Performance of EuroSCORE II in patients who have undergone heart valve surgery: a multicentre study in a Chinese population

Lv Wang†a, Qing-Qi Han†a, Fan Qiao†a, Chong Wanga, Xi-Wu Zhaonga, Lin Han†a and Zhi-Yun Xua,*

a Department of Cardiac and Thoracic Surgery, Changhai Hospital, Second Military Medical University, Shanghai, China
b Department of Cardiac and Thoracic Surgery, No. 252 Hospital of the Chinese People’s Liberation Army, Hebei, China

* Corresponding author. Department of Cardiac and Thoracic Surgery, Changhai Hospital, Second Military Medical University, Shanghai, China.
Tel: +86-21-31161741; fax: +86-21-64590979; e-mail: zhiyunx@hotmail.com (Z.-Y. Xu).

Received 9 January 2013; received in revised form 5 April 2013; accepted 16 April 2013

Abstract

OBJECTIVES: The EuroSCORE II is an updated version of the EuroSCORE. This multicentre study validated the EuroSCORE II and logistic EuroSCORE in Chinese patients who underwent heart valve surgery.

METHODS: A total of 11,170 adult patients underwent heart valve surgery from January 2008 to December 2011. Model discrimination and calibration were assessed for both EuroSCORE II and logistic EuroSCORE. The patients were divided into three subgroups according to the weight of the procedures, and the performance of EuroSCORE II for each group was assessed. A correlation analysis was performed for operative complications and EuroSCORE II.

RESULTS: The in-hospital mortality of this series was 2.02% (226 of 11,170), and the predicted mortality rate was 2.62 ± 5.75% by EuroSCORE II and 2.55 ± 6.51% by logistic EuroSCORE (LES). The C-statistics of EuroSCORE II and LES were 0.72 [95% confidence interval (CI) 0.69–0.75] and 0.67 (95% CI 0.63–0.70), respectively. Both models failed the Hosmer–Lemeshow goodness-of-fit test, with a P < 0.05. According to the weight of the procedure, the isolated non-CABG subgroup had the best discrimination (C-statistics: 0.76 in the non-CABG group, 0.67 in the 2 procedures group and 0.73 in the 3+ procedures group). The complication ratio was strongly related to the EuroSCORE II-predicted mortality (Pearson correlation coefficient: 0.90 for ARDS, 0.97 for acute renal failure, 0.97 for prolonged ventilation and 0.94 for a prolonged ICU stay).

CONCLUSIONS: EuroSCORE II was an improvement upon its original logistic model for Chinese patients who underwent heart valve surgery, particularly for a single-valve procedure. The EuroSCORE II-predicted mortality correlated with the operative complications.

Keywords: Risk factor • EuroSCORE • Heart valve surgery • Validation

INTRODUCTION

The European System for Cardiac Operative Risk Evaluation (EuroSCORE) was a logistic regression model for cardiac operative mortality. It was initially published in 1999 as an additive model (additive EuroSCORE, AES) for calculating the bedside operation risk [1]. It was then used in some more sophisticated fields, such as the study of risk, quality control and interinstitutional comparisons. To fulfil the requirements mentioned above, the logistic EuroSCORE (LES) was published in 2003 [2]. The EuroSCORE has since become a standard cardiac surgery risk prediction tool used worldwide.

With improved outcomes in cardiac surgery, the EuroSCORE was recently reported to overestimate the cardiac surgery risk [3–8]. EuroSCORE II, an updated model of the original EuroSCORE, was developed to fit contemporary cardiac surgical practices. The EuroSCORE II was proven to have highly satisfactory internal validation results, in both the model calibration and the discrimination

[9]. Recent studies for the external validation of EuroSCORE II confirmed that EuroSCORE II greatly improved upon the original AES and LES [10, 11]. Other studies demonstrated that EuroSCORE II had good or excellent accuracy in generic cardiac surgery, but was not well suited for isolated CABG [11–13]. The performance of the EuroSCORE II for heart valve surgery has not been well validated.

The purpose of this study was to estimate the performance of EuroSCORE II for patients who underwent heart valve surgery in four cardiac surgical units in China and to compare its prediction performance with that of the logistic EuroSCORE.

MATERIALS AND METHODS

Patients

This multicentre retrospective study was performed in patients who underwent heart valve surgery at four cardiac surgical units in the period from January 2008 to December 2011. The cardiac surgical units participating in this study included Fu Wai Hospital in
Beijing; Changhai Hospital in Shanghai; Zhongshan Hospital of Fudan University in Shanghai; and the Guangdong Cardiovascular Institute in Guangzhou. The patient preoperative characteristics and operation types were collected from the patients’ medical records to fit the EuroSCORE II. As the EuroSCORE II is a cardiac surgery risk model that could be applied for all types of adult cardiac surgery, all patients who underwent valve surgery were included in this study. To ensure the reliability of the study results, those patients who underwent congenital surgeries were excluded. We also excluded patients with missing data for sex, age, valve surgery type or status when discharged from the hospital.

Data collection

Information on patients and the procedural risk factors for all patients were collected retrospectively from their medical records, resulting in a completely retrospective dataset stored in a local cardiac valve database. The database contained all the risk factors that were included in the final models of the EuroSCORE II and EuroSCORE. Percentage of missing value of the original EuroSCORE II risk factor was <1%. The EuroSCORE II- and EuroSCORE-predicted mortality risks were calculated for every patient in the final dataset by the published logistic algorithm [2, 9].

The Canadian Cardiovascular Society (CCS) angina class were not presented in the original medical files of patients with no symptoms of angina, and we therefore assumed that any patients without a record of angina in the original medical files were CCS level I, as the majority of our patients with valve surgery were relatively young and rarely had coronary artery disease as well.

Creatinine clearance was calculated according to the Cockcroft-Gault equation [14], which is based on the patient’s age, weight, gender and serum creatinine value (mg/dl). Acute renal failure was defined as a new requirement for dialysis or an increase in the serum creatinine to >2.0 mg/dl and double the most recent preoperative serum creatinine level. ARDS was defined as a postoperative PaO₂/FiO₂ ratio ≤200 within 48 h, with diffuse bilateral infiltrates of the lungs upon examination of chest X-rays. The EuroSCORE II was designed for almost all the adult cardiac surgical procedures as it weights the surgical procedures as CABG, 1 non-CABG, 2 major procedure and 3+ major procedures. The whole patient set was divided into three subgroups according to the weight of the procedure as EuroSCORE II did, to evaluate the performance of EuroSCORE II for type of procedures.

The prediction accuracy of the EuroSCORE II was evaluated not only for in-hospital mortality but also for other major postoperative complications, including adult respiratory distress syndrome (ARDS), acute renal failure, prolonged ventilation (ventilation > 48 h) and prolonged ICU stay (> 72 h). Next, the patient sets were divided into deciles according to the EuroSCORE II-predicted mortality to estimate the relationship between the observed complication rate and the EuroSCORE II score.

Statistical analysis

The model performances of the EuroSCORE II and EuroSCORE were assessed for both discrimination and calibration. The model discrimination was assessed by the C-statistic (index of concordance) with a 95% confidence interval (CI). The C-statistic was the proportion of predictions that were concordant with the observations and was calculated as the area under the receiver operating characteristic curve. The C-statistic values ranged from 0.5 (no ability to discriminate) to 1.0 (full ability to discriminate). The discriminative power of the model was considered reasonable when the C-statistic exceeded 0.7 and strong when the C-statistic exceeded 0.8 [15]. The model calibration was evaluated by the Hosmer–Lemeshow goodness-of-fit (H–L) test. The calibration plot shows the mean predicted mortality against the observed mortality for 10 equally sized groups based on the ranked predicted risks calculated by the EuroSCORE II. A P-value < 0.05 was considered to indicate statistical significance. The correlation between the EuroSCORE II and the operative complications was analysed with the C-statistic. Univariate regression analysis was performed to estimate the relationship between the observed complication rate and the EuroSCORE II score. All statistical analyses were performed using SPSS statistical software package version 19.0 (IBM/SPSS, Inc., Chicago, IL, USA).

RESULTS

Patient population

The original dataset included 12 170 patients; 924 patients were excluded for congenital heart surgery; 49 were excluded for missing gender, age and surgery type information; and 27 were excluded for missing the discharge status. The final dataset of 11 170 patients comprised the study patient population. The study population at our institutions was younger, had more females and fewer preoperative complications than that of the EuroSCORE II population. The baseline patient characteristics of our study population and those of the EuroSCORE II are presented in Table 1. The operative details of the study population are listed in Table 2.

Operative mortality

The in-hospital mortality of this series was 2.02% (226 of 11 170), and the predicted mortality rate was 2.62% by the EuroSCORE II and 2.55% by the LES. The O/E ratios of our study population with the EuroSCORE II and LES were 0.77 and 0.80, respectively. The EuroSCORE II performed better than the LES in predicting the operative mortality, with a C-statistic of 0.72 (95% CI 0.69–0.75) using the EuroSCORE II and 0.67 (95% CI 0.63–0.70) using the LES (Fig. 1). Both models failed the Hosmer–Lemeshow goodness-of-fit (H–L) test. The calibration plot of the EuroSCORE II was very close to the 45° line (indicating perfect prediction) until the 9th and 10th deciles, whereas the plot of the LES was scattered far from the 45° line. As demonstrated by the calibration plots, the EuroSCORE II predicted mortality corresponded well with the observed mortality except for the last two deciles, whereas the LES had poor calibration over the entire range of probabilities.

Model performance by weight of the procedure

The whole series was divided into three subgroups according to the weight of the procedure (isolated non-CABG procedure, 2...
The isolated non-CABG subgroup had the best discrimination using the EuroSCORE II (C-statistics: 0.76 in the single non-CABG group, 0.67 in the 2-procedures group and 0.73 in the 3+ procedure group). The H-L test suggested good calibration in all three subgroups ($\chi^2 = 15.52$ and $P = 0.11$ in the isolated non-CABG procedure group, $\chi^2 = 12.10$ and $P = 0.28$ in the 2 procedures group, and $\chi^2 = 15.73$ and $P = 0.11$ in the 3+ procedures group). Among the three subgroups, the 3+ procedures subgroup seriously overestimated mortality (O/E ratio: 0.81 in the single non-CABG group, 0.83 in the 2 procedures group and 0.67 in the 3+ procedures group). The EuroSCORE II had both good discrimination and good calibration for a single, non-CABG procedure, but the concordance was not as good for the 2 and 3+ procedures groups. Table 4 shows the details of the EuroSCORE II performance in the three subgroups.

**EuroSCORE II and operative complications**

The model discrimination of the EuroSCORE II for major postoperative complications, such as ARDS, acute renal failure,
prolonged ventilation and prolonged ICU stay, was evaluated. The EuroSCORE II showed reasonable discrimination for postoperative ARDS (C-statistics 0.75) and prolonged ventilation (C-statistics 0.70) and showed poor discrimination for acute renal failure (C-statistics 0.65) and prolonged ICU stay (C-statistics 0.66). Table 5 shows the C-statistics values for the end-points of these complications.

The patient dataset was divided into 10 groups according to the EuroSCORE II-predicted mortality. The univariate regression and correlation analyses between the EuroSCORE II-predicted mortality and the risk of complications showed that the complication ratio was strongly correlated with the EuroSCORE II-predicted mortality (Pearson correlation coefficient: 0.90 for ARDS, 0.97 for acute renal failure, 0.97 for prolonged ventilation and 0.94 for prolonged ICU stay; see Table 5). Fig. 3 demonstrates the relationship between the operative complications and the EuroSCORE II.

**DISCUSSION**

This study revealed the great improvement of the EuroSCORE II over its original logistic model in predicting hospital mortality in Chinese patients who underwent heart valve surgery. The EuroSCORE II had both good discrimination and good calibration for the entire data series and had the best discriminatory power in the subgroup of patients who underwent an isolated, non-CABG procedure (e.g. AVR/MVR/MVR repair) compared with those patients who underwent more complex surgical procedures. The EuroSCORE II correlated with other postoperative complications, such as acute renal failure, ARDS, prolonged ventilation and prolonged ICU stay.

For the patients in this study, the EuroSCORE II showed good prediction accuracy with a C-statistic of 0.72 in the whole series, whereas the LES only had a C-statistic of 0.66. This result suggests that the EuroSCORE II was more suitable than the LES for the prediction of valve surgery mortality in our study population. Other validation studies using the LES in the Chinese population demonstrated that the EuroSCORE did not have a strong predictive power for CABG or valve surgery [8, 16–18]. A validation study of the EuroSCORE and EuroSCORE II in an Italian population [19] revealed that both models had excellent discrimination for patients who underwent adult cardiac surgery (C-statistic of 0.82 for both models) but did not provide accurate probabilities (both failed the H–L test). Another study in an Italian population demonstrated that the EuroSCORE II had better calibration and similar discrimination compared with the old version of the EuroSCORE in patients who underwent CABG [11]. In a large sample study in the UK, the EuroSCORE II demonstrated better discrimination and calibration [12]. These studies showed the excellent discrimination of the EuroSCORE II, with C-statistics of 0.82 [19], 0.81 [11] and 0.81 [12]. The EuroSCORE remained excellent in the two studies in the Italian population, with C-statistics of 0.78–0.8. The model performance of both the EuroSCORE II and EuroSCORE in the Chinese population was not as good as that for the Italian population, which may reflect the different demographics and varied medical care of the patients in the two countries. The calibration plots also demonstrated that the EuroSCORE II was more suitable than the LES for the prediction of operative mortality in a Chinese population with valve disease, particularly for patients with low risk. The overestimation of mortality by the EuroSCORE II for the last two deciles of patients may have led to the failure of the model calibration.
There are huge differences in the aetiology of heart valve disease between the EuroSCORE II population and our study population. It was well demonstrated that degenerative etiologies were the most frequent in aortic VHD and mitral regurgitation population. It was well demonstrated that degenerative etiologies lead to the majority of VHD in the European population. Valve degeneration leads to the majority of VHD in the European population. For the patients with mitral valve disease, left atrial and pulmonary arterial pressures increase over time and longstanding pressure overload to the right ventricle induces right ventricular enlargement, tricuspid annular dilatation and leaflet tethering. These pathological changes on TV geometry cause functional TR, with concomitant changes on TV geometry cause functional TR, with concomitant

d術 repairment in our study population, which makes up about 60% of our patient set. As the reason mentioned above, our study populations have high incidences of mitral valve disease. For the patients with mitral valve disease, left atrial and pulmonary arterial pressures increase over time and longstanding pressure overload to the right ventricle induces right ventricular enlargement, tricuspid annular dilatation and leaflet tethering. These pathological changes on TV geometry cause functional TR, with concomitant tricuspid valve repairment when undergoing MV surgery. Overall, the high incidence of concomitant tricuspid valve repairment in our study population reflects the baseline character difference between the Chinese population and the EuroSCORE II population, which may be part of the reason why EuroSCORE II did not have as good predictive power for our study population as for the European population.

Adult cardiac surgery includes numerous surgical procedures and combinations of procedures. The type of procedures or interventions played an important role in predicting the cardiac surgery risk. Complex surgical interventions may lead to a high risk of mortality, as they are always associated with longer operation and CBP times, greater iatrogenic trauma and a more severe cardiac pathological state. The EuroSCORE II was suitable for almost all adult cardiac procedures after cataloguing the cardiac interventions by the weight of the procedures. The inclusion of the risk factor of the impact of the surgical treatment on the operative mortality was a great improvement with the use of the EuroSCORE II as the LES did not take the weight of the procedure into account. However, is this simple classification accurate enough for all cardiac surgical procedures? In this study of Chinese patients undergoing valve surgery, the EuroSCORE II performed the best for an isolated, non-CABG procedure, with a C-statistic of 0.76, which was higher than the 0.67 for 2 procedures and 0.73 for 3+ procedures. The validation study in the UK had a similar result of model discrimination according to the weight of the procedure (the C-statistics for a single non-CABG procedure, 2 procedures and 3+ procedures were 0.78, 0.73 and 0.73, respectively) [12]. The results of the UK and our studies both suggested that the EuroSCORE II had the best discrimination for an isolated, non-CABG procedure, followed by 3+ procedures and then 2 procedures. To achieve an equally satisfactory model performance for all surgical procedures, establishing multiple models, such as the STS2008 cardiac surgical risk model [21-23], may be an option. However, that needs sufficient patient volume for each surgical procedure, which may be impossible for some relatively uncommon surgical procedures. At present, EuroSCORE II still one of the best choices for adult cardiac surgical risk prediction.

As the most popular test method for model calibration of a logistic regression model in the most recent 10 years, the H–L test was used in almost all of the EuroSCORE validation studies. In this study, the EuroSCORE II failed the H–L test in the entire series. The calibration plot demonstrated that the predicted mortality highly correlated with the observed mortality in low- to middle-risk patients. When dividing the study population into three subgroups, the H–L test of each subgroup by the EuroSCORE II did not reject the null hypothesis for any of the subgroups and showed good calibration. The UK validation study also demonstrated that the EuroSCORE II failed the H–L test for the analysis of their entire study population, but showed good calibration for their analysis of the subgroups according to the weight of the procedure [12]. The H–L test appears to fail more often in larger samples than in relatively small samples. Consequently, Nashef et al. [9] suggested using the risk-adjusted mortality ratio (RAMR = observed/predicted mortality) instead of the H–L test to evaluate the calibration of a risk model. A RAMR of 1 indicates that the predicted mortality is consistent with the observed mortality and that the risk prediction model is well calibrated. In this study, the RAMR of the overall population by the EuroSCORE II was 0.77, suggesting that the EuroSCORE II overestimated mortality in the whole series.

The EuroSCORE was used to predict not only operative mortality but also other complications, such as operative morbidities

<table>
<thead>
<tr>
<th>Complication</th>
<th>C-statistics (95% CI)</th>
<th>Pearson correlation coefficient (P)</th>
<th>Regression coefficient (95% CI)</th>
<th>Constant term (95% CI)</th>
</tr>
</thead>
<tbody>
<tr>
<td>ARDS</td>
<td>0.75 (0.67–0.82)</td>
<td>0.90 (&lt;0.001)</td>
<td>6.05 (3.65, 8.45)</td>
<td>0.01 (-0.00, 0.02)</td>
</tr>
<tr>
<td>RF</td>
<td>0.65 (0.61–0.68)</td>
<td>0.97 (&lt;0.001)</td>
<td>1.82 (1.44, 2.20)</td>
<td>-0.01 (-0.02, -0.00)</td>
</tr>
<tr>
<td>Vent</td>
<td>0.70 (0.68–0.72)</td>
<td>0.97 (&lt;0.001)</td>
<td>0.54 (0.43, 0.64)</td>
<td>-0.01 (-0.01, 0.00)</td>
</tr>
<tr>
<td>ICU</td>
<td>0.66 (0.65–0.67)</td>
<td>0.94 (&lt;0.001)</td>
<td>3.97 (2.77, 5.18)</td>
<td>0.11 (0.07, 0.15)</td>
</tr>
</tbody>
</table>

ARDS: adult respiratory distress syndrome; RF: acute renal failure; Vent: prolonged ventilation; ICU: prolonged ICU stay.
and recovery time [24, 25]. We also assessed the relationship between the EuroSCORE II-predicted mortality and operative complications. When the patients were divided into 10 groups according to the EuroSCORE II-predicted mortality, the mean EuroSCORE II-predicted mortality also correlated well with the ratio of other operative complications, such as ARDS, acute renal failure, prolonged ventilation and prolonged ICU stay, with Pearson correlation coefficients ranging from 0.90 to 0.97. For operative complications, the EuroSCORE II only showed good discrimination for ARDS and prolonged ventilation, with C-statistics of 0.75 and 0.70, respectively. Therefore, the EuroSCORE II can be used for approximate predictions of operative complications, particularly for ARDS and acute renal failure, using a linear regression equation.

A limitation of this study is that the end-point was defined as hospital death recorded in the original medical record; however, the true mortality after valve surgery would be higher. The EuroSCORE II defined the end-point as death either prior to discharge from the hospital or within 30 or 90 days of the surgery, but used the standard of death in the hospital where the operation took place. This definition would lead to an underestimation of the observed mortality in this study, which may be one of the reasons that the mortality in our study population was low. Second, the proportions of each specific valve surgery type differed widely. We were not able to evaluate the predictive power of the EuroSCORE II in each valve surgery type due to an insufficient number of cases. Third, as a retrospective study, the incorrect and missing data in the original medical records could not be re-collected. These abovementioned limitations may lead to some bias in our study results.

ACKNOWLEDGEMENTS

The consulting surgeons were Wang Wei, Wang Chun-Sheng and Zhuang Jian. We thank all members of Department of Cardiothoracic Surgery of Fuwai Hospital, Zhongshan Hospital and Guangdong Cardiovascular Institute who contributed data to the study for investing their time and effort for the data collection, without which this work would not have been possible.

Funding

This work was supported by the Specialty Fund of Health Ministry of China: standardized treatment of heart valve disease (grant number 200802096).

Conflict of interest: none declared.

REFERENCES