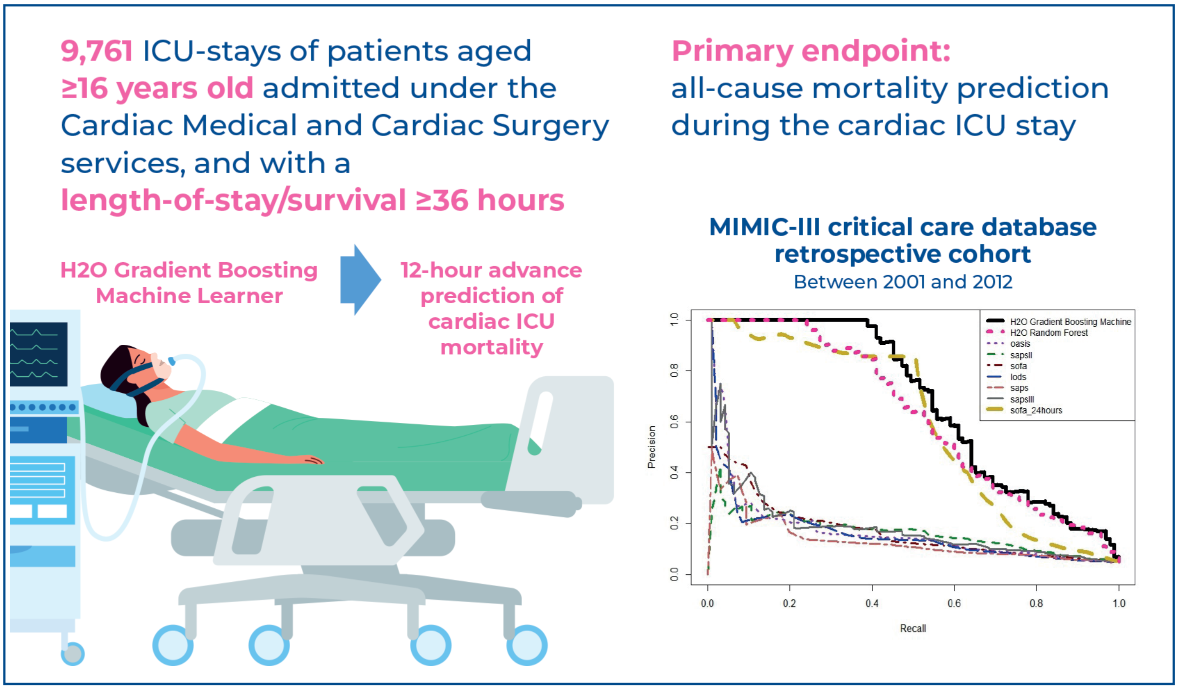
ORIGINAL ARTICLE

einstein
Official Publication of the Instituto Israelita
de Ensino e Pesquisa Albert Einstein

Comparing ensemble learning algorithms and severity of illness scoring systems in cardiac intensive care units: a retrospective study



9,761 ICU-stays of patients aged **≥16 years old** admitted under the Cardiac Medical and Cardiac Surgery services, and with a **length-of-stay/survival ≥36 hours**

Primary endpoint: all-cause mortality prediction during the cardiac ICU stay

H2O Gradient Boosting Machine Learner → **12-hour advance prediction of cardiac ICU mortality**

MIMIC-III critical care database retrospective cohort
Between 2001 and 2012

Author

Beatriz Nistal-Nuño

Correspondence

E-mail: nistalnunobeatriz7@gmail.com

DOI

DOI: 10.31744/einstein_journal/2024AO0467

In Brief

Beatriz Nistal-Nuño designed a machine learning system type of ensemble learning for patients undergoing cardiac surgery and intensive care unit cardiology patients, based on sequences of cardiovascular physiological measurements and other intensive care unit physiological measurements in addition to static features, which generates a score for prediction of mortality of cardiac intensive care unit patients.

Highlights

- Gradient Boosting Machine and Random Forest models were built for prediction of mortality at cardiac intensive care units.
- A total of 9,761 intensive care unit stays of patients admitted under a Cardiac Surgery and Cardiac Medical services were studied.
- The AUROC and AUPRC values were significantly superior to seven conventional systems compared.
- The machine learning models' calibration curves were substantially closer to the ideal line.

How to cite this article:

Nistal-Nuño B. Comparing ensemble learning algorithms and severity of illness scoring systems in cardiac intensive care units: a retrospective study. einstein (São Paulo). 2024;22:eAO0467.

einstein
Official Publication of the Instituto Israelita
de Ensino e Pesquisa Albert Einstein

e-ISSN: 2317-6385

ORIGINAL ARTICLE

Comparing ensemble learning algorithms and severity of illness scoring systems in cardiac intensive care units: a retrospective study

Beatriz Nistal-Nuño[1]

[1] Complexo Hospitalario Universitario de Pontevedra, Pontevedra, PO, Spain.

DOI: 10.31744/einstein_journal/2024AO0467

ABSTRACT

Objective: Logistic Regression has been used traditionally for the development of most predictor tools of intensive care unit mortality. The purpose of this study is to combine shared risk factors between patients undergoing cardiac surgery and intensive care unit cardiology patients to develop a risk score for prediction of mortality in cardiac intensive care unit patients, using machine learning. **Methods:** Gradient Boosting Machine and Distributed Random Forest models were developed based on 9,761 intensive care unit-stays from the MIMIC-III database. Sequential and static features were collected. The primary endpoint was intensive care unit mortality prediction. Discrimination, calibration, and accuracy statistics were evaluated. The predictive performance of traditional scoring systems was compared. **Results:** Machine learning models' AUROC and AUPRC were significantly superior to all conventional systems for the primary endpoint ($p<0.05$), with AUROC of 0.9413 for Gradient Boosting Machine and 0.9311 for Distributed Random Forest. Sensitivity was 0.6421 for Gradient Boosting Machine, 0.6 for Distributed Random Forest, and <0.3 for all conventional systems except for serial SOFA (0.6316). Precision was 0.574 for Gradient Boosting Machine, 0.566 for Distributed Random Forest, and <0.5 for all conventional systems. Diagnostic odds ratio was 58.8144 for Gradient Boosting Machine, 51.2926 for Distributed Random Forest and <34 for all conventional systems. Brier score was 0.025 for Gradient Boosting Machine and 0.028 for Distributed Random Forest, being worse for the traditional systems. Calibration curves of Gradient Boosting Machine and Distributed Random Forest were substantially closer to the ideal line. **Conclusion:** The machine learning models showed superiority over the traditional scoring systems compared, with Gradient Boosting Machine having the best performance. Discrimination and calibration were excellent for Gradient Boosting Machine, followed by Distributed Random Forest. The machine learning methods exhibited better capacity for most accuracy statistics.

Keywords: Cardiac surgery procedures; Ensemble learning; Mortality; Intensive care units; Risk factors; Calibration

How to cite this article:
Nistal-Nuño B. Comparing ensemble learning algorithms and severity of illness scoring systems in cardiac intensive care units: a retrospective study. einstein (São Paulo). 2024;22:eAO0467.

Associate Editor:
Helder Takasgi Imoto Nakaya
Hospital Israelita Albert Einstein, São Paulo, SP, Brazil.
ORCID: https://orcid.org/0000-0001-5297-9108

Corresponding author:
Beatriz Nistal-Nuño
Department of Anesthesiology, Complexo Hospitalario Universitario de Pontevedra
Mourente s/n
Zip code: 36071 - Pontevedra, PO, Spain
Phone: +34 981 295 899
E-mail: nistalnunobeatriz7@gmail.com

Received on:
Jan 29, 2023

Accepted on:
Apr 15, 2024

Conflict of interest:
none.

INTRODUCTION

Mortality risk prediction scores have an important function in intensive care units (ICUs). Patients in ICUs have more severe physiologic derangement and consequently a higher risk of mortality.[1] Implementation of these tools may help identify high-risk patients and reduce preventable deaths,[2,3] potentially resulting in improvements in ICU care and outcomes.

Despite improvements in surgical techniques and perioperative care, cardiac surgery operations still have a risk of mortality. Patients' responses to cardiac surgery and cardiopulmonary bypass (CPB) are particular to the postoperative cardiac surgery ICU patients, showing temporary pathophysiological changes which may influence many variables captured by the general ICU scoring systems. However, most of these changes have a limited effect on outcomes.[4]

These patients are susceptible to multiple interventions that are also very common in the cardiology ICUs, in addition to the shared primary physiological and clinical manifestations of predominant heart diseases among cardiac surgery and cardiology patients. Appropriate mortality risk assessment of this subgroup of patients is essential for monitoring the health evolution of these patients in the ICU, where primary cardiac derangements can lead to imminent severe critical events and death.

Several general ICU scoring algorithms are widely established and are used for evaluating the severity of critical illness and anticipating mortality. These include the SOFA,[5] SAPS,[6] SAPS II,[7] SAPS III,[8] LODS,[9] and the OASIS.[10] Although postoperative cardiac surgery patients or specifically cardiac ICU patients were not included during their creation, these scoring systems are widely used for risk estimation after cardiac surgery and in cardiac ICUs.[11]

Specific risk models have been created to estimate the mortality risk of cardiac surgery, mainly using preoperative data.[11] The EuroSCORE and the STS risk stratification algorithms are the two main preoperative risk models in cardiac surgery.[12] EuroSCORE II, created using logistic regression (LR), was better calibrated than EuroSCORE, preserving powerful discrimination.[13]

Nilsson et al. contrasted 19 risk score models regarding their ability to predict 30-day and 1-year mortality following cardiac surgery. Discrimination was significantly higher for logistic[14] and additive[15] EuroSCORE models, followed by Cleveland Clinic[16] and Magovern[17] systems.[18]

Disease-specific risk scores for the population of cardiac surgery patients have also been developed. For example, Ariyaratne et al. created a multi-variable LR model using preoperative data to anticipate early mortality following aortic valve replacement (AVR) in adults. Their model (AVR-Score) showed good discrimination and calibration.[19] Wang et al. compared four different risk scores for predicting in-hospital mortality following heart valve surgery (EuroSCORE II, Ambler risk score, NYC risk score, and STS). The four risk scores gave an imprecise prediction for individual risk in patients undergoing multiple valve surgery.[20-22]

Other risk algorithms have been created for cardiac surgery patients using postoperative variables, such as the Vasoactive-Inotropic Score (VIS), which was developed initially to predict mortality and morbidity following pediatric cardiac surgery.[23] It has been examined recently in adults.[11] The discrimination for unfavorable outcome of VISmax was better than APACHE II, SAPS II, and similar to SOFA. Calibration revealed a good fit.[11]

Lamarche et al. developed a score that incorporated both preoperative and intraoperative features building a multiple LR model that estimated 30-day mortality after adult cardiac surgery. Their model's AUROC decreased when rerun using only preoperative variables. They argued that mortality risk also needs to be assessed immediately after surgery because the surgery constitutes a turning point.[24]

An ICU scoring system specifically developed for general adult cardiac surgical patients to estimate mortality risk using postoperative data is the CASUS.[4] The initial CASUS, an additive algorithm, was created as a tool for daily mortality risk classification in ICU patients admitted following cardiac surgery.[25] The logistic variant of the additive CASUS (Log-CASUS) showed a clear superiority to the logistic EuroSCORE and the additive CASUS.[4] Doerr et al. did not find any improvement by merging a preoperative and a postoperative scoring model. Therefore, they recommended a separate computation of the two scores.[2]

Disease-specific scores have been created for several subsets of cardiology patients. A multi-variable LR model was developed for in-hospital mortality risk estimation by Granger et al., for patients with acute coronary syndromes with and without ST-segment elevation. This algorithm can be utilized as a simple nomogram to assess risk in individual patients.[26] The GWTG-HF Program was developed by Peterson et al. for individual prediction of risk of in-hospital mortality in patients hospitalized with heart failure (HF). The model had good discrimination and calibration.[27] The ADHERE study recognized BUN, serum creatinine, and systolic blood pressure (SBP) as the best predictors of in-hospital mortality in patients with HF. This model stratifies patients as low, intermediate, or high risk.[28]

The only score developed for hospital mortality risk prediction for general adult ICU cardiology patients is the model developed by Jentzer et al. This score identified risk predictors available at ICU admission to develop the M-CARS using LR. The M-CARS showed discrimination and calibration superior to conventional ICU risk scores.[29]

Many of the scoring systems mentioned above can allow clinicians to estimate preoperatively the mortality risk following cardiac surgery. However, very few are designed to calculate the mortality risk after undergoing cardiac surgery. Preoperative risk stratification may help in the choice among cardiac surgery and other therapeutic modalities.(18,22) However, it omits the surgery results failing to consider intraoperative and postoperative variables.(22)

In addition, most of the available cardiac ICU outcome prediction models, as well as the established severity-of-illness scoring systems, collect static risk factors around the time of ICU admission ignoring changes in patient status, which is considered in the current work by measuring the variation of variables over time.

Most of the predictor tools developed to date that estimate mortality following cardiac surgery or in cardiology patients use LR to compute the score, as well as the majority of the conventional severity-of-illness systems. There is increasing evidence that machine learning (ML) models can provide a more accurate outcome prediction than LR in the ICU.(30-32) Ensemble modeling has been widely used in surgery literature.(33) Machine learning approaches can overcome the statistical limitations of LR regarding the assumptions of independence of observations, such as in repeated measurements, and the absent multicollinearity among the independent features.(34) Physiologically, it is likely that intercorrelations exist between many homeostatic and clinical variables.

This research is built upon previous work by the author,(35) seeking to refine the previous models and trying to further improve the predictive performance.

OBJECTIVE

This research aimed to create a predictive tool using ensemble machine learning techniques for intensive care unit mortality for individual adult patients admitted to the intensive care unit under a Cardiac Surgery or a Cardiac Medical service.

METHODS

Patient cohort

The MIMIC-III critical care database version v1.4. was utilized, a large dataset containing de-identified clinical data of over forty thousand patients admitted to ICUs between 2001 and 2012 at the tertiary care hospital Beth Israel Deaconess Medical Center (BIDMC) in the United States. It contains high temporal resolution ICU data, collected from Metavision and CareVue bedside monitors.(36,37)

A retrospective study was performed. A final cohort of 9,761 ICU-stay patient records was used, selected as shown in the CONSORT Flow Diagram of table 1. Code in PostgreSQL language created for the selection of the ICU-stays is accessible at.(38)

Table 1. CONSORT Flow Diagram used to plot the flow of data selection of the patient cohort

All distinct ICU-stays in MIMIC-III
n=61,532
↓
Exclude ICU-stays of patients of age <16 years old
n=8,109
↓
Exclude ICU-stays of patients admitted to the BIDMC under a service different from the Cardiac Medical or Cardiac Surgery services
n=38,373
↓
Exclude ICU-stays with a length-of-stay and survival <36 hours following ICU admission
n=5,289
↓
Study cohort of distinct ICU-stays n=9,761

Patients with multiple re-admissions to the ICU were included as distinct ICU-stays.
ICU: intensive care unit.

Study design

Data were extracted during a 6-hours window to produce 12-hours advance predictions by the ML models developed. The ML models were contrasted to the severity-of-illness scores of OASIS, SAPS, SAPS II, SAPS III, LODS, SOFA and serial SOFA. The threshold of 36-hours length-of-stay was selected to permit 24-hours of data gathering in the ICU for computation of the conventional scores, as the calculation interval for these scores is the first 24-hours of ICU stay (except for 1-hour for SAPS III).

This threshold allows then to make a prediction at 12-hours before patient death/discharge from the ICU by the ML models and conventional systems. The ML methods were also contrasted to the serial SOFA, which is computed also at 12-hours before patient death/discharge, calculated from the immediately preceding 24-hours of data.(39,40) The established systems' scores were calculated using open-source code complementing the MIMIC-III database.(41)

The primary endpoint was all-cause mortality prediction during the ICU stay. The secondary endpoint was all-cause in-hospital mortality prediction for the same hospital admissions of the corresponding ICU stays.

Data collection

Variables collected were of type sequential and static. The 1-hour time-resolution physiological and laboratory variables extracted during 6 consecutive hours in the ICU consisted of heart rate (HR), pH, pulse pressure, respiratory rate, blood oxygen saturation, SBP, mean blood pressure, temperature, hemoglobin, lactate and white blood cell count. These relevant variables are commonly available in the ICU (Table 2).

The cardiovascular evaluation is of great interest in cardiac ICU patients, that is why several of the sequential features were selected. Blood lactate levels have previously been identified as a predictor of mortality in ICU patients[4,42] and therefore lactate was included. Hemoglobin was selected as in.[34]

Static features were collected as demographic variables at the time of ICU admission comprising type of ICU admission, gender, ethnicity, and age. The static clinical features consisted of the need of vasopressors during the ICU stay, the use of dialysis, and several ICD-9 procedures and diagnoses documented on discharge (Table 2).

The demographic features were selected because they have demonstrated to be essential mortality risk factors. Ethnicity has an empiric association with outcomes and has been included in many mortality predictor tools.[12,27,43] Admission type was selected based on [3,11-15,24] (Table 2).

Use of inotropes and vasopressors indicates severity of patient condition and therefore was selected based on.[5,11-13,24] The diagnosis of cardiac arrest was reported as an independent risk factor for mortality in previous studies.[26,29,44] Primary pulmonary hypertension, a well-known prognostic factor in cardiac surgery, is associated with augmented morbidity and mortality and therefore was selected based on.[3,13-15,22,24,43] Myocardial infarction was selected based on,[12-15,34,43] and atrial fibrillation/flutter based on.[3,12,26] Renal function is important in predicting ICU outcomes and therefore the need for dialysis was selected based on.[3,4,12,13,34,43,44] Ventricular assist device (VAD) implantation was selected based on,[4,24] and extracorporeal membrane oxygenation (ECMO)/CPB based on.[24,44] Pressure ulcers were selected based on[29,45] (Table 2).

For a single missing hourly value, a last observation carried forward imputation was used. For a missing value in the first hourly measurement of the 6-hours window, mean imputation was used.

Table 2. Patient variables extracted for developing the ML models. Diagnoses and procedure variables were obtained based on ICD-9 codes

Sequential variables (Time window=6 hours)		Static variables	
Physiologic	HR, pulse pressure, respiratory rate, blood oxygen saturation, SBP, mean blood pressure, Temperature	ICU admission kind (admission_ type)	elective, urgent, or emergency
Laboratory*	pH, hemoglobin, white blood cell count, blood lactate levels	Race (ethnicity_ grouped)	white, black, hispanic, asian, native, unknown, other
		gender	
		age at ICU entrance (patient_age)	
		Angioplasty or stent(s) (angio_stent)	PTCA, Status-post PTCA, Open chest coronary artery angioplasty, Insertion of non-drug-eluting or drug-eluting coronary artery stent(s)[35]
		VAD (vad)	Implant of Single Ventricular (Extracorporeal) External Heart Assist System, Insertion of temporary non-implantable extracorporeal circulatory assist device, Insertion of (percutaneous external) heart assist device[35]
		vaso_flag	norepinephrine, epinephrine, phenylephrine, vasopressin, dopamine, or isoprenaline
		ppulm_ hypert	Primary pulmonary hypertension
		dialysis	Renal dialysis status, Peritoneal dialysis, Hemodialysis
		myocardial_ infarction	Acute myocardial infarction of anterolateral wall, other anterior wall, inferolateral wall, inferoposterior wall, other inferior wall, other lateral wall, other specified sites, unspecified site; Old myocardial infarction
		cardiac_ arrest	Cardiac arrest, Personal history of sudden cardiac arrest
		cpb_ecmo	Percutaneous cardiopulmonary bypass, Extracorporeal circulation auxiliary to open heart surgery, ECMO, History of ECMO
		pressure_ ulcer	Pressure ulcer of buttock, unspecified site, elbow, upper back, lower back, hip, ankle, heel, other site, unspecified stage, stage I, stage II, stage III, stage IV, unstageable
		atrial_fib_flu	Atrial fibrillation, atrial flutter

* As these laboratory features were quantified less frequently than hourly, their interval for extraction was extended 23 hours backward for their first hourly measurement of the 6-hour window.

HR: heart rate; ICU: intensive care unit; VAD: ventricular assist device; PTCA: percutaneous transluminal coronary angioplasty; ECMO: extracorporeal membrane oxygenation.

Machine learning models

Data were loaded into KNIME (KNIME AG, Zurich, Switzerland)[46] to build the ML models. The input dataset was split randomly into two partitions, 80% for training and 20% for testing. The same testing data (n=1,953) were used to assess the performance of all ML models and conventional systems compared.

The ML models created were the Distributed Random Forest (DRF) and Gradient Boosting Machine (GBM). These models are provided by H2O®, implemented in KNIME.[46] These models are of the type of ensemble learning. The DRF classification model is built with Bagging. Gradient Boosting Machine for classification is a forward learning ensemble algorithm, built with Boosting.[46] A parameter optimization loop was used in order to locate the optimal parameters for the ML methods.[46]

Performance evaluation

Discrimination was evaluated using the AUROC and ROC curves, together with the precision recall curve (PRC) and AUPRC,[47] which ignore the amount of true negatives (TNs) and can be useful for problems with class imbalance. The AUROC and AUPRC were calculated for all models for both outcomes.

It was hypothesized that the ML models created would be superior to the conventional systems in discrimination. This hypothesis was tested for both outcomes. The difference among the AUROCs was calculated with the method of DeLong et al.[48] The comparison of paired PRCs provided the 95% bootstrap CIs for the AUPRC differences.[49] MedCalc® Statistical Software version 20.111 (MedCalc Software Ltd, Ostend, Belgium; https://www.medcalc.org; 2022) was used for statistical analyses.

Calibration of the models was evaluated with the Brier score. This was calculated for both outcomes for the ML methods, OASIS, SAPS II and SAPS III. Calibration curves were created by Moving average algorithm for these models for the primary endpoint.

Sensitivity, specificity, positive predictive value (PPV), negative predictive value (NPV), and diagnostic odds ratio (DOR) were assessed for the primary endpoint for all models. The optimal criterion values to estimate these statistics were calculated taking into account not only sensitivity and specificity, but also the ICU mortality prevalence, and the costs of false positive (FP), true positive (TP), false negative (FN), and TN.

Ethics approval

The Institutional Review Boards of the BIDMC and the Massachusetts Institute of Technology (USA) approved the MIMIC-III database development. Individual consent to participate was waived because the project did not influence clinical care and all protected health information was anonymized. (Beatriz Nistal-Nuño was formally approved to access this database).

RESULTS

Out of the 9,761 ICU-stays, 570 patients died during the ICU stay. That represents a prevalence of 5.83% of ICU mortality. A total of 750 patients died during the same hospital admission of the corresponding ICU stays, representing a prevalence of 7.68% for in-hospital mortality.

Figure 1 displays the ROC curves for all models evaluated for ICU mortality prediction, which shows an AUROC of 0.9413 for GBM, 0.9311 for DRF, and ≤0.778 for all conventional systems except for serial SOFA (0.8722) (Table 3).

Table 3 shows the AUROC for all models evaluated for in-hospital mortality prediction, with an AUROC of 0.892 for DRF, 0.897 for GBM, and ≤0.7486 for the traditional systems except for serial SOFA (0.8237).

Figure 2 shows the PRCs for the ML methods and all severity-of-illness systems for ICU mortality prediction. The AUPRC was 0.62 for DRF, 0.671 for GBM, and ≤0.184 for all the traditional systems except for serial SOFA (0.587) (Table 3).

Table 3 shows the AUPRC for the ML methods and all severity-of-illness systems for in-hospital mortality prediction. The AUPRC was 0.534 for DRF, 0.599 for GBM, and ≤0.211 for the traditional systems except for serial SOFA (0.519).

The sensitivity was 0.6421 for GBM, 0.6 for DRF, and <0.3 for all the traditional systems except for serial SOFA (0.6316). The specificity was ≥0.9499 for all systems. The PPV was 0.574 for GBM, 0.566 for DRF, and ≤0.476 for all the traditional systems. The NPV was >0.94 for all systems. DOR was <34 for all the traditional systems. However, DOR reached 58.8144 for GBM, and 51.2926 for DRF (Table 3).

The Brier score for the primary endpoint was 0.025 for GBM, 0.028 for DRF, 0.059 for OASIS, 0.05 for SAPS III, and 0.101 for SAPS II. The Brier score for in-hospital mortality prediction was 0.0408 for GBM, 0.0445 for DRF, 0.072 for OASIS, 0.062 for SAPS III, and 0.107 for SAPS II (Table 3).

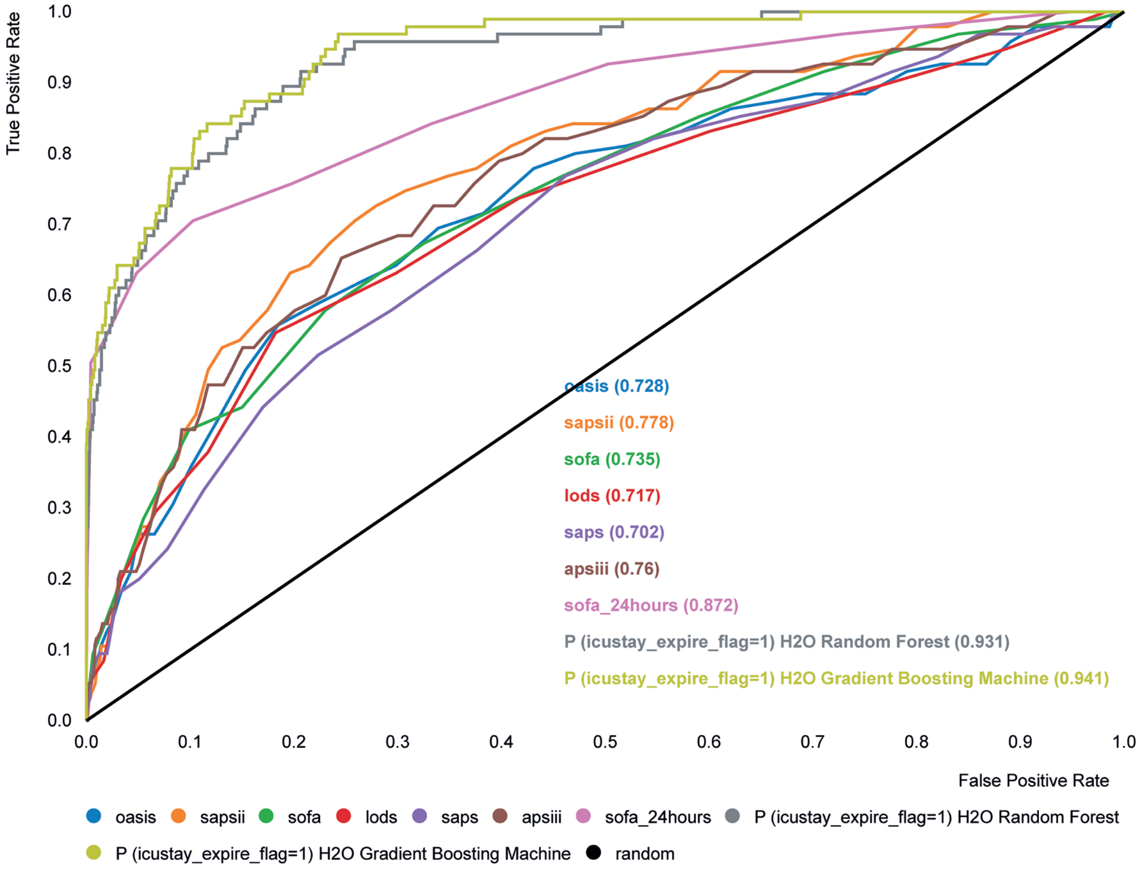


Figure 1. ROC curves for mortality prediction in the cardiac intensive care units for the machine learning methods and severity-of-illness systems

Table 3. Performance of the machine learning methods and severity-of-illness systems for prediction of mortality in the cardiac intensive care units, and in-hospital mortality*

	DRF	GBM	OASIS	SAPS III	SAPS II	LODS	SOFA	Serial SOFA (sofa_24hours)	SAPS
Optimal criterion value[#]	>0.2405	>0.1401	>44	>81	>63	>9	>12	>6	>27
AUROC for ICU mortality (95%CI)[†]	0.9311 (0.907 -0.956)	0.9413 (0.920 -0.963)	0.7282 (0.672 -0.785)	0.7599 (0.709- 0.811)	0.778 (0.730- 0.826)	0.7172 (0.660- 0.775)	0.735 (0.682-0.788)	0.8722 (0.829- 0.915)	0.702 (0.647- 0.757)
AUROC for in-hospital mortality (95%CI)[†]	0.892 (0.865 -0.919)	0.897 (0.869 -0.924)	0.6848 (0.634-0.736)	0.7486 (0.705-0.792)	0.7473 (0.703- 0.792)	0.6901 (0.641- 0.740)	0.6862 (0.638-0.735)	0.8237 (0.782- 0.866)	0.6461 (0.595- 0.697)
AUPRC for ICU mortality (95%CI)[‡]	0.620 (0.519 -0.712)	0.671 (0.571 -0.758)	0.169 (0.106 -0.258)	0.184 (0.119 -0.275)	0.161 (0.0998 -0.249)	0.159 (0.0981 -0.246)	0.171 (0.108- 0.260)	0.587 (0.486- 0.681)	0.132 (0.0777- 0.216)
AUPRC for in-hospital mortality (95%CI)[‡]	0.534 (0.449 -0.617)	0.599 (0.514 -0.679)	0.176 (0.120 -0.250)	0.211 (0.150 -0.288)	0.186 (0.129 -0.262)	0.179 (0.123 -0.254)	0.191 (0.133- 0.267)	0.519 (0.435- 0.603)	0.145 (0.0951- 0.216)
Sensitivity	0.6	0.6421	0.2632	0.2105	0.1895	0.2	0.0947	0.6316	0.1789
Specificity	0.9715	0.9704	0.9499	0.9672	0.9699	0.9666	0.9935	0.9516	0.9699
PPV	0.566	0.574	0.246	0.285	0.281	0.271	0.476	0.447	0.269
NPV	0.975	0.978	0.954	0.952	0.951	0.951	0.947	0.977	0.95
DOR	51.2926	58.8144	6.7435	7.817	7.488	7.2168	16.12	33.4358	6.9882
Brier score** for ICU mortality	0.028	0.025	0.059	0.05	0.101				
Brier score** for in-hospital mortality	0.0445	0.0408	0.072	0.062	0.107				

*Results shown were calculated from testing set (n=1953); [#]The thresholds shown were used to calculate the accuracy statistics, calculated taking into account prevalence of mortality (5.83%) and estimated costs (cost FP=1, cost FN=4, cost TP=0, cost TN=0); [†]The 95%CI was calculated as AUROC ± 1.96 Standard Error; [‡]The 95%CI was calculated with the method of Boyd et al;[49] **The Brier score was calculated as the average squared error of the prediction.
DRF: Distributed Random Forest; GBM: Gradient Boosting Machine; ICU: intensive care unit; PPV: positive predictive value; NPV: negative predictive value; DOR: diagnostic odds ratio.

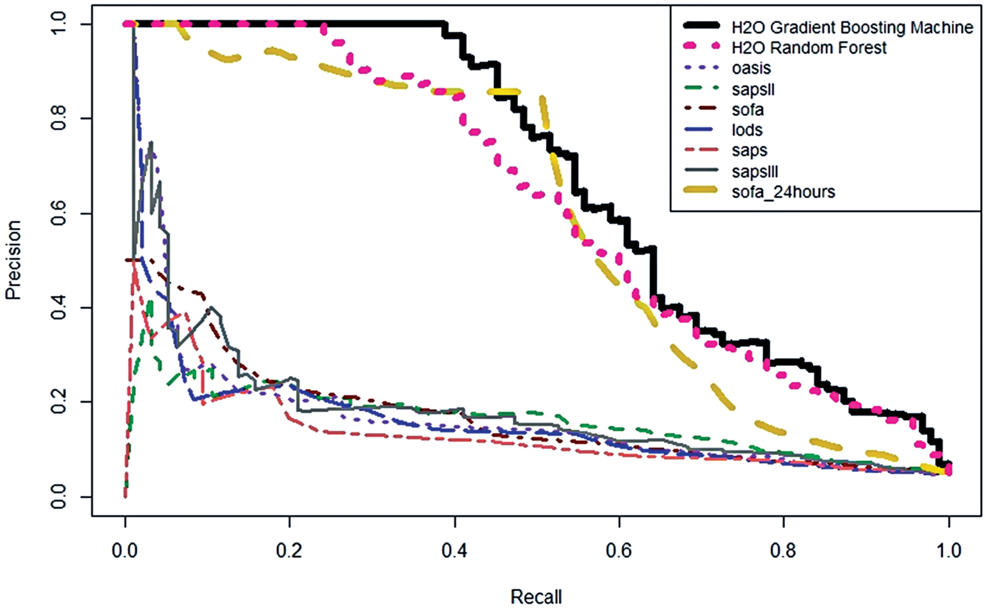


Figure 2. Precision recall curves for the machine learning methods and severity-of-illness systems for cardiac intensive care unit mortality prediction

Calibration curves are displayed in figure 3. The diagonal line represents the ideal calibration curve. The curves for GBM and DRF were substantially closer to this ideal line.

The ML models' AUROCs and AUPRCs were significantly superior ($p<0.05$) to all conventional systems for the primary and secondary endpoints, except for DRF over serial SOFA for the secondary outcome for the AUPRC (Table 3).

In view of the application of these ML models, it is essential for clinicians to know the reasons behind predictions. The SHAP algorithm was applied, which explains patient-specific predictions. The Shapley Value of a feature for a specific patient denotes how much the feature has conferred to the divergence of the actual prediction from the mean prediction.(46,50)

The SHAP algorithm was applied to GBM**.** The ICU stay whose prediction of ICU mortality was chosen to be explained corresponded to a male patient of 71 years old, admitted as elective to the ICU, who did not survive to ICU discharge. This patient had a myocardial infarction and cardiac arrest and received vasopressors and dialysis. He had atrial fibrillation/flutter presenting HR >100 bpm and SBP <88 mmHg. This patient was correctly classified by GBM, assigning him a probability of mortality of 0.8661 (Figure 4).

DISCUSSION

From the results for ICU mortality prediction, GBM showed the highest AUROC followed by DRF. The traditional systems showed much lower values except for serial SOFA. The same pattern was noticed for the secondary outcome. The superiority of the ML models' AUROCs was statistically significant for both outcomes.

For the primary outcome, the AUPRCs were higher for GBM and DRF, and much lower for the traditional systems except for serial SOFA. A similar pattern was obtained for the secondary outcome. The superiority of the ML models' AUPRC was statistically significant for both outcomes, except for the comparison between DRF and serial SOFA for the secondary outcome.

The high values of AUROC and AUPRC for the GBM and DRF indicate that the discriminatory power of these two models for predicting cardiac ICU mortality was excellent, significantly surpassing the traditional systems. The slight drop in AUROC and AUPRC for in-hospital mortality prediction for the ML methods can be explained because the ML models were designed for prediction of ICU mortality.

The sensitivity was much higher for the ML methods over the traditional systems, except for serial SOFA. The specificity was very high for all models. The PPV was low for the traditional systems. However, PPV was intermediate for the ML methods. The difficulty to reach a high PPV is due to the very low frequency of ICU mortality in the cohort. The PPV of 0.476 for SOFA should be interpreted as being lower than that as the optimal cutoff value selected for SOFA turned out to be very high, yielding therefore a very low sensitivity of 0.0947 producing then much fewer FP. The higher PPV for the ML methods is important for a predictor in the ICU, indicating a low rate of FP. However, the cost of FN was established as much higher than FP for the calculation of the optimal threshold values because of the importance of not missing any patients that would end up dying in the ICU. The costs of FN and FP can be adjusted at the criteria of clinicians in order to calculate the optimal thresholds, but when using high thresholds, clinicians would be advised to take more caution in their decision-making for patients estimated at low risk by the model.(31)

Diagnostic odds ratio was substantially higher for the ML methods over the conventional systems, of which serial SOFA showed the best value. This is important as DOR is based on the positive and negative likelihood ratios that are independent of mortality prevalence, while PPV and NPV are highly dependent on mortality prevalence.

The results showed the superiority of serial SOFA over the static SOFA. This was to be expected because in SOFA, calculated from the first 24-hours of ICU admission, for example the cardiovascular measurement is performed based on the inotropes and vasopressors needed. However, extensive use of these medications in the early postoperative stage of cardiac surgery

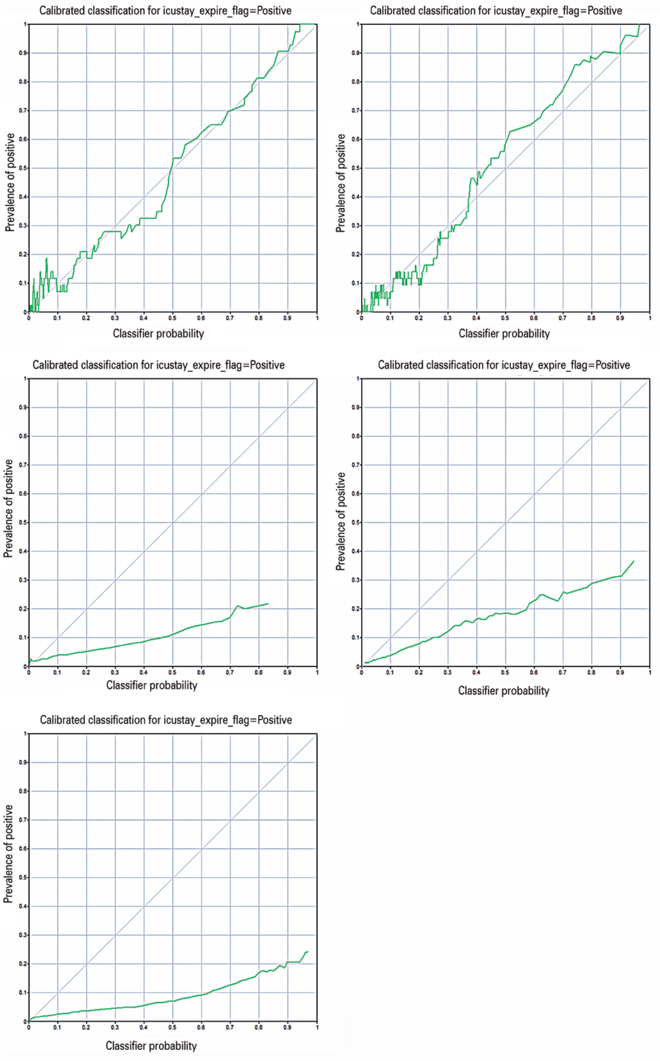
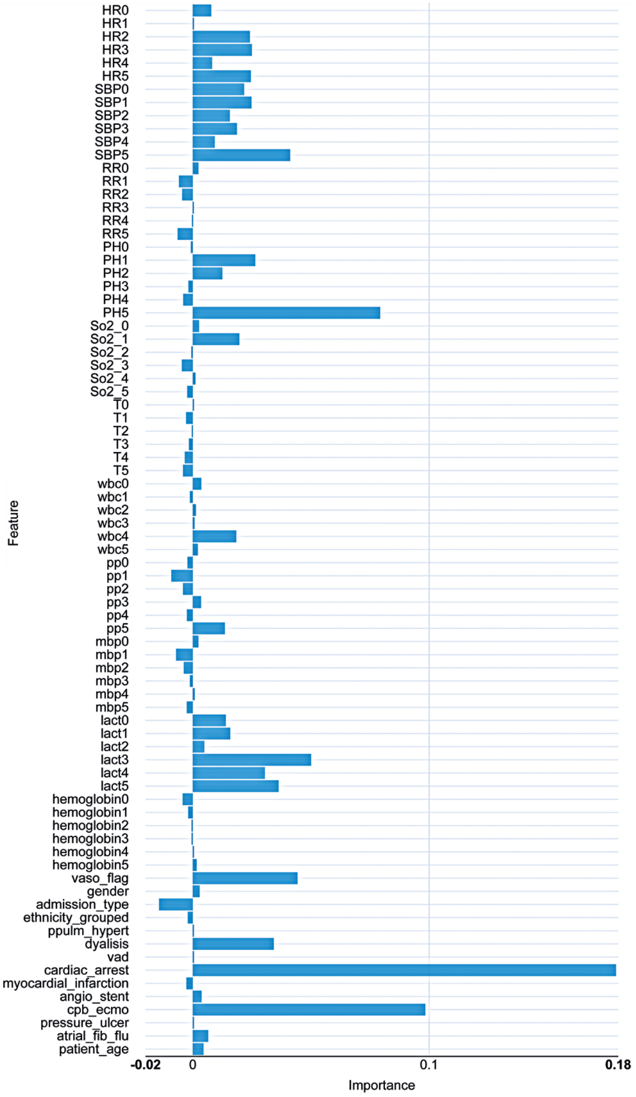


Figure 3. Calibration curves for intensive care unit mortality prediction. Top left: GBM. Top right: DRF. Medium left: OASIS. Medium right: SAPS III. Bottom left: SAPS II. For each probability yielded by the model for the endpoint variable, the plot shows the actual frequencies in the cohort observed for all cases for which the model yielded that probability



HR: heart rate; SBP: systolic blood pressure; RR: respiratory rate; So2: blood oxygen saturation; T: temperature; wbc: white blood cell count; pp: pulse pressure; mbp: mean blood pressure; lact: blood lactate levels; vad: ventricular assist device.

Figure 4. SHAP Algorithm for a correctly predicted non-survivor in the cardiac intensive care unit by the GBM. Shapley values are depicted on the horizontal axis. The cardiac arrest is the feature with the biggest Shapley value, having the greatest contribution towards mortality. Use of vasopressors and dialysis contributed also notably towards mortality. Atrial fibrillation/flutter and elevated age contributed also towards mortality. The elective admission contributed towards survival

might lead to unrealistic high scores for SOFA in all the cardiac surgery patients. The importance of given vasopressors likely increases over time in the ICU, then it can be captured by serial SOFA. The ML models further improve the serial SOFA by including more variables in order to assess the cardiovascular system.[40]

Lower values of the Brier score reflect better calibration. The Brier scores were stronger for the ML methods, which is visualized by the excellent calibration curves obtained for GBM and DRF.

The goal of this research was achieved, as better results were obtained in comparison to previous work.[35] This higher performance is mainly due to stricter patient selection criteria, a slightly wider time window, and the use of powerful ML algorithms in a workflow design that avoids overfitting.

In comparison to LR, the ML methods used have the practical strengths that they are adequate for nonlinear data and nonlinear relationships between features and log odds of outcomes, and they are robust to correlated features and feature distributions.

The lack of transparency is mitigated by the applied explainer algorithm. The SHAP algorithm provides understanding of how the ML algorithms arrived at the individual predictions (Figure 4), giving insight into the importance of features.

The cardiac ICU exemplifies the scenario where the dynamic interdependency of various risk factors on the survival clinical outcome is most prominent, where ML algorithms have unique strengths compared to LR or a weighted summation of scores. When prognoses are not immediately clinically apparent, these ML models could provide clinicians a more informed decision regarding potential short-term survival outcomes.[31,32] The models developed in this work provide the possibility to examine changes in mortality risk over time if predictions are generated at pre-defined intervals with updated patient data.

The utility of these ML models, containing over 60 variables, will likely depend on the ease with which they can be used. For this reason, future work should provide easy-to-use prediction support tools for clinicians. An application should be provided in the form of a user interface accessible in the ICU that would be integrated with available electronic medical records data.[31,32]

The single institution assessment is a limitation of this work. It would be recommended to conduct a multi-center based study to further validate these findings.

CONCLUSION

The excellent predictive abilities of these machine learning models further advance the science of cardiac intensive care unit risk modeling, suggesting they could be used for early recognition of cardiac patients at high risk of mortality so as to improve outcomes. Nonetheless, no scoring model can replace clinical assessment at a patient's bedside, they can only act as an objective instrument in decision making.

AUTHOR CONTRIBUTION

Beatriz Nistal-Nuño: conceptualization, data curation, formal analysis, investigation methodology, project administration, resources, software, supervision, validation, visualization, writing - original draft and writing - review & editing.

AUTHOR INFORMATION

Nistal-Nuño B: http://orcid.org/0000-0003-2210-0726

REFERENCES

1. LaPar DJ, Gillen JR, Crosby IK, Sawyer RG, Lau CL, Kron IL, et al. Predictors of operative mortality in cardiac surgical patients with prolonged intensive care unit duration. J Am Coll Surg. 2013;216(6):1116- 23.
2. Doerr F, Heldwein MB, Bayer O, Sabashnikov A, Weymann A, Dohmen PM, et al. Combination of European System for Cardiac Operative Risk Evaluation (EuroSCORE) and Cardiac Surgery Score (CASUS) to Improve Outcome Prediction in Cardiac Surgery. Med Sci Monit Basic Res. 2015;21:172- 8.
3. Rahmanian PB, Kröner A, Langebartels G, Özel O, Wippermann J, Wahlers T. Impact of major non-cardiac complications on outcome following cardiac surgery procedures: logistic regression analysis in a very recent patient cohort. Interact Cardiovasc Thorac Surg. 2013;17(2):319- 27.
4. Doerr F, Badreldin AM, Bender EM, Heldwein MB, Lehmann T, Bayer O, et al. Outcome prediction in cardiac surgery: the first logistic scoring model for cardiac surgical intensive care patients. Minerva Anestesiol. 2012;78(8):879- 86.
5. Vincent JL, Moreno R, Takala J, Willatts S, De Mendonça A, Bruining H, et al. The SOFA (Sepsis-related Organ Failure Assessment) score to describe organ dysfunction/failure. On behalf of the Working Group on Sepsis-Related Problems of the European Society of Intensive Care Medicine. Intensive Care Med. 1996;22(7):707-10.
6. Le Gall JR, Loirat P, Alperovitch A, Glaser P, Granthil C, Mathieu D, et al. A simplified acute physiology score for ICU patients. Crit Care Med. 1984;12(11):975- 7.
7. Le Gall JR, Lemeshow S, Saulnier F. A new Simplified Acute Physiology Score (SAPS II) based on a European/North American multicenter study. JAMA. 1993;270(24):2957- 63.
8. Poncet A, Perneger TV, Merlani P, Capuzzo M, Combescure C. Determinants of the calibration of SAPS II and SAPS 3 mortality scores in intensive care: a European multicenter study. Crit Care. 2017;21(1):85.
9. Le Gall JR, Klar J, Lemeshow S, Saulnier F, Alberti C, Artigas A, et al. The Logistic Organ Dysfunction system. A new way to assess organ dysfunction in the intensive care unit. ICU Scoring Group. JAMA. 1996;276(10):802-10.
10. Johnson AE, Kramer AA, Clifford GD. A new severity of illness scale using a subset of Acute Physiology And Chronic Health Evaluation data elements shows comparable predictive accuracy. Crit Care Med. 2013;41(7):1711- 8.
11. Koponen T, Karttunen J, Musialowicz T, Pietiläinen L, Uusaro A, Lahtinen P. Vasoactive-inotropic score and the prediction of morbidity and mortality after cardiac surgery. Br J Anaesth. 2019;122(4):428- 36.

12. Shahian DM, Jacobs JP, Badhwar V, Kurlansky PA, Furnary AP, Cleveland JC Jr, et al. The Society of Thoracic Surgeons 2018 Adult Cardiac Surgery Risk Models: Part 1-Background, Design Considerations, and Model Development. Ann Thorac Surg. 2018;105(5):1411- 8.

13. Nashef SA, Roques F, Sharples LD, Nilsson J, Smith C, Goldstone AR, et al. EuroSCORE II. Eur J Cardiothorac Surg. 2012;41(4):734- 45.

14. Roques F, Michel P, Goldstone AR, Nashef SA. The logistic EuroSCORE. Eur Heart J. 2003;24(9):881- 2.

15. Roques F, Nashef SA, 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(6):816- 22.

16. Higgins TL, Estafanous FG, Loop FD, Beck GJ, Blum JM, Paranandi L. Stratification of morbidity and mortality outcome by preoperative risk factors in coronary artery bypass patients. A clinical severity score. JAMA. 1992;267(17):2344- 8.

17. Magovern JA, Sakert T, Magovern GJ Jr, Benckart DH, Burkholder JA, Liebler GA, et al. A model that predicts morbidity and mortality after coronary artery bypass graft surgery. J Am Coll Cardiol. 1996;28(5):1147- 53.

18. Nilsson J, Algotsson L, Höglund P, Lührs C, Brandt J. Comparison of 19 pre-operative risk stratification models in open-heart surgery. Eur Heart J. 2006;27(7):867- 74.

19. Ariyaratne TV, Billah B, Yap CH, Dinh D, Smith JA, Shardey GC, et al. An Australian risk prediction model for determining early mortality following aortic valve replacement. Eur J Cardiothorac Surg. 2011;39(6):815- 21.

20. Ambler G, Omar RZ, Royston P, Kinsman R, Keogh BE, Taylor KM. Generic, simple risk stratification model for heart valve surgery. Circulation. 2005;112(2):224- 31.

21. Hannan EL, Wu C, Bennett EV, Carlson RE, Culliford AT, Gold JP, et al. Risk index for predicting in-hospital mortality for cardiac valve surgery. Ann Thorac Surg. 2007;83(3):921- 9.

22. Wang C, Tang YF, Zhang JJ, Bai YF, Yu YC, Zhang GX, et al. Comparison of four risk scores for in-hospital mortality in patients undergoing heart valve surgery: a multicenter study in a Chinese population. Heart Lung. 2016;45(5):423- 8.

23. Gaies MG, Gurney JG, Yen AH, Napoli ML, Gajarski RJ, Ohye RG, et al. Vasoactive-inotropic score as a predictor of morbidity and mortality in infants after cardiopulmonary bypass. Pediatr Crit Care Med. 2010;11(2):234- 8.

24. Lamarche Y, Elmi-Sarabi M, Ding L, Abel JG, Sirounis D, Denault AY. A score to estimate 30-day mortality after intensive care admission after cardiac surgery. J Thorac Cardiovasc Surg. 2017;153(5):1118- 25.e4.

25. Hekmat K, Kroener A, Stuetzer H, Schwinger RH, Kampe S, Bennink GB, et al. Daily assessment of organ dysfunction and survival in intensive care unit cardiac surgical patients. Ann Thorac Surg. 2005;79(5):1555- 62.

26. Granger CB, Goldberg RJ, Dabbous O, Pieper KS, Eagle KA, Cannon CP, Van De Werf F, Avezum A, Goodman SG, Flather MD, Fox KA; Global Registry of Acute Coronary Events Investigators. Predictors of hospital mortality in the global registry of acute coronary events. Arch Intern Med. 2003;163(19):2345-53.

27. Peterson PN, Rumsfeld JS, Liang L, Albert NM, Hernandez AF, Peterson ED, Fonarow GC, Masoudi FA; American Heart Association Get With the Guidelines-Heart Failure Program. A validated risk score for in-hospital mortality in patients with heart failure from the American Heart Association get with the guidelines program. Circ Cardiovasc Qual Outcomes. 2010;3(1):25-32.

28. Fonarow GC, Adams KF Jr, Abraham WT, Yancy CW, Boscardin WJ; ADHERE Scientific Advisory Committee, Study Group, and Investigators. Risk stratification for in-hospital mortality in acutely decompensated heart failure: classification and regression tree analysis. JAMA. 2005;293(5):572- 80.

29. Jentzer JC, Anavekar NS, Bennett C, Murphree DH, Keegan MT, Wiley B, et al. Derivation and Validation of a Novel Cardiac Intensive Care Unit Admission Risk Score for Mortality. J Am Heart Assoc. 2019;8(17):e013675.

30. Johnson AE, Mark RG. Real-time mortality prediction in the Intensive Care Unit. AMIA Annu Symp Proc. 2018;2017:994- 1003.

31. Kanwar MK, Lohmueller LC, Kormos RL, Teuteberg JJ, Rogers JG, Lindenfeld J, et al. A Bayesian Model to Predict Survival After Left Ventricular Assist Device Implantation. JACC Heart Fail. 2018;6(9):771- 9.

32. Loghmanpour NA, Kormos RL, Kanwar MK, Teuteberg JJ, Murali S, Antaki JF. A Bayesian Model to Predict Right Ventricular Failure Following Left Ventricular Assist Device Therapy. JACC Heart Fail. 2016;4(9):711- 21.

33. Zhang Z, Chen L, Xu P, Hong Y. Predictive analytics with ensemble modeling in laparoscopic surgery: a technical note. Laparosc Endosc Robot Surg. 2022;5(1):25- 34.

34. Ranucci M, Castelvecchio S, Menicanti L, Frigiola A, Pelissero G. Risk of assessing mortality risk in elective cardiac operations: age, creatinine, ejection fraction, and the law of parsimony. Circulation. 2009;119(24):3053- 61.

35. Nistal-Nuño B. Machine learning applied to a Cardiac Surgery Recovery Unit and to a Coronary Care Unit for mortality prediction. J Clin Monit Comput. 2022;36(3):751- 63.

36. Johnson AE, Pollard TJ, Shen L, Lehman LW, Feng M, Ghassemi M, et al. MIMIC-III, a freely accessible critical care database. Sci Data. 2016;3(1):160035.

37. Johnson A, Pollard T, Mark R. MIMIC-III Clinical Database (version 1.4). PhysioNet. 2016 [cited 2019 July 29]. Available from: https://doi.org/10.13026/C2XW26

38. Nistal-Nuño B. Replication Data for: Ensemble learning algorithms predicting patient mortality at Cardiac Intensive Care Units. Harvard Dataverse, version 1; 2022 [cited 2022 Nov 19]. Available from: https://doi.org/10.7910/DVN/WPZIV1

39. Ferreira FL, Bota DP, Bross A, Mélot C, Vincent JL. Serial evaluation of the SOFA score to predict outcome in critically ill patients. JAMA. 2001;286(14):1754- 8.

40. Pätilä T, Kukkonen S, Vento A, Pettilä V, Suojaranta-Ylinen R. Relation of the Sequential Organ Failure Assessment score to morbidity and mortality after cardiac surgery. Ann Thorac Surg. 2006;82(6):2072- 8.

41. Johnson AE, Stone DJ, Celi LA, Pollard TJ. The MIMIC Code Repository: enabling reproducibility in critical care research. J Am Med Inform Assoc. 2018;25(1):32- 9.

42. Hayashi Y, Endoh H, Kamimura N, Tamakawa T, Nitta M. Lactate indices as predictors of in-hospital mortality or 90-day survival after admission to an intensive care unit in unselected critically ill patients. PLoS One. 2020;15(3):e0229135.

43. Parsonnet V, Dean D, Bernstein AD. A method of uniform stratification of risk for evaluating the results of surgery in acquired adult heart disease. Circulation. 1989;79(6 Pt 2 Suppl I):I3- 12.

44. Suarez-Pierre A, Fraser CD, Zhou X, Crawford TC, Lui C, Metkus TS, et al. Predictors of operative mortality among cardiac surgery patients with prolonged ventilation. J Card Surg. 2019;34(9):759- 66.

45. Kaewprag P, Newton C, Vermillion B, Hyun S, Huang K, Machiraju R. Predictive models for pressure ulcers from intensive care unit electronic health records using Bayesian networks. BMC Med Inform Decis Mak. 2017;17(S2 Suppl 2):65.

46. Berthold MR, Cebron N, Dill F, Gabriel TR, Kötter T, Meinl T, et al. KNIME: The Konstanz Information Miner. In: Preisach C, Burkhardt H, Schmidt-Thieme L, Decker R, editors. Data Analysis, Machine Learning and Applications. Studies in Classification, Data Analysis, and Knowledge Organization. Berlin: Springer; 2008. pp. 319- 26.

47. Davis J, Goadrich M. The relationship between precision-recall and ROC curves. In: Proceedings of the 23rd international conference on Machine learning (ICML '06). New York, NY: Association for Computing Machinery; 2006. pp. 233- 40.

48. 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(3):837- 45.

49. Boyd K, Eng KH, Page CD. Area under the Precision-Recall Curve: Point Estimates and Confidence Intervals. In: Blockeel H, Kersting K, Nijssen S, Železný F, editors. Machine Learning and Knowledge Discovery in Databases. ECML PKDD 2013. Lecture Notes in Computer Science. Vol 8190. Berlin, Heidelberg: Springer; 2013. pp. 451- 66.

50. Štrumbelj E, Kononenko I. Explaining prediction models and individual predictions with feature contributions. Knowl Inf Syst. 2014;41(3):647- 65.