92

Published by Elsevier Inc.
0003-4975/$36.00
https://doi.org/10.1016/j.athoracsur.2021.02.052

Cardiac Operative Risk in Latin America: A Comparison of Machine Learning Models vs EuroSCORE-II

Raúl Santiago Molina, Eng, María Alejandra Molina-Rodríguez, MD, Francisco Mauricio Rincón, MD, and Javier Dario Maldonado, MD

Cardiovascular Surgical Department, Clinica Universitaria Colombia, Bogotá, Colombia

ABSTRACT

BACKGROUND Machine learning is a useful tool for predicting medical outcomes. This study aimed to develop a machine learning–based preoperative score to predict cardiac surgical operative mortality.

METHODS We developed various models to predict cardiac operative mortality using machine learning techniques and compared each model to European System for Cardiac Operative Risk Evaluation-II (EuroSCORE-II) using the area under the receiver operating characteristic (ROC) and precision-recall (PR) curves (ROC AUC and PR AUC) as performance metrics. The model calibration in our population was also reported with all models and in high-risk groups for gradient boosting and EuroSCORE-II. This study is a retrospective cohort based on a prospectively collected database from July 2008 to April 2018 from a single cardiac surgical center in Bogotá, Colombia.

RESULTS Model comparison consisted of hold-out validation: 80% of the data were used for model training, and the remaining 20% of the data were used to test each model and EuroSCORE-II. Operative mortality was 6.45% in the entire database and 6.59% in the test set. The performance metrics for the best machine learning model, gradient boosting (ROC: 0.755; PR: 0.292), were higher than those of EuroSCORE-II (ROC: 0.716, PR: 0.179), with a *P* value of .318 for the AUC of the ROC and .137 for the AUC of the PR.

CONCLUSIONS The gradient boosting model was more precise than EuroSCORE-II in predicting mortality in our population based on ROC and PR analyses, although the difference was not statistically significant.

(Ann Thorac Surg 2022;113:92-9)

Cardiac surgery has become an important tool in the treatment of cardiovascular disease. Mortality scores aim to precisely predict cardiac mortality for a given procedure (even multiple procedures) in each patient.[1]

Cardiovascular disease is the leading cause of death worldwide, and in low-income countries approximately 80% of deaths are related to cardiovascular disease. By 2018, approximately 20% of the general population fell below the threshold for multidimensional poverty in Colombia.[2] There are no scores specifically validated in Colombia, except for the European System for Cardiac Operative Risk Evaluation (EuroSCORE) validation by Figueredo and associates.[3]

Risk assessment models in cardiac surgery have been developed since the 1980s. As the number of procedures grew and databases became a widespread tool to evaluate patient characteristics, it became possible to build statistical models with preoperative variables to predict operative mortality. The Society of Thoracic Surgeons has created multiple risk assessment models for different cardiac procedures with more than 700,000 patients.[4]

The most widely used model is EuroSCORE-II, which is believed by many to be the gold standard for operative mortality prediction.[5] For the Latin American population, the first model was created in Argentina in 2009. It consisted of multivariate logistic regression and surpassed EuroSCORE using the area under the curve (AUC) of the receiver operating characteristic (ROC).[6,7]

The Supplemental Tables and Supplemental Figure can be viewed in the online version of this article [https://doi.org/10.1016/j.athoracsur.2021.02.052] on http://www.annalsthoracicsurgery.org.

Accepted for publication Feb 25, 2021.

Address correspondence to Mr Molina, Cardiovascular Surgical Department, Clinica Universitaria Colombia, Bogotá, Colombia; email: rs.molina10@uniandes.edu.co.

Recent literature considers machine learning a subset of artificial intelligence that refers to the ability of machines to learn independently from data and make accurate predictions.[8] Mortality prediction is a task that falls under a machine learning subfield called supervised learning, where a machine is trained to learn an input-output function with a collection (data set) of such pairs.[9,10] Furthermore, we consider 2 classes (death or no death), so it is a binary classification problem.[10]

Algorithms developed in the machine learning field have been used to analyze large databases, and it has been demonstrated that they are superior to simpler models such as logistic regression in mortality prediction after cardiac surgery.[11] There is a tradeoff, however, between complex models and their interpretability.[12] For predictive algorithms in medicine, it tends to be preferable to have complex relationships between variables, resulting in better performance in practice, than simple models with easy interpretations of risk factors.[13]

The objective of this study was to develop a machine learning model to predict operative cardiac mortality using preoperative variables from retrospective patient data.

PATIENTS AND METHODS

STUDY POPULATION. We collected data from all adult patients undergoing cardiovascular surgical procedures for acquired cardiomyopathies between July 2008 and April 2018 at the Clinica Universitaria Colombia in Bogotá, Colombia. Our Cardiovascular Surgical Department was created in 2008, and since then, we have collected an in-house 351-variable database. This database contains diagnostic, preoperative, intraoperative, and in-hospital postoperative variables.

CLINICAL OUTCOMES. This study's only outcome was operative mortality, defined as death during the same hospitalization as surgery or after discharge but within 30 days of the procedure.[14] After signing informed consent forms, patients granted access to electronic health records; this was approved by our hospital ethics committee. Sample size calculations were unnecessary for this study, and we used all available data. The data were collected from electronic health records by a physician tasked with data entry. The database was password-protected, and only members of the research group had access to it.

MODEL DEVELOPMENT. The database was separated into 2 independent, randomly sampled sets at a ratio of 4:1 between the training and test sets. All further analyses were performed in the training set unless explicitly stated. Feature selection was made using the χ^2 test[15] for nominal and categorical variables, with operative mortality as the response variable; only variables with $P < .05$ were selected. The same process was repeated with continuous and ordinal variables using the Mann-Whitney U test.[16] Sex, body mass index, and aortic insufficiency were not significantly different between outcome groups but were included because of their clinical significance. Multiple machine learning models were trained with the selected features and validated on the test set using the ROC AUC.[17] Given that there was a high imbalance between the predicted classes (30-day mortality was an uncommon occurrence in the data set), precision-recall (PR) AUC was also used.[18] Optimal thresholds for each curve were obtained by maximizing Youden's J statistic and the maximum F measure (F1), respectively.[19,20] To assess model calibration, the expected calibration error and maximum calibration error were used and reported as supplemental material.[21] For groups considered high risk in the test set, observed mortality, model confidence, and calibration error were calculated and compared to low-risk groups.[21] These included patients with age greater than 75 years, a left ventricular ejection fraction (LVEF) of less than 50%, arrhythmias, heart failure, dialysis, emergency and urgency upon admission, cardiogenic shock. or resuscitation.[22-24]

Confidence intervals and statistical significance tests for metrics in the test set were performed using the bootstrap method with 1000 bootstrapped samples, whenever applicable.[25] Furthermore, differences in the ROC AUCs were also validated analytically using DeLong's method.[26]

Each model was trained 10 times with different random seeds (whenever applicable), and the average of the probability outputs was taken as the model's prediction, as described by Allyn and coworkers.[11]

In this study, 6 different models were used:

- Logistic regression models the log-odds of the outcome using a weighted combination of the input features and outputs a probability using the logistic function.[10]
- Naive Bayes fits a distribution for each feature (assuming conditional independence) and applies Bayes' theorem to output a prediction.[10]
- Multilayer perceptron is a low-depth neural network that can approximate complex decision boundaries by using activations between neuron layers.[27]
- Support vector machine (SVM) is a classification algorithm that best separates classes based on data points (support vectors) lying close to each class's boundaries.[28]
- Random forest is a model that consists of simple decision trees (collections of logical rules to separate the population into smaller homogeneous

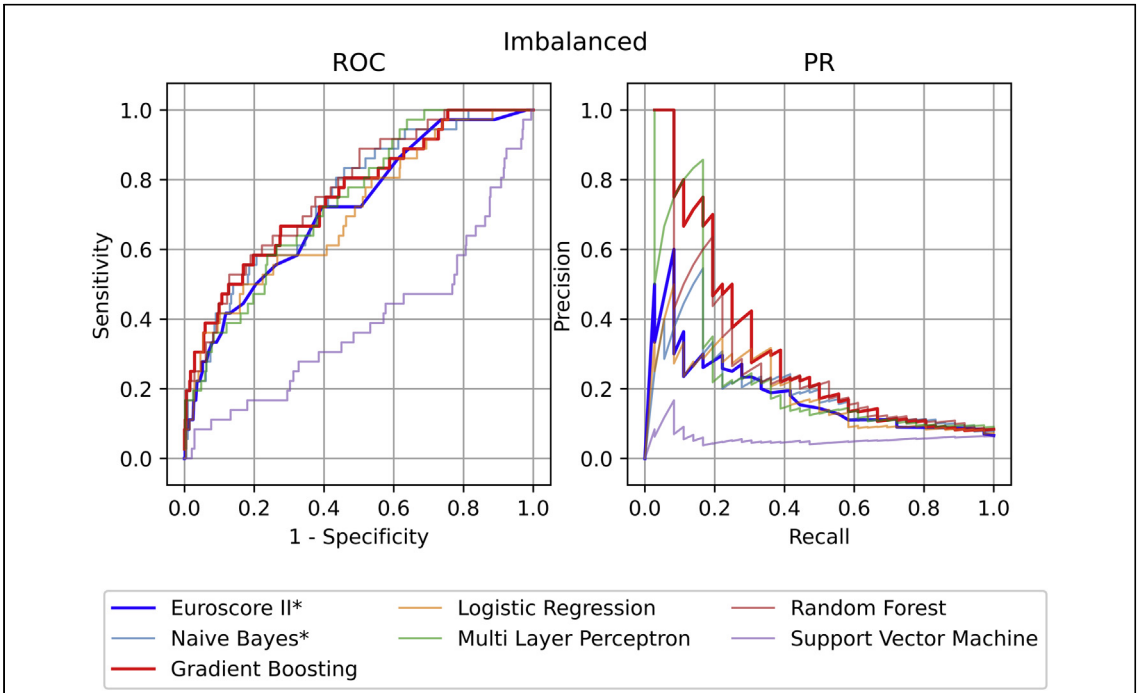

FIGURE 1 Comparison between receiver operating characteristic (ROC) and precision-recall (PR) curves for each of the trained models without class balance. *: Class balancing for naive Bayes models was not implemented, and European System for Cardiac Operative Risk Evaluation II (EuroSCORE-II) is provided as a reference.

subgroups); each takes a random subsample of both the data and the possible features to make a prediction.[29]

- Gradient boosting (boosted trees), a strong classifier, is built from an ensemble of decision trees. Leveraging a differentiable cost function, it is possible to greedily improve the model's predictions by fitting each consecutive tree with the last one's residuals (gradients).[30]

Inverse-weighted coefficients proportional to class frequency were used in the logistic regression model to give equal importance to both outcomes during training (class balance). As such, misclassification in rare events (death) incurs a greater cost for optimization problems.

Inverse-weighted cost functions were used to achieve class balance in the multilayer perceptron, SVM, and gradient boosting models. Additionally, the random forest model used weighted subsampling with replacement (each decision tree was trained with a 1:1 alive:dead ratio). Outputs of the SVM were transformed into class probabilities using Platt scaling, adjusted on the training data.[31]

All model implementations and statistical analyses were performed in Python using the SciPy, XGBoost, Keras, and Sci-kit learn libraries.[30,32-34] Hyperparameter tuning was performed using a grid search on the test set. The code for the project is available at https://github.com/santiag0m/cardio.

RESULTS

Out of an initial sample of 2960 patients, 174 (5.88%) were excluded due to missing data in 1 or more variables. This left a total sample of 2786 patients split into training (n = 2228; 80%) and test (n = 558; 20%) sets. We had 179 deaths among the 2786 patients (mortality rate of 6.42%). A total of 23 features were selected based on the criteria described earlier. Summary statistics for the selected variables can be found in Supplemental Table S1. The year in which the procedure was performed was not statistically significant ($P = .51$), so it was excluded from the analysis.

For each model, the ROC AUC and PR AUC were calculated and compared to those of EuroSCORE-II. The resulting curves for models trained on imbalanced and balanced data are shown in Figures 1 and 2.

Most models surpassed the EuroSCORE-II ROC AUC, the best being random forest trained on imbalanced data (ROC AUC of 0.716 vs 0.771, $P = .181$) (Table 1). In terms of the PR AUC, the best model was gradient boosting trained on imbalanced data, which was also superior to EuroSCORE-II (0.179 vs 0.292, $P = .137$) (Table 2).

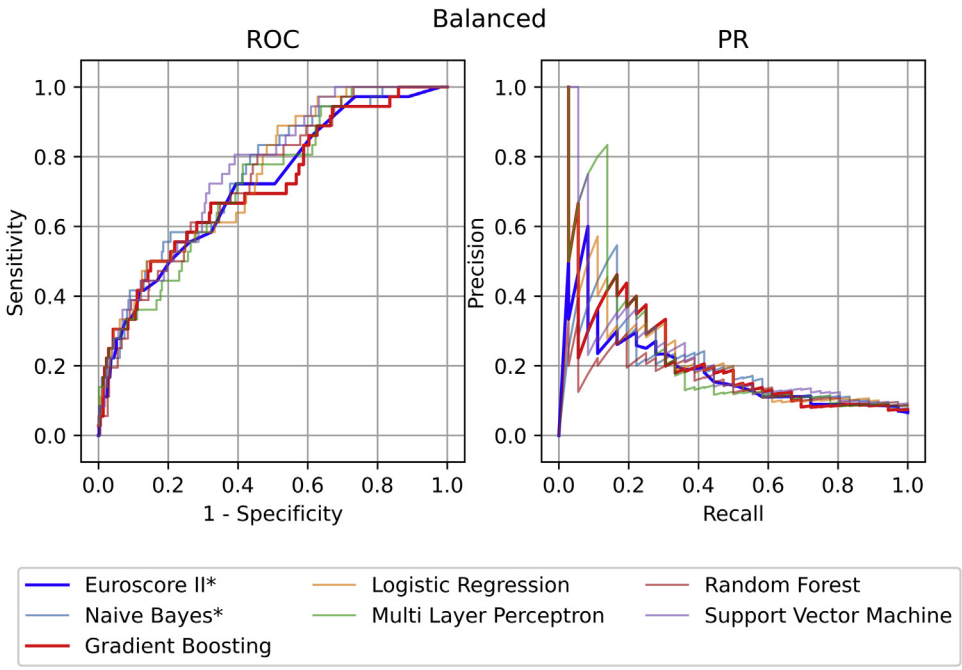

FIGURE 2 Comparison between receiver operating characteristic (ROC) and precision-recall (PR) curves for each of the trained models with class balance. *: Class balancing for naive Bayes models was not implemented, and European System for Cardiac Operative Risk Evaluation II (EuroSCORE-II) is provided as a reference.

For the ROC curve, the optimal threshold was calculated using Youden's J statistic and the F1 measure for the PR curve (Tables 1, 2). The closest model to the ideal classifier in the ROC curve was SVM trained on balanced data (J = 0.415), which was superior to EuroSCORE-II (J = 0.328, $P = .276$). For the PR curve, the best classifier was gradient boosting trained on imbalanced data (F1 = 0.355), which was also superior to EuroSCORE-II (F1 = 0.274, $P = .137$).

It is important to note that although using an optimal ROC threshold is an intuitive estimate (closer to the mortality mean in the test set, 6.00%), it is highly conservative, as it assigns more than one third of patients as high risk for EuroSCORE-II. The PR threshold is less intuitive but achieves a closer estimate of the high-risk population, as shown in Figure 3.

Calibration metrics for each of the trained models are presented in Supplemental Table S2. The model with the

TABLE 1 Metrics Related to the ROC Curve for Each of the Models Trained

Method	ROC AUC (95% CI)	*P* Value	*P* Value (DeLong)	Youden's J Statistic (95% CI)	Threshold, %	*P* Value
EuroSCORE-II	0.716 (0.630-0.800)	…	…	0.328 (0.235-0.498)	6.00	…
Naive Bayes	0.749 (0.668-0.829)	.419	.450	0.376 (0.299-0.563)	0.494	.563
Logistic regression	0.706 (0.614-0.796)	.834	.823	0.331 (0.238-0.520)	9.62	.959
SVM	0.390 (0.302-0.479)	< .001 [a]	< .001 [a]	0.055 (0.006-0.172)	15.4	< .001 [a]
Random forest	**0.771 (0.690-0.845)**	.160	.181	0.399 (0.319-0.584)	9.84	.276
MLP	0.739 (0.662-0.813)	.558	.564	0.347 (0.305-0.526)	5.73	.746
Gradient boosting	0.755 (0.666-0.838)	.369	.318	0.391 (0.285-0.578)	7.15	.401
Logistic regression (balanced)	0.745 (0.666-0.819)	.466	.448	0.375 (0.314-0.537)	31.3	.393
SVM (balanced)	0.761 (0.686-0.831)	.226	.223	**0.415** (0.327-0.570)	5.24	.174
Random forest (balanced)	0.731 (0.651-0.808)	.719	.729	0.350 (0.275-0.526)	33.4	.754
MLP (balanced)	0.725 (0.643-0.805)	.824	.813	0.364 (0.275-0.515)	3.10	.588
Gradient boosting (balanced)	0.717 (0.626-0.805)	.979	.976	0.351 (0.257-0.538)	57.6	.765

[a]$P \leq .001$. The values in bold indicate the best performing model for each metric EuroSCORE, European System for Cardiac Operative Risk Evaluation; MLP, multilayer perceptron; ROC, receiver operating characteristic; SVM, support vector machine.

TABLE 2 Metrics Related to the PR Curve for Each of the Models Trained

Method	PR AUC (95% CI)	*P* Value	F1 (95% CI)	Threshold, %	*P* Value
EuroSCORE-II	0.179 (0.095-0.290)	…	0.274 (0.187-0.420)	19.0	…
Naive Bayes	0.193 (0.107-0.321)	.795	0.306 (0.211-0.444)	2.26	.591
Logistic regression	0.181 (0.097-0.299)	.971	0.338 (0.211-0.475)	19.1	.340
SVM	0.056 (0.035-0.095)	< .001 [b]	0.122 (0.094-0.233)	2.74	.003 [a]
Random forest	0.260 (0.123-0.387)	.249	0.311 (0.229-0.483)	9.84	.545
MLP	0.225 (0.109-0.356)	.475	0.280 (0.196-0.453)	16.1	.936
Gradient boosting	**0.292** (0.136-0.430)	.137	**0.355** (0.242-0.517)	30.7	.252
Logistic regression (balanced)	0.198 (0.103-0.323)	.748	0.314 (0.211-0.460)	83.3	.526
SVM (balanced)	0.212 (0.105-0.326)	.580	0.306 (0.211-0.452)	16.5	.612
Random forest (balanced)	0.154 (0.089-0.245)	.550	0.283 (0.186-0.408)	51.7	.878
MLP (balanced)	0.218 (0.104-0.356)	.543	0.295 (0.190-0.457)	31.8	.764
Gradient boosting (balanced)	0.193 (0.101-0.314)	.798	0.319 (0.210-0.469)	77.2	.496

[a]$P \leq .01$; [b]$P \leq .001$. EuroSCORE, European System for Cardiac Operative Risk Evaluation; MLP, multilayer perceptron; PR, precision-recall; SVM, support vector machine.

lowest expected calibration error and maximum calibration error was the random forest trained on imbalanced data (0.016 and 0.030, respectively), which was better than EuroSCORE-II (0.029 and 0.051, respectively). Although bootstrap confidence intervals and significance values were calculated for both metrics, they were highly biased for some of the models (expected calibration error and maximum calibration error expected values in the bootstrap distribution were different from those in the sample by a large margin), so they were not considered suitable for the study. The results for calibration error in high-risk groups in the test sample are presented in Table 3 for the EuroSCORE-II and gradient boosting models. We found that conditions with larger samples, such as age, ejection fraction, and urgency upon admission, had lower calibration errors for the high- and low-risk groups in both models. Smaller-sized groups such as reanimation (n = 12) or arrhythmia (n = 73) displayed broad variations between the risk groups.

Using an additive attribution model[35] on the input variables for gradient boosting (Supplemental Figure S1), we found that the variable behaviors can be divided according to their level of measurement (binary, categorical, or continuous) and depending on the operative mortality outcome.

For some binary variables (3 or more procedures and cardiogenic shock), being positive implies greater risk in all cases. Other binary variables also attribute high risk to positive values, but some patients with negative values may present a higher risk; this was the case for arrhythmia, diabetes, dialysis, heart failure, peripheral artery disease, and stroke. Finally, some variables did not attribute greater or lesser risk to the outcome, such as hypertension or sex.

In the categorical variables, we see some variables in which greater severity carries greater risk, such as chronic obstructive pulmonary disease, tricuspid regurgitation, and urgency upon admission. Quantitative variables such as age and LVEF had values that limit regions of positive and negative risk (75 years for age and 45% for LVEF) and presented stable ranges with low risk where it neither increased nor decreased (20-50 years for age and 55% or greater for LVEF). In contrast, in the creatinine and hematocrit groups, there was only a negative risk range (0.5-1.1 mg/dL for creatinine and 40%-47% for hematocrit). A more complicated case was body mass index, as the risk varied for each case, and there was not a clear trend to explain its impact; however, lower values of body mass index (<20 kg/m^2) were more likely to have positive risk.

COMMENT

This model predicts operative mortality in cardiac surgery in Colombia. Our cohort's operative mortality was 6.42%, which is similar to that reported worldwide and in the Colombian population but higher than that reported in the EuroSCORE-II cohort.[5] When we compared our population's demographics with those from EuroSCORE-II, we found older patients with a higher prevalence of diabetes, endocarditis, renal failure, and chronic obstructive pulmonary disease. We are certain that this is due to a late referral to cardiac surgery caused by failings in our health system, as has been the case in other Latin American countries.

When evaluating the models using AUC ROC and PR curves (Tables 1 and 2), it is important to note that the ROC's optimal classifier tends to overestimate the risk of operative mortality, assigning more value to false-negative cases. In contrast, an optimal classifier in the PR curves will equally value false positives and false negatives (Figure 3). These results highlight an issue regarding model metrics and clinical utility. The ROC

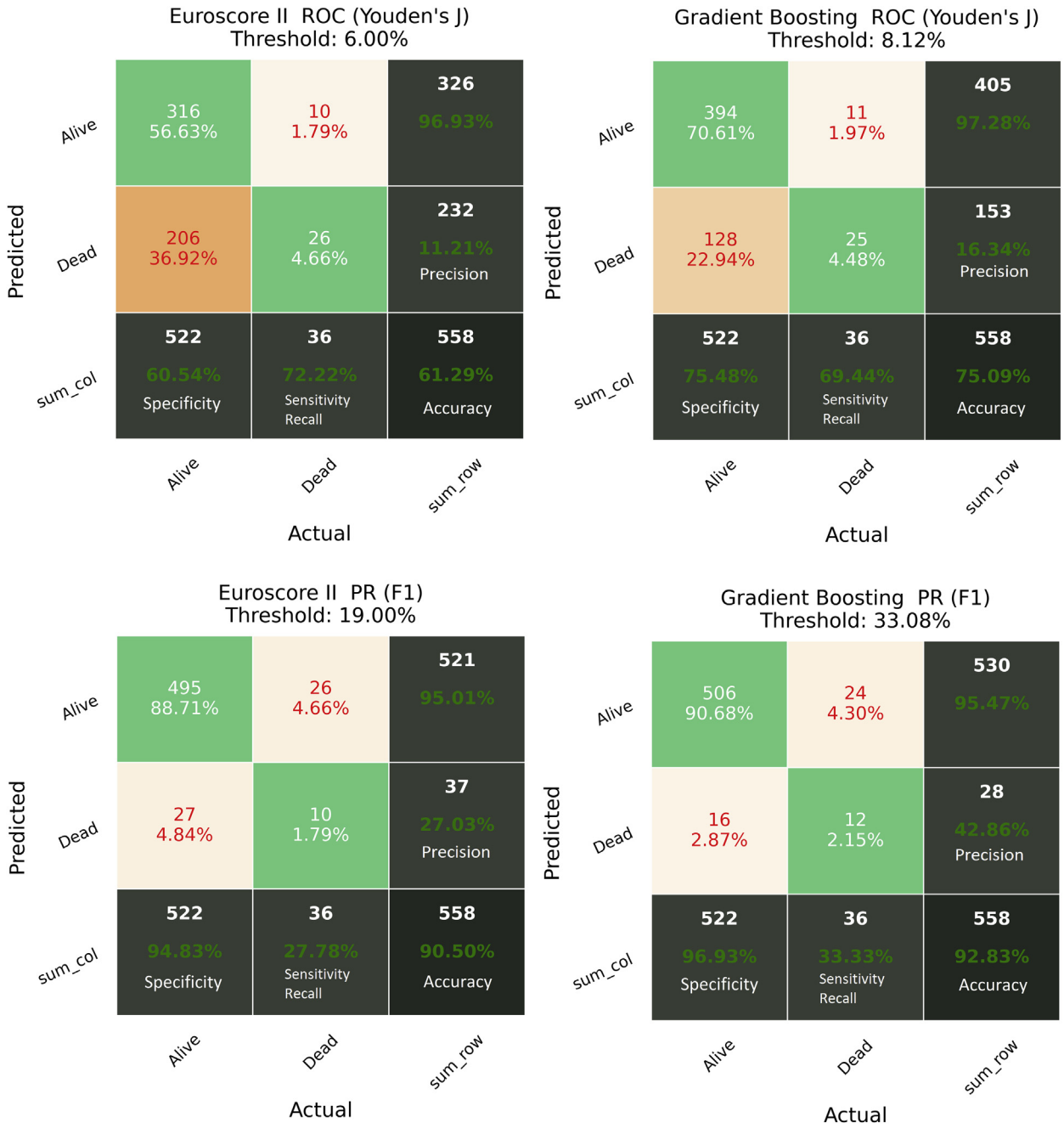

FIGURE 3 Confusion matrices obtained in the test set with optimal thresholds that maximize Youden's J statistic (top), which prioritizes sensitivity and specificity, and F1 (bottom), which prioritizes precision and recall. Shown for the European System for Cardiac Operative Risk Evaluation II (EuroSCORE-II) (left) and gradient boosting (right). (PR, precision-recall; ROC, receiver operating characteristic.)

metric privileges sensitivity and the optimal gradient boosting classifier underestimates death risk in only 1.97% of patients but overestimates it for 22.9%. The optimal classifier under the PR metric, however, which sets precision as its priority, overestimates death risk in only 2.87% of patients but underestimates it in 4.30%.

Regarding model calibration, the model with the lowest calibration error was the random forest, outperforming both EuroSCORE-II and gradient boosting. We found that the calibration error for different risk groups is likely to be dependent on the number of patients with each condition (as defined in Table 3). It is

TABLE 3 Patient Count and Calibration Metrics for Low- and High-Risk Groups in the Test Data Set

				EuroSCORE-II		Gradient Boosting	
Variable	Risk	n (%)	Observed Mortality, %	Model Confidence, %	Calibration Error, %	Model Confidence, %	Calibration Error, %
Age	High (> 75 y)	173 (31.0)	10.4	10.4	0.000	10.8	0.428
	Low (≤ 75 y)	385 (69.0)	4.68	5.37	0.691	6.01	1.34
Ejection fraction	High (< 50%)	164 (29.4)	8.54	9.01	0.476	11.2	2.65
	Low (≥ 50%)	394 (70.6)	5.58	6.06	0.477	5.98	0.39
Urgency upon admission	High (Nonelective)	352 (63.1)	7.39	7.81	0.420	8.24	0.852
	Low (Elective)	206 (36.9)	4.85	5.43	0.573	6.26	1.40
Arrhythmia	High (Yes)	73 (13.1)	12.3	10.9	1.41	21.4	9.10
	Low (No)	485 (86.9)	5.57	6.33	0.761	5.41	0.156
Cardiogenic shock	High (Yes)	18 (3.23)	38.9	19.7	19.2	35.3	3.55
	Low (No)	540 (96.8)	5.37	6.50	1.13	6.58	1.21
Reanimation	High (Yes)	12 (2.15)	33.3	19.7	13.7	23.7	9.67
	Low (No)	546 (97.9)	5.86	6.65	0.788	7.15	1.29
Heart failure	High (Yes)	124 (22.2)	10.5	9.40	1.08	14.8	4.32
	Low (No)	434 (77.8)	5.30	6.22	0.922	5.42	0.122
Dialysis	High (Yes)	55 (9.86)	12.7	10.7	2.07	12.0	0.728
	Low (No)	503 (90.1)	5.77	6.52	0.755	7.02	1.25

A comparison is made between the EuroSCORE-II and gradient boosting models. EuroSCORE, European System for Cardiac Operative Risk Evaluation.

important to consider that the models we trained may not have the best performance in risk groups with small sample sizes, such as resuscitation and arrhythmia.

We decided to use the gradient boosting algorithm trained without class balance as our prediction model; it had the highest AUC of all models for the PR curve and competitive performance for the ROC curve.

Our model's complex relationships (Supplemental Figure S1) are consistent with many clinical heuristics and cannot be expressed with traditional models such as logistic regression. It is also essential to consider that feature attribution is personalized and offers a possible explanation of why some binary variables fluctuate between positive and negative risk depending on the patient. We hypothesize that the presence of comorbidities, such as heart failure or dialysis, explains abnormal values in other variables (eg, LVEF or creatinine) that otherwise would be unaccounted for and could indicate other underlying conditions.

Although our cohort is smaller than that used for EuroSCORE-II and other models in the machine learning literature,[36] the addition of new patients to our data set in the future can help improve the performance of our model and increase sample sizes for high-risk groups.

Due to the limited quantity of available data, hyperparameter search and model selection (not training) were performed using the held-out test set. A prospective study is underway in our center to validate the model and ensure that such selection is correct.

Another purpose of this study was to create an application that Latin American cardiac surgeons can use to calculate cardiac operative mortality risk. This study's expanded aim was to introduce our model to the community and begin testing it as part of a future multicentric study in Latin America. Our model was tested in our population, and a multicentric study will help us determine its external validity.

With this first report's publication, we want to invite other Latin American centers to become a part of this project. We hope to provide cardiac surgeons and anesthetists with a useful specialized tool with which to better inform the decision of whether to operate on a cardiac patient.

Our calculator is available at http://cardiorisk.ml.

REFERENCES

1. Welke KF, Ferguson TB, Coombs LP, et al. Validity of The Society of Thoracic Surgeons National Adult Cardiac Surgery Database. *Ann Thorac Surg*. 2004;77:1137-1139.

2. DANE. *Boletín Técnico: Pobreza Multidimensional En Colombia*. Bogotá D.C., Colombia; 2019.

3. Figueredo A, Díaz F, Murcia AS, Gómez JC, Figueredo M. Utilidad del EuroSCORE en la predicción de mortalidad intrahospitalaria en una institución de enfermedades cardiovasculares de Colombia. *Revista Colombiana de Cardiología*. 2013;20:164-169.

4. Jacobs JP, Shahian DM, Prager RL, et al. Introduction to the STS National Database Series: Outcomes Analysis, Quality Improvement, and Patient Safety. *Ann Thorac Surg*. 2015;100:1992-2000.

5. Nashef SAM, Roques F, Sharples LD, et al. EuroSCORE II dagger. *Eur J Cardiothorac Surg*. 2012;41:734-745.

6. 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;16:9-13.

7. Carosella VC, Navia JL, Al-Ruzzeh S, et al. The first Latin-American risk stratification system for cardiac surgery: can be used as a graphic pocket-card score. *Interact Cardiovasc Thorac Surg*. 2009;9:203-208.

8. Kilic A. Artificial intelligence and machine learning in cardiovascular health care. *Ann Thorac Surg*. 2020;109:1323-1329.

9. LeCun Y, Bengio Y, Hinton G. Deep learning. *Nature*. 2015;521:436-444.

10. Bishop CM. Pattern Recognition and Machine Learning. New York: Springer-Verlag, *2006*.

11. Allyn J, Allou N, Augustin P, et al. A comparison of a machine learning model with EuroSCORE II in predicting mortality after elective cardiac surgery: a decision curve analysis. *PLOS ONE*. 2017;12:e0169772.

12. Caruana R, Lou Y, Gehrke J, Koch P, Sturm M, Elhadad N. Intelligible models for healthcare: predicting pneumonia risk and hospital 30-day readmission. Paper presented at: Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. Sydney, Australia; 2015.

13. Goldstein BA, Navar AM, Carter RE. Moving beyond regression techniques in cardiovascular risk prediction: applying machine learning to address analytic challenges. *Eur Heart J*. 2016;38:1805-1814.

14. The Society of Thoracic Surgeons. Performance Measure Descriptions | STS. Available at: https://www.sts.org/quality-safety/performance-measures/descriptions. Accessed July 17, 2020.

15. Pearson K. On the criterion that a given system of deviations from the probable in the case of a correlated system of variables is such that it can be reasonably supposed to have arisen from random sampling. *London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science*. 1900;50(302):11-28.

16. Mann HB, Whitney DR. On a test of whether one of two random variables is stochastically larger than the other. *Ann Math Stat*. 1947;18:50-60.

17. Hanley JA, McNeil BJ. The meaning and use of the area under a receiver operating characteristic (ROC) curve. *Radiology*. 1982;143:29-36.

18. Davis J, Goadrich M. The relationship between Precision-Recall and ROC curves. Paper presented at: Proceedings of the 23rd International Conference on Machine Learning; Pittsburgh, Pennsylvania;2006.

19. Youden WJ. Index for rating diagnostic tests. *Cancer*. 1950;3:32-35.

20. Chinchor N. MUC-4 evaluation metrics. Paper presented at: MUC4 '92 Proceedings of the 4th conference on Message understanding. McLean, Virginia; June 1992.

21. Naeini MP, Cooper GF, Hauskrecht M. Obtaining well calibrated probabilities using bayesian binning. Paper presented at: AAAI'15 Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence. Austin, Texas; January 2015.

22. Afilalo J, Steele R, Manning WJ, et al. Derivation and validation of prognosis-based age cutoffs to define elderly in cardiac surgery. *Circ Cardiovasc Qual Outcomes*. 2016;9:424-431.

23. Sepehri A, Beggs T, Hassan A, et al. The impact of frailty on outcomes after cardiac surgery: A systematic review. *J Thorac Cardiovasc Surg*. 2014;148:3110-3117.

24. Yancy CW, Jessup M, Bozkurt B, et al. 2013 ACCF/AHA Guideline for the Management of Heart Failure: A Report of the American College of Cardiology Foundation/American Heart Association Task Force on Practice Guidelines. *J Am Coll Cardiol*. 2013;62:e147-e239.

25. Efron B. Bootstrap methods: another look at the jackknife. *Ann Stat*. 1979;7:1-26.

26. DeLong ER, DeLong DM, Clarke-Pearson DL. Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach. *Biometrics*. 1988;44:837-845.

27. Yan H, Jiang Y, Zheng J, Peng C, Li Q. A multilayer perceptron-based medical decision support system for heart disease diagnosis. *Expert Syst Appl*. 2006;30:272-281.

28. Boser BE, Guyon IM, Vapnik VN. A training algorithm for optimal margin classifiers. Paper presented at: Proceedings of the Fifth Annual Workshop on Computational Learning Theory, 1992.

29. Breiman L. Random forests. *Mach Learn*. 2001;45:5-32.

30. Chen T, Guestrin C. XGBoost: A scalable tree boosting system. Paper presented at: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2016.

31. Platt JC. Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods. Advances in Large Margin Classifiers. 1999.

32. Virtanen P, Gommers R, Oliphant TE, et al. SciPy 1.0: fundamental algorithms for scientific computing in Python. *Nature Meth*. 2020;17:261-272.

33. Chollet F. Keras: The Python deep learning library. Astrophysics Source Code Library. *2018*.

34. Pedregosa F, Varoquaux G, Gramfort A, et al. Scikit-learn: machine learning in Python. *J Mach Learn Res*. 2011;12:2825-2830.

35. Lundberg SM, Lee SI. A unified approach to interpreting model predictions. Paper presented at: NIPS'17 Proceedings of the 31st International Conference on Neural Information Processing Systems, 2017.

36. Kilic A, Goyal A, Miller JK, et al. Predictive utility of a machine learning algorithm in estimating mortality risk in cardiac surgery. *Ann Thorac Surg*. 2020;109:1811-1819.