Accuracy of EuroSCORE II in patients undergoing minimally invasive mitral valve surgery

Marco Moscarelli*a,b,*, Giacomo Bianchi*, Rafik Margaryan*, Alfredo Cerillo*, Pierandrea Farneti*, Michele Murzi* and Marco Solinas*

*a Pasquinucci Hospital, Ospedale del Cuore, Fondazione Monasterio, Massa, Italy
b Honorary Research Fellow, National Heart and Lung Institute (NHLI), Imperial College, London, UK
* Corresponding author. Department of Cardiac Surgery, Pasquinucci Hospital, Via Aurelia Sud, 54100 Massa (MS), Italy. Tel: +39-0503-152216; fax: +39-0503-152166; e-mail: m.moscarelli@imperial.ac.uk (M. Moscarelli).

Received 6 May 2015; received in revised form 18 August 2015; accepted 20 August 2015

Abstract

OBJECTIVES: EuroSCORE II has been implemented with the view to providing better performance than the previous logistic EuroSCORE. However, until now, no external validations have been carried out in the minimally invasive context. Therefore, we sought to validate the accuracy of EuroSCORE II in a retrospective series of consecutive patients undergoing minimally invasive mitral valve surgery.

METHODS: Data of 1609 consecutive patients who underwent minimally invasive mitral valve surgery in our institution were retrospectively reviewed. The accuracy of EuroSCORE II was assessed in terms of discrimination and calibration. Discrimination was tested via analysis of the area under the curve of receiver operator characteristic; calibration was achieved by calculating the observed versus expected mortality ratio and the Hosmer–Lemeshow test for test probability; global accuracy was assessed by using Brier’s score; results were compared with the previous logistic EuroSCORE version. EuroSCORE II performance was also tested for discrimination of postoperative complications. Discrimination subgroup analysis was carried out for single surgeon results, and for high-risk patients those outliers were defined after boxplot analysis (EuroSCORE II ≥6%).

RESULTS: EuroSCORE II showed good discrimination power (area under the curve 0.846), and was statistically superior to logistic EuroSCORE (P = 0.01). In terms of calibration, both EuroSCORE II and logistic over-predicted mortality; with regard to adverse events, the discrimination of EuroSCORE II was adequate for acute renal failure, low-output syndrome and increased intensive care unit stay; area under the curve of receiver operating characteristic for high-risk patients with EuroSCORE ≥6% was suboptimal (0.654); single surgeon results did not influence the discrimination of EuroSCORE II.

CONCLUSIONS: EuroSCORE II showed good discrimination power in our series of minimally invasive mitral valve patients; however, it over-predicted mortality. Individual performance did not influence discrimination. Performance was suboptimal for prediction of complications and for high-risk subgroup in-hospital mortality.

Keywords: Mitral valve • Minimally invasive surgery • Statistics • Risk analysis/modelling

INTRODUCTION

Having a reliable tool of risk stratification remains of paramount importance for two main reasons: firstly, our population is ageing and physicians have to deal with an increased number of patients with comorbidities; secondly, procedures such as mitral-clip [1], valve-in-valve [2] and transapical neo-chordae [3] are now emerging, and a clear risk cut-off above which patients would be better served with these catheter-based techniques is advocated. The two most used algorithms of risk prediction are the Society of Thoracic Surgeons Predicted Risk of Mortality (STS-PROM) [4] and EuroSCORE II (ES II); the latest ES II was launched in 2011 and published in 2012 [5], replacing the previous logistic version (LES).

However, ES II may need further validations for different reasons. ES II was based on a lower number of patients and included fewer variables than STS-PROM; among the total number of procedures screened (n = 22 381), only 3984 (17.8%) were mitral procedures (8.7% repair and 9.2% replacement) and there was no specification on minimally invasive surgery [5]; some authors, particularly for specific subgroups with comorbidities such as aged patients [6], have questioned the validity of ES II; therefore ad hoc analysis of the model is advocated [7]. Lastly, because of the steep learning curve, there is still reluctance in offering minimally invasive surgery in high-risk patients [8]. On the other hand, STS-PROM cannot be considered superior to ES II; in fact it does not support mitral valve-associated procedures such as
tricuspid valve surgery; moreover, its discrimination for isolated mitral valve repair and replacement has been similar to ES II [9].

We therefore sought to undertake an external validation of the accuracy of ES II in a retrospective series of patients who consecutively underwent minimally invasive mitral valve surgery ± tricuspid surgery in our institution, and compare it with the previous logistic version.

**MATERIALS AND METHODS**

This was a retrospective observational study of data from 1609 consecutive patients who underwent video-assisted minimally invasive mitral valve surgery ± tricuspid valve repair ± atrial fibrillation (AF) ablation via right minithoracotomy between November 2003 and August 2013 in our institution. We started our minimally invasive mitral valve surgery programme in early 2003 and since then, it has been our standard approach for mitral disease with very few exclusion criteria (e.g. right chest previous major surgery/salvage operation). From November 2003, we started inserting data in the electronic database. Our data collection works on an integrated database that merges five sections, each filled by different physicians, nurses and technicians of different specialties. Patients with concomitant tricuspid valve surgery, AF ablation, atrial septal defect and left appendage closure carried out through right minithoracotomy were included in the study as well. Since our institutional database inputs only LES and additive EuroSCORE, the newer ES II was calculated a posteriori.

**Surgical techniques**

The surgical approach has been described before [10]. In brief, a 5-Fr catheter introducer sheath is placed percutaneously in the right femoral vein. Once the sheath is in place and secure, a 5–6 cm incision is performed in the third or fourth intercostal space. Once thoracotomy is performed, two working ports are established. The pericardium is opened, keeping the incision approximately 3–4 cm above the phrenic nerve. When possible we prefer direct cannulation of the ascending aorta. Femoral artery cannulation is used when central cannulation is either contraindicated or not performable, such as with re-do patients with patent saphenous vein grafts, previous ascending aortic replacement and/or severe adhesions around the ascending aorta. Once the ascending aorta is purse-stringed and the adventitia prepared, heparin is given. A guidewire is passed through the introducer sheath and under transoesophageal echocardiographic (TOE) guidance is advanced up to the superior vena cava. The sheath is removed; dilators are used to tunnel the cannula path and then the venous cannula is advanced over the guidewire till its final position is at the superior vena cava level. The ascending aorta is then cannulated under direct vision. In patients who required opening of the right atrium, we used a single two-stage femoral venous cannula (RAP Venous Cannula, Estech, Inc., San Ramon, CA, USA), which allows the superior and the inferior vena cava to drain simultaneously. Our routine approach includes direct clamping of the ascending aorta with a flexible transthoracic clamp and antegrade delivery of cold crystalloid or haematic cardioplegia. When this not achievable, endoaortic balloon clamping with antegrade cardioplegia is used. Hypothermic ventricular fibrillation (28–30°C core temperature) or beating-heart surgery is used in re-do patients with patent coronary artery bypass grafts or ascending aortic calcifications [11]. Intervention proceeds in a conventional manner. Surgical ablation of AF is performed with standard endocardial left atrial lesions using a monopolar radiofrequency probe concomitant with left atrial appendage exclusion or creating a box lesion with Cobra fusion (Estech).

**End-points and definitions**

The primary accuracy end-point was in-hospital mortality. Secondary end-points were: intraoperative conversion to sternotomy (either intra- or postoperatively), reopening for bleeding, stroke, transient ischaemic attack, acute renal failure (defined as the requirement for haemodialysis or an increased creatinine level >200 mmol/l), low cardiac output syndrome (LCOS), prolonged intensive care unit stay (>48 h), postoperative atrioventricular block (AF of new onset) and combined end-points. Since the ES II cut-off for high-risk patients is not well defined, and since our population in relation to ES II followed a non-parametric distribution that is heavily right-skewed, we defined as ‘high-risk’ those minor and major outliers identified after boxplot analysis (ES II ±6%). For surgeons’ subgroup analysis, we identified three surgeons who did the largest number of operations; Surgeon A’ was the most experienced with the largest volume of cases; Surgeon B’ and ‘C’ were the second and third most senior surgeons already proctored by Surgeon A’; the last block ‘Others’ included six surgeons in training phases who had performed less than 100 operations per head. Operations performed by ‘Surgeons A–C’ were proctored, while operations performed by those in the group ‘Others’ were proctored or with a senior in theatre, and represent the training block.

**Statistical analysis**

Patients’ data were summarized as mean ± standard deviation, or median and interquartile range for asymmetrically distributed continuous variables. Categorical variables were expressed as percentages or prevalence, as appropriate. The Kolmogorov–Smirnov test was used to check for the normality of data in groups before further analysis. Analysis of the area under the curve (AUC) of receiver operator characteristic (ROC) was carried out to assess the discrimination power of LES and ES II on predicting in-hospital mortality and for assessing goodness-of-fit in logistic regression. An ROC c-statistic greater than 0.7 and a non-significant Hosmer–Lemeshow test probability (P > 0.05) were considered acceptable. For in-hospital mortality, comparison of ES II and LES curves was accomplished using three different methods, namely Delong, bootstrapping and Venkatraman, the first two comparing the AUC and the last the ROC curves themselves [12]; Somers’ Dxy rank correlation was also used; a Dxy = 1 indicates a perfect discrimination, while Dxy = 0 indicates that the model is performing random predictions; for each model, the comparison of the actual slope and intercept with the ideal value of 1 and 0 was performed with the U-statistic (unreliability test) and tested against a χ² distribution with 2 degrees of freedom [13]; agreement between expected and observed outcomes and prediction was depicted as ‘calibration’: in perfectly calibrated models, the measured prediction is close to a 45° line in the calibration plot; the intercept of this calibration slope depicted the amount of deviation, i.e. the over- or under-estimation, of the measured outcome. Calibration was also tested with the Hosmer–Lemeshow goodness-of-fit test and the test
probability ($P > 0.05$) was considered acceptable. The observed/expected mortality ratio ($O/E$) was calculated, and a value less than 1 was considered as risk model overestimation of mortality (whereas a value more than 1 as underestimation). General accuracy of the model was also tested calculating the Brier score, mathematically the mean of the squared residuals between the predicted probability and the actual outcome for each patient; a Brier score within 0.25 was set as an acceptable upper cut-off. Subgroup analysis was carried out in high-risk patients considered as outliers after boxplot analysis (ES II ≥ 6%). Statistical analysis was performed with R software (R Project for Statistical Computing, http://www.r-project.org/, version 3.0.3).

RESULTS

Preoperative characteristics and in-hospital mortality

Preoperative characteristics are presented in Table 1. Among the consecutive series of 1609 patients considered in the analysis, 28 died while in hospital (1.74%). Mean ES II and LES were 3.2 and 5.5%, respectively. Areas under ROC curve for ES II and LES were 0.846 [95% confidence interval (CI); 0.772–0.920] and 0.692 (95% CI; 0.603–0.780), respectively ($P = 0.01$) (Fig. 1A and B; Table 2). As shown in Fig. 2A and B, the patterns of calibration between the two models were very different; LES starts diverging from the ideal diagonal over 10%-predicted probability and further more for higher estimated values, showing a tendency to over-prediction; while ES II was closer to the ideal diagonal up to 30%-predicted probability; this indicates better performance of ES II than LES with regard to high-risk profile patients (Fig. 2A and B; Table 2). Furthermore, the Hosmer–Lemeshow test failed by the LES ($P < 0.001$), indicating unreliable probabilities, whereas it passed by the ES II ($P = 0.11$). The ES II observed/expected mortality ratio was 0.54. Brier’s scores along with summary statistics for accuracy are presented in Table 2.

Postoperative complications

Nine postoperative complications were analysed. Conversion to sternotomy was included as well. ES II and LES areas under the ROC curves and their comparison are illustrated in Table 3. ES II showed AUC above 0.7 for acute renal failure, LCOS and prolonged ICU stay; LES showed good discrimination for stroke and acute renal failure.

High-risk patients’ subgroup analysis

One hundred and ninety-nine patients had ES II ≥ 6% (mean 12.9%). The in-hospital mortality rate was 9% (18 patients). The $O/E$ was 0.69. The area under the ROC curve was 0.654 (95% CI;

![Figure 1](https://example.com/figure1.png)

**Figure 1:** (A) Logistic EuroSCORE and (B) EuroSCORE II ROC statistic. Areas under ROC curve for ES II and LES were 0.846 (95% CI; 0.772–0.920) and 0.692 (95% CI; 0.603–0.780), respectively ($P = 0.01$).
Surgeons’ related mortality analysis

Surgeons’ mortality subgroup analysis was carried out, and ES II and LES were tested. The three surgeons (A–C) performing the bulk of the operations were reviewed separately from a fourth group (others: trainees) that included six surgeons who did less than 100 operations per head, and were considered to be in minimally invasive training. Areas under the ROC curves are presented in Table 5.

DISCUSSION

Since less-invasive mitral procedures such as mitral-clip [1], valve-in-valve [2], transapical neo-chordae [3] are emerging as alternatives to surgery, having a model of prediction of risk with good accuracy would be of crucial importance to identify which patients might be better served with conventional surgery and with catheter-based intervention. Two major components of accuracy are calibration and discrimination [14]. Calibration is defined as the agreement between model prediction and observed event incidences (observed/expected ratio), and discrimination is a measure of how well a model can separate those who will experience the event from those who will not [15]. Analysis of the area under the ROC is a valid method for detecting if a system has acceptable sensitivity and specificity to discriminate between false positive and false negative [16], and simply tells us that a patient who will experience the event has a higher risk score than a patient who will not. An AUC less than 0.5 indicates no discrimination power at all and an AUC of 1.0 indicates perfect prediction, while an AUC above 0.7 is considered clinically acceptable [14]. The Hosmer–Lemeshow test is a commonly used procedure for assessing goodness-of-fit in logistic regression and, although imperfect, has been widely used for evaluation of risk-scoring models [17]. The Brier score is a valid method to assess global accuracy and should be as close as possible to 0 [7].

The two most used risk model predictions in cardiac surgery are STS-PROM and ES II. ES II was launched in 2011 and published in 2012 as an improvement over the previous logistic version [5]. Since then, different external validations [7, 13, 18, 19] and meta-analysis [20] have been performed, but no evaluation of accuracy on patients undergoing minimally invasive mitral valve surgery has been carried out. This specific validation of the ES II risk model in a minimally invasive context is relevant because, on the one hand, it contributes to give strength to the score in examinations, and on the other, it may work as a quality control of surgical performance for a technique that involves a steep learning curve. No other studies have assessed so far the accuracy of a risk model in patients undergoing minimally invasive mitral valve surgery, and we therefore sought to internally validate the ES II calibration and discrimination in our series of patients who have undergone minimally invasive mitral valve surgery from 2003 to 2013. Similar to the works of other authors [9], the discrimination of ES II was very good, with an AUC of 0.846 and significantly superior to LES (0.692, \( P = 0.001 \)); however, ES II over-predicted

| Table 2: EuroSCORE II and logistic summary statistic for discrimination and calibration (patients, \( n = 1609 \)) |
|-----------------------------------------------|---------------|
|                                | LES           | ES II          |
|-----------------------------------------------|---------------|
| Brier score                               | 0.017         | 0.016          |
| Discrimination AUC                        | 0.69          | 0.85           |
| DeLong’s test \( P \)-value               | <0.001        | <0.001         |
| Somers’ Dxy                                | 0.38          | 0.69           |
| Calibration Slope                         | 0.166         | 0.713          |
| Intercept                                 | 0.008         | -0.005         |
| U-statistic \( P \)-value                 | <0.001        | <0.001         |
| Hosmer–Lemeshow P-value                   | <0.001        | 0.11           |

AUC: area under the curve; ES II: EuroSCORE II; LES: logistic EuroSCORE.
invasive mitral valve surgery in case of operative adverse outcome; hence, a benchmark is missing. The possible relation with ES II.

Among postoperative complications we included mortality (O/E 0.54). Moreover, the ES II Brier score was 0.016, which is highly significant; furthermore, the Hosmer-Lemeshow test failed by the Logistic EuroSCORE (P < 0.001), indicating unreliable probabilities, while it passed by ES II (P = 0.11). Like other authors [7], we wished to assess the discrimination of ES II and LES on adverse events including conversion, and as a result, only acute renal failure, prolonged CICU stay and LCOS, had an AUC above 0.7 for ES II, while LES had good discrimination for acute renal failure.

Table 3: Postoperative adverse events after minimally invasive mitral valve surgery

<table>
<thead>
<tr>
<th>Event</th>
<th>Total = 1609, n (%)</th>
<th>LES (AUC, 95% CI)</th>
<th>EuroSCORE II (AUC, 95% CI)</th>
<th>P-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Conversion to sternotomy</td>
<td>34 (2.1%)</td>
<td>0.555 (0.468–0.641)</td>
<td>0.648 (0.566–0.729)</td>
<td>0.21</td>
</tr>
<tr>
<td>Reopening for bleeding</td>
<td>78 (4.8%)</td>
<td>0.611 (0.551–0.671)</td>
<td>0.608 (0.542–0.674)</td>
<td>0.95</td>
</tr>
<tr>
<td>Stroke</td>
<td>33 (2.0%)</td>
<td>0.715 (0.629–0.801)</td>
<td>0.650 (0.542–0.758)</td>
<td>0.41</td>
</tr>
<tr>
<td>TIA</td>
<td>23 (1.42%)</td>
<td>0.496 (0.338–0.604)</td>
<td>0.545 (0.433–0.657)</td>
<td>0.56</td>
</tr>
<tr>
<td>Acute renal failure</td>
<td>24 (1.49%)</td>
<td>0.728 (0.681–0.776)</td>
<td>0.778 (0.676–0.881)</td>
<td>0.54</td>
</tr>
<tr>
<td>&gt;48 h ICU stay</td>
<td>212 (13.1%)</td>
<td>0.687 (0.650–0.787)</td>
<td>0.750 (0.714–0.787)</td>
<td>0.03</td>
</tr>
<tr>
<td>Postop pacemaker insertion</td>
<td>54 (3.36%)</td>
<td>0.696 (0.634–0.724)</td>
<td>0.696 (0.634–0.757)</td>
<td>0.91</td>
</tr>
<tr>
<td>AF</td>
<td>559 (34.7%)</td>
<td>0.612 (0.547–0.677)</td>
<td>0.636 (0.609–0.664)</td>
<td>0.76</td>
</tr>
<tr>
<td>Combined end-point</td>
<td>760 (47.2%)</td>
<td>0.657 (0.630–0.683)</td>
<td>0.699 (0.674–0.723)</td>
<td>0.002</td>
</tr>
</tbody>
</table>

Combined end-point: conversion to sternotomy, reopening of bleeding, stroke, TIA, acute renal failure, LCOS, postop pacemaker implantation or AF.

High-risk patients defined as per EuroSCORE II ≥6%.

Table 4: Analysis of AUC for EuroSCORE II and logistic EuroSCORE for high-risk patients

<table>
<thead>
<tr>
<th>Event</th>
<th>Total = 199, n (%)</th>
<th>LES (AUC, 95% CI)</th>
<th>EuroSCORE II (AUC, 95% CI)</th>
<th>P-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>High-risk in-hospital mortality</td>
<td>18 (9%)</td>
<td>0.412 (0.312–0.511)</td>
<td>0.654 (0.504–0.803)</td>
<td>0.001</td>
</tr>
</tbody>
</table>

AUC: area under the ROC curve; ES II: EuroSCORE II; LES: logistic EuroSCORE; CI: confidence interval.

Table 5: Analysis of EuroSCORE II and LES in relation to surgeons’ performance

<table>
<thead>
<tr>
<th>Surgeon</th>
<th>ES II mean (median/IQR)</th>
<th>Death, n (%)</th>
<th>LES (AUC, 95% CI)</th>
<th>EuroSCORE II (AUC, 95% CI)</th>
<th>P-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Surgeon A (n = 604)</td>
<td>3.7% (1.6–2.3)</td>
<td>14 (2.3%)</td>
<td>0.706 (0.594–0.817)</td>
<td>0.848 (0.775–0.921)</td>
<td>0.17</td>
</tr>
<tr>
<td>Surgeon B (n = 345)</td>
<td>3.6% (1.6–2.9)</td>
<td>4 (1.15%)</td>
<td>0.730 (0.554–0.916)</td>
<td>0.885 (0.812–0.958)</td>
<td>0.41</td>
</tr>
<tr>
<td>Surgeon C (n = 155)</td>
<td>3.0% (1.6–2.6)</td>
<td>3 (1.9%)</td>
<td>0.891 (0.822–0.961)</td>
<td>0.906 (0.798–1.00)</td>
<td>0.92</td>
</tr>
<tr>
<td>Others (n = 505)</td>
<td>2.4% (1.6–2.0)</td>
<td>7 (1.4%)</td>
<td>0.537 (0.354–0.720)</td>
<td>0.704 (0.497–0.911)</td>
<td>0.19</td>
</tr>
</tbody>
</table>

Surgeons A, B, C defined as seniors having performed more than 150 procedures per head, while group ‘Other’ included six surgeons in the minimally invasive mitral valve surgery training phase with less than 100 procedures per head.

AUC: area under the ROC curve; ES II: EuroSCORE II; LES: logistic EuroSCORE; CI: confidence interval; IQR: interquartile range.

The authors [7] thought that it would have been statistically appropriate to consider high risk those outliers identified after boxplot analysis: outliers started from ES II ≥6% (n = 199; mean 12.9%); however, discrimination was low for both ES II and LES, 0.654 (95% CI; 0.504–0.803) and 0.412 (95% CI; 0.312–0.511), respectively. The accuracy of ES II still remains to be validated in larger and more defined high-risk populations. Since 2013, surgeons have been continuously trained...
at different times, thereby validation of ES II at different periods has not been carried out. In fact, analysis of performance of ES II was rather tested for individual surgeons. Mortality rates for all the three surgeons who performed more than 150 cases was lower than predicted, as well as the mortality of the group ‘Others’ that included six surgeons who performed less than 100 operations per head (Table 5). An ad hoc analysis evaluating the single surgeons’ performance was moved from previous findings that showed variation of sensitivity and specificity of ES II because of difference surgeons’ performance volume related [21, 22]. In our unit, we observed that the discrimination power of ES II did not differ according to the surgical volume of single surgeons; however, the operations performed in the group ‘Others’ were proctored, with a senior either scrubbed or in theatre. Lastly, it may be speculated that continuous improvements in terms of postoperative care may have a significant impact in reduction of complications, and that may contribute to the fact that ES II overestimates mortality.

Study limitations

This study has several limitations. Although it includes a relative large number of patients (1609) for whom data were inserted prospectively in the database, it is based on a retrospective analysis; moreover, this is a single-centre experience with all the limitations that are involved. ‘Poor mobility’ was inserted a posteriori by reviewing the electronic notes. We considered uniquely the in-hospital mortality and no analysis on mid- or long-term survival was carried out. Moreover, we did not distinguish between repair and replacement. Another limitation may be the definition of high risk, which although may follow a statistical rationale, ultimately remains to be considered arbitrary. Another limitation is that ES II performance was not tested specifically in the strata of low- or medium-risk patients.

CONCLUSIONS

With this study, we concluded that ES II showed good discrimination power for patients who underwent minimally invasive mitral valve surgery, better than that of the previous logistic version. Neither individual performance nor that of surgeons in training significantly disrupted the accuracy of ES II. However, ES II performance was suboptimal in predicting postoperative complications. Further data are needed for the validation of ES II in high-risk subgroups.

Conflict of interest: none declared.

REFERENCES


