The validity of the original EuroSCORE and EuroSCORE II in patients over the age of seventy

Michael Poullis*, Mark Pullan, John Chalmers and Neeraj Mediratta

Department of Cardiac Surgery, Liverpool Heart and Chest Hospital, Liverpool, UK

* Corresponding author. Consultant Cardiothoracic Surgeon, Liverpool Heart and Chest Hospital, Thomas Drive, Liverpool L14 3PE, UK. Tel: +44-151-2281616; fax: +44-151-2932254; e-mail: mpoullis@hotmail.com (M. Poullis).

Received 21 March 2014; received in revised form 4 August 2014; accepted 7 August 2014

Abstract

OBJECTIVES: EuroSCORE II, despite improving on the original EuroSCORE system, has not solved all the calibration and predictability issues. We investigated the sensitivity, specificity and predictability of original EuroSCORE and EuroSCORE II system in elderly patients.

METHODS: The original logistic EuroSCORE and EuroSCORE II were assessed via receiver operator characteristic (ROC) and Hosmer-Lemeshow test probability analysis with regard to accuracy of predicting in-hospital mortality. Analysis was performed on isolated coronary artery bypass graft (CABG) \(n = 2913\), aortic valve replacement (AVR) \(n = 814\), mitral valve surgery (MVR) \(n = 340\), combined AVR and CABG cases \(n = 517\) and the above cases combined \(n = 4584\). Elderly was defined as \(\geq 70\) years old. Age <70 was used for comparative purposes.

RESULTS: Institutional mortality was 2.9%, for all isolated CABG, AVR, MVR and combined AVR and CABG cases. In all patients aged \(\geq 70\) neither the original EuroSCORE nor EuroSCORE II had a ROC c-statistic above 0.7. For isolated CABG, the ROC c-statistic was not acceptable in patients \(\geq 70\) years of age, but was fine for patients under the age of 70 years. For isolated AVR the ROC c-statistic was good for patients aged less than 70 years of age for both risk models; however, the ROC was unacceptably low in patients aged \(\geq 70\) for both models. For isolated MVR, the ROC c-statistic and Hosmer–Lemeshow test probability was good for all patients regardless of age. For combined AVR and CABG, the ROC c-statistic was unacceptably low for all patients, regardless of age group using the original EuroSCORE, and in those aged \(\geq 70\) using the EuroSCORE II risk model. The original EuroSCORE had no issues with the Hosmer–Lemeshow test probability; however, EuroSCORE II had poor model predictability for all patients, \(P < 0.0001\), and for isolated CABG, \(P = 0.05\) and AVR, \(P = 0.06\).

CONCLUSIONS: The original EuroSCORE and the EuroSCORE II risk models should be used with caution in patients aged 70 or older undergoing cardiac surgery in the modern era. Below the age of 70, both models are sensitive, specific and have good predictive power. Our work needs validation by other large groups.

Keywords: Risk modelling • EuroSCORE

INTRODUCTION

The original EuroSCORE system was an important step in risk assessment in adult cardiac surgery [1]. Over the subsequent decade after its introduction, issues with regard to calibration and predictability arose. EuroSCORE II improved on the original EuroSCORE system but has not solved all the calibration and predictability issues [2, 3].

Any surgeon will testify that aortic calcification and diffuse coronary disease are more important in determining mortality than any of the risk factors used to compute the EuroSCORE, original or II. Neither of these factors are possible to quantify easily and no accepted quantification system has been widely adopted, partially explaining why they are not part of the EuroSCORE risk prediction system. Age, diabetes and renal impairment are important risk factors for calcification, the latter two are part of EuroSCORE II. Previous work evaluating/comparing EuroSCORE and EuroSCORE II did not sub-analyse by age, limiting model assessment by operation type only [2].

We investigated if elderly patients, defined as aged 70 years or older, who are more likely to have aortic calcification and diffuse coronary disease may not be adequately modelled by the EuroSCORE system, original or II.

METHODS

Patients

The characteristics of the patients in this study are tabulated in Table 1. Patients who had undergone consecutive isolated coronary artery bypass graft (CABG), isolated aortic valve replacement (AVR), isolated mitral valve surgery (MVR), combined AVR and CABG and other (aortic, left ventricular aneurysmectomy, acquired ventricular septal defect, atrial myxoma, pulmonary embolectomy, cardiac trauma, pericardiectomy, atrial septal closure etc.) were included, total \(n = 5576\), between January 2006 and
March 2010. In-hospital mortality was defined as death at the institution where surgery was performed. All patients were operated on at the Liverpool Heart and Chest hospital. The original and EuroSCORE II risk scores were calculated as previously described [2, 4]. Pulmonary artery pressures were not available for the majority of the patients, and poor mobility was not available for any of the patients (see Limitations section).

**Statistical analysis**

Qualitative variables were expressed as percentages and quantitative variables as mean ± standard deviation or median (interquartile range) according to variable distribution and were compared by use of the \( \chi^2 \) or Fisher’s exact test and unpaired t-test or Mann–Whitney U-test as appropriate.

The performances of the original EuroSCORE and the EuroSCORE II system with regard to our institute outcomes were assessed via receiver operator characteristics (ROC) for sensitivity and specificity and the Hosmer–Lemeshow test probability, for predictive power. An ROC c-statistic greater than 0.7 and a non-significant Hosmer–Lemeshow test probability (\( P > 0.05 \)) were considered clinically acceptable with regard to sensitivity, specificity and calibration.

Analyses were performed for all operations, isolated CABG, isolated AVR, isolated MVR and combined AVR and CABG.

**Elderly** was defined as age ≥70 years old. All patients and patients aged less than 70 years old were used as a comparator group.

**Benchmarking**

We benchmarked our in-hospital mortality figures against the UK national results (http://www.scts.org).

**Statistical software**

All statistical analyses were performed with MedCalc for Windows, version 13.1.1 (MedCalc Software, Mariakerke, Belgium).

**RESULTS**

Benchmarking of our institutional mortality rates compared with the UK, mortality in 2012, 2.9%, \( n = 34,174 \), did not reveal any differences, \( P = 0.1 \) (part of the continuous UK cardiac surgery quality assessment programme by the society of cardiothoracic surgeons, www.scts.org).

The logistic EuroSCORE had a median of 3.7 for patients under the age of 70, and a median of 6.9 for patients aged 70 or older, EuroSCORE II had a median of 2.5 for patients under the age of 70 and a median of 5.0 for patients aged 70 or older (Table 1).

**Baseline risk model assessment**

The ROC c-statistic and Hosmer–Lemeshow test probability for all age groups for all operations, isolated CABG, AVR, MVR and
combined AVR and CABG is shown in Fig. 1A and B, respectively. The original EuroSCORE had a ROC below 0.7 for isolated AVR and combined AVR and CABG, indicating suboptimal risk prediction for clinical usage. The original EuroSCORE had no issues with the Hosmer–Lemeshow test probability; however, the EuroSCORE II system had a highly significant statistic, $P < 0.0001$, for all patients and a $P$ value of 0.05 and 0.06 for isolated CABG and AVR, respectively, implying poor model predictability.

All patients

Figure 2 demonstrates the ROC c-statistic and Hosmer–Lemeshow test statistic for all patients analysed by age group. In patients aged $\geq 70$ the original EuroSCORE had a ROC c-statistic below 0.7, indicating suboptimal risk prediction for clinical usage. The original EuroSCORE had no issues with the Hosmer–Lemeshow test probability; however, the EuroSCORE II system had a highly significant statistic, $P < 0.0001$, for all patients and a $P$ value of 0.05 and 0.06 for isolated CABG and AVR, respectively, implying poor model predictability.

Procedure breakdown

The ROC c-statistic and Hosmer–Lemeshow test statistic plots broken down by procedure and age are shown in Figs 3A–D and 4A–D, respectively.

Isolated coronary artery bypass graft

The ROC c-statistic was acceptable for patients regardless of age (Fig. 3A). The original EuroSCORE was, however, only just acceptable in patients aged $\geq 70$. Subanalysis, with patients aged over 80 removed, confirmed the finding, indicating that no interaction of the group aged over 80 was skewing the result.

The Hosmer–Lemeshow test probability was satisfactory for both models regardless of age.

Isolated aortic valve replacement

The ROC c-statistic was unacceptably low in patients aged above 70 for both models. The Hosmer–Lemeshow test probability was satisfactory for all age groups regardless of risk model. Room for improvement existed with regard to Hosmer–Lemeshow in patients less than 70 years of age (Fig. 4B).

Isolated mitral valve surgery

The ROC c-statistic and Hosmer–Lemeshow test probability were acceptable for all patients. Room for improvement existed with regard to Hosmer–Lemeshow for both models in patients $\geq 70$ years of age (Fig. 4C).

Combined aortic valve replacement and coronary artery bypass graft

The ROC c-statistic was unacceptably low for patients $\geq 70$ years of age for both the original EuroSCORE and EuroSCORE II risk models. The ROC c-statistic was just acceptable in patients less than 70 with the original EuroSCORE system (Fig. 3D).

The Hosmer–Lemeshow test probability was satisfactory for both risk models regardless of age.

Age distribution of adult cardiac surgery population

It can be seen from Fig. 5A that 40% of our patients are aged $\geq 70$ years old, and 22% are above 75 years old, indicating that a significant number of patients are affected by the issues with regard to
model accuracy documented above. The difference between predicted and observed mortality by age for our institution is shown in Fig. 5B.

**Number of risk factors by age group**

It can be seen from Fig. 6 that as patients increase in age, the number of risk factors they have increases. The imbalance of risk factor number may explain the model prediction imbalance between patients aged above or below 70.

**DISCUSSION**

The original EuroSCORE and the EuroSCORE II risk models do not have adequate ROC and Hosmer–Lemeshow test probability to allow accurate assessment of patients aged ≥70 years undergoing……

The original EuroSCORE and EuroSCORE II were developed with a patient population mean age of 62.5 years old and 64.9 years old respectively; our study had a mean age of 69.3 years of age [1, 3]. The development of a risk model in patients substantially younger than the current mean age of patients undergoing cardiac surgery may help to explain the loss of risk model calibration.

It is well known that the average age of patients undergoing cardiac surgery is increasing, who frequently have more comorbid risk factors, and with an increasing proportion undergoing combined valve and graft procedures [5, 6]. The finding that neither of the risk models is accurate in this increasingly important patient and operation subgroup is of concern, as 40% of our patients are aged ≥70 years. Previous work has identified issues with both the original EuroSCORE and EuroSCORE II in octogenarian and high-risk patients [7, 8].

The original and EuroSCORE II prediction models rely on a logarithmic formula involving the risk covariates. Each covariate has a confidence interval. This confidence interval despite being described in both the original and EuroSCORE II risk model papers is never utilized to provide a confidence interval for the predicted risk score. The larger the number of factors, the more covariates, the larger the confidence intervals will be for the predicted mortality. We demonstrated that older patients have more covariates as risk factors.

We also suggest that aortic calcification and diffuse coronary disease are more important in determining mortality than many of the risk factors used to compute the EuroSCORE, original or II. Neither of these factors are included in any risk models, but both are more common in elderly patients, and we speculate, but offer no direct evidence, this may also help to explain why the models do not work well in patients over the age of 70 years of age.

The original and EuroSCORE II risk models were devised utilizing mortality defined as death within 30 days of operation or within the same hospital admission. We utilized in-hospital mortality regardless of duration as our definition as this is the quality metric utilized in the UK to assess cardiac surgeons’ risk-adjusted outcomes. Re-analysis, based on the EuroSCORE definition of death, identified no significant differences with regard to ROC and Hosmer-Lemeshow values regardless of method of defining death post surgery (data not shown).

The strength of our analysis is that our series is modern, consecutive and large. Our database has been independently validated, and benchmarking did not identify any significantly differences between our institutional results and those from the Great Britain and Ireland National Adult Surgery Cardiac Database Report 2008, so we feel our data are robust and applicable to other units in the UK [2].

Risk models for death can be quantified via the use of the c-statistic, and the Hosmer-Lemeshow test. The c-statistic is the area under a ROC characteristic, and represents the probability that predicting the outcome is better than chance. It compares the goodness of fit of logistic regression model. A value of 0.5 indicates that the model is no better than chance at predicting the outcome, and a value of 1.0 indicates that the model perfectly identifies the outcome. Models are typically considered clinically usable when the C-statistic is higher than 0.7 and strong when C exceeds 0.8 [2, 9]. The Hosmer-Lemeshow assesses goodness of fit for logistic regression models, and assesses whether or not the observed event rates match expected event rates in deciles of fitted risk values of the model population. Models, which have expected and observed event rates in the subgroups that are similar are called well calibrated, and have a non-significant Hosmer-Lemeshow test probability. The more non-significant the Hosmer-Lemeshow test probability the better the model calibration (ability to predict mortality in differing risk patients).

The finding of ROC scores between 0.5 and 0.6 for patients aged ≥70 undergoing isolated AVR, means the models are only a little better than guess work with regard to risk prediction for this procedure. The finding that EuroSCORE does not accurately predict mortality after AVR well has been previously described in the elderly [10].

Both risk models were well calibrated for all age groups as assessed by the Hosmer-Lemeshow test probability. We suggest that this is due to the effect of age, as the risk of cardiac surgery is known to increase exponentially with age, which is the mathematical basis of logistic regression, which the EuroSCORE system relies upon. Older patients tend to have higher number of risk factors, so the model accuracy decreases. This underlies the theory of the Akaike information criterion.
Issues with EuroSCORE accuracy have been addressed in the UK by a process called recalibration [11]. This consists in multiplying the EuroSCORE predicted risk by a fixed scalar quantity so that the cumulative sum curve fit is improved. The nonlinear relationship with age between observed and predicted risk demonstrates why a scalar multiplier is methodologically inaccurate.

The original EuroSCORE and EuroSCORE II relies on logistic regression modelling, which is a logarithmic mathematical technique. The use of neural networks [12], Bayesian techniques [13, 14] and covariate risk factor matching [15] may however offer a better modelling technique, due to the nonlinear interaction of the known risk factors. Despite numerous attempts at remodelling, using logistic regression has failed to produce a model that is sensitive, specific and well-calibrated. Changing the modelling technique may be overdue, and in an era of large computerized databases, covariate risk factor matching is theoretically advantageous [15].

Despite being a very large institution, we did not have enough patients to evaluate the numerous other combinations and permutations of CABG and mitral valve, other cardiac and aortic surgery interventions. We did not perform separate subset analysis of these cases to avoid drawing conclusion on a statistically underpowered analysis.

LIMITATIONS

EuroSCORE II has introduced the variable, poor mobility, as a risk factor. This is not recorded in our database, so all patients were analysed as not scoring for this risk factor. The number of patients with poor mobility in our institution, the definition of which is difficult to define exactly, that would affect our analysis is very low. The majority of CABG patients did not have pulmonary artery pressures measured preoperatively. For purposes of analysis they were assumed to be normal and so did not score. This is a singleinstitution study and results thus have limitations and may not represent national and international practice and outcomes, and thus may represent an important source of potential bias.

CONCLUSIONS

The original EuroSCORE and the EuroSCORE II risk models should be used with caution in patients aged ≥70 years undergoing cardiac surgeries in the modern era. Below the age of 70, both models are sensitive, specific and have good predictive power. Our work needs validation by other large groups.

Conflict of interest: none declared.

REFERENCES