual coronary artery bypass grafting (CABG) production of nearly any cardiothoracic centre, and simulates thereby the number of patients a centre

Table 3

The comparison between observed and predicted hospital mortality by risk grouping after correction for sample size of the included risk categories

<table>
<thead>
<tr>
<th>EuroSCORE risk grouping</th>
<th>No. at risk</th>
<th>No. of observed deaths</th>
<th>% observed deaths</th>
<th>No. of predicted deaths</th>
<th>% predicted deaths</th>
<th>% difference</th>
<th>P value of difference</th>
</tr>
</thead>
<tbody>
<tr>
<td>0–8</td>
<td>1746</td>
<td>29</td>
<td>1.66</td>
<td>67.92</td>
<td>3.89</td>
<td>57</td>
<td>&lt;0.0001</td>
</tr>
<tr>
<td>9–11</td>
<td>211</td>
<td>21</td>
<td>9.95</td>
<td>20.63</td>
<td>9.78</td>
<td>−1.8</td>
<td>1</td>
</tr>
<tr>
<td>12–22</td>
<td>94</td>
<td>31</td>
<td>32.98</td>
<td>13.25</td>
<td>14.1</td>
<td>−133</td>
<td>0.003</td>
</tr>
</tbody>
</table>
would submit to a regional quality control centre using the EuroSCORE as a tool for correction in patient variability.

This single institution dataset carries the risk that any inference can be reduced to a centre effect but avoids inter-institutional strata in the analysis, thereby possibly limiting the applicability of the inferences.

The dataset starts in January 1997 and includes all patients operated on in a similar timeframe (September to November 1995) as the EuroSCORE dataset.

One of the weaknesses of the study sample is the inclusion of only CABG patients in the analysis, but CABG patients represented 63% of the EuroSCORE dataset [7]. An unstable prediction in this domain would exclude stable predictions in aggregated adult cardiac surgery datasets, including patients with rare diseases and large risk at surgery. This dataset is a consecutive departmental series and does therefore not include the bias possibly associated with patient selection.

The current study dataset with an average value of 5.0% risk identifies this dataset as a higher risk dataset of consecutive patients. The national average [8] of the EuroSCORE patients varied from 3.7% risk for the 4277 German patients to 4.7% risk for the 2422 Spanish patients. The high risk within the study dataset is confirmed when compared with the CABG subset within the original EuroSCORE dataset, where the average EuroSCORE value was 3.3% risk. The presence of these high-risk patients, including those in cardiac arrest at the start of surgery, is an additional challenge for the scoring system. The EuroSCORE publications are unclear about the numbers of patients at risk in these highest risk categories.

4.3. Strengths of the EuroSCORE

The EuroSCORE dataset is a cross-section of contemporary European cardiac surgery and the applicability of the system in most European countries should be one of its strengths. The definitions are well-described and the selected variables are not influenced by surgical technique. The simplicity of the calculation, on paper or via an information technology tool, is one of the biggest assets of this system.

The EuroSCORE is an open system; its coefficients are in the public domain. A risk-adjustment system with undisclosed coefficients (e.g. Society of Thoracic Surgery system) is unacceptable.

4.4. Weaknesses of the EuroSCORE

The first limitation that becomes obvious is the limitation of the spectrum. A 0% risk prediction is unrealistic in any medical intervention. The upper limitation of the prediction to 22% risk prediction for the highest risk patients is similarly unrealistic. An average 30 day mortality of 35% [9] was observed in a previous study for all patients in cardiopulmonary resuscitation undergoing CABG. The number of patients with high operative risk in the EuroSCORE dataset is unavailable. The information about the distribution of risk in the original EuroSCORE dataset is limited to the statement that 29% of the developmental dataset had a score above or equal to 6. The low risk of the original EuroSCORE dataset expressed by the low average EuroSCORE.

Fig. 3. The % difference between the observed and predicted hospital mortality in samples increasing in risk and sample size in a stepwise manner.
value but also by the standard deviation of only 2.5, versus our own standard deviation of 4, is an additional indication. The overall comparison gives an impression of overall good prediction but the cumulative approach (Fig. 3) indicates that the overall prediction would have been statistically incorrect, if the study dataset would have been limited to patients with a risk under 11%. Additional proof can be found in the observation that only 20% of the total variability in hospital mortality is explained by the EuroSCORE.

The ROC test used gives information about the discriminatory power, thereby indicating the relation between the scale and the identification of the survivors from the non-survivors, but gives no information about the actual calculated individual values. The obtained ROC value of 0.83 is higher than the 0.76 obtained in the original validation of the EuroSCORE system for the validation dataset. The ROC analysis was also performed with the variable ‘age’ instead of the EuroSCORE. The ROC value was 0.64 for this single variable. The obtained ROC value of 0.83 is only borderline adequate since it is only similar to the discriminatory power of the prediction of rain [10,11] in major weather datasets. This seems a low level of discriminatory power but it might not be. It could be an observation that predicting death after cardiac surgery is as uncertain as predicting rain. We have intentionally not compared the EuroSCORE with any of the other existing (outdated?) scoring systems, and wanted to evaluate this as a stand-alone project.

The maximum obtainable sensitivity and specificity (64% sensitivity and 87% specificity) precludes that the EuroSCORE will misclassify many patients, even at peak performance. This is essential information for any quality control audit.

The most important limitation of the EuroSCORE lies in the over score in the low-risk and the under score in the very high-risk domains with the pivot at the 12% margin. This observation of under scoring in low-risk patients and over scoring in high-risk patients is frequently observed in scores based on logistic regression methodologies in order to prevent distortion of the model in the medium risk groups.

### 4.5. Use of the EuroSCORE in quality control and informed consent

Quality of care is the optimal balance between costs and benefits of a procedure. The early peri-procedural timeframe is only one aspect of the medical cost, extends for several months and is therefore an incomplete interval for the evaluation of quality of care. The EuroSCORE scores only the mortality during the hospital stay [12], which is a short part of the early peri-procedural timeframe and therefore even less of an appropriate interval.

The possibility of using the EuroSCORE as a risk-severity score of datasets for CUSUM [13] (cumulative sum) plotting is obvious. The CRAM [14] (cumulative risk-adjusted mortality) plotting technique credits or debits a surgeon or centre according to the predicted risk of the operation. Most of the scoring systems can be used but the used system needs to be accurate for the complete spectrum of patients. The observed under scoring of the hospital mortality for the high-risk patients, if the EuroSCORE is used, will penalize the centres or surgeons more than appropriately if a high-risk patient dies.

The neutralization of all scores above a certain value (e.g. 11 or 12) in quality control systems such as the CRAM might be advisable. The higher risk patients would thereby not influence the direction of these plots. Fig. 4 shows a CRAM plot of the study population with (CCRAM or corrected cumulative risk plot) and without (CRAM) a

![Fig. 4. The cumulative risk-adjusted mortality without (CRAM) and with (CCRAM) neutralization of the EuroSCORE values above 12.](https://academic.oup.com/wjcts/article-abstract/20/6/1176/468509)
neutralization of the EuroSCORE. The CCRAM plot does not penalize the analyzed institution for the very high-risk patients and shows from the first patients a positive performance versus the scoring system. A certain prerequisite of CRAM analysis, whether or not using the EuroSCORE, is clearly an analysis of the spectrum and distribution of risk in the studied dataset.

The use of the EuroSCORE in informed consent is definitely an improvement above an intuitive value or an average value of an institutional experience, but the legal authorities have to understand the uncertainty of the prediction in court.

4.6. Conclusion

The EuroSCORE is an easy scoring system including most of the usual risk factors. The EuroSCORE allows a common European language in cardiac surgical risk-adjustment, but any use of a language should be accompanied with the knowledge of its descriptive limitations. The overall acceptable prediction of the hospital mortality using the EuroSCORE is the consequence of an overestimation of the low risk and an underestimation of the high risk. A centre effect can be the cause of the considerable overestimation of the lower risk. This centre effect is part of the variability in outcome describing the quality of delivered care. In this particular analysis it is difficult to reduce the variability between observed and predicted only to a centre effect since this would indicate an over performance of the centre in the low-risk and an under performance of that same centre in the high-risk populations.

Some quality control systems such as the CUSUM and CRAM plots using this scoring system might require adaptations for the higher risk patients. Similar studies are needed before a final evaluation of the EuroSCORE can be made, but a critical warning in the high-risk domain is appropriate. At this time, we have adopted the EuroSCORE for most of our quality control systems.

References