European Journal of Cardio-Thoracic Surgery 41 (2012) 734–745
doi:10.1093/ejcts/ezs043

ORIGINAL ARTICLE

EuroSCORE II†

Samer A.M. Nashef[a,*], François Roques[b], Linda D. Sharples[c], Johan Nilsson[d], Christopher Smith[a], Antony R. Goldstone[e] and Ulf Lockowandt[f]

[a] Papworth Hospital, Cambridge, UK
[b] University Hospital Centre (CHU), Fort de France, Martinique, France
[c] Medical Research Council, Biostatistics Unit, Cambridge, UK
[d] Division of Cardiothoracic Surgery, Skåne University Hospital, Lund, Sweden
[e] Department of Radiology and Nuclear Medicine, Castle Hill Hospital, Hull, UK
[f] Karolinska Hospital, Stockholm, Sweden

* Corresponding author. Papworth Hospital, Cambridge CB23 3RE, UK. Tel: +44-1480-364299; e-mail: sam.nashef@papworth.nhs.uk (S.A.M. Nashef).

Received 13 October 2011; received in revised form 5 January 2012; accepted 6 January 2012

Abstract

OBJECTIVES: To update the European System for Cardiac Operative Risk Evaluation (EuroSCORE) risk model.

METHODS: A dedicated website collected prospective risk and outcome data on 22 381 consecutive patients undergoing major cardiac surgery in 154 hospitals in 43 countries over a 12-week period (May–July 2010). Completeness and accuracy were validated during data collection using mandatory field entry, error and range checks and after data collection using summary feedback confirmation by responsible officers and multiple logic checks. Information was obtained on existing EuroSCORE risk factors and additional factors proven to influence risk from research conducted since the original model. The primary outcome was mortality at the base hospital. Secondary outcomes were mortality at 30 and 90 days. The data set was divided into a developmental subset for logistic regression modelling and a validation subset for model testing. A logistic risk model (EuroSCORE II) was then constructed and tested.

RESULTS: Compared with the original 1995 EuroSCORE database (in brackets), the mean age was up at 64.7 (62.5) with 31% females (28%). More patients had New York Heart Association class IV, extracardiac arteriopathy, renal and pulmonary dysfunction. Overall mortality was 3.9% (4.6%). When applied to the current data, the old risk models overpredicted mortality (actual: 3.9%; additive predicted: 5.8%; logistic predicted: 7.57%). EuroSCORE II was well calibrated on testing in the validation data subset of 5553 patients (actual mortality: 4.18%; predicted: 3.95%). Very good discrimination was maintained with an area under the receiver operating characteristic curve of 0.8095.

CONCLUSIONS: Cardiac surgical mortality has significantly reduced in the last 15 years despite older and sicker patients. EuroSCORE II is better calibrated than the original model yet preserves powerful discrimination. It is proposed for the future assessment of cardiac surgical risk.

Keywords: Risk assessment • EuroSCORE • Cardiac surgery • Mortality

INTRODUCTION

The European System for Cardiac Operative Risk Evaluation [1] (EuroSCORE) is a cardiac risk model for predicting mortality after cardiac surgery. It was published in 1999 and derived from an international European database [2] of patients who had undergone cardiac surgery by the end of 1995. The system has been highly successful and used worldwide both for the measurement of risk and as a benchmark for the assessment of the quality of cardiac surgical services, with more than 1300 formal citations in the medical literature.

Over the last few years, several professionals from many parts of the world [3–9] have published evidence that the model now overpredicts risk as the results of cardiac surgery have substantially improved with a sustained reduction of risk-adjusted mortality, so that the model may now be inappropriately calibrated for current cardiac surgery.

Despite the calibration problem, both additive [1] and logistic [10] versions of the model have remained powerfully discriminatory with an area under the receiver operating characteristic (ROC) curve of around 0.75–0.8. Nevertheless, there is some evidence that discrimination may be improved further by refining and modifying some of the risk factors and the way the model handles them, such as renal dysfunction [11, 12].

The purpose of this study was to renew EuroSCORE in order to maintain and optimize its usefulness in contemporary cardiac surgical practice.

METHODS

Recruitment

Using journals, conferences, articles and presentations as well as the existing www.EuroSCORE.org website, cardiac surgical units

†Presented at the 25th Annual Meeting of the European Association for Cardio-Thoracic Surgery, Lisbon, Portugal, 1–5 October 2011

Table 1: Participating countries (43) and number of units (154)

Argentina	1	Denmark	2	Israel	1	Russia	3	Switzerland	2
Austria	2	Finland	4	Italy	15	Saudi Arabia	2	Syria	1
Belarus	1	France	16	Japan	3	Serbia	4	Taiwan	1
Belgium	8	Germany	9	Lithuania	1	Slovenia	1	Turkey	1
Bosnia	1	Greece	2	Montenegro	1	South Africa	1	UAE	1
Brazil	4	Holland	6	New Zealand	1	Spain	19	UK	12
Canada	2	Hungary	1	Norway	1	Sudan	1	Uruguay	1
China	2	India	4	Poland	1	Sweden	5	USA	3
Croatia	2	Ireland	1	Portugal	4				

worldwide were invited to participate in data collection for the project. Expressions of interest were received on the EuroSCORE website from 520 individuals, all of whom were given further detailed information on participation and invited to register. This resulted in 214 units registering to take part. Of these, 160 units from 44 countries submitted data and 154 successfully completed data collection. The distribution of the units by nation is given in Table 1.

The data set

Review of the literature and of feedback received from many users of the logistic EuroSCORE identified the following areas for potential improvement:

- Creatinine clearance (CC) is a better predictor than absolute serum creatinine.
- Hepatic function is not represented.
- Defining unstable angina by the use of intravenous nitrates is out of date.
- Some continuous variables are treated as dichotomous (number of previous heart operations, serum creatinine, pulmonary artery pressure).
- The model is not sufficiently sensitive to the 'weight' of the intervention.

A new set of risk factors was assembled to include the original EuroSCORE variables modified or complemented to take account of the above areas. The risk factor information collected is given in Table 2.

Table 2: Data set

- Patient-related factors
  - Age and sex
  - Height and weight
  - Pulmonary disease
  - Diabetes status
  - Extracardiac arteriopathy
  - Neurological or musculoskeletal dysfunction
  - On dialysis
  - Last serum creatinine
  - Brain-natriuretic peptide
  - Serum albumin
- Cardiac-related factors
  - Symptomatic status
    - NYHA
    - CCS
  - LV function
  - Recency and size of last myocardial infarct
  - Systolic PA pressure
  - Active endocarditis
- Previous cardiac surgery
- Operation-related factors
    - Urgency
    - Elective
    - Urgent
    - Emergency
    - Salvage
  - Type of procedure(s) performed in detail
  - Times of
    - Bypass
    - Cross-clamp
    - Deep hypothermic arrest
    - Selective cerebral perfusion

Data collection

Data were collected from consecutive patients operated over a 12-week period (3 May–25 July 2010, inclusive) and entered into the web database. Records could be opened at the beginning of the data collection period and held as 'pending' while additional information on procedures and outcomes was obtained. Most units submitted data on line, either contemporaneously or after completing a paper dataform. Three units sent data as spreadsheets, and the EuroSCORE project team entered these on their behalf. Ninety days after the last eligible operation, units with pending records were urged repeatedly to complete their data. The last did so by May 2011 and the data set was finally closed to data entry.

Once data collection was completed, the responsible person for every unit was asked to confirm on the website itself and by email that the data provided were an accurate and complete representation of the unit's entire cardiac surgical activity during the study period. Where there was a major discrepancy between a unit's projected activity on registration and actual data received, the responsible officer was asked to provide an explanation. Units which could not provide a satisfactory explanation or those with major flaws in inadequate or missing data were removed from the study.

Data preparation

The data set was downloaded directly from the database into a spreadsheet and contained 24 385 records. Centres with few observations, incomplete or duplicate data entry were dropped, leaving 23 451 cases. Cases from the same country with same date of birth, sex, date of operation, height and weight were assumed to be duplicate records and only one of the records was

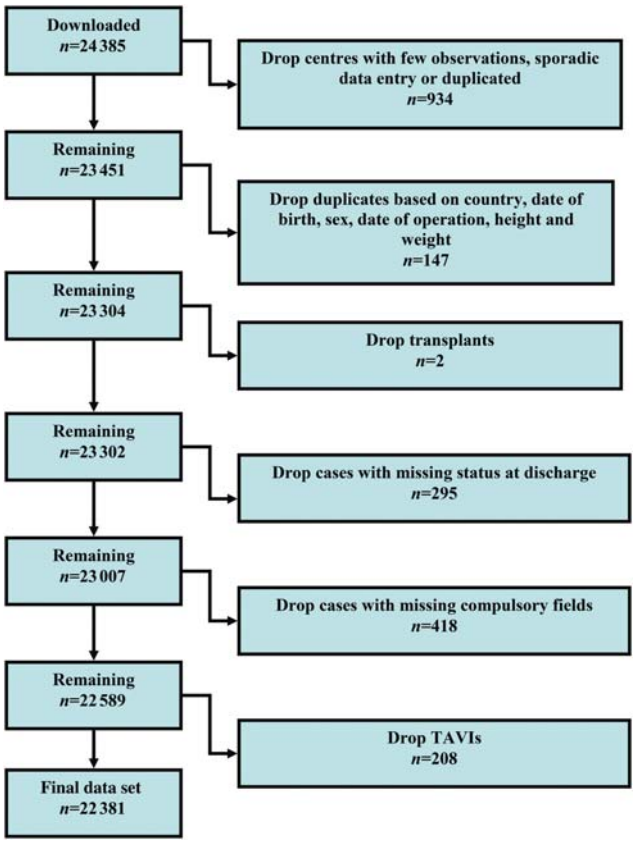

Figure 1: Downloaded data and final data set for analysis.

retained in the final data set. Duplicates were identified for 136 patients, resulting in 147 records being removed (129 were entered twice, 4 were entered three times, 2 were entered four times and 1 was entered five times). Two cases were known to be transplants and were removed. Of the 23 302 cases that remained, 295 (1.2%) did not have survival status at discharge or 90 days and were dropped. Additionally, because we had pending records in the data set there were 418 with at least one missing required field, mostly musculoskeletal dysfunction ($n = 159$, all one centre), Canadian Cardiovascular Society (CCS) angina class ($n = 204$, all in two centres) and New York Heart Association (NYHA) class ($n = 30$, all in one centre). After checking that these were either in a small number of centres or related to a time period rather than sporadically missing, we concluded that they could be assumed 'missing completely at random' and so excluded these cases from the analysis; this left 22 589 cases. Finally, we decided to exclude 208 transcatheter aortic valve implant (TAVI) procedures from the analysis with a view to a separate study in the future. Thus, the final data set contained 22 381 cases (Fig. 1) from 154 centres in 43 countries (Table 1).

At the completion of data collection, logic checks were carried out on all data. The data set prepared for the analysis was therefore excellent in quality and completeness, rivalling if not exceeding the standards set by the original EuroSCORE database of 1995.

Analysis

Tables 3 and 4 summarize patient profile and procedures performed. Body mass index (BMI) was calculated using the formula

Table 3: EuroSCORE II demographics and comorbidity ($n = 22\,381$)

Variable	Frequencies (%) or mean (SD) [range]
Patient-related factors	
Age (years)	64.6 (12.5) [18–95]
Female	6919 (30.9%)
Weight (kg)	77.9 (15.9) [30–182]
Height (cm)	168.5 (9.6) [100–213]
BMI (calculated) (kg/m^2)	27.4 (4.8) [9.6–82.6]
Body surface area (calculated)	1.87 (0.21) [1.04–2.90]
Diabetes—no	16 783 (75.0%)
Diet only	803 (3.6%)
Oral therapy only	3103 (13.9%)
Insulin	1705 (7.6%)
Pulmonary disease	2384 (10.7%)
Neurological dysfunction	713 (3.2%)
Serum creatinine (μmol/l)	96.4 (57.1)
Serum creatinine (mg/dl)	1.13 (0.92)
Serum creatinine > 200 μmol/l	562 (2.6%)
CC (calculated)	83.6 (50.9)
On dialysis	244 (1.1%)
Serum albumin (g/l)	31.6 (19.0)
Active endocarditis	497 (2.2%)
Critical preoperative state	924 (4.1%)
Pre-op VT/VF or aborted sudden death	137 (0.6%)
Pre-op cardiac massage	94 (0.4%)
Pre-op ventilation	251 (1.1%)
Pre-op inotropes	475 (2.1%)
Pre-op IABP	384 (1.7%)
Pre-op acute renal failure	108 (0.5%)

VT: ventricular tachycardia; VF: ventricular fibrillation; IABP: intra-aortic balloon pump.

BMI = weight (kg)/ height^2 (m^2). CC as an estimate of glomerular filtration rate was calculated using the Cockcroft–Gault formula:

$$\mathrm{CC\,(ml/min)} = \frac{(140 - \mathrm{age\,(years)}) \times \mathrm{weight\,(kg)} \times 0.85\,(\mathrm{if\ female})}{72 \times \mathrm{serum\ creatinine\,(mg/dL)}}$$

Our aim was to construct a model that could be applied very widely and that could be incorporated into local data collection and management systems. Therefore, a parsimonious approach to model fitting was adopted. To this end, the data were divided into a developmental data set (16 828 patients) and a validation data set (5553 patients) using random sampling from a binomial distribution with a probability 0.25. Initially, a series of single-variable logistic regression models was fitted to the developmental data set in order to identify variables that were associated with mortality. The variables are listed in Table 5. These models were used to assess the nature of associations between predictors and risk of death, such as linear associations and threshold models, as well as the optimal grouping of categorical variables. Of these variables, 14 were considered compulsory for the final model [age, sex, extracardiac arteriopathy, chronic lung disease, poor mobility, previous cardiac surgery, CC, active endocarditis, critical preoperative state, left ventricular (LV) function, systolic pulmonary artery pressure, urgency and weight of procedure].

Additional variables were included in the model using forward selection based on likelihood ratio statistics comparing models

Table 4: EuroSCORE II: types of procedure (n = 22 381)

Urgency of operation	
Elective	17 165 (76.7%)
Urgent	4135 (18.5%)
Emergency	972 (4.3%)
Salvage	109 (0.5%)
CABG (isolated)	10 448 (46.7%)
Valve procedures	10 353 (46.3%)
Aortic valve	
Repair	269 (1.2%)
Replacement	6753 (30.2%)
Regurgitation and stenosis	971 (13.8%)
Mostly regurgitation	1534 (21.7%)
Mostly stenosis	4545 (64.4%)
Mitral valve	
Repair	1935 (8.7%)
Replacement	2049 (9.2%)
Regurgitation and stenosis	534 (13.3%)
Mostly regurgitation	2901 (72.4%)
Mostly stenosis	568 (14.2%)
Tricuspid valve	
Repair	1031 (4.6%)
Replacement	79 (0.4%)
Pulmonary valve	
Repair	10 (0.04%)
Replacement	46 (0.2%)
Thoracic aortic surgery	1636 (7.3%)
Ascending aortic replacement	1100 (4.9%)
Root replacement with coronary reimplantation	492 (2.2%)
Partial aortic arch replacement	141 (0.6%)
Total aortic arch replacement	47 (0.2%)
Descending aortic replacement	42 (0.2%)
Thoracoabdominal aortic replacement	22 (0.1%)
Other procedure on the thoracic aorta	106 (0.5%)
Pericardiectomy	64 (0.3%)
Other major heart procedure	130 (0.6%)

with and without each variable and Akaike's information criterion (AIC), until well-fitting models were developed. A likelihood ratio statistic with $P < 0.05$ or a change in AIC of at least 10 led to inclusion of a new variable. Due to the colinearity between some covariates, there were several models that fitted the data and gave similar prediction. The final model, named EuroSCORE II, was chosen on the basis of clinical face validity (reflecting current knowledge in the field of cardiac surgery) and predictive accuracy (maintaining the area under the ROC curve at 80% or more). The logistic equation used was

$$\text{predicted mortality} = \frac{e^{(\beta_0 + \sum \beta_i X_i)}}{1 + e^{(\beta_0 + \sum \beta_i X_i)}},$$

where β_0 is the constant of the logistic regression equation = −5.324537, β_i the coefficient of the variable X_i, for age, $X_i = 1$ if patient age ≤60; X_i increases by one point per year thereafter (age 60 or less $X_i = 1$; age 61 if $X_i = 2$; age 62 if $X_i = 3$ and so on). EuroSCORE II risk factors and their coefficients are detailed in Table 6.

The model was then tested on the validation data set for calibration (by comparing the observed and predicted mortality) and for discrimination (using the area under the ROC curve). Goodness of fit of the final model was tested using the Hosmer–Lemeshow statistic. In addition, a range of model diagnostics were employed to assess the validity of the model.

(i) In order to accommodate the hierarchical structure of the data (patients are clustered within hospitals), hospitals were incorporated into the model as random intercepts. Although there was some evidence of inter-hospital heterogeneity, it was small and the random effects model did not improve prediction substantially. Therefore, we retained the fixed effects model for prediction.

(ii) Centres with missing outcome data were studied in depth to assess mechanisms and the effect on risk predictions. No mechanism for missing outcomes was identified but 179 of 295 cases (61%) came from seven centres; no other centre had more than nine cases with missing outcomes. Exclusion of these centres did not affect predictions.

(iii) We conducted 10-fold cross-validation by dividing the data set into 10 equally sized samples at random, refitting the model to each of the 10 sets comprising 90% of the data, calculating the area under the ROC curve for the unused 10% in each case and averaging over 10 areas under the ROC curves. The resulting areas ranged from 0.77 to 0.83, with an average of 0.80, very similar to that of the validation set in our original analysis.

(iv) Graphical methods included examination of the effects of leaving out cases with particular covariate patterns on the model coefficients, the model χ^2-statistics and the influence statistics. On the basis of these plots, we were not able to identify those variables that further discriminate between cases with the same covariate pattern in our final model.

RESULTS

Definition of mortality

For such a clear-cut binary outcome measure, the definition of an early or operative death remains contentious. It could be defined as any of the following:

- death in the same hospital as the operation took place, before discharge from hospital;
- death in the same hospital or at another hospital but before discharge from hospital;
- death within 30 days of surgery regardless of location;
- death within 90 days of surgery regardless of location;
- a compound of some or all of the above.

The selection of an appropriate and practical definition is an important part of this project and that was the rationale for seeking the status of patients on discharge from the base hospital, at 30 days and at 90 days when this information is available. All units were able to supply data on status at discharge, but not all units were able to provide data on 30-day and 90-day status, so that we only have data on 56.6% of the patients at 30 days and 44.4% of patients at 90 days.

An analysis of units which were able to provide 30-day status data shows that of 12 673 patients, 12 164 patients were discharged alive, of whom 79 patients had died by 30 days. Of 12 160 patients who are alive at 30 days, 75 went on to die before hospital discharge. This means that in this group, hospital mortality (4.015%) and 30-day mortality (4.048%) are virtually identical although the overlap is incomplete. Combining the two raises the mortality to 4.63% so that the additional post-discharge 'drop-off' rate at 30 days is 0.615%.

Table 5: Variables associated with mortality

Variable	Univariable logistic model coefficients	AIC, *P*-value[a]
Patient-related factors		
Age[b]	0.0486477	5427.836, $P < 0.0001$
Female	0.3951562	5498.874, $P < 0.0001$
Extracardiac arteriopathy	0.7637420	5465.051, $P < 0.0001$
Pulmonary disease	0.4544856	5506.220, $P = 0.0001$
Neurological or musculoskeletal dysfunction	0.7644773	5499.414, $P < 0.0001$
Previous cardiac surgery	1.2818960	5402.522, $P < 0.0001$
Serum creatinine > 200 ($n = 16\,201$)	1.5384690	5171.137, $P < 0.0001$
Serum creatinine (per μmol/l up to 200) ($n = 16\,201$)	0.0138048	5095.174, $P < 0.0001$
Serum creatinine > 90–110 μmol/l	0.2218056	5079.254, $P < 0.0001$
Serum creatinine > 110–130 μmol/l	0.7177771	
Serum creatinine > 130–200 μmol/l	1.2135250	
Serum creatinine > 200 μmol/l ($n = 16\,201$)	1.8226770	
CC ≤ 50	1.6887740	5015.454, $P < 0.0001$
CC > 50–85 ($n = 16\,201$)	0.6674962	
On dialysis	1.2033870	5501.826, $P < 0.0001$
Active endocarditis	1.4029890	5465.037, $P < 0.0001$
Critical preoperative state	2.1827250	5189.438, $P < 0.0001$
Cardiac-related factors		
CCS angina class = 4	0.8217379	5470.678, $P < 0.0001$
NYHA class II	0.0777918	5250.462, $P < 0.0001$
NYHA class III	0.7037355	
NYHA class IV	1.9128670	
LVEF 30–50%	0.4626558	5382.789, $P < 0.0001$
LVEF < 30%	1.4371450	
LVEF 30–50%	0.4626558	5266.459, $P < 0.0001$
LVEF 20–29%	1.5041660	
LVEF < 20 ($n = 16\,614$)	1.6481420	
Myocardial infarct in previous 4–91 days	0.2863484	5458.959, $P < 0.0001$
Myocardial infarct in previous 0–72 h	1.4105750	
Systolic pulmonary pressure > 60 mmHg	0.7201059	5507.010, $P = 0.0001$
Systolic pulmonary pressure 20–60 mmHg	0.1647881	5506.203, $P = 0.0002$
Systolic Pulmonary Pressure > 60 mmHg	0.7566437	
Operation-related factors		
Urgent operation	0.8295933	5216.267, $P < 0.0001$
Emergency	1.8999760	
Salvage	2.9450770	
Other than isolated coronary surgery	0.7193801	5447.145, $P < 0.0001$
Thoracic aortic surgery	0.8267812	5477.658, $P < 0.0001$
Aortic arch surgery	1.1779710	5481.659, $P < 0.0001$
Postinfarct ventricular septal rupture	Insufficient cases	
Isolated CABG	Baseline	5429.399, $P < 0.0001$
Thoracic aortic surgery	1.1809770	
Everything else	0.6245959	
Isolated CABG	Baseline	5327.226, $P < 0.0001$
Single non-CABG procedure	0.2216732	
Two procedures	0.8473152	
Three or more procedures	1.2831780	

[a]AIC is Akaike's information criterion and assesses the fit of the model; lower values indicate better fit. AICs are only comparable if models use the same data, i.e. the same cases. *P*-values are from likelihood ratio tests.
[b]Age is less continuous in the logistic model and is set to 1 for ages ≤60, increasing by 1 thereafter; in the additive model score 1 point per 5 years or part thereof above 60 years.

An analysis of units which were able to provide 90-day status data shows that of 9939 patients, 9464 were discharged alive, of whom 155 had died by 90 days. Of 9309 patients who are alive at 90 days, 31 went on to die before hospital discharge. This means that 90-day mortality (6.023%) is higher than hospital mortality (4.779%). Combining the two raises the mortality to 6.34%, so that the additional post-discharge 'drop-off' rate at 90 days is 1.56%, or a further 0.946% mortality over and above the 30-day rate.

In summary, therefore, when hospital mortality is around 4%, adding 30-day mortality increases it by ∼0.6% and adding 90-day mortality increases it further by ∼0.9%.

It can be safely assumed that this level of data availability from units participating in the EuroSCORE project is at least representative and may even exceed that of cardiac surgical units globally so that, as things stand, we can expect only about half of the units to have ready access to 30-day and 90-day survival status data. Thus, despite the well-known advantages of 30-day and 90-day criteria [13], the only practical outcome measure that can be used in the current status of data availability to participating units must pragmatically be death in the hospital where the operation took place. This will therefore be the outcome measure used for the

Table 6: Final risk factors by multivariate regression for the model

Risk factor	Coefficient	Standard error	*z*	$P \geq \|z\|$	[95% confidence interval]
NYHA					
II	0.1070545	0.1463849	0.73	0.465	[−0.1798547, 0.3939637]
III	0.2958358	0.141466	2.09	0.037	[0.0185674, 0.5731042]
IV	0.5597929	0.1697565	3.30	0.001	[0.2270763, 0.8925095]
CCS4	0.2226147	0.1462888	1.52	0.128	[−0.0641061, 0.5093356]
IDDM	0.3542749	0.145863	2.43	0.015	[0.0683887, 0.6401611]
Age	0.0285181	0.0065954	4.32	0.000	[0.0155914, 0.0414448]
Female	0.2196434	0.0953505	2.30	0.021	[0.0327599, 0.4065269]
ECA	0.5360268	0.1106046	4.85	0.000	[0.3192458, 0.7528079]
CPD	0.1886564	0.1232126	1.53	0.126	[−0.0528358, 0.4301486]
N/M mob	0.2407181	0.1729494	1.39	0.164	[−0.0982564, 0.5796927]
Redo	01.118599	0.1226272	9.12	0.000	[0.8782539, 1.3589440]
Renal dysfunction					
On dialysis	0.6421508	0.3083468	2.08	0.037	[0.0378021, 1.2464990]
CC ≤ 50	0.8592256	0.1446758	5.94	0.000	[0.5756663, 1.1427850]
CC 50–85	0.303553	0.1240518	2.45	0.014	[0.0604159, 0.5466901]
AE	0.6194522	0.2046001	3.03	0.002	[0.2184433, 1.0204610]
Critical	1.086517	0.147657	7.36	0.000	[0.797115, 1.3759200]
LV function					
Moderate	0.3150652	0.1036182	3.04	0.002	[0.1119773, 0.5181530]
Poor	0.8084096	0.1498233	5.40	0.000	[0.5147614, 1.1020580]
Very poor	0.9346919	0.2917754	3.20	0.001	[0.3628227, 1.5065610]
Recent MI	0.1528943	0.136257	1.12	0.262	[−0.1141646, 0.4199531]
PA systolic pressure					
31–55 mmHg	0.1788899	0.1266713	1.41	0.158	[−0.0693812, 0.4271611]
≥55	0.3491475	0.1676641	2.08	0.037	[0.0205318, 0.6777632]
Urgency					
Urgent	0.3174673	0.1174178	2.70	0.007	[0.0873326, 0.5476020]
Emergency	0.7039121	0.1719835	4.09	0.000	[0.3668306, 1.0409940]
Salvage	1.362947	0.33706	4.04	0.000	[0.7023221, 2.0235730]
Weight of procedure					
1 non-CABG	0.0062118	0.1463574	0.04	0.966	[−0.2806434, 0.2930670]
2	0.5521478	0.1268137	4.35	0.000	[0.3035975, 0.8006980]
3+	0.9724533	0.1463969	6.64	0.000	[0.6855206, 1.2593860]
Thoracic aorta	0.6527205	0.221183	2.95	0.003	[0.2192097, 1.0862310]
Constant	−5.324537	0.1682446	−31.65	0.000	[−5.65429, −4.9947830]

NYHA: New York Heart Association; CCS: Canadian Cardiovascular Society; IDDM: insulin-dependent diabetes mellitus; ECA: extracardiac arteriopathy; CPD: chronic pulmonary dysfunction; N/M mob: neurological or musculoskeletal dysfunction severely affecting mobility; Redo: previous cardiac surgery; CC: creatinine clearance; AE: active endocarditis; Critical: critical preoperative state; LV: left ventricle; MI: myocardial infarction; PA: pulmonary artery; CABG: coronary artery bypass grafting. Weight of procedure '1 non-CABG': single major cardiac procedure which is not isolated CABG; 2: two major cardiac procedures; 3+: three or more major cardiac procedures. For age, $X_i = 1$ if patient age ≤60; X_i increases by one point per year thereafter (age 60 or less $X_i = 1$; age 61 if $X_i = 2$; age 62 if $X_i = 3$ and so on).

remainder of this article and for applications of the risk model in the foreseeable future.

Risk-adjusted mortality

A total of 22 381 patients were included in the study from 154 units in 43 countries. In comparison with the original 1995 EuroSCORE database (in brackets), the mean age was up at 64.7 (62.5) with 31% females (28%) and more patients had NYHA class IV, extracardiac arteriopathy and renal and pulmonary dysfunction. Overall mortality was 3.9% (4.6%), thus mortality is currently lower than in 1995 despite a worsening risk profile. When applied to the current data set, the original additive EuroSCORE predicted a mortality of 5.8% and the logistic 7.57%. This means that the current risk-adjusted mortality ratio (RAMR = observed/predicted) for the previous additive model is 0.67 and for the previous logistic model is 0.53. This confirms that the original EuroSCORE is now no longer appropriately calibrated and shows the substantial net improvement in cardiac surgical outcomes since 1995 with risk-adjusted mortality falling by nearly half.

Despite this, both the old logistic and additive EuroSCORE models retain very good discrimination, with an area under the ROC curve of 0.7896 for the logistic model and 0.7894 for the additive model (Fig. 2A and B).

Risk factors for the new model

Univariate regression analysis demonstrated that a number of risk factors are associated with increased mortality. These are detailed in Table 5. The rationale for the final selection of risk factors for the model and some salient facts about these risk factors are addressed below.

Age remains a significant predictor of mortality from 60 years onwards, but its impact has reduced when compared with 1995, with the β-coefficient dropping from over 0.06 to 0.0486 in univariate analysis. Multivariate analysis reduced this even further

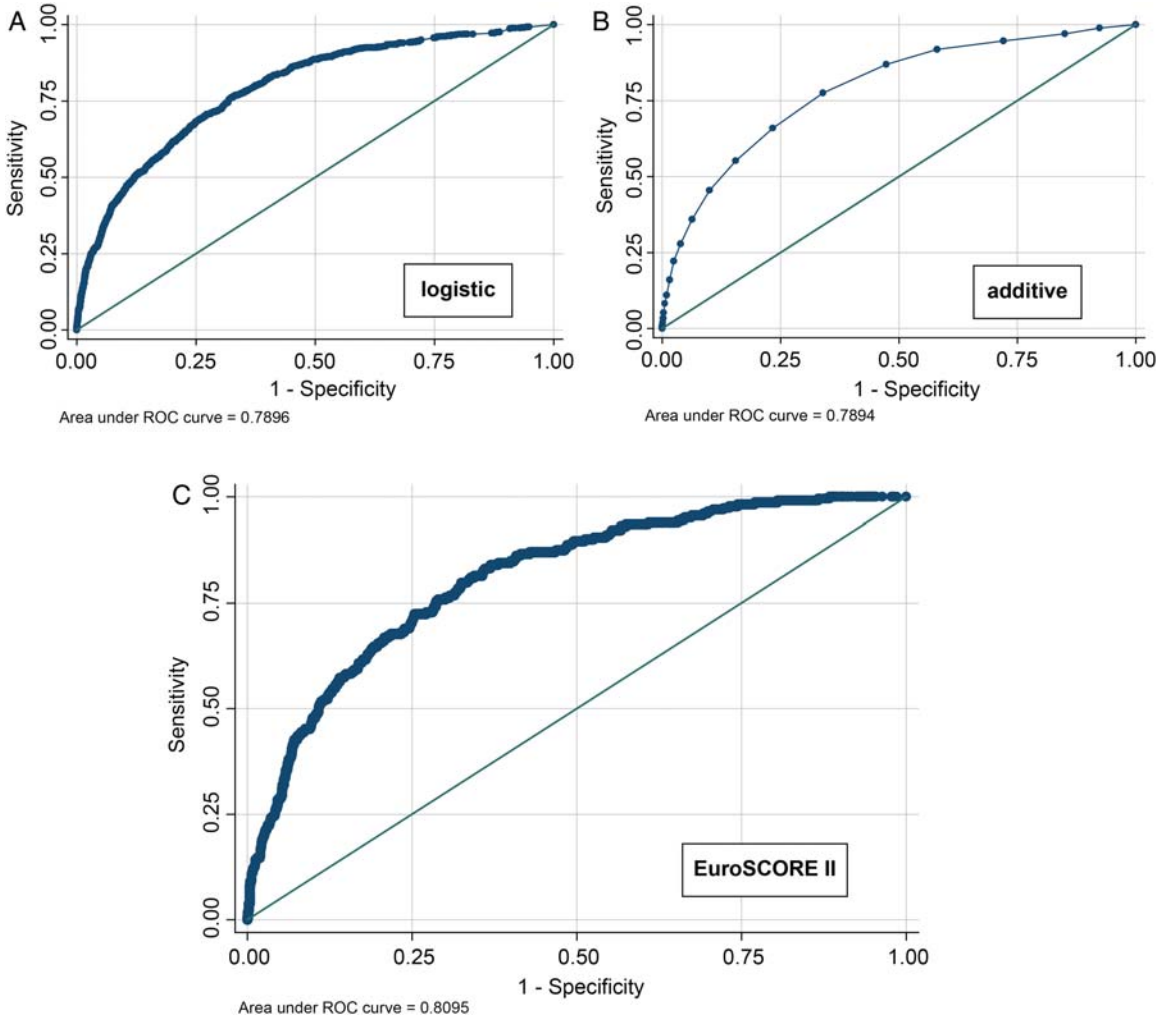

Figure 2: Areas under the ROC curve for the previous additive and logistic models applied to current data, and the new logistic EuroSCORE II model applied to the validation data set of 5553 patients.

(to 0.0286) when some of the other risk factors are taken into account, especially the new measure of renal function which relies on CC and that includes age in its calculation. It is noteworthy that of over 22 381 patients in the EuroSCORE database, only 21 patients (0.093%) were aged over 90. The oldest patient in the study was aged 95.

Females have a higher mortality than males. Most of the well-established EuroSCORE risk factors (extracardiac arteriopathy, pulmonary disease, critical preoperative state, etc.) continue to have an impact on mortality. After multivariate regression analysis and comparison of the coefficients between the original and updated prediction algorithms, specific areas were identified in which the new model differs substantially from the old one. These areas and the reasons for these differences are given below.

Symptomatic status is associated with increased risk. In the case of angina, only CCS angina class 4 was associated with poor outcome, whereas there was an increasing risk with an increasing NYHA class. Thus, the final decision was to incorporate NYHA classes II, III and IV but only angina CCS class 4 into the model. This has the advantage of both replacing the outdated definition of unstable angina and taking into account congestive cardiac failure, a significant risk factor in the original model which was sacrificed due to collinearity with other risk factors.

BMI is weakly associated with mortality. Low BMI appears to increase the risk of hospital death but high BMI does not. The relationship between BMI and risk was very weak ($P = 0.0845$) and this will not be considered for the final model.

Diabetes, which was not a feature of the original model, was revisited. Insulin-dependent diabetes was associated with mortality; orally treated diabetes less so, and diet-controlled diabetics actually had better outcomes than non-diabetics. Insulin-dependent diabetes features in the new model.

Reduced mobility has an effect whether due to neurological dysfunction or to musculoskeletal dysfunction.

CC, calculated using the Cockcroft–Gault formula, is a better predictor of mortality than absolute serum creatinine. Renal function is thus defined by calculated CC as follows:

- normal (>85 ml/min)
- moderately impaired (50–85 ml/min)
- severely impaired (<50 ml/min)
- virtually absent (on dialysis).

It is interesting to note that in renal dysfunction, the highest risk of mortality is in patients with severely impaired renal function who are not on established dialysis (Table 5).

It is known that liver failure increases the mortality of cardiac surgery [14, 15] and yet this risk factor is not usually represented in risk models. Of the various indicators of hepatic dysfunction, serum albumin concentration was selected as the one that is least affected by cardiac therapy and is the most objective and widely available test. Disappointingly, the relationship between serum albumin and risk was practically zero. There is some doubt about the measurement of serum albumin concentration, and it is possible that different centres have used different units and assay techniques.

Any previous cardiac surgery increases the risk, but the effect of multiple previous operations on outcomes is not significantly different from the effect of one previous operation. This risk factor is therefore retained without modification.

We explored the effect of the size and recency of myocardial infarction (MI) by requesting data on troponin levels and the temporal separation between infarct and operation. Unfortunately, the measuring of two types of troponin (I and T), the difficulty of identifying an easy and consistent conversion between the two, the multiplicity of assays available and the very wide variation between the 'normal' ranges from many hospitals meant that there is no practical, uniformly acceptable method of measuring infarct size consistently across units. As for recency, the most useful categorization was a three-level factor; MI in previous 72 h, MI 4 days to 3 months ago, no MI in last 3 months. However, this correlated very closely with urgency of operation (see below), and the effect was largely lost when urgency was appropriately taken into account. This risk factor (MI within 90 days) therefore remains unchanged.

Many units provided LV ejection fraction (LVEF) as a percentage in addition to subjective categorization of LV function. By dividing those with poor LV function into 'poor' (LVEF 21–30%) and 'very poor' (LVEF 20% or less), a slightly better fit could be obtained. Because of this, and despite some missing data on LVEF in the 'poor' LV group, we believe that adding the category 'very poor' is clinically indicated and may help reduce risk-averse behaviour. LV function is therefore divided into four categories: good, moderate, poor and very poor.

The predictive value of urgency was improved by subclassification. The previous model had 'emergency' as the only factor. The new model recognizes elective, urgent, emergency and salvage as urgency categories, and is therefore more predictive and in harmony both with other risk models and the subjective view of clinicians.

An important factor to receive attention in this study is the weight and nature of the intervention. The original model was criticized for the same risk to an isolated aortic valve replacement (AVR) as to a double valve replacement with triple coronary artery bypass grafting (CABG). In harmony with the original model, the lowest risk operation was found to be isolated on-pump CABG (off-pump surgery was associated with higher mortality, and this finding requires further study beyond the scope of this paper). We identified four classes of intervention 'weight' associated with an incremental effect on mortality:

- isolated CABG;
- single major cardiac procedure other than isolated CABG;
- two major cardiac procedures;
- three or more major cardiac procedures.

Pulmonary artery (PA) systolic pressure was treated as a dichotomous variable in the old model (>60 mmHg). We found an increasing risk associated with rising PA pressure from 30 to 55 mmHg, followed by a plateau. Pulmonary hypertension is therefore subdivided into two categories:

- PA pressure 30–55 mmHg and
- PA pressure 56 mmHg and above.

Despite evidence that brain natriuretic peptide (BNP) is an independent predictor of cardiac surgical outcomes [16], data on BNP were only available for 1638 patients (7.3%). This factor, though it may be useful in the future, is therefore not included in the model due to poor availability of data.

Finally, surgery on the thoracic aorta remains associated with higher mortality and therefore features in the risk model, but post-infarction ventricular septal rupture only appeared twice in the database with no deaths. This risk factor is therefore removed. Surgeons are reassured that the high-risk nature of this procedure continues to be recognized through other risk factors (weight of intervention, urgency, recent MI, critical preoperative state, PA pressure, etc).

Calibration and discrimination of the new model

Using the above risk factors, the final logistic model was constructed from the developmental data set and applied to both the developmental and validation data sets with very satisfactory results.

Calibration was tested by applying the final model to the validation data set which contained 5553 patients of whom 232 died in hospital (4.18%). The model-predicted mortality for this data set is 3.95%, a slight but acceptable underprediction.

Discrimination was tested by measuring the area under the ROC curve (Fig. 2C). When applied to the validation data set, the area under ROC curve was 0.8095 (95% confidence interval 0.7820–0.8360), indicating very good discrimination and a trend towards slightly but not significantly better discrimination than the old models. Goodness-of-fit test results are in Table 7.

Table 7: Goodness-of-fit data for EuroSCORE II logistic model

Group	Prob	obs0	exp0	obs1	exp1	total
1	0.0069	10	9.7	1607	1607.3	1617
2	0.0092	17	12.9	1558	1562.1	1575
3	0.0117	5	16.6	1590	1578.4	1595
4	0.0146	17	20.9	1577	1573.1	1594
5	0.0187	19	26.5	1585	1577.5	1604
6	0.0242	40	33.8	1547	1553.2	1587
7	0.0323	45	44.4	1551	1551.6	1596
8	0.0466	71	61.6	1523	1532.4	1594
9	0.0798	101	95.6	1494	1499.4	1595
10	0.8609	280	283.0	1315	1312.0	1595

Number of observations = 15 952; Hosmer–Lemeshow $\chi^2(8) = 15.48$; Prob > $\chi^2 = 0.0505$ (data collapsed into 10 quantiles of estimated probabilities (prob); obs: observed: exp: expected; 0: death; 1: survival).

Definitions and explanations of the risk factors

NYHA class. NYHA classification for dyspnoea:

- I: no symptoms on moderate exertion;
- II: symptoms on moderate exertion;
- III: symptoms on light exertion;
- IV: symptoms at rest.

CCS class 4. CCS class 4 angina (inability to perform *any* activity without angina or angina at rest).

IDDM. Insulin-dependent diabetes mellitus.

Extracardiac arteriopathy. One or more of the following:

- claudication;
- carotid occlusion or >50% stenosis (North American Symptomatic Carotid Endarterectomy Trial criteria);
- amputation for arterial disease;
- previous or planned intervention on the abdominal aorta, limb arteries or carotids.

Poor mobility. Severe impairment of mobility secondary to musculoskeletal or neurological dysfunction.

Previous cardiac surgery. One or more previous major cardiac operation involving opening the pericardium.

Renal dysfunction. This is assessed by CC as estimated using the Cockcroft–Gault formula and falls into three categories:

- CC 51–85
- on dialysis (regardless of serum creatinine)
- CC ≤ 50.

Active endocarditis. Patients still on antibiotic treatment for endocarditis at the time of surgery.

Critical preoperative state. Any one or more of the following occurring preoperatively in the same hospital admission as the operation:

- ventricular tachycardia or fibrillation or aborted sudden death;
- cardiac massage;
- ventilation before arrival in the anaesthetic room;
- inotropes;
- intra-aortic balloon counterpulsation or ventricular-assist device before arrival in the anaesthetic room;
- acute renal failure (anuria or oliguria <10 ml/h).

LV function or LVEF.

- good (LVEF 51% or more);
- moderate (LVEF 31–50%);
- poor (LVEF 21–30%);
- very poor (LVEF 20% or less).

Urgency of procedure.

- elective: routine admission for operation;
- urgent: patients not electively admitted for operation but who require surgery on the current admission for medical reasons and cannot be discharged without a definitive procedure;
- emergency: operation before the beginning of the next working day after decision to operate;
- salvage: patients requiring cardiopulmonary resuscitation (external cardiac massage) en route to the operating theatre or before induction of anaesthesia. This does not include cardiopulmonary resuscitation after induction of anaesthesia.

Recent MI. Within 90 days before operation.

Weight of procedure. This measures the extent or size of the intervention. The baseline is isolated CABG: operations 'heavier' than the baseline are in three categories:

- isolated non-CABG major procedure (e.g. single valve procedure, replacement of ascending aorta, correction of septal defect, etc.);
- two major procedures (e.g. CABG + AVR), or CABG + mitral valve repair (MVR), or AVR + replacement of ascending aorta, or CABG + maze procedure, or AVR + MVR, etc.);
- three major procedures or more (e.g. AVR + MVR + CABG, or MVR + CABG + tricuspid annuloplasty, etc.), or aortic root replacement when it includes AVR or repair + coronary reimplantation + root and ascending replacement).

Only major cardiac procedures count towards to the total. Examples of procedures which *do not* qualify are: sternotomy, closure of sternum, myocardial biopsy, insertion of intra-aortic balloon, pacing wires, closure of aortotomy, closure of atriotomy; removal of atrial appendage, coronary endarterectomy as part of CABG, etc.

DISCUSSION

Limitations

EuroSCORE II was constructed from an international, contemporaneous and highly accurate, validated database and should therefore be a robust risk model for use in cardiac surgery worldwide. There are, of course, limitations to this study and these are dictated by the restrictions imposed by the methodology and logistics of constructing the study.

With only 21 patients over the age of 90 in the data set, the risk model may not be accurate in these patients. The oldest patient in the EuroSCORE database was 95, and the model is therefore not validated in patients over this age.

Participating units were volunteers, and this introduces an element of bias in self-selection. However, there is no mechanism to force all units to participate or even to force a randomly selected sample of units to do so. Even if that were possible, any coercive element in such a study would have resulted in potentially greater bias introduced by reduced willingness, ability or both to provide data. We believe that voluntary participation improves the chances of obtaining high-quality data.

A further bias may be introduced by the simple possibility that units with the ability and willingness to provide data may have better outcomes than those without such facilities. To militate against such a bias, every effort was made to encourage units of all types to participate. Multimedia promotional literature clarified that we were not seeking only 'centres of excellence'. Furthermore, we believe that many units recognized that, if only centres of excellence participated in the study, the resultant model may set a standard that would be hard for many units to meet. We believe, but cannot prove, that both the number

and range of participating units are fairly representative of current cardiac surgery.

Validation is necessarily limited by the resources available to the project. We are unable directly to employ researchers to gather and validate data in such a large number of units on a global scale. Nevertheless, validation was facilitated by web-based data collection where multiple mandatory fields with error and range checks were employed throughout. These features, when added to the subsequent logic checks on received data, resulted in an overall data set of high quality. This is supported by the fact that the number of units which had to be disregarded in the analysis due to faulty data was indeed very small.

Finally, the model is based on logistic regression, taking account of multiple risk factor interactions. Artificial neural networks (ANNs) may surpass logistic regression and we are currently working on developing ANN versions of the model to determine what additional advantages may be obtained by this approach [17].

The new model

The old model is no longer appropriately calibrated. Risk-adjusted mortality has fallen by around a half in comparison with what could be expected in the 1990s. This is a powerful testament to the substantial improvement in quality of care that has been achieved in cardiac surgery. The new model is far better calibrated. It continues to rely on a relatively small number of risk factors, most of which featured in the original model. The modifications to the risk factors in the new model are modest but they are both evidence-based and intuitive. The new risk factors and definitions should better reflect current practice with improvement in discrimination as well as calibration, but that can only be gauged when the model is in use.

We set out to produce a global cardiac surgical risk model. Others believe that procedure-specific models are superior, and there have been and doubtlessly will be many models designed to predict risk in very specific circumstances. Procedure-specific risk models are better for narrowly defined procedures, but become problematic in complex double and triple procedures. There are inherent difficulties in data collection and procedure definition when developing such models for every conceivable type of operation and combination of operations.

Using EuroSCORE II

Like a scalpel or a needle holder, a risk model is an instrument in the cardiac surgeon's toolkit. Used judiciously, it can enhance the quality of cardiac surgical care by facilitating better decision-making and providing a benchmark for quality control. Ill-conceived and misguided application can damage both patients and surgeons. We highlight those areas where the application of risk modelling calls for special caution and the exercise of judgment.

(i) The intrinsic imperfection of modelling

No risk model predicts the outcome for an individual patient. Any prediction of percentage mortality is for a population of patients: an individual patient will either survive or die from an operation regardless of the predicted risk. However, for an individual patient, the knowledge of the predicted mortality for a group of similar patients undergoing the same procedure is an important part of decision-making and informed consent.

No risk model is perfect. The selection of risk factors in a model is a necessary compromise between what is practical and what is feasible. All surgeons know that not all risk factors appear in all models, nor would it be possible to devise a model which includes every conceivable risk factor and every rare medical syndrome. This is for valid reasons: such a model would be too complex for clinical use, and the database from which it can be derived is virtually impossible to assemble. The selection of risk factors to include in the model is necessarily a compromise in which risk factors compete for inclusion on the basis of four features:

- availability
- objectivity
- resistance to falsification
- credibility to users.

Where rare risk factors are present, the clinician must make a value judgment about the absolute applicability of the model to the patient in question.

It can be argued that a global risk model is unsuitable for the fine distinction between risk levels in different procedures, and this is of course true. The Society of Thoracic Surgeons risk model has a handful of broad categories, but risk modelling can even further be refined by additional subclassification based on procedure, pathology, different combinations of procedures and multiple combinations of the all of the above. However, such refinement has its price in that the resultant modelling becomes complex and there will be categories of patients and procedures that are not supported by the model. We deliberately set out to achieve a global risk model for cardiac surgery in adults and believe that we achieved this to a good standard, but that of course does not preclude the creation of separate procedure-specific risk models by other workers if this is perceived as desirable. The global nature of EuroSCORE is undoubtedly one of its limitations, but it has also been one of its great strengths in achieving user acceptability.

(ii) The variation in outcomes between centres and surgeons

A risk model sets a standard, and units and surgeons will perform at a level equal, below or above that standard. Once the model is in use, all units and surgeons should calculate their RAMR by dividing their actual (observed) mortality by their predicted mortality according to the risk model. The RAMR can then used in a number of risk assessment situations, two of which are addressed below:

- Informed consent and the evaluation of risk for an individual patient: the most accurate and scientific predicted mortality to quote to a patient is the predicted mortality for the procedure as calculated by EuroSCORE II multiplied by the unit's or the individual surgeon's RAMR. The RAMR can only be stated with meaningful confidence when sufficient observations have been made of predicted versus observed mortality in a large enough patient population representing at least a year's workload. It should be derived from recent experience and updated at regular intervals.
- Determining the level of risk at which a new or experimental procedure is justified: if it is decided that an experimental

procedure should be offered to patients whose risk from conventional cardiac surgery exceeds, for example, 20%, then the experimental procedure in a particular unit should only be offered to those whose predicted mortality is >20% after division by the unit's RAMR. If, for example, a unit is outperforming EuroSCORE II and its RAMR is 0.5, then the experimental procedure can be considered in patients whose score is 20% divided by 0.5, i.e. 40%.

(iii) The use of risk models as performance indicators

Some risk models, including EuroSCORE, have been used or adapted for use as performance indicators to evaluate the quality of a clinical service. EuroSCORE has been developed to deal with a global cardiac surgical practice, and this approach is appreciated by many cardiac surgeons with a mixed practice who require a single model for risk evaluation, but it is possible that any model designed for global cardiac surgical risk assessment may work better in some categories of patients than others. We plan to sub-analyse the data from the EuroSCORE database to determine whether there are indeed areas where such differences are important, and we have no doubt that other workers will carry out similar analyses in their units and databases. In the meantime, we advise that caution be exercised before applying the model to sub-groups of isolated procedures rather than to a global practice, and to surgeons whose practice consists overwhelmingly or exclusively of one type of heart operation or another. As ever, the determination of the performance threshold should be the responsibility of bodies and individuals who are able to assess the relevance of any risk model to their data and to select the appropriate threshold on the basis of careful and sound professional and clinical judgment.

(iv) The future

No model is future-proof. Work has already begun on the EuroSCORE III project, in which we plan to collect continuous, prospective contemporaneous data from specially selected units to determine when and how the model will require an update. This will ensure that the model will remain relevant as cardiac surgical results hopefully continue to improve.

CONCLUSION

EuroSCORE II, an update of the logistic EuroSCORE model, uses similar methodology but is derived from a more current data set and refined to incorporate evidence-based improvements and to reflect better current cardiac surgical practice. It is recommended for assessing risk in general adult cardiac surgery.

ACKNOWLEDGEMENTS

We are grateful to the *Journal of Heart Valve Disease* for promoting the project and to Sandra Church for help in designing promotional material. We thank all participating centres for investing time and effort into data collection, without which this project would not have been possible.

Funding

This study was supported by Edwards Laboratories, Karolinska Hospital, Stockholm and Papworth Hospital, Cambridge.

Conflict of interest: none declared.

REFERENCES

[1] Nashef SAM, Roques F, Michel P, Gauducheau E, Lemeshow S, Salamon R; the EuroSCORE Study Group. European System for Cardiac Operative Risk Evaluation (EuroSCORE). Eur J Cardiothorac Surg 1999;16:9–13.
[2] Roques F, Nashef SAM, Michel P, Gauducheau E, de Vincentiis C, Baudet E *et al*. Risk factors and outcome in European cardiac surgery: analysis of the EuroSCORE multinational database of 19 030 patients. Eur J Cardiothorac Surg 1999;15:816–23.
[3] Basraon J, Chandrashekhar YS, John R, Agnihotri A, Kelly R, Ward H *et al*. Comparison of risk scores to estimate perioperative mortality in aortic valve replacement surgery. Ann Thorac Surg 2011;92:535–40.
[4] Parolari A, Pesce LL, Trezzi M, Loardi C, Kassem S, Brambillasca C *et al*. Performance of EuroSCORE in CABG and off-pump coronary artery bypass grafting: single institution experience and meta-analysis. Eur Heart J 2009;30:297–304.
[5] Qadir I, Perveen S, Furnaz S, Shahabuddin S, Sharif H. Risk stratification analysis of operative mortality in isolated coronary artery bypass graft patients in Pakistan: comparison between additive and logistic EuroSCORE models. Interact Cardiovasc Thorac Surg 2011;13:137–41.
[6] Lebreton G, Merle S, Inamo J, Hennequin JL, Sanchez B, Rilos Z *et al*. Limitations in the inter-observer reliability of EuroSCORE: what should change in EuroSCORE II? Eur J Cardiothorac Surg 2011;40:1304–8.
[7] Shih HH, Kang PL, Pan JY, Wu TH, Wu CT, Lin CY *et al*. Performance of European System for Cardiac Operative Risk Evaluation in Veterans General Hospital Kaohsiung cardiac surgery. J Chin Med Assoc 2011;74:115–20.
[8] Akar AR, Kurtcephe M, Sener E, Alhan C, Durdu S, Kunt AG *et al.*; The working Group for the Turkish Society of Cardiovascular Surgery and Turkish Ministry of Health. Validation of the EuroSCORE risk models in Turkish adult cardiac surgical population. Eur J Cardiothorac Surg 2011;40:730–5.
[9] Yap CH, Reid C, Yii M, Rowland MA, Mohajeri M, Skillington PD *et al*. Validation of the EuroSCORE model in Australia. Eur J Cardiothorac Surg 2006;29:441–6.
[10] Roques F, Michel P, Goldstone A, Nashef SAM. The logistic EuroSCORE. Eur Heart J 2003;24:881–2.
[11] Van Gameren M, Klieverik LM, Struijs A, Venema AC, Kappetein AP, Bogers AJ *et al*. Impact of the definition of renal dysfunction on EuroSCORE performance. J Cardiovasc Surg 2009;50:703–9.
[12] Walter J, Mortasawi A, Arnrich B, Albert A, Frerichs I, Rosendahl U *et al*. Creatinine clearance versus serum creatinine as a risk factor in cardiac surgery. BMC Surg 2003;3:4.
[13] Sergeant P, de Worm E, Meyns B, Wouters P. The challenge of departmental quality control in the reengineering towards off-pump coronary artery bypass grafting. Eur J Cardiothorac Surg 2001;20:538–43.
[14] Shaheen AA, Kaplan GG, Hubbard JN, Myers RP. Morbidity and mortality following coronary artery bypass graft surgery in patients with cirrhosis: a population-based study. Liver Int 2009;29:1141–51.
[15] Suman A, Barnes DS, Zein NN, Levinthal GN, Connor JT, Carey WD. Predicting outcome after cardiac surgery in patients with cirrhosis: a comparison of Child-Pugh and MELD scores. Clin Gastroenterol Hepatol 2004;2:719–23.
[16] Cuthbertson BH, Croal BL, Rae D, Gibson PH, McNeilly JD, Jeffrey RR *et al*. N-terminal pro-B-type natriuretic peptide levels and early outcome after cardiac surgery: a prospective cohort study. Br J Anaesth 2009;3:647–53.
[17] Nilsson J MD, Ohlsson M, Thulin L, Höglund P, Nashef SAM, Brandt J. Risk factor identification and mortality prediction in cardiac surgery using artificial neural networks. J Thorac Cardiovasc Surg 2006;132:12–9.

APPENDIX. CONFERENCE DISCUSSION

Dr J. Takkenberg *(Rotterdam, The Netherlands)*: The new EuroSCORE, EuroSCORE II, has good calibration and excellent discrimination with an area

under the curve of 0.81. That's great. I have numerous questions, as you can imagine, but I was told to restrict myself to two. So my first question is a more technical question related to the modelling process. An area under the curve of 0.81 is excellent, but perhaps could be even better if the data set would be utilized to the fullest. I would use imputation techniques for those data that can be assumed missing completely at random and which you have now excluded. And I would actually apply bootstrapping techniques as a method of validation to prevent overfitting. Additionally, one could take into account potential interaction of variables. You state in your presentation that you looked at that. So, for example, the effect of female sex on mortality may be stronger or weaker in older patient groups compared to younger patient groups. So did you find any interaction and consider interaction terms when you were building EuroSCORE II? That would be my first question.

My second question pertains to the use of hospital mortality as the primary endpoint of your model. The hazard function of death after cardiac surgery shows an early phase of rapidly falling risk in the first three post-operative months. It underlines that surgical mortality is time-related rather than related to the location of the patient, and particularly in western countries where patients are discharged really early to general hospitals. Realizing the practical barriers to measure 30- and 90-day mortality, I nevertheless find that we should aim to report time-related outcomes in our models as part of our endeavour to continuously improve cardiac surgical care. So do you plan to build a EuroSCORE II model with time-related outcomes as well?

Dr. Nashef: First, may I say I fully agree with everything that you have said. In answer to your first question, there is no doubt that logistic regression modelling does offer you the opportunity to explore interactions between variables, but this has to be initiated in that you have to look for them yourself. And we have done our best to look for as many interactions as possible. In fact, one of the very interesting interactions we found is the interaction between renal function and age, and some of you, when you try the new model, are going to be appalled to find that when you enter a patient age 89, the predicted risk is very small, but that will increase as soon as you enter the creatinine clearance. So we have looked as much as we could for clinical interactions of risk factors even beyond what is immediately apparent. However, regression modelling by its nature is limited in its ability to do that. And as you say, bootstrap techniques and the use of artificial neural networks offer substantial possibilities in doing this, and I can assure you that we are currently working on an artificial neural network model to see if it can actually improve on what the regression model can offer.

And in response to your second question, of course, there is mortality after going from hospital and, of course, it doesn't plateau until 90 days have passed. However, this is a suboptimal situation. I would put it to you that those units that actually voluntarily participated are probably better at getting data than those units that did not participate. And even if you participate in the EuroSCORE project, and only half of you can give me 90-day mortality, then it is simply not practical to use that in a model that will be used for quality of performance measurement. Because the problem is those who have the data will report higher mortalities; those who don't have the data will report lower mortalities. So we have to accept that this is not ideal. But in the current status of data availability to units, this is what we are stuck with. In the future, we hope units will start looking at their 30-day and especially 90-day mortality. And when that information is available to all, then, of course, it should be included.

Dr Takkenberg: So you will promise to do that for EuroSCORE III then?

Dr Nashef: Work on EuroSCORE III has started already.

Dr A. Badreldin *(Jena, Germany)*: We participated in the data collection. I have just two quick questions. First of all, the benefit of any preoperative scoring system has been thoroughly discussed in the literature over the last 15 years so as to compare the performance of different centres on the National Registry level. Moreover, in clinical practice for us, for surgeons most importantly, it is a basis for the preoperative consents and preoperative discussion with the patient. You included the operative procedure in the new EuroSCORE. Should we drop this second rule of the scoring system (preoperative consents based on the score value), especially since we know that we would change our strategy, maybe often intraoperatively, according to TOE diagnostics or any unexpected surprise?

The second question very quickly. A scoring system should not be used in any centre unless it has already been validated internally for its reliability. You did not validate this score system, EuroSCORE II, yet. Should we add this note for any end user on the website to avoid any drawbacks that we have already experienced with the original EuroSCORE due to lack of this validation?

Dr Nashef: In answer to your first question, which I think was about incorporating data about the operative procedure, I would say that the amount of data that is requested in terms of the operative procedure itself is fairly limited in EuroSCORE and it also tends to be very, very objective data. All models are subject to gaming. And if somebody wants to be very clever and devious in gaming, then you can upgrade a procedure relatively easily. We hope that no surgeons do that. But I think we really have to have some measure of the weight of the intervention, because for every comment like yours, I have received 10 complaining that in the old EuroSCORE, an AVR scores the same as an AVR with five grafts and that this is really unfair.

Now, in answer to your second question—you have to remind me what your second question was.

Dr Badreldin: Validation, eventually the customization, first degree or second degree.

Dr Nashef: A model sets a standard. Units will perform as well as, better, or worse. And the data and the information you give to your patients should take into account how your unit performs and how you yourself perform against that model. In the next few months, an opportunity to calculate your own risk-adjusted mortality ratio will appear on the website so that the model can be tailor-made for your own institution. But, of course, until we have data from you on that, we cannot really do it yet.

You have criticized the old model for causing problems in that respect, particularly in relation to the TAVI selection. But, of course, if you're going to use a model, a model can discriminate very powerfully, but its calibration must be adjusted to what you yourself can achieve. Many units will hover around the average, some units will have a mortality that is much higher than predicted, and some units will have a mortality that is much lower than predicted. And clinicians should use their reasoning and their insights in order to adjust things for their own patients. This is when it comes to informed consent. In terms of the comparison of the quality of performance, the standard is there and it can still be used.