Survival after Major Cardiac Surgery: Performance and Comparison of Predictive Ability of EuroSCORE II and Logistic EuroSCORE in a Sample of Mediterranean Population

María Elena Arnáiz-García[1] Jose María González-Santos[1] Javier López-Rodríguez[1] María José Dalmau-Sorlí[1] María Bueno-Codoñer[1] Adolfo Arévalo-Abascal[1]

[1] Department of Cardiac Surgery, University Hospital of Salamanca, Salamanca, Spain

Thorac Cardiovasc Surg 2014;62:298–307.

Address for correspondence María Elena Arnáiz-García, MD, Department of Cardiac Surgery, University Hospital of Salamanca, Paseo San Vicente, 58-182 Salamanca, Salamanca 37007, Spain (e-mail: elearnaiz@hotmail.com).

Abstract

Background The European System for Cardiac Operative Risk Evaluation (EuroSCORE) II has been recently introduced to improve mortality prediction in cardiac surgery. We compare the predictive ability of the new EuroSCORE II with that of the original logistic EuroSCORE and we made an evaluation of a sample of our population submitted to major cardiac surgery in the context of a Mediterranean country.

Materials and Methods Predicted and observed mortality were recorded in 1,200 consecutive patients undergoing major cardiac surgery at our institution with both logistic EuroSCORE and EuroSCORE II. Patients were grouped according to type of surgery: isolated valvular ($n = 538$), isolated coronary ($n = 322$), combined ($n = 192$), and miscellaneous ($n = 148$). Predictive capacity of both scales was compared for overall population and for each group in terms of calibration and discrimination using the observed by expected mortality rate, Hosmer–Lemeshow test, and C-statistic.

Results Overall mortality was 6.8%, whereas that predicted by logistic EuroSCORE and EuroSCORE II was 9.7 and 3.7%, respectively. Mortality in our population was higher than mortality expected according to the original EuroSCORE II database. For all groups included in our population, logistic EuroSCORE overestimated mortality and EuroSCORE II underestimated the outcome even more. However, EuroSCORE II showed better calibration than logistic EuroSCORE for overall, valvular, and combined surgery. In contrast, logistic EuroSCORE demonstrated better calibration for coronary surgery. Discrimination capacity was good for both risk scores, but it was superior for logistic EuroSCORE than for EuroSCORE II in all considered subgroups unless combined surgery.

Conclusion Mortality in our population was higher than the mortality that would have been expected by the new EuroSCORE II analysis. Although EuroSCORE II has good calibration and discrimination capacity, both are worse than those demonstrated by logistic EuroSCORE. Forthcoming evaluations are necessary when the new model will be widely used.

Keywords
- cardiac
- surgery
- mortality
- risk assessment

received
July 15, 2013
accepted after revision
February 24, 2014
published online
May 28, 2014

DOI http://dx.doi.org/10.1055/s-0034-1376255.
ISSN 0171-6425.

Introduction

Currently, outcome predictions are necessary for physicians and patients to make decisions on treatment options. In addition, early mortality is a well-recognized indicator of quality of care in surgical patients. Risk scales are important tools to both indicate surgery and assess the quality of perioperative management. However, the value of the risk scales depends on its ability to predict prognosis in a particular patient or in patient groups.

The EuroSCORE (European System for Cardiac Operative Risk Evaluation) scale is a risk model for predicting 30-day mortality after cardiac surgery.[1,2] It was developed from a multinational European database with 14,799 patients who have undergone cardiac surgery by the end of 1995. The first EuroSCORE scoring system was described in two versions. Initially, it was based on an additive system (the additive EuroSCORE) and then logistic EuroSCORE was proposed as a more sound approach to risk prediction. Logistic EuroSCORE was first published in 1999 and gained wide popularity in most European countries.[1–3] Since then, the use of both risk scores has spread all over the world and has become a standard for measurement of risk in many European cardiac surgery units and also in other continents.[4–7]

The original EuroSCORE has shown a good prediction capacity for many years.[8,9] However, concerns about its calibration and discriminative power have appeared during last decade. Several investigators have suggested that calibration of the model could be unadjusted resulting in mortality risk overestimation, especially when applied to high-risk patients.[10–14] This has been related with the old date of the registry, as patients underwent surgery more than 15 years ago.[15] It has also been argued that certain predicting variables included in logistic EuroSCORE are highly correlated. Hence, original EuroSCORE may be inappropriately calibrated for application in current cardiac surgery practice.

Currently, despite older age and sicker conditions of the patients submitted to cardiac surgery, a reduction in early mortality has been clearly noted.[11,16,17] Better outcomes have been related to technical advances and improvements in perioperative and postoperative care introduced during last decade. These changes forced a revision of prediction models and the necessity to develop an updated mortality risk scale.[18,19]

Accordingly, logistic EuroSCORE was amended and the new EuroSCORE II came and was first presented in October 2011.[16] The EuroSCORE II was constructed from an international current data collection of 22,381 patients, to reflect a more current view about cardiac surgical management and practice. First experiences with this new risk scale have been recently reported, mostly showing a better calibration and discrimination capacity than that obtained by logistic EuroSCORE.[11,14,20–22]

The aim of this study is to validate the predictive power of the EuroSCORE II in terms of calibration and discrimination in comparison with the original logistic EuroSCORE in a population with high proportion of aged patients and complex procedures.

Materials and Methods

The EuroSCORE II

The new EuroSCORE is indicated to assess surgical risk in general adult cardiac surgery. Main differences between EuroSCORE II and original logistic EuroSCORE remain on predictors. New evidence-based risk factors have been incorporated and some others have been modified or excluded. Besides, its impact on risk model has been updated using a standard logistic regression approach. Among patient-related factors, gender, age, previous cardiac surgery, chronic pulmonary disease, active endocarditis, and critical preoperative state remain with some changes on definitions. Diabetes on insulin therapy is now included. Extracardiac arteriopathy now considers previous amputation and neurological dysfunction is replaced by poor mobility. Also, serum creatinine value has been replaced by creatinine clearance (CC) and renal dysfunction is now stratified according to its severity: moderate if CC is 50 to 85 mL/min and severe when CC is < 50 mL/min. In addition, left ventricular ejection fraction is now classified in four categories. The New York Heart Association classification has been included and pulmonary hypertension is now stratified in two categories, according to the systolic pulmonary arterial pressure (moderate from 31 to 55 mm Hg and severe if more than 55 mm Hg).

Operation-related risk factors have also been modified. Priority of surgery has been amended and now includes four categories: elective, urgent, emergent, and salvage procedures. Surgery on the thoracic aorta remains as a risk factor, but ventricular septal rupture has been excluded. However, one of the main changes in EuroSCORE II is that this scale takes into account the complexity or weight of the surgical procedure with four possible categories: isolated coronary artery bypass graft (CABG), single non-CABG, two procedures, or three procedures.

Population of the Study

We performed a retrospective study including 1,200 consecutive patients undergoing mayor cardiac surgery at our institution from January 1 2009 to July 31 2012. Logistic EuroSCORE[1–3] was calculated as data were prospectively entered in our database (SICCS, Informatics System for Cardiac Surgery, Biomenco, Barcelona, Spain). The new EuroSCORE II[16] was retrospectively calculated in the same patients using the free online calculator available in the official Web site of the EuroSCORE project (www.EuroSCORE.org). Main demographic, clinical, and operative characteristic were also retrospectively obtained from our database. All patients were stratified on subgroups according to the surgical category:

1. Isolated valve surgery: repair or replacements of the aortic, mitral, tricuspid, or pulmonary valve.
2. Isolated coronary surgery, including on-pump and off-pump procedures.
3. Combined valve and coronary surgery.
4. Miscellaneous procedures included surgery for congenital cardiopathy in adult, aortic root, ascending aorta or aortic arch surgery, tumoral surgery, ventricular surgery, surgery

of the mechanical complications of infarction, surgery for cardiac traumas, surgical ablation of arrhythmias, and pulmonary embolectomy associated or not to other procedures.

The goodness of fit of both logistic scales and discriminative capacity were analyzed in the estimation of operative mortality. Operative mortality was defined as that occurring during hospital stay or in the 30 days following surgery.

Calibration/Discrimination and Statistical Analysis

The predictive power of the new EuroSCORE II was analyzed and compared with that of the logistic EuroSCORE. Accuracy of prediction in both risk scales was evaluated in terms of calibration and discrimination.

Calibration of both scales was first assessed calculating the observed/expected mortality ratio (O/E ratio) obtained by dividing observed by expected mortality. This reflects the ability of the risk score to estimate the real outcome. Thus, an O/E ratio above 1 means underestimation of mortality and an O/E ratio below 1 reflects that actual mortality is overestimated. Furthermore, calibration of both risk scales was further analyzed with the "goodness-of-fit" test of Hosmer-Lemeshow (H-L) for logistic regression models. This test determines how much the predicted incidence of the event match the observed incidence of events along a range of scores grouped by increasing risk deciles. A p value < 0.05 indicates lack of fit of the risk model reflecting a poor calibration.[14–23]

Discrimination was assessed using the C-statistic. This was calculated by measuring the area under the receiving operating characteristic (ROC) curves. It represents the probability that predicting the outcome is better than chance alone. It is used to differentiate between the individuals of a sample who suffer an event (death) and those who do not. This analysis allows comparison of logistic regression models. A C-statistic of 0.5 denotes a null ability to discrimination (the model is not better than chance at predicting the outcome). A C-statistic between 0.7 and 0.9 is a reasonable value for contemporary models and a C-statistic value over 0.9 denotes an excellent discriminative power. A C-statistic of 1 indicates a perfect discrimination, when the model perfectly identifies the outcome.[14–24]

All statistical analyses were performed with the statistical package SPSS software version 20.0 (SPSS, Inc., Chicago, IL, United States) for Windows. Comparison between ROC curves were performed using the MedCalc 12.3 statistical package. A p-value < 0.05 was considered significant. Data are presented as the mean ± standard deviation for continuous variables and as percentages for discrete variables.

Results

A summary of the population characteristics is shown in ►**Table 1**. The median age was 73 years (interquartile range, 64–77), and 63.6% were males. Patients were distributed according to the type of surgery as follows: isolated valve surgery ($n = 538$; 44.8%), isolated coronary surgery ($n = 322$; 26.8%), combined valve and coronary surgery ($n = 192$; 16%), and miscellaneous surgical procedures ($n = 148$; 12.3%).

Main risk factors and characteristics define the high-risk profile of our patients. Population had medium age of 73 years: 142 (11%) had peripheral arteriopathy and 70 (5.8%) had chronic pulmonary obstructive disease. Renal function was moderately decreased in 597 patients (49.7%) and severely decreased in 272 patients (22.7%). Seventy-six patients (6.3%) had previous cardiac surgery and 193 (16.1%) had recently suffered a myocardial infarction. One hundred and seventy patients (14.2%) had severe pulmonary artery hypertension. The surgical procedure was other than isolated CABG in 884 patients (73.4%) and 136 (11.3%) underwent a highly complex procedure. Surgery was performed under critical condition in 76 patients (7.7%).

Calibration

Overall 30-day operative mortality was 6.8% (81 patients). The observed mortality was compared with that predicted with both EuroSCORE risk models for the overall population and for all subgroups analyzed (►**Fig. 1**). When applied to the whole population, the mortality predicted by logistic EuroSCORE was 9.7 ± 11.1%, thus overestimating observed mortality in 2.9% as reflected by an O/E mortality ratio of 0.70. In contrast, the new EuroSCORE II predicted a mortality of 3.7 ± 4.6%, underestimating actual mortality in 3.1%, as reflected by an O/E ratio of 1.83. The expected and observed mortality calculated with both scores and the O/E ratio for the overall population and for each surgical subgroup are shown in ►**Table 2**.

Mortality in our population was higher than the mortality that would have been expected by the experience collected in the population included in the original database of EuroSCORE II. According to our small sample of population belonging to a Mediterranean country, logistic EuroSCORE significantly overestimated mortality in almost all analyzed subgroups but combined surgery subset; logistic EuroSCORE–predicted mortality for this subgroup had a perfect concordance with observed mortality. In contrast, a significant underestimation in predicted mortality was observed for all surgical categories with the EuroSCORE II. Again, the mortality actually observed in every subgroup was approximately the average of the mortality calculated by both scales. The O/E ratio ranged between 0.49 and 1.0 for logistic EuroSCORE and between 1.42 and 2.09 for EuroSCORE II. The logistic EuroSCORE showed a perfect calibration in combined surgery and a poor one in miscellaneous procedures, while these positions were reversed with the EuroSCORE II.

For better quality assurance, the observed and expected mortalities were also calculated with patients stratified by risk groups: < 2, 2–4.9, 5–9.9, and > 10% (►**Fig. 2**). The logistic EuroSCORE showed good calibration capacity in medium and very high risk groups. Generally, EuroSCORE II showed an underestimation of mortality for all subgroups, but superior for medium and very high groups, just the same subgroups where logistic EuroSCORE showed better calibration.

Table 1 Population characteristics according to logistic EuroSCORE and EuroSCORE II variables

	EuroSCORE	Both scales	EuroSCORE II
Patient-related factors			
Age (y)		73 (64–77)	
Sex			
Male		763 (64.6)	
Female		437 (36.4)	
Diabetes on insulin			71 (5.9)
CPOD		70 (5.8)	
Peripheral arteriopathy		142 (11.8)	
Neurological dysfunction	6 (0.5)		
Poor mobility			10 (8.3)
Renal dysfunction			
Cr > 200 µmol/L	25 (2.1)		
CC ≥ 85 mL (min)			324 (27.0)
CC 50–85 mL/min			597 (49.7)
CC ≤ 50 mL/min			272 (22.7)
On dialysis			7 (0.6)
Active endocarditis	23 (1.9)		
Previous cardiac surgery		76 (6.3)	
Critical preoperative state		92 (7.7)	
Cardiac-related factors			
Recent AMI		193 (16.1)	
Unstable angina			
CCS class IV			
NYHA			
I			132 (11.0)
II			548 (45.7)
III			384 (32.0)
IV			136 (11.3)
	EuroSCORE	**Both scales**	**EuroSCORE II**
Left ventricular function (%EF)			
> 50	975 (81.2)		975 (81.2)
31–50	186 (15.3)		186 (15.3)
≤ 30	39 (3.2)		34 (2,8)
≤ 20			5 (0.4)
Pulmonary artery pressure (mm Hg)			
SPAP ≤ 30			773 (64.4)
SPAP 31–55			257 (21.4)
SPAP ≥ 55			170 (14.2)
SPAP ≥ 60	168 (14.0)		
Operation-related factors			
Previous cardiac surgery		76 (6.3)	
Priority			
Elective			781 (65.1)

(*Continued*)

Table 1 (*Continued*)

	EuroSCORE	Both scales	EuroSCORE II
Urgent			374 (31.2)
Emergent	45 (3.7)		45 (3.7)
Salvage			0
Other than isolated CABG	884 (73.7)		
Surgery of the aorta		125 (0.4)	
Post-AMI VSD	1 (0.1)		
Weight of the procedure			
Other than isolated CABG	881 (73.4)		
0 Isolated CABG			319 (26.6)
1 Isolated non-CABG			317 (26.4)
2			428 (35.7)
3			113 (9.4)
4			22 (1.8)
5			1 (0.1)

Abbreviations: CABG, coronary artery bypass grafting; CC, creatinine clearance; CPOD, chronic pulmonary obstructive disease; Cr, creatinine; EF, ejection fraction; NYHA, New York Heart Association; Post-AMI, acute myocardial infarction; SPAP, systolic pulmonary artery pressure; VSD, ventricular septal defect.
Note: Data are presented as the median (interquartile range) or as absolute frequencies (percentage) depending on variable type.

The calibration was analyzed in more detail with the test of H-L. Results of this analysis are shown in ►**Table 3**. With regard to the total of our population analyzed, the calibration of the EuroSCORE II was better than that of the logistic EuroSCORE, although with both scales prediction significantly diverged from the observed incidence of the event. Calibration of the logistic EuroSCORE was better adjusted in coronary surgery and performed quite well in miscellaneous and combined surgery, while the EuroSCORE II was better fitted in the combined and miscellaneous surgery. Both scales demonstrated an acceptable calibration, except the logistic EuroSCORE on valvular surgery.

Discrimination

Discriminative capacity of logistic EuroSCORE and EuroSCORE II was assessed by measuring the area under the curve (AUC)–ROC for both the overall population and for each type of surgery subgroup (►**Fig. 3**). Results of this analysis are shown in ►**Table 4**. Accuracy of both risk models was good for the entire population, being superior for logistic EuroSCORE than for EuroSCORE II. The original logistic EuroSCORE also showed superior discriminative capacity than EuroSCORE II for all surgical groups except combined surgery. Discriminative capacity of both scales was good for coronary surgery, fairly good for miscellaneous surgery, and only modest for valvular and combined surgery, the only group in which the EuroSCORE II showed more accurate results.

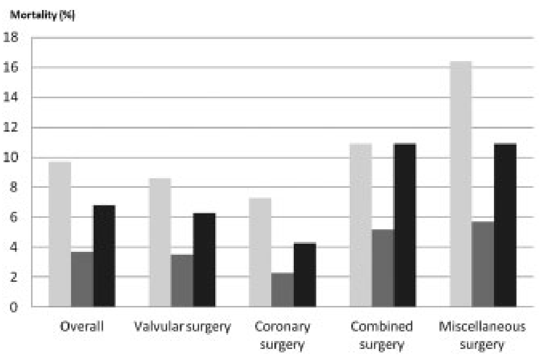

Fig. 1 Logistic EuroSCORE, EuroSCORE II–predicted mortality, and observed mortality. EM LES, logistic EuroSCORE–expected mortality; EM EII, EuroSCORE II–expected mortality; OM, observed mortality (overall population, valvular surgery, coronary surgery, combined surgery, miscellaneous surgery).

Discussion

Risk overestimation of logistic EuroSCORE has been reported for several years and it is also corroborated in our analysis. Different publications have highlighted this assertion.[9,12,17] To the sight of our outcomes, mortality observed was higher than mortality expected by the experience collected in the main population included in the original EuroSCORE II analysis. That is the reason why we conclude that in our single-institution experience, logistic EuroSCORE overestimated the observed mortality. Besides, in contrast to some recent publications,[11,16,20,21] EuroSCORE II underestimates mortality even in a more pronounced grade for both, the whole population, and in any subgroup it was divided. Observed by expected mortality ratio was approximately 0.7 of that predicted by logistic EuroSCORE and 1.83 of that estimated by EuroSCORE II. Differences in model accuracy can be justified by the high-risk profile of our population, clearly superior to that usually reported in previous publications.[16–20] As it is shown in ►**Table 1**, our patients are older than those included in other series, and had more frequently severe renal

Table 2 Calibration of logistic EuroSCORE and novel EuroSCORE II

	No. of patients	Observed mortality rate (%)	Predicted mortality rate, LES (%)	O/E mortality ratio LES	Predicted mortality rate, EII (%)	O/E mortality ratio, EII
Total population	1,200	6.8%	9.7 ± 11.1	0.70	3.7 ± 4.6	1.83
Valvular surgery	538	6.3%	8.6 ± 8.6	0.73	3.5 ± 4.1	1.80
Coronary surgery	322	4.3%	7.3 ± 10.3	0.59	2.3 ± 3.0	1.86
Combined surgery	192	10.9%	10.9 ± 10.6	1.0	5.2 ± 5.6	2.09
Miscellaneous surgery	148	8.1%	16.4 ± 17.5	0.49	5.7 ± 6.4	1.42

Abbreviations: EII: Euroscore II; LES, logistic EuroSCORE; O/E mortality ratio, observed by expected ratio.
Notes: Predicted and observed mortality for the overall population and for all types of surgery subgroups. Expected mortality by the observed and the expected mortality ratio (O/E ratio) calculated for both logistic EuroSCORE (LES) and EuroSCORE II (EII).

dysfunction, pulmonary hypertension, or other preoperative critical conditions. Besides, our patients are more often operated on urgent or emergent basis and are submitted to more complex procedures. In consequence, the mortality estimated by both scales is greater than that reported in similar studies.

Cardiac surgery risk models are important tools for cardiologist and cardiac surgeons to assess operative risk and advise patient about cardiac procedures. The original scores have been used for more than two decades with undeniable usefulness, but its accuracy has been recently questioned.[11,19,25,26] The main reason argued is that mortality related to cardiac surgery has decreased despite sicker and more complex patients are nowadays submitted to cardiac surgery. Besides, old risk scores are based on populations with a small proportion of patients older than 80 years and a reduced number of valvular surgery procedures.[1–3,16] This profile has changed as a consequence of demographic variations, the increasing role of percutaneous procedures, and changes on surgical indications. Improvements in quality of perioperative care have also influenced in outcomes and mortality associated to cardiac operations.[16,19,27]

Logistic EuroSCORE was the model used to date in most European countries.[3,5] However, it is well known that this scale, created from a database of patients operated in 1995, nowadays overestimates mortality following cardiac surgery as the model is outdated.[10,12,14] Thus, efforts in creation of new accurate risk models and attempts to improve accuracy of existing ones have been recently undertaken. Aimed to improve the accuracy of the old model, the new EuroSCORE II emerged.[16] To reach that purpose, the new EuroSCORE was developed from a largest database of patients collected from diverse countries all over the world and included different predictors and a new classification of the previously existing. Based on 23,000 patients from 150 institutions of 43 countries, the updated EuroSCORE II focused on increasing both calibration and discrimination. Calibration refers to the grade of concordance between predicted mortality risk, or probability to die, and the actual observed mortality. Thus, well-calibrated models are those with similar observed and expected event rates. Discrimination is the capacity of risk models to identify probability to die in each particular patient. It defines the score ability to point a difference between postoperative survivors and nonsurvivors.

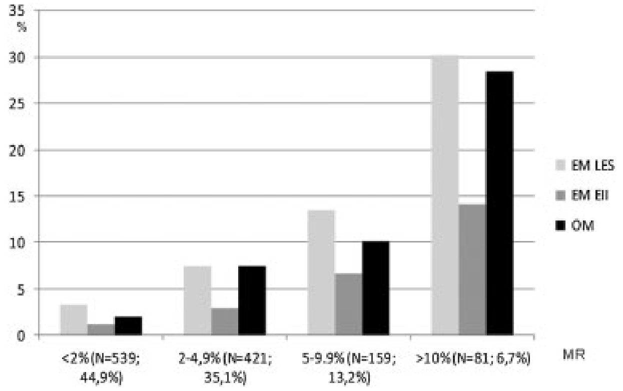

Fig. 2 Observed and expected mortalities for the population stratified by risk groups. Observed mortality (OM), expected mortality by logistic EuroSCORE (EMLES), and expected mortality by EuroSCORE II (EMEII). MR, mortality risk. Low risk, < 2%; medium risk, 2–4.9%; high risk, 5–9.9%; and very high risk, < 10%.

To improve performance of the new score in valvular surgery, changes in risk factors and weight impact of major cardiac procedures performed were incorporated in the new risk scale.[16] However, prognostic factor addition must be done carefully. It is well known how increasing the number of variables increases the possibility of errors because of difficulty in variables interpretation. Thus, prediction risk scales with only a few parameters tend to have enough calibration and to be quite stable. Logistic EuroSCORE can be considered in this way.[3]

Although EuroSCORE II does not significantly increase the number of risk factors to be considered, a more precise definition of several items have been introduced. In addition, the methodology used to develop EuroSCORE II, especially data recruitment and timing, has also been criticized. Some authors argued that EuroSCORE II recruited data in a time period when observed mortality is usually lesser than expected and thus can underestimate the risk of death for patient operated outside this seasonal period.[28,29] This supports that changes in EUROSCORE II are needed[30,31] and a more extensive validation too.[22,32]

Table 3 Results of Hosmer-Lemeshow test applied for logistic EuroSCORE and EuroSCORE II

	No. of patients	χ^2 logistic EuroSCORE	*p*	χ^2 EuroSCORE II	*p*
Overall population	1,200	24.666	0.002	16.501	0.036
Valvular surgery	538	22.837	0.004	10.199	0.257
Coronary surgery	322	5.266	0.733	8.492	0.387
Combined surgery	192	6.023	0.645	4.843	0.774
Miscellaneous surgery	148	5.499	0.703	5.558	0.697

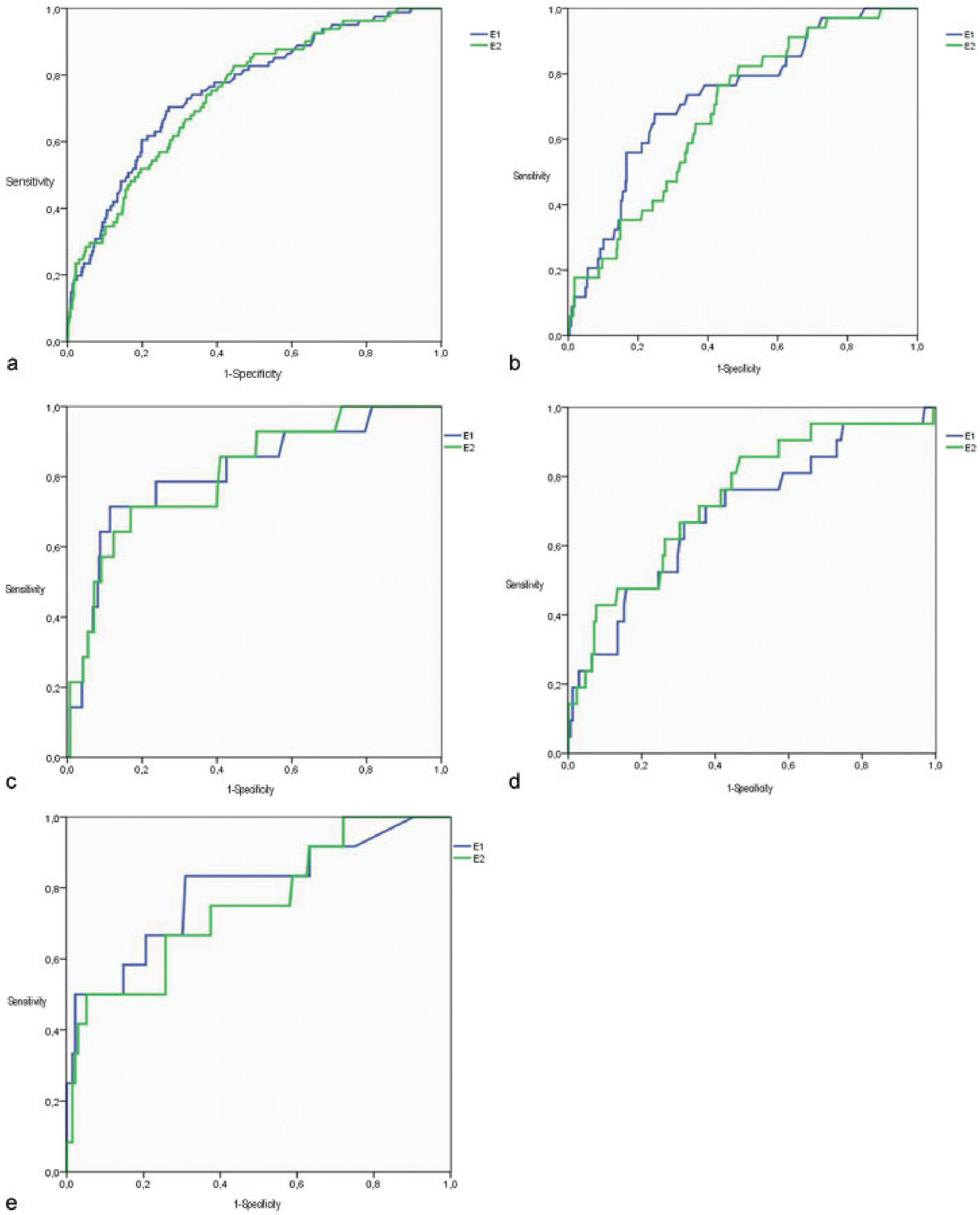

Fig. 3 Discriminative power of logistic EuroSCORE and EuroSCORE II. ROC analysis of logistic EuroSCORE and EuroSCORE II for each of pathology surgical subgroups: (a) overall population, (b) valvular surgery, (c) coronary surgery, (d) combined surgery, and (e) miscellaneous surgery.

Table 4 Discriminative capacity for both risk scales (logistic EuroSCORE and EuroSCORE II)

	AUC LES (CI)	AUC EII (CI)	p
Overall population ($n = 1,200$)	0.758 (0.732–0.782)	0.745 (0.719–0,769)	0.4928
Valvular surgery ($n = 538$)	0.731 (0.691–0.768)	0.688 (0.646–0.727)	0.2536
Coronary surgery ($n = 322$)	0.813 (0.765–0.854)	0.809 (0.761–0.850)	0.8345
Combined surgery ($n = 192$)	0.698 (0.628–0.762)	0.740 (0.672–0.800)	0.2756
Miscellaneous surgery ($n = 148$)	0.793 (0.718–0.854)	0.754 (0.676–0.821)	0.4194

Abbreviations: AUC, area under the curve revealing the discrimination power of logistic EuroSCORE and EuroSCORE II; CI, confidence interval.

In our experience, when calibration was analyzed according to the results of the H-L statistic, EuroSCORE II is better calibrated than logistic EuroSCORE for the overall population and for valvular and combined surgery. On the contrary, the logistic EuroSCORE retains better calibration than the new scale for coronary surgery. Both scales had good and similar calibration for miscellaneous procedures. However, it has been recently suggested that the H-L test is no longer valid to determine calibration, and it should be replaced with the observed by expected mortality ratio.[23]

In terms of discrimination, both scales have a quite good predictive capacity. However, discriminative power of logistic EuroSCORE was better than EuroSCORE II in overall population and in all subgroups except that of combined surgery. Especially for coronary and miscellaneous surgery, an AUC value near or greater than 0.8 for logistic EuroSCORE reflects an excellent discrimination power for the original model. Nevertheless, improvements in definition of surgical complexity of surgery have probably improved the diagnostic accuracy of the EuroSCORE II. In addition, we consider that factors reflecting patient frailty should also be included in the model to improve predictive accuracy especially in the very old patients.

Limitations

The main limitations of our study are the limited sample size of the population, especially in some subgroups, and the fact that this is a single-center study and our patient cohort belongs to a unique institution.

Conclusion

In conclusion, EuroSCORE II represents an update of logistic EuroSCORE and it is considered an acceptable contemporary cardiac surgery risk model. Nonetheless, in contrast with the findings of previous publications, it does not seem to significantly improve the performance of older version when applied to our small population. Although a mortality overestimation of logistic EuroSCORE is patent, we have found that EuroSCORE II underestimates expected mortality in even greater magnitude. Moreover, although the new EuroSCORE II has good discrimination capacity, applied to our population it is worse than that demonstrated by logistic EuroSCORE for the whole population and for most patient subgroups analyzed. A larger validation trial should better allow for evaluation of the predictive capacity of the new scale, especially for patients with high comorbidity.

References

1 Nashef SAM, Roques F, Michel P, Gauducheau E, Lemeshow S, Salamon R. European System for Cardiac Operative Risk Evaluation (EuroSCORE). Eur J Cardiothorac Surg 1999;16(1):9–13

2 Roques F, Nashef SAM, Michel P, et al. Risk factors and outcome in European cardiac surgery: analysis of the EuroSCORE multinational database of 19030 patients. Eur J Cardiothorac Surg 1999;15(6): 816–822, discussion 822–823

3 Roques F, Michel P, Goldstone AR, Nashef SA. The logistic EuroSCORE. Eur Heart J 2003;24(9):881–882

4 Nashef SA, Roques F, Hammill BG, et al; EurpSCORE Project Group. Validation of European System for Cardiac Operative Risk Evaluation (EuroSCORE) in North American cardiac surgery. Eur J Cardiothorac Surg 2002;22(1):101–105

5 Roques F, Nashef SAM, Michel P, Pinna Pintor P, David M, Baudet E; EuroSCORE Study Group. Does EuroSCORE work in individual European countries? Eur J Cardiothorac Surg 2000;18(1):27–30

6 Nilsson J, Algotsson L, Höglund P, Lührs C, Brandt J. Early mortality in coronary bypass surgery: the EuroSCORE versus The Society of Thoracic Surgeons risk algorithm. Ann Thorac Surg 2004;77(4): 1235–1239, discussion 1239–1240

7 Yap CH, Reid C, Yii M, et al. Validation of the EuroSCORE model in Australia. Eur J Cardiothorac Surg 2006;29(4):441–446, discussion 446

8 Ngaage DL. The EuroSCORE has served us well. Eur J Cardiothorac Surg 2010;38(1):114, author reply 114–115

9 Siregar S, Groenwold RHH, de Heer F, Bots ML, van der Graaf Y, van Herwerden LA. Performance of the original EuroSCORE. Eur J Cardiothorac Surg 2012;41(4):746–754

10 Nissinen J, Biancari F, Wistbacka JO, et al. Is it possible to improve the accuracy of EuroSCORE? Eur J Cardiothorac Surg 2009;36(5): 799–804

11 Biancari F, Vasques F, Mikkola R, Martin M, Lahtinen J, Heikkinen J. Validation of EuroSCORE II in patients undergoing coronary artery bypass surgery. Ann Thorac Surg 2012;93(6):1930–1935

12 Basraon J, Chandrashekhar YS, John R, et al. Comparison of risk scores to estimate perioperative mortality in aortic valve replacement surgery. Ann Thorac Surg 2011;92(2):535–540

13 Gogbashian A, Sedrakyan A, Treasure T. EuroSCORE: a systematic review of international performance. Eur J Cardiothorac Surg 2004;25(5):695–700

14 Chalmers J, Pullan M, Fabri B, et al. Validation of EuroSCORE II in a modern cohort of patients undergoing cardiac surgery. Eur J Cardiothorac Surg 2013;43(4):688–694

15 Choong CK, Sergeant P, Nashef SA, Smith JA, Bridgewater B. The EuroSCORE risk stratification system in the current era: how accurate is it and what should be done if it is inaccurate? Eur J Cardiothorac Surg 2009;35:59–61

16 Nashef SA, Roques F, Sharples LD, et al. EuroSCORE II. Eur J Cardiothorac Surg 2012;41(4):734–744, discussion 744–745

17 Carnero-Alcázar M, Silva Guisasola JA, Reguillo Lacruz FJ, et al. Validation of EuroSCORE II on a single-centre 3800 patient cohort. Interact Cardiovasc Thorac Surg 2013;16(3):293–300

18 Collart F, Feier H, Kerbaul F, et al. Valvular surgery in octogenarians: operative risks factors, evaluation of Euroscore and long term results. Eur J Cardiothorac Surg 2005;27(2):276–280

19 Qadir I, Salick MM, Perveen S, Sharif H. Mortality from isolated coronary bypass surgery: a comparison of the Society of Thoracic Surgeons and the EuroSCORE risk prediction algorithms. Interact Cardiovasc Thorac Surg 2012;14(3):258–262

20 Di Dedda U, Pelissero G, Agnelli B, De Vincentiis C, Castelvecchio S, Ranucci M. Accuracy, calibration and clinical performance of the new EuroSCORE II risk stratification system. Eur J Cardiothorac Surg 2013;43(1):27–32

21 Barili F, Pacini D, Capo A, et al. Does EuroSCORE II perform better than its original versions? A multicentre validation study. Eur Heart J 2013;34(1):22–29

22 Grant SW, Hickey GL, Dimarakis I, et al. How does EuroSCORE II perform in UK cardiac surgery; an analysis of 23 740 patients from the Society for Cardiothoracic Surgery in Great Britain and Ireland National Database. Heart 2012;98(21):1568–1572

23 Nezic D, Borzanovic M, Spasic T, Vukovic P. Calibration of the EuroSCORE II risk stratification model: is the Hosmer-Lemeshow test acceptable anymore? Eur J Cardiothorac Surg 2013; 43(1):206

24 Hosmer DW, Lemeshow S. Assessing the Fit of the Model, Applied Logistic Regression, 2nd ed. New Jersey: Wiley-Blackwell; 2012: 143–203

25 Choong CK, Sergeant P, Nashef SA, Smith JA, Bridgewater B. The EuroSCORE risk stratification system in the current era: how accurate is it and what should be done if it is inaccurate? Eur J Cardiothorac Surg 2009;35(1):59–61

26 Sergeant P, Meuris B, Pettinari M. EuroSCORE II, illum qui est gravitates magni observe. Eur J Cardiothorac Surg 2012;41(4): 729–731

27 Hickey GL, Grant SW, Bridgewater B. Validation of the EuroSCORE II: should we be concerned with retrospective performance? Eur J Cardiothorac Surg 2013;43(3):655

28 Hickey GL, Bridgewater B. How well calibrated is EuroSCORE II? Eur J Cardiothorac Surg 2013;43(1):208

29 Poullis M, Fabri B, Pullan M, Chalmers J. Sampling time error in EuroSCORE II. Interact Cardiovasc Thorac Surg 2012;14(5): 640–641

30 Lebreton G, Merle S, Inamo J, et al. Limitations in the inter-observer reliability of EuroSCORE: what should change in EuroSCORE II? Eur J Cardiothorac Surg 2011;40(6):1304–1308

31 Collins GS, Altman DG. Design flaws in EuroSCORE II. Eur J Cardiothorac Surg 2013;43(4):871

32 Nashef SA, Sharples LD, Roques F, Lockowandt U. EuroSCORE II and the art and science of risk modelling. Eur J Cardiothorac Surg 2013; 43(4):695–696

Invited Commentary: The Meaning of Differences between Observed and Expected Scored Outcomes

The increasing availability of data processing computers in the past few decades led to the registration and evaluation of more and more data, also in medicine. In the field of risk scoring, this promoted the replacement of simple addition of risk components by Cox models. A Cox model is a formula that weights the influence of factors for the target event. The formula permits the calculation of estimated relative risks or odds ratios for risk factor combinations using a uniform mathematical function. The combination of observed mortalities of possibly large samples and the associated factor weights permits the calculation of expected death rates (death risk) instead of only risk proportions. But what does a difference between observed and expected event rates mean?

Model construction means simplification. To construct a model, the innumerable circumstances that lead to the target event have to be reduced to a reasonable number of conditions of importance.

Differences between observed and expected outcomes may be caused/explained by the following:

1. Several unconsidered factors of minor importance
2. Single important factors that are not present in the score construction population
3. Considered factors whose weight is different in the score construction population and the score application population

In other words, disregarded or inappropriately weighted factors provoke outcomes diverging from the model's prediction. Such factors might be patient related (more severe manifestation of a condition, or occurrence of a severe comorbidity not considered at all in the applied score) or treatment related (e.g., more or less well performing surgeons, or different surgical policies/instruments, or any other difference in circumstances influencing the outcome). It is not the fault of a score if estimated (expected) and observed results do not coincide—such differences give reason to look for their cause.

Quality assurance is one of the promoters of the increasing use of scores. Better performing units might find out (and publish) their reasons for better results, and underperformers might have a critical look at themselves or the circumstances under which they are supposed to work. If funding cuts induce decreasing performance, scores can show when an acceptable grade of deterioration comes close or gets exceeded.

Sites having at their disposal data of large populations with observed-to-expected scoring differences might consider to calculate their own models to determine which factor weights differ from the factor weights of the applied score, or to introduce site-specific factors helping to better explain their results. Thus, scoring and the comparison of expected and observed outcomes might help improve the treatment of our patients.

Dietmar Boethig, MD
Hannover Medical School, 30625 Hannover, Germany
boethig.dietmar@mh-hannover.de
Editorial Board, Statistical Advisor
The Thoracic and Cardiovascular Surgeon