Prospective validation of EuroSCORE II in patients undergoing cardiac surgery in Argentinean centres

Raúl A. Borracci*, Miguel Rubio, Leonardo Celano, Carlos A. Ingino, Norberto G. Allende and Rodolfo A. Ahuad Guerrero

Hospital de Clínicas, School of Medicine, Buenos Aires University, Buenos Aires, Argentina

* Corresponding author. La Pampa 3030, 1428 Buenos Aires, Argentina. Tel: +54-11-47845917; fax: +54-11-49616027; e-mail: raborracci@gmail.com (R.A. Borracci).

Received 13 August 2013; received in revised form 28 October 2013; accepted 12 November 2013

Abstract

OBJECTIVES: The European System for Cardiac Operative Risk Evaluation II (EuroSCORE II) is an updated version of the original EuroSCORE that must be extensively validated. The objective was to prospectively evaluate the efficacy of EuroSCORE II in predicting the immediate results of cardiac surgery in Argentinean centres.

METHODS: A prospective consecutive series of 503 adults who underwent cardiac surgery between January 2012 and April 2013 was studied. EuroSCORE II discrimination and accuracy were assessed in the overall cohort and in two surgically defined subgroups: isolated coronary artery bypass graft (CABG) surgery and non-CABG surgery. Additionally, a risk-adjusted cumulative sum control chart analysis was performed.

RESULTS: In-hospital overall mortality rate was 4.17%, while the mortality rate predicted by the EuroSCORE II was 3.18% ($P = 0.402$). Receiver operating characteristic curve analysis demonstrated a good overall (area 0.856) and non-CABG subgroup (area 0.857) discrimination ($P = 0.0001$), while discrimination in the CABG subgroup was poorer (area 0.794, $P = 0.014$). The model showed good calibration in predicting in-hospital mortality, both overall (Hosmer–Lemeshow, $P = 0.082$) and for each subgroup (non-CABG, $P = 0.308$, and CABG, $P = 0.150$).

CONCLUSIONS: EuroSCORE II reflects a better current surgical performance and offers a new quality standard to evaluate local outcomes. EuroSCORE showed an overall good discriminative capacity and calibration in this local population; nevertheless, the model performed optimally in non-CABG surgery and in highest-risk patients, underestimating in-hospital mortality in lowest-risk cases. The latter finding may be interpreted as an inadequate behaviour of the model, as a poor performance of surgeons or both. Larger prospective studies will elucidate this hypothesis.

Keywords: European System for Cardiac Operative Risk Evaluation II • Validation • Operative risk • Cardiac surgery • Argentina

INTRODUCTION

The European System for Cardiac Operative Risk Evaluation II (EuroSCORE II) [1] has been recently shown to improve the performance of the original EuroSCORE [2], as the latter has been criticized for overestimating the operative risk [3, 4]. A retrospective validation of EuroSCORE II in patients undergoing coronary artery bypass surgery in Finland [5] showed that the score is particularly accurate in predicting operative mortality risk of high-risk patients, a weakness of the original EuroSCORE. However, this new score could not perform equally well in predicting mortality in low-risk patients. Prospectively collected multicentre clinical data from the UK [6] suggested that EuroSCORE II is an acceptable contemporary generic risk model, though it is poorly calibrated for isolated coronary artery bypass surgery in both the lowest- and highest-risk patients. Additionally, some concern remains over its accuracy for aortic valve replacement, aortic surgery and miscellaneous procedures [7]. Based on these reported results, other authors agree that additional prospective validation of EuroSCORE II is necessary in larger populations and especially in other geographic regions [8, 9].

The aim of this study was to prospectively evaluate the efficacy of EuroSCORE II in predicting the immediate operative results of patients undergoing cardiac surgery in a local population of Argentina.

MATERIALS AND METHODS

This study included a prospective consecutive series of 503 adult patients who underwent cardiac surgery at the Buenos Aires University Hospital in Argentina and its associated clinics between January 2012 and April 2013. All types of surgery were included, except acute aortic dissection, transplant and transcatheter aortic valve implantation (TAVI). Baseline and operative data were prospectively collected in a clinical registry, which included the variables needed to estimate the operative risk mortality based on...
EuroSCORE II [1]. Definitions proposed by the EuroSCORE were fully adopted, and the risk score was determined with an online interactive calculator (http://www.euroscore.org/calc.html). Creatinine clearance was calculated with the Cockcroft–Gault formula [10]. Additional information on other risk factors, comorbidities, major operative complications and in-hospital operative mortality was added in the computerized database. The primary end point was in-hospital all-cause mortality. EuroSCORE II discrimination and accuracy were assessed in the overall cohort and in two surgically defined subgroups: isolated coronary artery bypass graft (CABG) surgery and non-CABG surgery. Surgical techniques included intermittent antegrade St Thomas cardioplegia and routine haemofiltration during extracorporeal circulation, proximal anastomoses after cross-clamping in coronary artery bypass grafting, Octopus System® stabilizer (Medtronic, Minneapolis, MN, USA) and intracoronary shunts for off-pump coronary artery bypass and ultra-fast-track anaesthesia.

Statistical analysis

Continuous variables were expressed as mean ± standard deviation (SD). Frequency distribution of scores was complemented with quartiles. Kolmogorov–Smirnov (K–S) goodness-of-fit test was used to analyse normal distributions. Univariate comparison of dichotomous variables was performed using \( \chi^2 \) test. Yates’ corrected \( \chi^2 \) test was used when cell expected values were between 3 and 5, while two-tailed Fisher’s exact test was used when values were <3. Model calibration was assessed by the Hosmer–Lemeshow goodness-of-fit test, and receiver operating characteristics (ROCs) curve analysis was used to estimate the performance of the EuroSCORE II for predicting in-hospital mortality. The area under the ROC curve with its 95% confidence interval (CI) was calculated. Accuracy was assessed by the Brier score [11], which is the average square difference between the predicted probability and the true occurrence of mortality, the score being close to 0 when accuracy is better. Predictive accuracy of individual Brier scores was assessed with Spiegelhalter’s method [12]. Comparison between Brier scores was done with a modified method described by Redelmeier et al. [13], assuming a standard error of the difference between two independent observed Brier scores as the square root of \( \frac{V}{n_1} + \frac{V}{n_0} \), being \( V \) variance. The expected-to-observed operative mortality ratios were calculated to further estimate the performance of the risk score, grouped by quartiles. A ratio of 1 indicates an optimal performance of the scoring method. Finally, we used an expected-to-observed probability of death chart for individual risk [14] and the variable life-adjusted (VLAD) display for monitoring the results. The VLAD plot points out the difference between the expected and observed cumulative mortality of consecutive cardiac surgical cases, showing whether an operative performance is above or below what might be expected. VLAD monitoring is similar to the risk-adjusted cumulative sum control chart (CUSUM) analysis [15]. The scoring system functions by accumulating penalties for each death and rewards for every survivor, in this case, based on the inherent risk calculated by EuroSCORE II [16]. The 95% two-sided boundary limits for the VLAD chart were calculated as proposed by Grunkemeier et al. [17]. Statistical analysis was performed with SPSS Statistics for Windows, Version 17.0 (SPSS, Inc., Chicago, IL, USA), and a \( P \)-value of ≤0.05 was considered statistically significant.

### RESULTS

The characteristics of the population and the immediate postoperative adverse events are given in Table 1. In-hospital overall mortality rate was 4.17% (21 observed deaths), while the mortality rate predicted by EuroSCORE II was 3.18% (±0.920% (16 mean expected deaths, \( \chi^2 \) 0.70 \( P = 0.402 \)) CABG surgery mortality rate was 2.20% (6 observed deaths), while the predicted mortality rate was 1.47% (±1.49%) (4 mean expected deaths, Yates’ corrected \( \chi^2 \) 0.10, \( P = 0.749 \)). Non-CABG surgery mortality rate was 6.52% (15 observed deaths) and the predicted mortality by the model was 5.22% (±8.42%) (12 mean expected deaths, \( \chi^2 \) 0.35 \( P = 0.552 \)). The histogram of expected mortality based on EuroSCORE II showed a positively skewed distribution, since 75% of patients \((n = 377)\) had a score under 2.86 (Fig. 1). ROC curve analysis demonstrated a good discrimination, both overall (area 0.856, 95% CI 0.792–0.920, \( P < 0.001 \)) and in the non-CABG subgroup (area 0.857, 95% CI 0.787–0.928, \( P < 0.001 \)), whereas discrimination in the CABG

<table>
<thead>
<tr>
<th>Table 1: Preoperative population characteristics and immediate postoperative results ( (n = 503) )</th>
</tr>
</thead>
<tbody>
<tr>
<td>Variables</td>
</tr>
<tr>
<td>Preoperative and intraoperative</td>
</tr>
<tr>
<td>Age (mean ± SD) (range)</td>
</tr>
<tr>
<td>Female gender</td>
</tr>
<tr>
<td>Insulin-dependent diabetes</td>
</tr>
<tr>
<td>Heart failure</td>
</tr>
<tr>
<td>Stroke</td>
</tr>
<tr>
<td>Pulmonary disease</td>
</tr>
<tr>
<td>Extracardiac arteriopathy</td>
</tr>
<tr>
<td>Renal failure (dialysis)</td>
</tr>
<tr>
<td>Unstable angina*</td>
</tr>
<tr>
<td>Recent myocardial infarction</td>
</tr>
<tr>
<td>Active endocarditis</td>
</tr>
<tr>
<td>Prior cardiac surgery (redo)</td>
</tr>
<tr>
<td>Moderate-to-severe LV dysfunction</td>
</tr>
</tbody>
</table>

*Defined as nitrate infusion at operating theatre arrival.

**Miscellaneous surgery included cardiac neoplasm, ventricular and ascending aortic aneurysm and internal or external cardiac rupture.

*Calculated exclusively for coronary surgery \((n = 273)\).
The model showed good calibration in predicting overall in-hospital mortality (Hosmer-Lemeshow, $P = 0.082$) and for each subgroup of patients (non-CABG, $P = 0.308$ and CABG, $P = 0.150$). The accuracy assessed with the Brier score was 0.039 for the overall cohort ($Z = 1.697$, $P = 0.090$), 0.021 for the CABG subgroup ($Z = 0.704$, $P = 0.482$) and 0.056 for the non-CABG subgroup ($Z = 1.455$, $P = 0.146$). In short, data showed that all individual Brier scores had a good predictive accuracy according to Spiegelhalter’s method (K-S for the overall cohort: $Z = 10.7$, $P < 0.001$; modified Redelmeier’s method to compare CABG vs non-CABG subgroups: $Z = 2.311$, $P = 0.021$).

In the overall series, the expected-to-observed operative mortality ratio was 0.76, in the CABG surgery subgroup was 0.69 and in the non-CABG surgery subgroup was 0.75. Figure 2 summarizes the expected and observed overall mortality according to EuroSCORE II and the expected-to-observed operative mortality ratios. Each red bar indicates the observed mortality for each quartile of EuroSCORE II, ranging from 2.42 to 7.87% in the highest quartile. The inner line shows the expected-to-observed mortality ratios, obtained by dividing for each quartile, the mean expected mortality of EuroSCORE II by the observed mortality. The line indicates that the optimal performance of the score was in the highest quartile, where the ratio was near to 1. On the contrary, the worst performance of EuroSCORE II was in its two lowest quartiles, since mortality ratios were 0.3 and 0.4, respectively.

Figure 3 demonstrates the expected risk minus the observed outcome for each patient. The top half of the charts demonstrates the risk of death for each survivor on a probability scale between 0 and 1. The bottom half of the charts represents patients who died, scaled according to their expected mortality −1. Most deaths occurred in patients assigned as having low risk by EuroSCORE II, especially in the CABG subgroup. Finally, the VLAD plot in Figure 4 displays the difference between expected and real cumulative mortality over time. The line of the series fluctuated slightly above the lower CI limit.
DISCUSSION

Local validation and implementation of updated operative risk scores are necessary to ameliorate the quality of patient care. Improvement of surgical techniques and perioperative medical attention have significantly reduced cardiac surgery mortality in the past 10 years, despite older and sicker patients. At present, the original additive and logistic EuroSCORE is poorly calibrated for predicting in-hospital mortality in contemporary cardiac surgery, overestimating the real risk. Recently, EuroSCORE II was proposed as a better assessment of operative risk; however, some clinical inconsistencies have emerged in external validation cohorts [6–9].

The aim of this study was to prospectively evaluate the EuroSCORE II risk model in contemporary clinical use. EuroSCORE II performed well in the overall cohort of this local population of Argentina and could be considered an adequate updated generic cardiac surgery risk score to predict in-hospital mortality. However, the system only performed optimally in its highest quartile and underestimated mortality in the lowest-risk patients. EuroSCORE II especially seems to be too demanding with the first-quartile lowest-risk cases. Discrimination of the model was better assessed in the non-CABG clinical subgroup than in the CABG surgery cohort. As a logical consequence of the lower calculated expected risk, an important influence on the evaluation of the results using the VLAD analysis was observed. The VLAD plot showed that surgeon performance was near the higher accepted level of mortality (95% inferior confidence limit). Figure 3 demonstrates that all the patients who died in the CABG subgroup had the lowest estimated risk, while only half of the patients who died in the non-CABG subgroup had the lowest-expected mortality. These results are consistent with a better performance of EuroSCORE II in discriminating the non-CABG subgroup.

In short, the ability of EuroSCORE II to predict 30-day mortality in high-risk patients is of particular importance, a weakness of the original model. Nevertheless, this new risk score did not perform equally well in predicting mortality in low-risk patients. Though low-risk cases were sufficiently represented in our validation dataset, EuroSCORE II seemed to underestimate the risk. Nevertheless, the external validation of a risk model involves two problems. A score that overestimates the individual patient risk may lead to a false sense of reassurance, while underestimation may be considered as a sign of poor calibration or discrimination of the model, instead of being considered a poor surgical performance.

At present, the majority of published studies have analysed the external validity of EuroSCORE II in retrospective series of large institutional databases [18, 19]. A limitation of this approach is that most of those registers have incomplete data to estimate correctly the EuroSCORE II, especially due to the lack of information about creatinine clearance, left ventricular function and systolic pulmonary artery pressure, all of these items being included in the new version of the model. Two retrospective validation studies performed in Chinese patients undergoing heart valve surgery concluded that EuroSCORE II had a good discriminative power and calibration in single-valve surgery [20, 21]. Another retrospective analysis in Spain demonstrated good discriminative capacity, but poor calibration in all types of major cardiac surgery [22]. Prediction of mortality in high-risk patients was assessed in two other retrospective studies and conclusions are still controversial [23, 24]. Noyez et al. [25] compared the performance of the initial EuroSCORE with that of the new version as a tool to evaluate risk in 200 patients undergoing CABG surgery combined with aortic or mitral valve replacement. They concluded that EuroSCORE II decreases the overestimation associated to the initial model, with ~50% risk reduction.

Despite being a prospective consecutive study, a limitation of our work is that the analysis was done in a relatively small number of patients, and the sample represents only a portion of patients undergoing major cardiac surgery annually in Argentina.

CONCLUSIONS

EuroSCORE II could reflect a better current surgical performance by offering a new quality standard to evaluate local surgical outcomes. In our study, EuroSCORE II showed an overall good discriminative capacity and calibration in this local population. However, the model performed optimally in non-CABG surgery and in highest-risk patients, underestimating in-hospital mortality in lowest-risk cases. Nonetheless, the latter findings may be interpreted as an inadequate behaviour of the model, as a poor performance of surgeons or both. Larger prospective studies will elucidate this hypothesis in the future.

Conflicts of interest: none declared.
REFERENCES