Validation of EuroSCORE II on a single-centre 3800 patient cohort

Manuel Carnero-Alcázar^{a,*}, Jacobo Alberto Silva Guisasola^a, Fernando José Reguillo Lacruz^a, Luis Carlos Maroto Castellanos^a, Javier Cobiella Carnicer^a, Enrique Villagrán Medinilla^a, Teresa Tejerina Sánchez^b and José Enrique Rodríguez Hernández^a

^a Department of Cardiac Surgery, Hospital Clínico San Carlos, Madrid, Spain

^b Department of Pharmacology, Medicine School, Universidad Complutense de Madrid, Madrid, Spain

* Corresponding author. Department of Cardiac Surgery, Hospital Clínico San Carlos, Plaza Cristo Rey s/n, 28040 Madrid, Spain. Tel: +34-913303691; fax: +34-91-3303453; e-mail: manuelcarneroalcazar@hotmail.es (M. Carnero-Alcázar).

Received 18 June 2012; received in revised form 26 September 2012; accepted 3 October 2012

Abstract

OBJECTIVES: To compare and validate the new European System for Cardiac Operative Risk Evaluation (EuroSCORE) II with EuroSCORE at our institution.

METHODS: The logistic EuroSCORE and EuroSCORE II were calculated on the entire patient cohort undergoing major cardiac surgery at our centre between January 2005 and December 2010. The goodness of fit was compared by means of the Hosmer–Lemeshow (HL) chi-squared test and the area under the curve (AUC) of the receiver operating characteristic curves of both scales applied to the same sample of patients. These analyses were repeated and stratified by the type of surgery.

RESULTS: Mortality of 5.66% was observed, with estimated mortalities according to logistic EuroSCORE and EuroSCORE II of 9 and 4.46%, respectively. The AUC for EuroSCORE (0.82, 95% confidence interval [CI] 0.79–0.85) was lower than that for EuroSCORE II (0.85, 95% CI 0.83–0.87) without the differences being statistically significant ($P = 0.056$). Both scales showed a good discriminative capacity for all the pathologies subgroups. The two scales showed poor calibration in the sample: EuroSCORE ($\chi^2 = 39.3$, $P_{HL} < 0.001$) and EuroSCORE II ($\chi^2 = 86.69$, $P_{HL} < 0.001$). The calibration of EuroSCORE was poor in the groups of patients undergoing coronary ($P_{HL} = 0.01$), valve ($P_{HL} = 0.01$) and combined coronary valve surgery ($P_{HL} = 0.012$); and that of EuroSCORE II in the group of coronary ($P_{HL} = 0.001$) and valve surgery ($P_{HL} < 0.001$) patients.

CONCLUSIONS: EuroSCORE II demonstrated good discriminative capacity and poor calibration in the patients undergoing major cardiac surgery at our centre.

Keywords: Validation • EuroSCORE • EuroSCORE II

INTRODUCTION

The European System for Cardiac Operation Risk Evaluation (EuroSCORE) [1] is a risk model that permits predicting postoperative mortality after major cardiac surgery. This scale was first published in 1999 [2]. Since then, it has been used widely to predict postoperative mortality [3], and as benchmarking of the results in hospitals worldwide. For more than a decade, this scale has been validated in innumerable papers demonstrating excellent discriminative capacity for the different types of cardiac surgery [4]. However, in recent years, different publications have highlighted that this scale could be overestimating the postoperative risk in some subgroups such as octogenarian patients, aortic valve and off-pump coronary surgery [5–7]. This poor calibration can be explained by the technical and technological advances in cardiac surgery, anaesthesiology and perfusion, which have resulted in a decrease in risk-adjusted mortality.

For these reasons, EuroSCORE II [8] was recently published. This new predictive model of postoperative mortality was constructed based on the surgical results observed in more than 22 000 patients

operated on in hospitals all around the world (though most of them are from European countries). It has demonstrated a discriminative capacity similar to EuroSCORE ($AUC_{EuroSCORE II} = 0.81$ vs $AUC_{EuroSCORE} = 0.78$), and good calibration (Hosmer–Lemeshow chi-squared test, χ^2_{HL} [EuroSCORE II] = 15.48; $P = 0.0505$) [8]. As with the original model, it is necessary, at this stage, to verify the external validity of the new scale in order to guarantee its applicability.

The objective of this study was to compare the calibration and discriminative capacity of the postoperative risk estimate by means of the logistic regression models of both scales (EuroSCORE and EuroSCORE II) in patients who have undergone a major cardiac surgery procedure at our centre.

MATERIALS AND METHODS

Population of the study

The entire patient cohort who underwent major cardiac surgery in our centre between January 2005 and December 2010 (both

inclusive) was evaluated ambispectively and consecutively. All the adult subjects who underwent the following procedures were included:

- (i) Isolated coronary surgery: including on-pump and off-pump.
- (ii) Isolated valve surgery: including mitral, aortic, tricuspid and pulmonary surgery, with prosthetic replacements or repairs, through traditional or minimally invasive techniques.
- (iii) Combined valve and coronary surgery.
- (iv) Thoracic aortic surgery: aortic root, tubular ascending aorta, aortic arch or descending thoracic aorta.
- (v) Other procedures: congenital cardiomyopathy in adults, mechanical complications of infarction, surgical ablation of auricular or ventricular arrhythmias, pericardiectomies, tumour surgery, cardiac trauma, etc.

All these types of surgery were included regardless of the priority level (elective/emergency/emergent/salvage). Procedures not listed in the original EuroSCORE II manuscript [8], and transcatheter prosthesis implantations were excluded. All those patients who underwent any other cardiac surgical procedures or those for whom it was not possible to estimate any of the scores of the study (EuroSCORE and EuroSCORE II) were excluded from the study.

Variables and case study events

Pre-, intra- and postoperative information were gathered prospectively from a coded database of more than 200 items. The logistic EuroSCORE [3] was calculated on the entire patient sample prior to the time of surgery. Retrospectively, the logistic EuroSCORE II [8] was calculated on the same cohort of subjects. The calculation of the score of each patient was carried out by at least two surgeons from our centre through the online calculator available at www.euroscore.org.

The discriminative capacity and the goodness of fit of both logistic scales were analysed in the estimation of postoperative mortality (understood as hospital mortality or that occurring in the period of 30 days following surgery).

Afterwards, the analysis was repeated in the subgroups of surgical pathologies defined in the above section.

The patient's signed consent was obtained prior to the operation. Afterwards, the ethics committee of our Institution supplied the consent for the estimate of EuroSCORE II in the same patient cohort.

Statistical analysis

The quantitative variables were expressed by the mean and standard deviation or 95% CI except in distributions that were not normal, in which case, they were summarized by the median and interquartile range. The qualitative variables were expressed in absolute frequencies and percentages. The quantitative variables were compared by Student's *t*-test or the Wilcoxon test (depending on the normality of the distribution). The qualitative variables were compared by means of the chi-squared or Fisher's exact test.

The calibration of both scales was assessed via the use of the Hosmer–Lemeshow (HL) test, which compares the observed vs expected mortality by risk decile. Calibration is considered to be poor if the test is significant. The discrimination measures the

capacity of a model (in this case EuroSCORE and EuroSCORE II) to differentiate between the individuals of a sample that suffer an event (in this case, death) and those that do not. The discriminative capacity of the analysed events of the two scales was estimated by means of receiver operating characteristic (ROC) curves [9]. Their areas under the curve (AUC) were calculated and compared with the statistic *z* according to a normal law [*N* (0,1)] with the method proposed by Hanley and McNeil [10] using the statistical package MEDCALC 12.2.

For the rest of the analysis, the statistical package SPSS® 19.0 (SPSS, Inc., Chicago, IL, USA) for Windows® was used. A *P*-value <0.05 was considered significant.

RESULTS

A total of 4780 cardiac procedures were performed at our institution between January 2005 and December 2010. Only 4342 could be considered for the purpose of this study given the type of procedures for which EuroSCORE II was designed [8]. The value of the two risk scales could only be estimated on 3798 (87.5%) patients. For the 544 remaining patients, it was not possible to estimate the EuroSCORE II value due to the fact that not all the data referring to one or more variables necessary for its calculation had been collected at the time the patient was treated [we lacked left ventricular ejection fraction (LVEF) value in 78 (14.4%) patients, renal impairment in 162 (29.8%), New York Heart Association (NYHA) in 21 (3.9%), pulmonary artery pressures in 290 (53.3%), Poor mobility definition in 179 (32.9%), CCS4 in 91 (16.7%)].

The following procedures were performed: 1231 myocardial revascularizations (32.41%), 1727 isolated valve surgeries (45.47%), 301 combined valve and coronary procedures (10.63%), 416 surgeries of the thoracic aorta (10.95%) and 123 major cardiac surgery procedures not corresponding to any of the previous groups (3.23%).

Distribution of variables of EuroSCORE and EuroSCORE II

Table 1 summarizes the distribution of the different variables of both scales. Nine of the 17 variables of the original EuroSCORE [1] remained in EuroSCORE II [8] (age, sex, peripheral arteriopathy, chronic obstructive pulmonary disease (COPD), re-operation, active endocarditis, critical condition, recent infarction and thoracic aortic surgery). 37.73% of the sample was female. A mean age of 67.03 years (standard deviation, SD 10.15) was observed. Prevalence of some comorbidities were: peripheral arteriopathy (15.22%), COPD (9.48%), re-operation (12.24%) or critical condition (6.95%). The coding of other variables was modified, such as renal function, LVEF, pulmonary hypertension, surgical priority or the weight of the procedure in the new EuroSCORE II. With the new scale, a higher prevalence of patients with nephropathy (72%) was observed in EuroSCORE II compared with EuroSCORE (5.82%) (*P* < 0.001). In the new scale, the coding of the postinfarction septal rupture has disappeared (whose incidence in our sample was 0.16%). Finally, new variables were introduced: Diabetes mellitus in treatment with insulin (prevalence of 8.9%) and NYHA functional classification (with a high prevalence of functional classes III and IV: 33.36 and 5.21%, respectively).

Table 1: Distribution of the different variables of EuroSCORE and EuroSCORE II

EuroSCORE II Preprocedure		EuroSCORE	
Patient-related factors			
Age		67.03 (SD 10.15)	
Female		1433 (37.73%)	
Peripheral arteriopathy		578 (15.22%)	
COPD		360 (9.48%)	
Diabetes on insulin	338 (8.9%)		
Poor mobility	235 (6.19%)	Neurological dysfunction	201 (5.29%)
Renal impairment		Cr > 200 μmol/l	221 (5.82%)
Dialysis	88 (2.32%)		
CC ≤50	812 (21.38%)		
CC >50–85	1834 (48.29%)		
Cardiac-related factors			
Active endocarditis		123 (3.26%)	
Recent AMI		566 (14.9%)	
NYHA class			
II	1147 (30.2%)		
III	1267 (33.36%)		
IV	198 (5.21%)		
CCS4	193 (5.08%)	Unstable angina	681 (17.93%)
LVEF (%)			
>50	2773 (73.01%)	>50	2773 (73.01%)
31–50	544 (16.96%)	31–50	544 (16.96%)
21–30	238 (6.27%)	≤30	381 (10.03%)
≤20	143 (3.77%)		
Pulmonary artery pressure			
31–55 mmHg	659 (17.35%)	Pulmonary artery pressure ≥60 mmHg	418 (11.01%)
≥55 mmHg	427 (11.24%)		
Procedure			
Critical condition		264 (6.95%)	
Re-operation		465 (12.24%)	
Thoracic aorta		416 (10.95%)	
Emergency		Emergency	176 (4.63%)
Emergency	159 (4.18%)		
Emergent	176 (4.63%)		
Salvage	46 (1.21%)		
Weight of the procedure			
1 No CABG	1450 (38.18%)	Surgery other than isolated CABG	2567 (67.6%)
2	918 (24.17%)		
3+	157 (4.14%)		
		VSD post AMI	6 (0.16%)

Data expressed in *n* (%) and mean (SD). The definition of each of the items is taken as those previously published for EuroSCORE [1] and EuroSCORE II [8]. NYHA: New York Heart Association; CC: creatinine clearance; CCS4: angina at rest; COPD: chronic obstructive pulmonary disease; LVEF: left ventricle ejection fraction; AMI: acute myocardial infarction; VSD: ventricular septal defect.

Observed, estimated and adjusted mortalities

The global, estimated and adjusted mortalities for the entire patient cohort and pathology subgroups are summarized in Table 2 and Fig. 1. The mortality observed in our sample of patients was 5.66%. The mean mortalities predicted by EuroSCORE and EuroSCORE II were 9% (95% CI 8.67–9.33) and 4.46% (95% CI 4.2–4.61), respectively ($P < 0.001$). The risk-adjusted mortality index (RAMI) (observed/predicted mortality) was 0.63 and 1.27 for EuroSCORE and EuroSCORE II, respectively. A statistically significant decrease was observed in the mortality expected by EuroSCORE II with respect to EuroSCORE for all the pathology subgroups (Table 1): 3.28 vs 5.86% ($P < 0.001$) in the coronary surgery group; 4.42 vs 9.11% ($P < 0.001$) in the valve surgery subgroup; 6.58 vs 16.04% ($P < 0.001$) in thoracic aortic pathology, 5.83 vs 9.9% ($P = 0.003$) in the mixed surgery subgroup and 5.82 vs 12.74%

($P < 0.001$) in other major cardiac operations. This decrease in estimated mortality resulted in an increase of the RAMI_{EuroSCORE II}, so that in the coronary pathology subgroup it went from 0.53 to 0.94; in the isolated valve, aortic, mixed and other types of major cardiac surgery it went from <1 to indices between 1.24 and 1.82.

Receiver operating characteristic curves and goodness of fit

The two scales showed good discriminative capacity in the global patient sample, with the AUC being higher for EuroSCORE II (AUC 0.85, 95% CI 0.83–0.87) compared with EuroSCORE (AUC 0.82, 95% CI 0.79–0.85), although the difference was not statistically significant ($P = 0.056$; Fig. 2, Table 3). The goodness of fit was poor for both scales, with that

Table 2: Observed, estimated and adjusted mortalities for the overall cohort and subgroups of surgical pathologies

	<i>n</i> (%)	Observed mortality (% of the subgroup)	Mean ES (%) (95% CI)	Mean ESII (95% CI)	<i>P</i> -value	RAMI (ES)	RAMI (ESII)
Global	3798	215 (5.66)	9 (8.67–9.33)	4.46 (4.25–4.67)	<0.001	0.63	1.27
Coronary	1231 (32.41)	38 (3.09)	5.86 (5.43–6.3)	3.28 (2.97–3.59)	<0.001	0.53	0.94
Valve	1727 (45.47)	106 (6.14)	9.11 (8.65–9.58)	4.42 (4.11–4.72)	<0.001	0.67	1.39
Mixed	301 (7.92)	24 (7.97)	9.9 (8.75–11.04)	5.83 (5.09–6.56)	0.003	0.80	1.37
Aorta	416 (10.95)	34 (8.17)	16.04 (14.82–17.27)	6.58 (5.81–7.35)	<0.001	0.51	1.24
Others	123 (3.23)	13 (10.57)	12.74 (9.5–15.98)	5.82 (5.09–7.42)	<0.001	0.83	1.82

Observed mortality, predicted by EuroSCORE and EuroSCORE II and adjusted (quotient between observed and predicted mortalities for each scale). ES, proportion comparison: EuroSCORE; ESII: EuroSCORE II; 95% CI: confidence interval of 95%; RAMI: risk-adjusted mortality index; *P*: proportion comparison.

A *P*-value <0.05 is considered significant.

Figure 1: Observed, EuroSCORE and EuroSCORE II expected mortality.

corresponding to EuroSCORE II being worse: EuroSCORE ($\chi^2 = 39.3$, $df = 8$, $P_{HL} = <0.001$) and EuroSCORE II ($\chi^2 = 86.69$, $df = 8$, $P_{HL} < 0.001$; Table 3).

Table 3 shows the comparison of the AUC of the ROC curves and the χ^2_{HL} test for the pathology subgroups. EuroSCORE II showed greater discriminative capacity in all pathology subgroups except in that of the combined valve and coronary surgery. However, no difference was statistically significant ($P > 0.05$ in the Hanley and McNeil test for ROC curves comparison). Lower values of the χ^2_{HL} test of EuroSCORE were observed when compared with those of EuroSCORE II for all the subgroups of patients except in the group of mixed surgery, which indicates higher calibration of EuroSCORE for surgical pathologies. The calibration of EuroSCORE was particularly low in the groups of coronary ($P = 0.01$), valve ($P = 0.021$) and combined ($P = 0.029$) procedures. For EuroSCORE II, the calibration was very low in a statistically significant manner in the coronary ($P = 0.001$) and valve ($P < 0.001$) pathology groups. In the group of patients who underwent aortic surgery, the discrimination of EuroSCORE II was poor, without reaching a statistically significant level ($P = 0.058$). The calibration of EuroSCORE II was good in the set of other major surgical procedures and in the mixed pathology.

Figure 3a–e represents the ROC curves for the different groups of surgical pathologies. It was observed that all the lower limits of the 95% CI of the AUC of the ROC curves of the two scales

Figure 2: ROC curves for the global sample. *P* is the probability for $z \geq z_i$ [2]. $P < 0.05$ is considered significant.

for the different pathology groups and for the overall sample exceeded 0.7, which indicates good discriminative capacity. A higher discriminative capacity was observed of EuroSCORE II for the global sample and all the surgical groups, except for the mixed group, without these differences reaching a statistically significant level (Table 3).

DISCUSSION

EuroSCORE II has been recently published. The development of this new scale is due to many deficiencies observed over the years in the application of EuroSCORE to different patient samples such as: low prevalence of octogenarians (<2%) or valve surgery (<30%) in the cohort for which it was estimated, analysis of the impact of renal function in the estimation of mortality, decrease in the calibration as the results of cardiac surgery improved, etc. EuroSCORE II was calculated on a consecutive subcohort of 16 828 patients, and its validity estimated in another subcohort of 5553 subjects [8]. In the last, EuroSCORE II was capable of predicting hospital mortality after major cardiac surgery with an excellent discriminative capacity (AUC 0.81, 95% CI 0.78–0.83) [8]. The original EuroSCORE, applied to the same

Table 3: Discriminative capacity (AUC) and calibration of EuroSCORE and EuroSCORE II

	Discrimination (ROC curves and AUC)				Goodness of fit (HL)			
	AUC (95% CI)		Hanley and McNeil test (ROC difference)		ES		ESII	
	ES	ESII	z	P-value	χ^2	P-value	χ^2	P-value
Global	0.818 (0.791–0.846)	0.851 (0.827–0.874)	1.586	0.056	39.3	<0.001	86.69	<0.001
Coronary	0.884 (0.834–0.934)	0.9 (0.866–0.934)	0.529	0.298	20.1	0.01	26.58	0.001
Valve	0.779 (0.734–0.823)	0.827 (0.788–0.865)	1.575	0.059	19.67	0.012	50.43	<0.001
Mixed	0.786 (0.701–0.872)	0.769 (0.677–0.861)	–0.267	0.397	17.13	0.029	9.1	0.334
Aorta	0.813 (0.738–0.888)	0.85 (0.792–0.908)	0.764	0.224	8.5	0.29	15.03	0.058
Others	0.835 (0.7–0.97)	0.876 (0.77–0.982)	0.468	0.319	4.05	0.775	9.1	0.334

ES: EuroSCORE; ESII: EuroSCORE II; AUC: area under the curve; ROC: receiver operating characteristics.

The AUC are compared with the Hanley and McNeil test. The values z and the value of P corresponding to $P(z \geq z_i)$ are shown. Goodness of fit of the models: it is verified by the χ^2_{HL} test. A P -value <0.05 is considered significant.

validation cohort of 5553 subjects [8], also showed good discriminative capacity with an AUC of 0.78. In this study, with 3798 patients, we observed that both scales have good discriminative capacity with AUC of 0.82 (95% CI 0.79–0.85) and 0.85 (0.83–0.87) ($P=0.056$) for EuroSCORE and EuroSCORE II, respectively, with the second version being superior. These results coincide with those described by Nashef *et al.* [8].

By applying both logistic models on the surgical operation types (Tables 2 and 3), no statistically significant differences were observed comparing $AUC_{EuroSCORE}$ with $AUC_{EuroSCORE II}$. The worst discrimination of both models was observed in the patients with combined valve and coronary surgery, with AUC of 0.78 and 0.76 for EuroSCORE and EuroSCORE II, respectively ($P=0.397$). In the coronary surgery group (32.4% of the patients), the two scales discriminated quite precisely: $AUC_{EuroSCORE} = 0.88$ and $AUC_{EuroSCORE II} = 0.9$ ($P=0.298$). These AUC are comparable with those obtained by applying the two logistic models to the sample of 5553 patients of the paper by Nashef *et al.* [8]. In the valve surgery group, which represents 45.5% of the patients of our sample, the discrimination capacity ($AUC_{EuroSCORE} = 0.78$ and $AUC_{EuroSCORE II} = 0.83$) was less than that found in the global sample, in coronary surgery and in that described previously by Nashef *et al.* [8]. The AUC of EuroSCORE and EuroSCORE II for the global sample of this paper were similar to that found in a recent meta-analysis published by Siregar *et al.* [4]. In the last, which analysed the goodness of fit and discrimination of the EuroSCORE in more than 400 000 patients, the observed AUC varied between 0.7 and 0.8. In short, according to the results of this study and the study and analysis of Nashef *et al.* [8], EuroSCORE already showed excellent discriminative capacity that has been slightly improved with the new EuroSCORE II version.

The discrimination of the new EuroSCORE II improved in the subgroup of CABG, valve, aorta and other procedures, but not in the combined surgery subtype (Table 3). The good discriminative capacity of the new system in the patients who underwent isolated coronary surgery was already shown recently in a study by Banciari *et al.* [11] with 1027 subjects who received surgical myocardial revascularization, where EuroSCORE II obtained an AUC of 0.85 ($P=0.031$). When comparing AUCs stratified by the type of procedure, no statistically significant difference was detected.

In summary, we have observed that EuroSCORE II has improved its discrimination capacity in the overall sample and subtypes of surgery (but in the mixed valve and graft group). On

the other hand, this improvement seems to be very subtle (the greatest absolute AUC difference [corresponding to valve procedures] being only 0.048).

Many prior studies have detected an overestimation of hospital mortality when applying EuroSCORE to different patient subgroups [4–6, 12]. For example, in a meta-analysis published by Parolari *et al.* [12] in 26 621 patients who underwent valve surgery, the RAMI ranked between 0.45 and 0.89. In the meta-analysis of Siregar *et al.* [4], EuroSCORE overestimated the mortality in patients who underwent coronary, valve and combined surgery, with RAMI between 0.43 and 0.62. When the logistic EuroSCORE was applied to 5553 patients of the subcohort of the EuroSCORE II validation [8], an expected mortality of 7.57% was observed when compared with an observed mortality of 3.9%. The poor calibration of the original EuroSCORE was the principal motive that drove the development of a new model [8]. However, upon analysing the calibration of the new model in the validation sample of EuroSCORE II [8], it was observed that, despite there not being a large discrepancy between the observed (3.95%) and the estimated mortalities (4.18%; difference in less than 10%), the HL test was nearly significant ($\chi^2 = 15.48$, $df=0$, $P=0.051$), which reveals differences (without attaining statistical significance) between the estimated and observed mortalities.

In our sample, EuroSCORE overestimated, and EuroSCORE II underestimated, mortality (Fig. 1 and Table 2). There was an observed mortality of 5.66% and expected/RAMI mortalities of 9%/0.63 and 4.46%/1.27 for EuroSCORE and EuroSCORE II, respectively. On the other hand, both scales demonstrated poor calibration (with significant results in the HL test), EuroSCORE II being worse than EuroSCORE. (EuroSCORE: $\chi^2 = 39.3$, $df=8$, $P_{HL} < 0.001$ and EuroSCORE II: $\chi^2 = 86.69$, $df=8$, $P_{HL} < 0.001$; Table 3). Although the significance of the tests could be explained by excessively large sample sizes, it seems clear that, in our population, the new EuroSCORE II showed a worse calibration than EuroSCORE.

While EuroSCORE predicted higher mortality than that observed in all the subtypes of procedures (Table 2), EuroSCORE II estimated mortality below that obtained except in the subjects who underwent coronary surgery where the adjusted mortality was close to 1 (0.94).

Upon analysing the calibration on the types of surgical pathologies (Table 3; with smaller n and HL tests less influenced by sample size), EuroSCORE showed lower χ^2_{HL} in all the subgroups, indicating worse calibration when applying EuroSCORE II (except for mixed

Figure 3: ROC curves for pathology subgroups: (a) coronary surgery; (b) valve surgery; (c) mixed surgery; (d) aortic surgery; (e) other major cardiac surgery. P is the probability for $z \geq z_1$ [2]. $P < 0.05$ is considered significant.

group). The HL test showed statistically significant differences between the observed and expected mortalities by applying both scales to the valve and isolated coronary pathology subgroups. EuroSCORE II estimated a significantly lower mortality than that observed in the patients subjected to coronary and valve surgery, which represented 77.88% of the patient sample (Tables 2 and 3).

In brief, EuroSCORE overestimated, while EuroSCORE II underestimated, mortality. Both scales have poor calibration, the second being apparently worse than the previous.

We observed differences between our patient sample and the sample of subjects on which EuroSCORE II [8] was designed that could explain the poor calibration of this scale in our sample. Namely, higher mean age (67.03 vs 64.6 years), greater prevalence of females (37.3 vs 30.9%), diabetes mellitus (8.9 vs 7.6%), neurological dysfunction (5.29 vs 3.2%), serum Cr >200 µmol/l (5.82 vs 2.6%), dialysis (2.3 vs 1.1%), active endocarditis (3.2 vs 2.2%), critical condition (6.95 vs 4.1%), lesser proportion of isolated coronary surgery (32.4 vs 46.7%), etc., which resulted in a mean mortality estimated by the logistic EuroSCORE II of 4.46 compared with 3.95% of the validation sample of 5553 patients of EuroSCORE II [8]. The distribution of the variables of EuroSCORE II in our sample differs widely from that used in the work on the design of the EuroSCORE II [8]. This could explain, to a significant extent, the poor calibration observed in our sample.

Another difference that could explain the limited validity of the new scale in our patient sample is that EuroSCORE II was only validated for the prediction of hospital mortality [8], while in this study, the case study event included hospital mortality as well as that occurring up to 30 days after the operation. In fact, in the paper of Nashef *et al.* referring to the design of the new system, the data referring to hospital mortality were gathered on nearly 100% of the subjects, while the percentage of missing values at 1 month in the follow-up was 43.4% [8]. Furthermore, in the patient sample of the present manuscript, subjects operated on since 2004 were included; and EuroSCORE II can lose external validity if applied to patients who underwent cardiac surgery before 2010 (EuroSCORE II was estimated from patients operated on between May and July 2010) [8].

Despite no significant differences existing between the estimated (3.95%) and observed (4.18%) mortalities in the validation cohort of EuroSCORE II in the original work by Nashef *et al.* [8], the result of the HL test was nearly significant ($P = 0.0505$). The problems in the calibration of the new model observed in our patient sample and in the one designed for the scale [8] can be due to various reasons:

- (i) Descriptive-predictive defects of the model. The absence of interaction variables (in the pursuit of parsimonious models), the classification of continuous variables, the inclusion of factors such as pulmonary hypertension or ventricular function in the model regardless of the type of surgery, etc. could have generated instabilities of the model in certain population subgroups [13, 14].
- (ii) Inclusion bias of the patients and selection of the centres: there is the possibility that the centres that voluntarily decided to participate have better results than those that did not participate [8]. Future validations of the model in participating and nonparticipating centres will be necessary to verify definitively its discrimination and calibration.
- (iii) The coding of certain variables: e.g. the greater weight in mortality of creatinine clearance under 50 ml/min ($b = 0.8592256$) compared with dialysis ($b = 0.6421508$) is

striking, which can be explained by the low prevalence of patients on dialysis (1.1%) in the paper by Nashef *et al.* [8]. Furthermore, the renal function was estimated by the creatinine clearance calculated by the Cockcroft-Gault formula, which signifies an important calibration improvement with respect to the previous scale [1, 15]. However, it has been shown that there are better estimators of the renal function such as the clearance calculated with the Modification of Diet in Renal Disease formula [15].

- (iv) Interobserver variability: a recent study [16] showed the existence of interobserver discrepancies in the calculation of EuroSCORE I in 26.3% of 1719 patients. In this study, it was shown that the majority of the variability was due to the score of 5 of the 17 variables. EuroSCORE II includes 1 variable and 9 more categories, which could further increase the risk of higher rates of interobserver discrepancies.
- (v) Seasonal effect: recently, a study showed how during the time in which the data were reported of EuroSCORE II, lower mortality was recorded than that of the rest of the year [17].

The new EuroSCORE II is a necessary update of the surgical risk-prediction model most extensively used worldwide: EuroSCORE. The included modifications have increased its calibration and maintained a fairly good discrimination with respect to the previous scale in evaluating the risk of the patients undergoing major cardiac surgery nowadays [8]. The results of this study show that, despite having very good discriminative capacity, the goodness of fit of EuroSCORE II is worse than that of its predecessor. It is necessary to apply this new model in larger patient samples and in many centres in order to investigate its external validity with greater precision.

LIMITATIONS

This study was conducted in a single centre. The patient cohort included subjects operated on more than 5 years ago, where the validation of EuroSCORE II could be questioned. We do not know that the impact of the subjects lost in the calibration and discrimination of the model.

Funding

This work was supported by Fondo de Investigaciones Sanitarias FIS PI080920 (Health Research Fund from the Spanish Ministry of Health) and by Red Temática de Investigación Cardiovascular RECAVA RD/06/0014/1007 (Instituto de Salud Carlos III, Spanish Ministry of Health). The funders had no role in the study design, data collection and analysis, decision to publish or preparation of the manuscript.

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 19030 patients. *Eur J Cardiothorac Surg* 1999;15:816–23.
- [3] Roques F, Michel P, Goldstone A, Nashef SAM. The logistic EuroSCORE. *Eur Heart J* 2003;24:881–2.
- [4] 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;41:746–54.
- [5] Basraon J, Chandrashekar 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.
- [6] 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.
- [7] Yap CH, Reid C, Yui M, Rowland MA, Mohajeri M, Skillington PD *et al.* Validation of the EuroSCORE model in Australia. *Eur J Cardiothorac Surg* 2006;29:441–6.
- [8] Nashef SAM, Roques F, Sharples L, Nilsson J, Smith C, Goldstone AR *et al.* EuroSCORE II. *Eur J Cardiothorac Surg* 2012;41:1–12.
- [9] Royston P, Moons KG, Altman DG, Vergouwe Y. Prognosis and prognostic research: developing a prognostic model. *BMJ* 2009;338:b604.
- [10] Hanley JA, McNeil BJ. A method of comparing the areas under receiver operating characteristic curves derived from the same cases. *Radiology* 1983;148:839–43.
- [11] Banciai F, Vasques F, Mikkola R, Martin M, Lahtinen J, Heikkien J. Validation of EuroSCORE II in patients undergoing coronary artery bypass surgery. *Ann Thorac Surg* 2012;93:1930–5.
- [12] Parolari A, Pesce LL, Trezzi M, Cavallotti L, Kassem S, Loardi C *et al.* EuroSCORE performance in valve surgery: a meta-analysis. *Ann Thorac Surg* 2010;89:787–93.
- [13] Miettinen OS. Confounding and effect-modification. *Am J Epidemiol* 1974;100:350–3.
- [14] Sergeant P, Meuris B, Pettinari M. EuroSCORE II, illum qui est gravitates magni observe. *Eur J Cardiothorac Surg* 2012;41:729–31.
- [15] 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.
- [16] Lebreton G, Merle S, Inamo J, Hennequin JL, Sanchez B, Rilos Z *et al.* Limitation in the inter-observer reliability of EuroSCORE: what should change in EuroSCORE II? *Eur J Cardiothorac Surg* 2011;40:1304–8.
- [17] Poullis M, Fabri B, Pullan M, Chalmers J. Sampling time error in EuroSCORE II. *Interact CardioVasc Thorac Surg* 2012;14:640–1.