Performance of the EuroSCORE Models in Emergency Cardiac Surgery

Stuart W. Grant, MRCS; Graeme L. Hickey, PhD; Ioannis Dimarakis, PhD, MRCS; Graham Cooper, MD, FRCS (CTh); David P. Jenkins, MS, FRCS (CTh); Rakesh Uppal, FRCS (CTh); Iain Buchan, MD, FFPH; Ben Bridgewater, PhD, FRCS (CTh)

Background—Accurate risk-adjustment models are useful for clinical decision making and are important for minimizing any tendency toward risk-averse clinical practice. In cardiac surgery, emergency patients are potentially at greatest risk of inappropriate risk-averse clinical decisions. UK cardiac surgery outcomes are currently risk-adjusted with EuroSCORE models. The objective of this study was to assess the performance of the EuroSCORE models in emergency cardiac surgery.

Methods and Results—The National Institute for Cardiovascular Outcomes Research database was used to identify adult cardiac surgery procedures performed in the United Kingdom between April 2008 and March 2011. A subset of procedures (July 2010–March 2011) was used for EuroSCORE II validation. The outcome measure was in-hospital mortality. Model calibration (Hosmer–Lemeshow test, calibration plots, calculation of calibration intercept and slope) and discrimination (area under receiver-operating characteristic curve [area under the curve]) were assessed. In total, 109,988 cardiac procedures at 41 hospitals were included, of which 3342 were defined as emergency procedures. Compared with performance in all cardiac surgery and nonemergency cardiac surgery, the logistic EuroSCORE and EuroSCORE II models had poorer discrimination (area under the curve, 0.703 and 0.690, respectively) and poorer calibration for emergency surgery. The EuroSCORE risk factors of female sex, chronic pulmonary disease, neurological disease, active endocarditis, unstable angina, recent myocardial infarction, and pulmonary hypertension were not identified as important risk factors for emergency cardiac surgery.

Conclusions—Both EuroSCORE models demonstrated poor calibration and comparatively poor discrimination for emergency cardiac surgery. This has important implications when these models are used for clinical decision making or to adjust governance analyses. (Circ Cardiovasc Qual Outcomes. 2013;6:0-0.)

Key Words: cardiac surgery • clinical governance • emergency surgery • EuroSCORE • risk prediction
WHAT IS KNOWN

- Risk models are used for clinical decision making and to characterize the quality of cardiac surgery.
- Any risk model used should be accurate so that risk-averse clinical decisions are minimized and inappropriate conclusions about quality are avoided; this is particularly important for emergency cardiac surgery.

WHAT THE STUDY ADDS

- Compared with performance in all cardiac surgery and nonemergency cardiac surgery, both the logistic EuroSCORE and EuroSCORE II demonstrate relatively poor discrimination for emergency cardiac surgery.
- Both the logistic EuroSCORE and EuroSCORE II are poorly calibrated for emergency cardiac surgery and tend to underpredict risk in the lower-risk patients and to overpredict risk in the higher-risk patients.
- A number of the original EuroSCORE risk factors do not seem to be important for emergency cardiac surgery.

Methods

Data Collection, Validation, and Cleaning

Prospectively collected data for cardiac surgery procedures performed at all NHS and some private hospitals in the United Kingdom between April 1, 2008, and March 31, 2011 (inclusive), were extracted from the National Institute for Cardiovascular Outcomes Research National Adult Cardiac Surgery Audit database. The data were cleaned by resolving transcriptional discrepancies, numeric irregularities, and clinical conflicts. Only procedures that were either the first or the only cardiac procedure within a single admission were included. The following records were removed: duplicate records, records with missing inhospital mortality data, records with unreliable date information, records with no indicator of operative urgency, and records with absent hospital identifier information. Because transplantations, trauma, or primary ventricular assist device procedures are not included in SCTS clinical governance analyses, these records were also excluded. To ensure that the performance of each model was assessed with data prospective to that used for model development, EuroSCORE II was validated only with the subset of procedures performed between July 26, 2010, and March 31, 2011 (inclusive). In addition to the above exclusion criteria, any patient older than 95 years of age was removed, in keeping with development of EuroSCORE II.10 Emergency cardiac procedures were defined as any unscheduled patient with ongoing refractory cardiac compromise with no delay in surgery regardless of the time of day (SCTS-defined emergency surgery) or any patient requiring cardiopulmonary resuscitation en route to the operating theater or before the induction of anesthesia (SCTS-defined salvage surgery). Other variable definitions are available at http://www.ucl.ac.uk/nicor/audits/Adultcardiacsurgery/datasets.

The outcome measure for the study was in-hospital mortality, defined as death attributable to any cause during admission to the base hospital for cardiac surgery. When there were irresolvable conflicts or missing in-hospital mortality data, record linkage to the Office for National Statistics database, which records details of all deaths in the United Kingdom, was used to backfill mortality status. Data were returned to each contributing hospital for local validation before analysis.

Statistical Analysis

The logistic EuroSCORE and EuroSCORE II were calculated for each relevant record.15,16 If risk factor data required for calculation of the models were missing for a record, they were imputed with the baseline value. The percentage of missing data was calculated for EuroSCORE and EuroSCORE II risk factors as an indicator of data quality. The minimum data set required for submission to the SCTS database does not include all EuroSCORE II risk factors, meaning that a number of risk factor assumptions were required. These risk factor assumptions have been justified and published previously.11 For each model and data set, model performance was evaluated in 3 groups: all cardiac surgery, nonemergency surgery, and emergency surgery. Model performance was summarized by evaluating model calibration and discrimination.

Model calibration was evaluated with the following range of methods: (1) calculation of the calibration intercept and slope parameters and testing of unreliability, (2) the Hosmer–Lemeshow test,16 (3) comparison of observed and expected mortality in groups stratified by clinically chosen thresholds, (4) calculation of the overall observed-to-expected (O:E) ratio, and (5) graphical inspection of calibration plots.

To calculate the calibration intercept and slope parameters, a logistic regression model was fit with the dependent variable set as the observed outcomes and the independent variable set as the log-odds–transformed model predictions. Two parameters were estimated from the model: an intercept and a slope parameter. If the model is perfectly calibrated, the intercept and the slope equal 0 and 1, respectively. The intercept is a measure of overall model calibration, which is whether predictions agree on average with observed probabilities.17 A χ² test (on 2 df) for unreliability (null hypothesis: intercept=0 and slope=1) was also performed.18

The Hosmer–Lemeshow test allows the detection of departures from calibration within different groups. Contributing χ² statistics from each decile group were summed. Because this is an external model validation, the overall test statistic was evaluated with a χ² distribution on 10 df to test the hypothesis of whether the model predictions match the observed probabilities within the subgroups.16

The Hosmer–Lemeshow test decile stratification is arbitrary. Alternative approaches are (1) to group data together and (2) to group data into groups based on clinically defined thresholds. For option 1, calibration was summarized overall by calculating the O:E ratio (the number of observed in-hospital deaths compared with the number of predicted deaths). If a model is well calibrated on average, then the O:E ratio should be close to 1; departures above or below are indicative of underprediction and overprediction, respectively. Option 2 was performed only for EuroSCORE II because it is the most contemporarily calibrated model. For each validation cohort, records were split into a low-risk group (bottom 50% of cohort) and the remaining 50% of records were divided into 2 equal groups, medium risk and high risk, based on their predicted probability of inhospital mortality. The mean observed and predicted mortality was then calculated in each of these categories and compared by use of a χ² test on 1 df.

Calibration plots are graphical representations of the Hosmer–Lemeshow test. The calibration plots presented show for each model the mean predicted probability of outcome against the observed proportion of outcomes for 10 approximate equally sized groups on the basis of the ranked predicted risks calculated according to the models. Overlaid is the line of equality that would indicate a model with
perfect calibration and a nonparametric locally weighted scatterplot smoothing curve for each model showing the general relationship.\(^27\) Approximately 95% confidence intervals (CIs) for the observed mortality proportions are shown as error bars and are estimated from the Agresti–Coull approximation.\(^26\)

Model discrimination was evaluated by deriving the receiver-operating characteristic curve and calculating the area under the curve (AUC) summary statistic.\(^21\) The AUC is interpretable as follows: If 2 patients are drawn at random from the group of interest, one who experiences an event and the other who does not, then the probability that the true risk factor data ranges from 0.0% (age, previous cardiac surgery) to 5.0% (recent myocardial infarction). With the exception of 5 risk factors (neurological disease, creatinine >200 \(\mu\)mol/L, active infective endocarditis, left ventricular function, and recent myocardial infarction), all had <1% missing data (Table 1).

For the overall cohort, 111,011 records were extracted across 41 hospitals, of which 110,476 corresponded to unique admissions. A total of 337 records were removed because of missing hospital identifier, outcome, or operative urgency status. A further 151 records that represented transplantations, trauma, or primary ventricular assist device procedures were removed. The percentages of missing risk factor data ranged from 0.0% (age, previous cardiac surgery, critical preoperative state, septal rupture) to 5.0% (recent myocardial infarction). With the exception of 5 risk factors (neurological disease, creatinine >200 \(\mu\)mol/L, active infective endocarditis, left ventricular function, and recent myocardial infarction), all had <1% missing data (Table 1).

Of the 109,988 records included, 3342 were emergency procedures and 106,646 were nonemergency. The mean age in the emergency subgroup was 62.7 years (SD, 14.2 years), and the majority of patients (69.2%) were men. The prevalence of each logistic EuroSCORE risk factor in the study population is shown in Table 1. In total, there were 3449 (3.1%; 95% CI, 3.0–3.2) in-hospital deaths, of which 2860 (2.7%; 95% CI, 2.6–2.8) were in the nonemergency cohort and 589 (17.6%; 95% CI, 16.4–19.0) were in the emergency cohort.

The EuroSCORE II validation subset included 23,624 records, of which 22,862 were nonemergency procedures.
and 762 were emergency procedures. Table 2 shows the prevalence and percentage of missing data for each risk factor in the EuroSCORE II model. In total, for this subset, there were 744 (3.1%; 95% CI, 2.9–3.4) in-hospital deaths: 622 (2.7%; 95% CI, 2.5–2.9) were in the nonemergency cohort and 122 (16.0%; 95% CI, 13.6–18.8) were in the emergency cohort.

**Model Performance for All Cardiac Surgery**

The calibration plots (Figure 1, left) demonstrate that the logistic EuroSCORE is poorly calibrated for all cardiac surgery, significantly overpredicting mortality, which is confirmed by the O:E ratio of 0.43. The EuroSCORE II is better calibrated in this group with an O:E ratio of 0.92. The calibration slopes were 0.876 and 0.968 for the logistic EuroSCORE.
and EuroSCORE II, respectively (Table 3). The intercept for the logistic EuroSCORE (−1.247) suggested an overall loss of average calibration compared with the EuroSCORE II (−0.187). The observed versus predicted mortality in each clinical group for EuroSCORE II was 0.6% versus 1.0% \((P<0.001;\) low risk), 2.7% versus 2.4% \((P=0.191;\) medium risk), and 8.7% versus 9.3% \((P=0.084;\) high risk). Despite EuroSCORE II being well calibrated for 2 of 3 clinical groups, the Hosmer–Lemeshow test rejected the null hypothesis of calibration across the decile subgroups for both models. The receiver-operating characteristic curves for the models are shown in Figure 2. Both models demonstrated excellent discrimination with AUCs of 0.806 and 0.811 for the logistic EuroSCORE and EuroSCORE II, respectively.

### Model Performance for Nonemergency Surgery

For nonemergency surgery, the calibration plots (Figure 1, center) and O:E ratios (Table 3) demonstrate that the logistic EuroSCORE is poorly calibrated but the EuroSCORE II is well calibrated. The calibration slopes were close to 1, with the intercept for the logistic EuroSCORE (−1.220) suggesting an overall loss of average calibration compared with the EuroSCORE II (−0.085). For EuroSCORE II, the observed versus predicted mortality in each clinical group was 0.6% versus 1.0% \((P<0.001;\) low risk), 2.3% versus 2.3% \((P=0.829;\) medium risk), and 7.3% versus 7.8% \((P=0.141;\) high risk). Despite EuroSCORE II being well calibrated for 2 of 3 clinical groups, the Hosmer–Lemeshow test rejected the null hypothesis of calibration across the decile subgroups for both models. Both models demonstrated good discrimination, with an AUC of 0.786 for the logistic EuroSCORE and 0.799 for EuroSCORE II.

### Model Performance for Emergency Surgery

Both the logistic EuroSCORE and EuroSCORE II demonstrated poor calibration for emergency surgery (Figure 1, right). For the emergency cohort, the slope parameters were similar for both the logistic EuroSCORE and EuroSCORE II models (0.52 and 0.54, respectively), indicating, along with the calibration plots, that the models are poorly fitted for this cohort. Both models tended to underpredict risk in the lower-risk patients and to overpredict risk in the higher-risk patients. The intercept parameters were also markedly different from the optimal value of 0 (−1.16 and −0.64, respectively), indicating miscalibration. The logistic EuroSCORE significantly overpredicted the mortality rate in the emergency cohort with a mean expected mortality rate of 31.0%, giving an O:E ratio of 0.57.

The EuroSCORE II closely predicted the overall mortality rate for emergency surgery with a mean score of 15.6% (O:E, 1.03). For EuroSCORE II, the observed versus predicted mortality in each clinical group was 8.9% versus 4.4% \((P<0.001;\)

### Table 3. Model Performance Summary Statistics for the Logistic EuroSCORE and EuroSCORE II Models (Applied to the Appropriate Data Subsets) Across the 3 Groups: All Cardiac Surgery, Nonemergency Cardiac Surgery, and Emergency Cardiac Surgery

<table>
<thead>
<tr>
<th>Group</th>
<th>Model</th>
<th>HL Test</th>
<th>Intercept</th>
<th>Slope</th>
<th>O:E</th>
<th>AUC</th>
<th>95% CI</th>
</tr>
</thead>
<tbody>
<tr>
<td>All cardiac surgery</td>
<td>EuroSCORE</td>
<td>(\chi^2) 3209.6</td>
<td>&lt;0.001</td>
<td>-1.247</td>
<td>0.876</td>
<td>&lt;0.001</td>
<td>0.43</td>
</tr>
<tr>
<td></td>
<td>EuroSCORE II</td>
<td>31.2</td>
<td>&lt;0.001</td>
<td>-0.187</td>
<td>0.968</td>
<td>0.022</td>
<td>0.92</td>
</tr>
<tr>
<td>Nonemergency</td>
<td>EuroSCORE</td>
<td>2876.7</td>
<td>&lt;0.001</td>
<td>-1.220</td>
<td>0.903</td>
<td>&lt;0.001</td>
<td>0.41</td>
</tr>
<tr>
<td></td>
<td>EuroSCORE II</td>
<td>31.8</td>
<td>&lt;0.001</td>
<td>-0.085</td>
<td>1.011</td>
<td>0.016</td>
<td>0.90</td>
</tr>
<tr>
<td>Emergency</td>
<td>EuroSCORE</td>
<td>671.3</td>
<td>&lt;0.001</td>
<td>-1.157</td>
<td>0.516</td>
<td>&lt;0.001</td>
<td>0.57</td>
</tr>
<tr>
<td></td>
<td>EuroSCORE II</td>
<td>40.0</td>
<td>&lt;0.001*</td>
<td>-0.643</td>
<td>0.535</td>
<td>&lt;0.001</td>
<td>1.03</td>
</tr>
</tbody>
</table>

AUC indicates area under the receiver-operating characteristic curve; CI, confidence interval; HL, Hosmer and Lemeshow; and O:E, observed-to-expected ratio.

*Because of fewer records in the EuroSCORE II cohort, 5 approximate equal groups were determined by the quintiles of the predictions to achieve sufficiently sized expected values.
Risk factors for in-hospital mortality after emergency cardiac surgery derived from an updated model fit of the logistic EuroSCORE are shown in Table 4. Patient age (transformed as per the EuroSCORE transformation), extracardiac arteriopathy, previous cardiac surgery, creatinine >200 μmol/L, critical preoperative state, left ventricular function (moderate and poor), surgery other than isolated coronary artery bypass graft, surgery on the thoracic aorta, and postinfarct septal rupture were all important risk factors. Female sex, chronic pulmonary disease, neurological disease, active endocarditis, unstable angina, recent myocardial infarction and pulmonary hypertension were not retained in the model selection routine for in-hospital mortality after emergency cardiac surgery.

Discussion

This study demonstrates that the performance of both EuroSCORE models in emergency cardiac surgery is inadequate compared with all cardiac surgery and nonemergency surgery. Both models demonstrated acceptable discrimination for all cardiac surgery and nonemergency procedures. However, the discrimination in the emergency cohort was substantially lower. As expected, the logistic EuroSCORE was poorly calibrated for contemporary cardiac surgery within the United Kingdom for both the overall cardiac surgery and the nonemergency cohort; therefore, it is not surprising that it was also poorly calibrated in the emergency cohort. The EuroSCORE II model, on the other hand, was well calibrated in the overall and nonemergency cardiac surgery cohorts but less well calibrated in the emergency surgery cohort. In general, for emergency cardiac surgery, the EuroSCORE models tended to underpredict risk in the lower-risk patients and to overpredict in the higher-risk patients.

Although the emergency surgery group contributes only a relatively small number of the total number of patients, they are clinically high risk and have much to gain from surgery because they will usually die without it. This study suggests that if any EuroSCORE model is used for clinical decision making in emergency cardiac surgery, model predictions are likely to be unreliable and should be interpreted with caution. Denying a patient emergency cardiac surgery simply because the patient has a high EuroSCORE would certainly be inappropriate, especially given the tendency of both models to overpredict risk in higher-risk emergency patients.

In addition to the implications for clinical decision making, this study has important implications for cardiac surgery governance analyses. Performance in emergency or high-risk surgery is likely to be an important indicator of quality; however, including emergency procedures in such analyses is potentially inappropriate because of inadequate risk adjustment. It has previously been demonstrated that inappropriate risk adjustment leads to flawed clinical governance analyses. This is unlikely to be a limitation of model performance exclusive to the EuroSCORE models and may apply to other widely used models such as those developed by the Society of Thoracic Surgeons. The Society of Thoracic Surgeons models could not be validated in this study because required...
risk factor data were not routinely collected by the SCTS.\textsuperscript{28} Risk prediction models specific to emergency surgery should therefore be developed for emergency cardiac surgery governance analyses.

This study is based on a large clinical data set that has been validated locally at each contributing hospital. The number of emergency cardiac surgical procedures performed in the United Kingdom not captured by this data set is negligible because almost all emergency cardiac surgery is performed in NHS hospitals. As with any study based on clinical data, some data were missing. However, patients excluded from the study because of missing data represented only 0.31\% of the overall cohort. The percentage of missing data for each risk factor individually was generally small, and there were no significant differences in missing data between the subgroups analyzed in this study. An imputation approach was applied to complete the remaining missing data, and this imputation approach is justified by an understanding of how clinicians complete data as part of the SCTS database program.

A thorough assessment of model performance has been performed. Multiple approaches to assess model performance, including clinical face validity, are required to determine the usefulness of a model. In this study, model calibration was assessed through the use of recalibration intercept and slopes, the Hosmer–Lemeshow test, and calibration plots. In this study, the Hosmer–Lemeshow test rejected the hypothesis in all models and clinical groups. However, despite its widespread use in clinical studies, it is no longer advocated for discerning model performance because of known shortcomings.\textsuperscript{29,30} The nonparametric calibration curves are a useful diagnostic tool for examining the nature of any miscalibration. The O:E ratio statistic is a single summary statistic and thus can average out opposing regions of miscalibration to suggest that a model is well calibrated. This was the case for the EuroSCORE II model in the emergency cohort in which the O:E ratio was close to 1, attributable to the averaging out of underpredicting and overpredicting. Model discrimination has been assessed with the widely accepted method of AUC; however, this must be taken into account with the other performance measures.\textsuperscript{31}

The EuroSCORE, which was published in 1999 as an additive model and then superseded by a logistic model in 2003, has been used for clinical governance analyses in the United Kingdom and has been widely accepted throughout Europe.\textsuperscript{7,23} Since publication of the EuroSCORE, the predicted mortality now significantly exceeds observed mortality for most patient groups, which is presumably attributable to overall improvements in quality of care.\textsuperscript{6} To the best of our knowledge, this is the first detailed assessment of the performance of EuroSCORE risk models in emergency cardiac surgery. EuroSCORE model performance, however, has previously been assessed in a number of different high-risk cardiac surgery populations.\textsuperscript{31–40}

Whereas the focus of this study was on the performance of existing risk models in emergency cardiac surgery, this study also represents one of the largest analyses to date of which logistic EuroSCORE risk factors are important for in-hospital mortality after emergency cardiac surgery. A number of logistic EuroSCORE risk factors were not found to be important for emergency surgery, and the weightings of the remaining risk factors were found to be different from the original EuroSCORE weightings. It is likely that the overall clinical state of the patient (including pathogenesis and clinical presentation) associated with emergency status is the most important determinant of postoperative outcome. For these reasons, a simple recalibration of the EuroSCORE to improve model performance for emergency cardiac surgery would be inappropriate.

Although the findings of this study relate directly to cardiac surgery, it is likely that governance analyses will become more common in other surgical and interventional medical specialties. There is an increasing public appetite for comparative clinical outcomes data, and this will require robust analyses to allow safe comparisons to be made.\textsuperscript{5} This study raises important generic issues about the use of risk prediction models for emergency patients, which may potentially be reproduced across different subspecialty patient groups.

Acknowledgments
We acknowledge all members of the Society for Cardiothoracic Surgery in Great Britain and Ireland who contribute data to the SCTS database. The National Institute for Cardiovascular Outcomes Research, University College London, London, UK, provided the data for this study. We would also like to thank Norman Stein for his work in cleaning the SCTS database.

Sources of Funding
Part of this research was funded by a grant from Heart Research United Kingdom (grant RG2583).

Disclosures
None.

References


Performance of the EuroSCORE Models in Emergency Cardiac Surgery
Stuart W. Grant, Graeme L. Hickey, Ioannis Dimarakis, Graham Cooper, David P. Jenkins, Rakesh Uppal, Iain Buchan and Ben Bridgewater

_Circ Cardiovasc Qual Outcomes._ published online March 5, 2013;
_Circulation: Cardiovascular Quality and Outcomes_ is published by the American Heart Association, 7272 Greenville Avenue, Dallas, TX 75231
Copyright © 2013 American Heart Association, Inc. All rights reserved.
Print ISSN: 1941-7705. Online ISSN: 1941-7713

The online version of this article, along with updated information and services, is located on the World Wide Web at:
http://circoutcomes.ahajournals.org/content/early/2013/03/05/CIRCOUTCOMES.111.000018

Permissions: Requests for permissions to reproduce figures, tables, or portions of articles originally published in _Circulation: Cardiovascular Quality and Outcomes_ can be obtained via RightsLink, a service of the Copyright Clearance Center, not the Editorial Office. Once the online version of the published article for which permission is being requested is located, click Request Permissions in the middle column of the Web page under Services. Further information about this process is available in the Permissions and Rights Question and Answer document.

Reprints: Information about reprints can be found online at:
http://www.lww.com/reprints

Subscriptions: Information about subscribing to _Circulation: Cardiovascular Quality and Outcomes_ is online at:
http://circoutcomes.ahajournals.org//subscriptions/