Reliability of new scores in predicting perioperative mortality after mitral valve surgery

Fabio Barili, MD, PhD, MStat,[a] Davide Pacini, MD,[b] Claudio Grossi, MD,[a] Roberto Di Bartolomeo, MD,[b] Francesco Alamanni, MD,[c] and Alessandro Parolari, MD, PhD[c]

Objective: The study was designed to validate euroSCORE II and ACEF (age, creatinine, and ejection fraction) scores in patients undergoing isolated or associated mitral valve surgery and compare them with logistic euroSCORE and Society of Thoracic Surgeons scores.

Methods: Data on 3441 consecutive patients undergoing isolated or associated mitral valve surgery in a 6-year period were retrieved from 3 prospective institutional databases. Discriminatory power was assessed with the C index. Calibration was evaluated with calibration curves and associated statistics.

Results: In-hospital mortality was 3.4%. Discriminatory power was uniformly good (for euroSCORE II: area under curve, 0.79; 95% confidence interval, 0.74-0.84; for logistic euroSCORE: area under the curve, 0.78; 95% confidence interval, 0.74-0.83; for ACEF: area under the curve, 0.73; 95% confidence interval, 0.69-0.79) but significantly higher in euroSCORE models ($P < .05$ for Delong, bootstrap, Venkatraman methods). Calibration pattern was slightly better for the ACEF score, although related summary statistics (unreliability, Hosmer-Lemeshow test, Spiegelhalter z-test for calibration accuracy) were not significant even for euroSCORE II. The euroSCORE II demonstrated a performance similar to Society of Thoracic Surgeons score. Logistic euroSCORE confirmed the progressive trend toward overprediction previously demonstrated in the general cardiac surgical population (summary statistics $P < .05$). Analysis of score performances in the surgical group studied showed results comparable to the global population.

Conclusions: The euroSCORE II and ACEF scores are good predictors of perioperative mortality in patients undergoing isolated or associated mitral valve surgery, with better discrimination for the first and better calibration for the second. No algorithm seems suitable for risk estimation in mid and high-risk patients. (J Thorac Cardiovasc Surg 2014;147:1008-12)

ACD

The estimation of perioperative risk has gained an increasingly important role in cardiac surgery, because perioperative mortality is considered one of the main quality indices of cardiac surgery. Its prediction can lead to better patient understanding of the risks associated with the procedures and may also aid decision-making behavior of clinicians and serve as a guide for hospital oversight to allocate resources efficiently and maximize care for high-risk patients.[1] Several tools have been developed in recent years, and among them the widest diffusion in the European countries has been reached by additive and logistic euroSCOREs.[2] These algorithms, however, have been demonstrated to be no more than adequate in the modern surgical population, with a high risk of overestimation.[3] The lack of performance has been especially observed in valve surgery subgroups, because these models were developed and validated in a surgical population composed mainly of those undergoing coronary artery bypass grafting (CABG).[4]

To overcome the performance limitations of older euroSCOREs, an updated version has been recently released, with a similar core of risk factors but a different categorization of operations.[5] Moreover, a new European score, the ACEF (age, creatinine, and ejection fraction) score, has been recently developed on the basis of a very limited number of risk factors.[6] Recent validation studies of these new scores on the general population and the aortic valve surgery subgroup have been already published; however, no data are available regarding patients who undergo mitral surgery, a developing field for new technologies. This study was designed to validate the euroSCORE II and ACEF scores in patients undergoing isolated and associated mitral valve surgery and to compare their performance with those of the logistic euroSCORE and Society of Thoracic Surgeons (STS) scores.

From the Department of Cardiac Surgery,[a] S Croce Hospital, Cuneo, Italy; the Department of Cardiac Surgery,[b] University of Bologna, Policlinico S Orsola-Malpighi, Bologna, Italy; and the Department of Cardiovascular Sciences,[c] University of Milan, Centro Cardiologico Monzino IRCCS, Milan, Italy.

Disclosures: Authors have nothing to disclose with regard to commercial support.

Presented at The 2013 American Association for Thoracic Surgery Mitral Conclave, New York, New York, May 2, 2013.

Received for publication April 3, 2013; revisions received June 12, 2013; accepted for publication June 26, 2013; available ahead of print Aug 29, 2013.

Address for reprints: Fabio Barili, MD, PhD, MStats, Department of Cardiac Surgery, S Croce Hospital, Via M Coppino 26, 12100 Cuneo, Italy (E-mail: fabarili@libero.it, barili.f@ospedale.cuneo.it).

0022-5223/$36.00

http://dx.doi.org/10.1016/j.jtcvs.2013.06.042

MATERIALS AND METHODS

The study population included all patients who underwent isolated or associated mitral valve surgery from January 2006 to April 2012

Abbreviations and Acronyms

ACEF = age, creatinine, and ejection fraction
AUC = area under the curve
CABG = coronary artery bypass grafting
CI = confidence interval
STS = Society of Thoracic Surgeons

(3441 patients enrolled) within the departments of cardiac surgery of 2 university hospitals and 1 regional hospital. The study population was extracted from a larger database that has been updated to April 2012.[7] Preoperative and demographic information, operative data, and perioperative mortality and complications for all patients were retrieved from the institutional databases, which are prospectively collected. The institutional review boards approved the data set's use for research. The institutional ethical committees approved the study, and the requirement for informed, written consent was waived on the condition that subjects' identities remained masked. Data from the 3 centers were matched and stored in a dedicated data set.

The scores were tested on the prediction of in-hospital mortality. For the evaluation of the performance, the scores were calculated for each patient in accordance with published guidelines with a dedicated software. The STS score can be applied only to isolated mitral surgery and mitral surgery associated with CABG, so the comparison of the new scores with the STS score was performed only in these subcategories.

The performance of the scores was analyzed with a focus on discrimination power and calibration, as previously described.[7,8] The discrimination performance indicates the extent to which the model distinguishes between patients who will die or survive in the perioperative period. Discrimination performance was evaluated by constructing receiver operating characteristic curves for each model and calculating the area under the curve (AUC) with 95% confidence intervals (CIs). An area of 1.0 indicates perfect discrimination power, whereas an area of 0.5 indicates no discrimination of the binary outcome. The comparison among curves was analyzed with Delong, bootstrap, and Venkatraman methods, the first 2 of which compare the AUCs and the last the receiver operating characteristics curves themselves. Another index used to evaluate the predictive abilities was the Somers D_{xy} rank correlation between predicted probabilities and observed responses. When $D_{xy} = 0$, the model is making a random prediction; when $D_{xy} = 1$, the predictions are perfectly discriminating.

Calibration refers to the agreement between observed outcomes and predictions, and it was evaluated by generating calibration plots. The perfectly calibrated predictions stay on the 45° line, whereas curves below and above the diagonal reflect overestimation and underestimation, respectively. For each model, the comparison of actual slope and intercept with the ideal values of 1 and 0 was performed with the U statistic (unreliability test) and tested against a χ^2 distribution with 2 df. For testing whether the calibration curve was ideal, we used even the 1-df Spiegelhalter *z*-test, with its 2-tailed *P* value for calibration accuracy. Moreover, calibration was tested with Hosmer-Lemeshow goodness-of-fit test, which compares observed with predicted values by decile of predicted probability. The accuracy of the models was also tested by calculating the Brier score (quadratic difference between predicted probability and observed outcome for each patient), an overall performance measure that is 0 when the prediction is perfect.

Two-sided statistics were performed with a significance level of 0.05. For all analyses the R 2.15.1 software was used (R Foundation for Statistical Computing, Vienna, Austria).

RESULTS

The mean age of the group was 66.3 ± 11.8 years, and 44.1% were female. Isolated mitral valve surgery was performed in 1239 patients (36.0%), 1 associated procedure was performed in 1461 (42.5%), and 2 or more associated procedures were performed in 741 (21.5%). CABG was performed in 1023 patients (29.7%), without other associated procedures in 613 (17.8%). Other major procedures included aortic valve surgery in 1036 patients (30.1%), tricuspid valve surgery in 595 (17.3%), surgery for ascending aorta in 108 (3.1%), and surgery for left ventricular aneurysm in 71 (2.1%). In-hospital mortality was 3.4% (115 patients). The discriminatory power was good for all algorithms (for euroSCORE II: AUC, 0.79; 95% CI, 0.74-0.84; for logistic euroSCORE: AUC, 0.78; 95% CI, 0.74-0.83; for ACEF: AUC, 0.73; 95% CI, 0.69-0.79) but significantly higher for the euroSCORE models ($P < .05$ for Delong, bootstrap, and Venkatraman methods; Table 1). The calibration pattern was better for the ACEF score, with a line closer to the ideal diagonal (Figure 1). In addition, related summary statistics were more favorable for the ACEF score, although thy were not significant even for the euroSCORE II. Logistic euroSCORE confirmed the progressive trend toward overprediction previously demonstrated in the general cardiac surgical population (summary statistics $P < .05$; Table 1).

The analysis of scores' performances in isolated and associated mitral valve surgery showed comparable results. In isolated mitral surgery subgroup, the discriminatory power was significantly superior in STS score and euroSCORE algorithms (for STS: AUC, 0.82; 95% CI, 0.75-0.90; for euroSCORE II: AUC, 0.81; 95% CI, 0.75-0.88; for logistic euroSCORE: AUC, 0.82; 95% CI, 0.76-0.89; for ACEF: AUC, 0.76; 95% CI, 0.68-0.83; $P < .05$ for all comparisons between ACEF and other scores and $P > .05$ for comparisons between STS and euroSCOREs). In associated mitral valve surgery, the superiority of discriminatory performance of euroSCORE models was less marked and not significant (for euroSCORE II: AUC, 0.75; 95% CI, 0.69-0.82; for logistic euroSCORE: AUC, 0.70; 95% CI, 0.67-0.80; for ACEF: AUC, 0.70; 95% CI, 0.64-0.77; $P > .05$ for all comparisons). In patients who underwent mitral surgery associated with CABG, the STS score but not euroSCOREs demonstrated a significantly higher discrimination power relative to the ACEF score (for STS score: AUC, 0.76; 95% CI, 0.65-0.88; for euroSCORE II: AUC, 0.74; 95% CI, 0.61-0.86; for logistic euroSCORE: AUC, 0.73; 95% CI, 0.60-0.85; for ACEF: AUC, 0.70; 95% CI, 0.56-0.80; $P < .05$ for all comparisons between ACEF and STS score and $P > .05$ for all comparisons between STS and euroSCOREs and between euroSCOREs and ACEF).

The pattern of calibration was similar for all scores in the 2 subgroups. Logistic euroSCORE showed a tendency toward progressive overprediction, which was confirmed even by the associated summary statistics that were significant ($P < .0001$ for unreliability test, Hosmer-Lemeshow test, and Spiegelhalter *z*-test in both isolated and associated