Online version at http://www.minervamedica.it

Minerva Anestesiologica 2020 July;86(7):719-26
DOI: 10.23736/S0375-9393.20.14280-9

ORIGINAL ARTICLE

Case-mix affects calibration of cardiosurgical severity scores

Anna ZAMPERONI [1], Carlotta ROSSI [2], Stefano FINAZZI [2] *, Paolo DEL SARTO [3], Matteo MONDINI [2], Giovanni NATTINO [4], Daniele POOLE [5], Guido BERTOLINI [2], Cardiac surgical intensive care writing committee (GiViTI) ‡

[1]Cà Foncello Hospital, Aulss2, Treviso, Italy; [2]IRCCS Mario Negri Institute for Pharmacological Research, Villa Camozzi, Ranica, Bergamo, Italy; [3]Department of Critical Care, Fondazione Toscana G. Monasterio, G. Pasquinucci Heart Hospital, Massa, Italy; [4]Division of Biostatistics, College of Public Health, The Ohio State University, Columbus, OH, USA; [5]Anesthesia and Intensive Care Operative Unit, San Martino Hospital, Belluno, Italy

‡Members are listed at the end of the paper

*Corresponding author: Stefano Finazzi, IRCCS Mario Negri Institute for Pharmacological Research, Laboratory of Clinical Epidemiology, Villa Camozzi, Via GB Camozzi 3, 24015 Ranica, Bergamo, Italy. E-mail: stefano.finazzi@marionegri.it

ABSTRACT

BACKGROUND: Prognostic models are often used to assess the quality of healthcare. Several scores were developed to predict mortality after cardiac surgery, but none has reached optimal performance in subsequent validations. We validate the most used scores (EUROSCORE I and II, STS, and ACEF) on a cohort of cardiac-surgery patients, assessing their robustness against case-mix changes.

METHODS: The scores were validated on 14,559 patients admitted to 16 Italian cardiosurgical ICUs participating to Margherita-Prosafe project in 2014 and 2015. Calibration was assessed through Hosmer-Lemeshow Test, standardized mortality ratio, and GiViTI calibration test and belt. Discrimination was measured by the area under the ROC curve.

RESULTS: The study included 10,317 patients who were eligible to the calculation of the STS Score (4156 isolated valve, 4681 isolated CABG and 1480 single valve and CABG) which calibrated well in these subgroups. The ACEF Score and EUROSCORE I and II were available for 14,139, and 14,071 patients, respectively. EUROSCORE I significantly overestimated mortality; EUROSCORE II calibrated well overall, but underestimated mortality of patients undergoing complex surgery and non-elective ones. The ACEF Score calibrated poorly in elective and non-elective patients. Discrimination was acceptable for all models (AUC>0.70), but not for the ACEF Score.

CONCLUSIONS: Cardiac surgery scores calibrate poorly when the case-mix of validation and development samples differs. To grant reliability for benchmarking, they should be validated in the clinical settings on which they are applied and updated periodically. Advanced statistical tools are essential for the correct interpretation and application of severity scores.

(*Cite this article as:* Zamperoni A, Rossi C, Finazzi S, Del Sarto P, Mondini M, Nattino G, *et al.*; GiViTI. Case-mix affects calibration of cardiosurgical severity scores. Minerva Anestesiol 2020;86:719-26. DOI: 10.23736/S0375-9393.20.14280-9)

KEY WORDS: Cardiac surgical procedures; Calibration; Anesthesia.

Prognostic models are used to assess the quality of healthcare provision in single centers, providing a standard to compare with. The measurement of how well a single center performs against a standard is called benchmarking, which is considered the first step to detect weak points in the healthcare delivery process and to monitor the effectiveness of corrective interventions.

Several prognostic models have been developed in the last decades to predict mortality after cardiac surgery. The most widespread scores are the European System for Cardiac Operative Risk Evaluation (EUROSCORE) I and II,[1-3] the Society

of Thoracic Surgeons (STS) Score[4-7] and the Age, Creatinine, and Ejection Fraction (ACEF) Score.[8]

Several validation studies of cardiac-surgery severity scores have been conducted in different countries and in different times, providing heterogeneous results in terms of discrimination and calibration.[9-15] Discrimination is the ability of a score to correctly differentiate patients who die from those who survive, while calibration is defined as the ability of the model to correctly estimate the probability of the event (*e.g.* mortality). The latter is usually assessed with the Hosmer-Lemeshow statistics and traditional calibration plots.[16] This approach, however, does not allow to identify the classes of risk in which the model significantly miscalibrates, hampering the possibility to systematically investigate the reasons of miscalibration and to provide clear indications on the use of the model in specific settings.[17]

The aim of our study was to validate the most commonly used severity scores on a cohort of cardiac-surgery patients admitted to Italian intensive care units (ICUs), assessing their reliability in the face of case-mix changes, using an advanced statistical approach that overcomes the limits of the Hosmer-Lemeshow statistics and the traditional calibration plots.[18, 19]

Materials and methods

Study design and population

The Margherita-Prosafe project has been approved by local Ethical Committees of participating hospitals. No informed consent was requested according to current regulation given the observational nature of the study.

All patients aged more than 16 admitted to GiViTI cardiosurgical ICUs in 2014 or 2015 after cardiac surgery were considered eligible for the analysis. In case of readmissions, only the first admission to the ICU was considered.

Data collection

Clinical information was collected by means of a software developed by the GiViTI Coordination Centre. Collected information consists in demographics, comorbidities, clinical condition and failures at the ICU admission, relevant details concerning the cardiosurgical interventions, main procedures, complications during the stay and both ICU and hospital outcome, as well as all the information required to calculate the EUROSCORE I and II, STS and ACEF scores. The STS Score was calculated on three subsets of patients on which the score was developed: isolated valve, isolated coronary artery bypass graft (CABG), and single valve and CABG. The e-CRF is reported in the supplementary material.

Data validity

Data validity was assessed at different stages to avoid selection biases and input error and to guarantee internal consistency of the records. We excluded all patients admitted in months where more than 10% of admitted patients had had incomplete records.

Outcome

Hospital mortality is the outcome of the study. In case of patients transferred to other ICUs, we considered the outcome of the last hospital of admission.

Statistical analysis

Categorical variables are reported as frequency and percentage, continuous variables as mean and standard deviation (SD) or as median and interquartile range (IQR), as appropriate. Comparisons among categorical variables were performed with χ^2 or Fisher's Exact tests, while differences in continuous variables were tested with the *t*-test or Wilcoxon Test. P values of 0.05 or less were considered as significant.

The EUROSCORE I, EUROSCORE II, STS and ACEF scores performance was assessed in terms of calibration and discrimination on all eligible patients. Calibration was evaluated through three approaches: Hosmer-Lemeshow C Test, overall standardized mortality ratio (SMR), and the GiViTI calibration test and belt.[16, 18, 20] The latter approach allows to evaluate the calibration of prognostic models by assessing the reliability of estimates as a continuous function of expected probability. In particular, the calibration belt represents the relationship between observed and expected risk, provided with the appropriate confidence region. Statistically significant deviations from the 95% confidence band (and, when

increasing sensitivity was regarded as relevant, also from the 80% confidence band) were further investigated with subgroup analyses. The GiViTI calibration test was used to summarize the information conveyed by the calibration belt.

Discrimination was investigated by measuring the Area Under the Curve (AUC) in the Receiver Operator Characteristics (ROC) analysis.[21]

Results

A total of 14,559 patients undergoing cardiac surgery were admitted during 2014 and 2015 in the 16 ICUs participating in the GiViTI Cardiosurgical project. We excluded 298 patients admitted in months with low quality data. After excluding patients aged 16 or less, 14,155 patients (97.2%) were considered for the analysis. Among them, ACEF and EUROSCORE I/II were available for 14,139 and 14,071 patients, respectively. Patients eligible for the calculation of the STS Score were 10,317 (4156 isolated valve, 4681 isolated CABG, and 1480 single valve and CABG).

Patients' demographics, their preoperative characteristics, the type of intervention, and their outcome, in terms of length of stay and mortality are described in Table I.

Discrimination was similar for all models (Table II), with an AUC always above 0.70 except for the ACEF Score. EUROSCORE II had the largest AUC (0.77, 95% CI: 0.75-0.79).

STS models calibration

The three STS models were applied only on patients satisfying the inclusion criteria of the respective development sample (isolated CABG surgery, isolated valve surgery, CABG and valve surgery). SMRs, calibration tests (Table II) and belts (Figure 1) confirmed that STS calibrates well in the three groups.

EUROSCORE I calibration

EUROSCORE I significantly overestimated mortality (SMR=0.48; 95% CI: 0.43-0.54; GiViTI calibration test and H-L P value <0.001, Table II) and it did so in the whole range of risk (the GiViTI calibration belt always lies under the bisector, Figure 2A).

EUROSCORE II calibration

EUROSCORE II seemed to calibrate acceptably (Figure 2B; GiViTI calibration test P value =0.11, H-L P value =0.26, Table II). However, the overall predicted mortality was smaller than observed mortality (SMR=1.09; 95% CI: 1.01-1.18). The discordance between the two tests (the latter statistically significant, the former barely not significant) was due to the different approximations in the calculation of the standard errors. This signal was not ignored and we further investigated the sample with the calibration belt using the 80% confidence band to improve sensitivity. This analysis showed that the model underpredicted mortality for patients with lower risk of death (expected mortality ranging between 3 and 10%, Figure 2B). To spot the patients' category for which the score miscalibrated, we proceeded investigating clinically meaningful subgroups.

First, we stratified patients according to whether the surgery was elective or non-elective. Figure 2D shows that mortality was significantly underestimated in non-elective patients (calibration tests P value <0.001) with a risk of death up to 34%. In Figure 2C, conversely, the EUROSCORE II was shown to slightly overestimate mortality in elective patients, even if not significantly (calibration tests P value =0.079).

Second, we defined subgroups according to the STS classes. Figure 3 reports on the good calibration of EUROSCORE II in isolated CABG, isolated valve, and CABG coupled with valve surgery. However, in patients excluded from the STS development cohort, such as those undergoing double valve surgery, surgery of thoracic aorta, or other complex surgery, a significant underestimation of mortality (calibration tests P value <0.001) was detected for risk of death 14% or less.

Finally, we combined the two characteristics obtaining a subset of patients undergoing non-elective complex procedures for which the model largely underestimated mortality and a complementary subgroup for which EUROSCORE II calibrated correctly (Figure 4).

ACEF Score calibration

The ACEF Score did not calibrate overall (Figure 5; SMR =1.16, 95% CI: 1.07-1.24, GiViTI

TABLE I.—*Patients demographics, preoperative characteristics, and outcomes.*

	CS patients	Isolated valve surgery	Isolated CABG surgery	CABG and valve surgery
N. (%)	14,155	4156 (29.4)	4681 (33.1)	1480 (10.5)
Age-median (IQR)	70 (62-76)	71 (62-77)	69 (62-75)	74 (68-79)
Gender-N. (%)				
Male	9528 (67.3)	2239 (53.9)	3895 (82.4)	1036 (70.0)
Female	4627 (32.7)	1917 (46.1)	822 (17.6)	444 (30.0)
BMI-N (%)				
Underweight	503 (3.6)	177 (4.3)	98 (2.1)	49 (3.3)
Normal	6247 (44.5)	1922 (46.2)	1984 (42.4)	623 (42.1)
Overweight	5197 (37.0)	1426 (34.3)	1874 (40.0)	581 (39.3)
Obese	2099 (14.9)	631 (15.2)	725 (15.5)	227 (15.3)
Comorbidities-N. (%)				
Hypertension	10,212 (72.1)	2780 (66.9)	3690 (78.8)	1187 (80.2)
NYHA---				
II-III	5158 (36.4)	1825 (43.9)	1227 (26.2)	652 (44.1)
IV	470 (3.3)	106 (2.6)	102 (2.2)	44 (3.0)
Left main disease	3845 (27.2)	132 (3.2)	2658 (56.8)	674 (45.5)
Type I diabetes	133 (0.9)	30 (0.7)	64 (1.4)	20 (1.4)
Type II diabetes				
without insulin treatment	1891 (13.4)	408 (9.8)	905 (19.3)	261 (17.6)
with insulin treatment	959 (6.8)	159 (3.8)	521 (11.1)	129 (8.7)
Pulmonary hypertension				
Moderate (PASP: 31-55 mmHg)	962 (6.8)	368 (8.9)	72 (1.5)	94 (6.4)
Severe (PASP>55 mmHg)	400 (2.8)	111 (2.7)	6 (0.1)	26 (1.8)
Extracardiac arteriopathy	1181 (8.4)	173 (4.2)	569 (12.2)	182 (12.3)
Chronic pulmonary disfunction	782 (5.5)	216 (5.2)	225 (4.8)	95 (6.4)
Cerebrovascular disease	813 (5.7)	190 (4.6)	302 (6.5)	114 (7.7)
Active endocarditis	272 (1.9)	158 (3.8)	0 (0.0)	6 (0.4)
N/M mob	129 (0.9)	29 (0.7)	32 (0.7)	10 (0.7)
Myocardial infarction	1660 (11.7)	52 (1.3)	1283 (27.4)	176 (11.9)
Critical preoperative state	449 (3.2)	69 (1.7)	135 (2.9)	38 (2.6)
Dialysis	219 (1.5)	45 (1.1)	93 (2.0)	18 (1.2)
Ejection fraction-N (%)				
<30%	418 (3.0)	53 (1.3)	143 (3.1)	46 (3.1)
30-50%	4464 (31.6)	988 (23.8)	1732 (37.0)	526 (35.5)
>50%	9261 (65.5)	3115 (75.0)	2806 (59.9)	908 (61.4)
Creatinine (mg/dL)-Median (IQR)	1.0 (0.8-1.1)	0.9 (0.8-1.1)	1.0 (0.8-1.1)	1.0 (0.8-1.2)
Creatinine Clearence [a] (mL/min)-Median (IQR)	72.2 (54.5-93.7)	70.4 (53.1-91.0)	77.1 (58.9-97.8)	64.6 (49.7-83.6)
Urgency of intervention-N (%)				
Elective	11,839 (83.7)	3961 (95.3)	3553 (75.9)	1336 (90.3)
Non-elective	2316 (16.3)	195 (4.7)	1128 (24.0)	144 (9.7)
Redo-N (%)	997 (7.0)	398 (9.6)	41 (0.9)	45 (3.0)
Intervention-N (%)				
Valve surgery	7849 (55.5)	4156 (100.0)	0 (0.0)	1480 (100.0)
Aortic repair	201 (1.4)	0 (0.0)	0 (0.0)	0 (0.0)
Aortic replacement	4932 (34.8)	2433 (58.5)	0 (0.0)	1017 (68.7)
Mitral repair	1696 (12.0)	924 (22.2)	0 (0.0)	272 (18.4)
Mitral replacement	1733 (12.2)	799 (19.2)	0 (0.0)	191 (12.9)
Tricuspid repair	481 (3.4)	0 (0.0)	0 (0.0)	0 (0.0)
Tricuspid replacement	37 (0.3)	0 (0.0)	0 (0.0)	0 (0.0)
CABG surgery-	6776 (47.9)	0 (0.0)	4671 (100.0)	1480 (100.0)
Thoracic aorta surgery [b]	1516 (10.7)	0 (0.0)	0 (0.0)	0 (0.0)
ICU length of stay (days)-Median (IQR)	1 (1-2)	1 (1-2)	1 (1-2)	1 (1-3)
Hospital length of stay (days)-Median (IQR)	11 (8-17)	10 (8-15)	11 (8-15)	13 (9-19)
ICU outcome-N (%)				
Alive	13,829 (97.7)	4106 (98.8)	4637 (99.1)	813 (97.4)
Dead	326 (2.3)	50 (1.2)	44 (0.9)	22 (2.6)
Hospital outcome-N (%)				
Alive	13,639 (96.4)	4061 (97.7)	4590 (98.1)	1447 (97.8)
Dead	516 (3.6)	95 (2.3)	91 (1.9)	33 (2.2)

CS: cardiac surgery; CABG: coronary artery bypass grafting; PASP: pulmonary artery systolic pressure; N: number; IQR: Inter-quartile range; ICU: Intensive Care Unit; NYHA: New York Heart Association; N/M mob: Neurological or Musculoskeletal dysfunction severely affecting mobility.
[a]creatinine clearence is calculated with the Cockcroft-Gault formula; [b]descending aorta endoprothesis are excluded.

TABLE II.—*Discrimination and calibration of the prognostic models.*

Prognostic model	H-L Ĉ test, p (statistic, df)	GiViTI calibration test, p (statistic, m)	SMR (95% CI)	Area under ROC curve (95% CI)
Euroscore I	<0.001 (346.0, 10)	<0.001 (415.1, 1)	0.48 (0.43-0.54)	0.74 (0.72-0.76)
Euroscore II	0.261 (12.4, 10)	0.110 (4.4, 1)	1.09 (1.01-1.18)	0.77 (0.75-0.79)
ACEF Score	<0.001 (47.4, 10)	<0.001 (80.2, 2)	1.16 (1.07-1.24)	0.67 (0.65-0.70)
STS-isolated valve surgery	0.330 (11.4, 10)	0.208 (9.3, 2)	0.85 (0.67-1.03)	0.72 (0.66-0.78)
STS-CABG surgery	0.111 (15.6, 10)	0.500 (1.4, 1)	1.12 (0.91-1.33)	0.74 (0.68-0.80)
STS-CABG and valve surgery	0.172 (14.0, 10)	0.666 (0.80, 1)	0.94 (0.71-1.17)	0.71 (0.64-0.78)

The GiViTI model has been developed on the same sample where it is evaluated. The STS models are evaluated only on the subgroups for which they have been developed, that are, isolated valve surgery (N.=2291), CABG surgery (N.=2769) and CABG and valve surgery (N.=835).
H-L: Hosmer-Lemeshow; Df: degree of freedom; SMR: standardized mortality ratio; ROC: receiver operating characteristic.

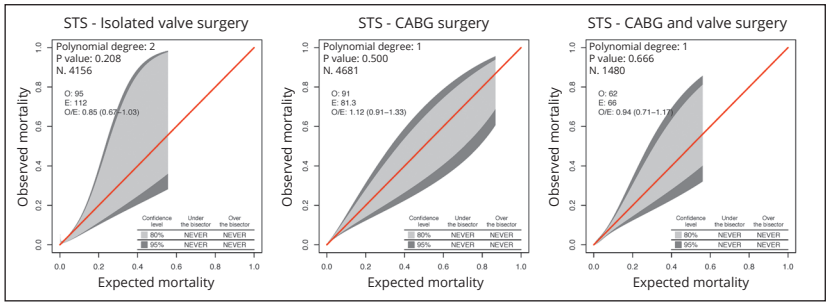


Figure 1.—Calibration of STS Score.

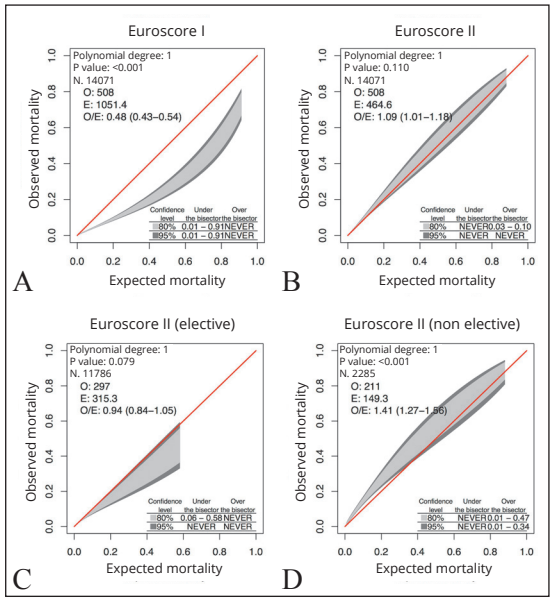


Figure 2.—A-D) Calibration of Euroscore I and Euroscore II overall and in elective and non-elective surgery (only Euroscore II).

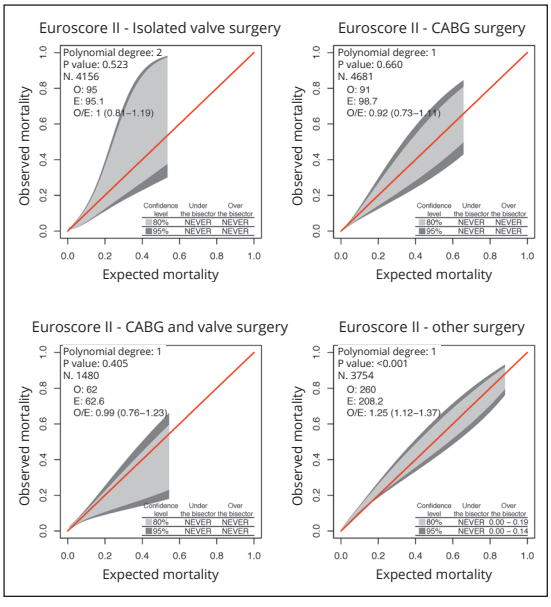


Figure 3.—Calibration of Euroscore II in STS subgroups and in complex surgeries not included in the STS categories.

calibration test and H-L P value <0.001). It overestimated mortality of medium- and high-risk elective patients (Figure 5B), whereas it underestimated mortality for low-risk non-elective patients and overestimated it for high-risk ones (Figure 5C).

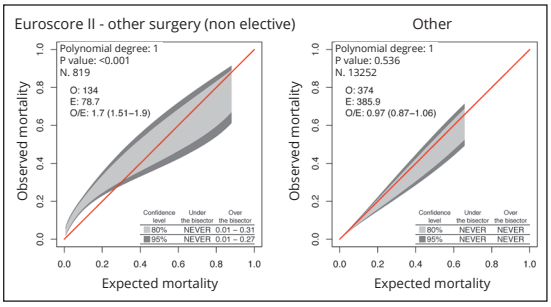


Figure 4.—Calibration of Euroscore II in non-elective complex surgeries and in the complementary set.

Discussion

In our study we have tested the validity of the main scores used to predict mortality in cardiosurgical patients, on a large cohort admitted to 16 ICUs representative of the Italian population, over a relatively short period of time.

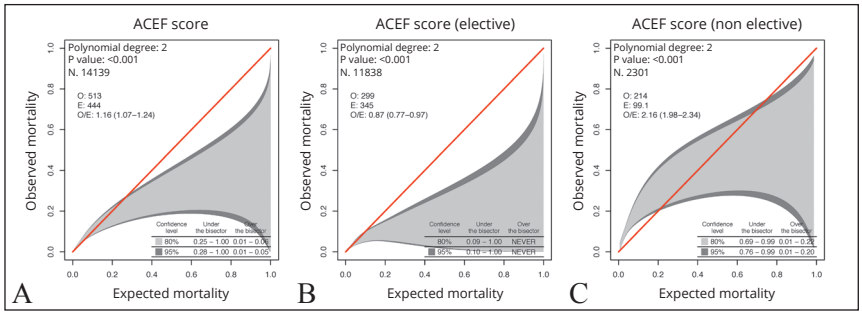
The main result of our analysis was that prognostic models calibrated poorly when there were differences in case-mix between the score development cohorts and our sample.

Actually, the STS models that were born for specific surgical patients without claims of generalizability, when tested on the appropriate surgical groups turned out to calibrate correctly.

Although the calibration of the EUROSCORE II seemed to be overall acceptable, the calibration belt suggested that in the lower classes of risk the score underpredicted (Figure 2). Further analysis confirmed this suspicion, showing that in the sample of patients submitted to non-elective more complex procedures (*e.g.* aortic surgery) the model miscalibrate significantly, predicting less deaths than observed for patients with expected mortality of 27% or less. This inability to predict in subsets typically occurs because of differences in case-mix between the cohort on which the model was developed and that on which it was applied.[22] The EUROSCORE II was developed prevalently on elective surgical patients (77%) undergoing CABG or valve surgery (93%). Not surprisingly the score calibrated fairly in patients undergoing elective procedures and poorly when surgery was urgent or emergent (Figure 2) and when procedures were other than CABG or valve surgery (Figure 3). The degree of miscalibration worsened when the two conditions were associated (Figure 4). Since the model does not account for the increased risk due to urgent/emergent and more complex interventions, it predicts fewer deaths than observed in this subset. Interestingly, mortality underprediction regarded only patients with lower risk-of-death (expected mortality ≤27%). However, for higher expected mortality rates imprecision increased (*i.e.* the confidence band broadened) because the number of patients decreased. Thus, miscalibration in the higher risks cannot be ruled out.

We, hence, hypothesize that in the development phase of the EUROSCORE II, the goodness-of-fit of the model was not tested on important subsets (what is called the assessment of uniformity of fit).[19, 23] Actually, if the score did not calibrate well in the subset of emergency patients in the development cohort, when applied to centers where the proportion of non-elective procedures is much higher, it will provide unreliable predictions. Our analysis shows that in the validation setting severity scores should be tested on important subsets to be sure that the model reliability will resist to case-mix variations. For this purpose, the calibration belt is an essential tool, since it helps delimiting the area of applicability of the score in a new context.

Figure 5.—A-C) Calibration of ACEF Score.



The ACEF Score seriously miscalibrated, probably because it is based on an excessively parsimonious model (it includes only three predictors: age, creatinine, and left-ventricular ejection fraction). In such underfitted models the prognostic weight of the predictors is biased, heavily affecting calibration when case-mix changes. For example, a patient with normal creatinine and normal cardiac inotropism, with severe diabetes and COPD, will obviously have a different risk of death compared to a patient with the same age without these comorbidities. Changes in the prevalence of these comorbidities (and of other risk factors) in the validation setting compared to the development cohort will skew the predictions of the model. Consistently, it was not surprising the STS and EUROSCORE II, which account for more prognostic variables, showed a much better predictive ability in our sample.

However, the EUROSCORE I, which is also less parsimonious than the ACEF, also performed poorly in our study. In this case miscalibration should be instead attributed to the aging of the score that was developed on a sample of patients collected in the late nineties (STS and EUROSCORE II are much more recent). Indeed, it is well known that the improvement of medicine over time increases the patients' chance of survival, and "aged" prognostic models tend to overestimate mortality,[24] which is exactly what the EUROSCORE I did (Figure 2).

Our study demonstrates the unreliability of the EUROSCORE I and the ACEF scores, which for different reasons provided seriously skewed predictions in our validation cohort and should not be used for mortality prediction at least in Italian ICUs. Although the STS and EUROSCORE II seem to be fair prognostic tools, it should be pointed out that these models should not be applied to centers with high rates of non-elective and/or complex operations.

In our study the availability of the GiViTI calibration belt was essential to investigate miscalibrating models, helping to spot the subsets for which the scores were not reliable. This statistical tool should also be used for prognostic models development to assess the uniformity of fit, a fundamental characteristic that allows models to maintain their reliability when they are tested in populations with different case-mix.

Conclusions

Healthcare systems operate with a progressive reduction of financial resources, imposing the improvement of efficiency and efficacy. In this context prognostic models are central for the assessment of the performance of single centers to detect inefficiencies, design corrective interventions, and monitor their efficacy. To grant their reliability these models should be validated in the clinical setting on which the will be applied, should be used cautiously, and updated periodically. Advanced tools as the GiViTI calibration belt are essential to guide clinicians in the correct interpretation and application of severity scores.

What is known

- Healthcare systems operate with a progressive reduction of financial resources, imposing the improvement of efficiency and efficacy. Prognostic models are used to assess the quality of healthcare.
- Several scores were developed to predict mortality after cardiac surgery, but none has reached optimal performance in subsequent validations.

What is new

- We assessed the calibration and discrimination of the most widespread scores. They calibrate poorly if validation and development case-mix differ. They should be updated and validated in the settings of application.
- STS & EUROSCORE II show good performance, EUROSCORE I overestimates mortality, ACEF poorly calibrates.

References

1. Nashef SA, Roques F, Michel P, Gauducheau E, Lemeshow S, Salamon R. European system for cardiac operative risk evaluation (EuroSCORE). Eur J Cardiothorac Surg 1999;16:9–13.

2. Roques F, Nashef SA, Michel P, Gauducheau E, de Vincentiis C, Baudet E, *et al.* Risk factors and outcome in Eu-

ropean cardiac surgery: analysis of the EuroSCORE multinational database of 19030 patients. Eur J Cardiothorac Surg 1999;15:816–22.

3. Nashef SA, Roques F, Sharples LD, Nilsson J, Smith C, Goldstone AR, *et al.* EuroSCORE II. Eur J Cardiothorac Surg 2012;41:734–44.

4. Shahian DM, O'Brien SM, Filardo G, Ferraris VA, Haan CK, Rich JB, *et al.*; Society of Thoracic Surgeons Quality Measurement Task Force. The Society of Thoracic Surgeons 2008 cardiac surgery risk models: part 3—valve plus coronary artery bypass grafting surgery. Ann Thorac Surg 2009;88(Suppl):S43–62.

5. O'Brien SM, Shahian DM, Filardo G, Ferraris VA, Haan CK, Rich JB, *et al.*; Society of Thoracic Surgeons Quality Measurement Task Force. The Society of Thoracic Surgeons 2008 cardiac surgery risk models: part 2—isolated valve surgery. Ann Thorac Surg 2009;88(Suppl):S23–42.

6. Shahian DM, O'Brien SM, Filardo G, Ferraris VA, Haan CK, Rich JB, *et al.*; Society of Thoracic Surgeons Quality Measurement Task Force. The Society of Thoracic Surgeons 2008 cardiac surgery risk models: part 1—coronary artery bypass grafting surgery. Ann Thorac Surg 2009;88(Suppl):S2–22.

7. Shahian DM, Edwards FH. The Society of Thoracic Surgeons 2008 cardiac surgery risk models: introduction. Ann Thorac Surg 2009;88(Suppl):S1.

8. Ranucci M. Minimal extracorporeal circulation: the real impact on postoperative outcome. Ann Thorac Surg 2009;87:352–3.

9. Howell NJ, Head SJ, Freemantle N, van der Meulen TA, Senanayake E, Menon A, *et al.* The new EuroSCORE II does not improve prediction of mortality in high-risk patients undergoing cardiac surgery: a collaborative analysis of two European centres. Eur J Cardiothorac Surg 2013;44:1006–11.

10. Ranucci M, Castelvecchio S, Conte M, Megliola G, Speziale G, Fiore F, *et al.* The easier, the better: age, creatinine, ejection fraction score for operative mortality risk stratification in a series of 29,659 patients undergoing elective cardiac surgery. J Thorac Cardiovasc Surg 2011;142:581–6.

11. 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;43:27–32.

12. Al-Ruzzeh S, Asimakopoulos G, Ambler G, Omar R, Hasan R, Fabri B, *et al.* Validation of four different risk stratification systems in patients undergoing off-pump coronary artery bypass surgery: a UK multicentre analysis of 2223 patients. Heart 2003;89:432–5.

13. Ranucci M, Di Dedda U, Castelvecchio S, La Rovere MT, Menicanti L; Surgical and Clinical Outcome Research (SCORE) Group. In search of the ideal risk-scoring system for very high-risk cardiac surgical patients: a two-stage approach. J Cardiothorac Surg 2016;11:13.

14. Paparella D, Guida P, Di Eusanio G, Caparrotti S, Gregorini R, Cassese M, *et al.* Risk stratification for in-hospital mortality after cardiac surgery: external validation of EuroSCORE II in a prospective regional registry. Eur J Cardiothorac Surg 2014;46:840–8.

15. Barili F, Pacini D, Rosato F, Roberto M, Battisti A, Grossi C, *et al.* In-hospital mortality risk assessment in elective and non-elective cardiac surgery: a comparison between EuroSCORE II and age, creatinine, ejection fraction score. Eur J Cardiothorac Surg 2014;46:44–8.

16. Lemeshow S, Hosmer DW Jr. A review of goodness of fit statistics for use in the development of logistic regression models. Am J Epidemiol 1982;115:92–106.

17. Poole D, Carlisle JB. Mirror, mirror on the wall… predictions in anaesthesia and critical care. Anaesthesia 2016;71:1104–9.

18. Finazzi S, Poole D, Luciani D, Cogo PE, Bertolini G. Calibration belt for quality-of-care assessment based on dichotomous outcomes. PLoS One 2011;6:e16110.

19. Nattino G, Finazzi S, Bertolini G. A new calibration test and a reappraisal of the calibration belt for the assessment of prediction models based on dichotomous outcomes. Stat Med 2014;33:2390–407.

20. Nattino G, Finazzi S, Bertolini G. A new test and graphical tool to assess the goodness of fit of logistic regression models. Stat Med 2016;35:709–20.

21. Zweig MH, Campbell G. Receiver-operating characteristic (ROC) plots: a fundamental evaluation tool in clinical medicine. Clin Chem 1993;39:561–77.

22. Poole D, Rossi C, Anghileri A, Giardino M, Latronico N, Radrizzani D, *et al.* External validation of the Simplified Acute Physiology Score (SAPS) 3 in a cohort of 28,357 patients from 147 Italian intensive care units. Intensive Care Med 2009;35:1916–24.

23. Poole D, Carrara G, Bertolini G. Intensive care medicine in 2050: statistical tools for development of prognostic models (why clinicians should not be ignored). Intensive Care Med 2017;43:1403–6.

24. Poole D, Bertolini G. Part II: Use and limitations of severity scores in critical care. In: Chiche JD MR, Putensen C, Rhodes A, editors. Patient Safety and Quality of Care in Intensive Care Medicine. Berlin: MWV & Co; 2009.

Conflicts of interest.—The authors certify that there is no conflict of interest with any financial organization regarding the material discussed in the manuscript.

Funding.—This work was supported by Fondazione Cariplo [2014-1962 to Stefano Finazzi].

Group name.—Members qualified as authors include the following: Roberto AGOSTINELLI (Ome), Andrea BALATA (Sassari), Giuseppe BUSCAGLIA (Genova), Mauro A. CALÒ (Mirano), Graziano CORTIS (Rozzano), Massimiliano GRECO (Rozzano), Matteo LUCCHELLI (Legnano), Ricardo MARTINEZ ESCOBAR (Brescia), Marco MAURELLI (Pavia), Carolina MONACO (Novara), Sandra NONINI (Milano), Alessandro RECH (Varese), Gianluigi REDAELLI (Monza), Alberto SENO (Trento), Andrea VARDENEGA (Mestre).

Comment in: Ranucci M. Calibration of risk scores: one model does not fit all. Minerva Anestesiol 2020;86:696–8. DOI: 10.23736/S0375-9393.20.14603-0.

History.—Article first published online: March 6, 2020. - Manuscript accepted: March 5, 2020. - Manuscript revised: March 2, 2020. - Manuscript received: November 12, 2019.