Original Article

Development of machine learning models for mortality risk prediction after cardiac surgery

Yunlong Fan[1,2#], Junfeng Dong[3#], Yuanbin Wu[1,2], Ming Shen[4], Siming Zhu[1,2], Xiaoyi He[1,2], Shengli Jiang[2], Jiakang Shao[1], Chao Song[1,2]

[1]Medical School of Chinese PLA, Beijing, China; [2]Department of Cardiovascular Surgery, the First Medical Centre of Chinese PLA General Hospital, Beijing, China; [3]Department of Organ Transplantation, Changzhen Hospital, Navy Medical University, Shanghai, China; [4]Department of Cardiology, The First Hospital of Hebei Medical University, Shijiazhuang, China

Contributions: (I) Conception and design: Y Fan, J Dong, C Song, J Shao; (II) Administrative support: S Jiang, C Song; (III) Provision of study materials or patients: Y Wu; (IV) Collection and assembly of data: S Zhu, J Shao; (V) Data analysis and interpretation: Y Fan, M Shen, X He; (VI) Manuscript writing: All authors; (VII) Final approval of manuscript: All authors.

[#]These authors contributed equally to this work and should be considered as co-first authors.

Correspondence to: Chao Song, MD. Department of Cardiovascular Surgery, Chinese PLA General Hospital, No. 28 Fuxing Road, Haidian District, Beijing 100853, China. Email: SongC301hospital@163.com; Jiakang Shao, MD. Medical School of Chinese PLA, No. 28, Fuxing Road, Haidian District, Beijing 100853, China. Email: ShaoJK301hospital@163.com.

Background: We developed machine learning models that combine preoperative and intraoperative risk factors to predict mortality after cardiac surgery.

Methods: Machine learning involving random forest, neural network, support vector machine, and gradient boosting machine was developed and compared with the risk scores of EuroSCORE I and II, Society of Thoracic Surgeons (STS), as well as a logistic regression model. Clinical data were collected from patients undergoing adult cardiac surgery at the First Medical Centre of Chinese PLA General Hospital between December 2008 and December 2017. The primary outcome was post-operative mortality. Model performance was estimated using several metrics, including sensitivity, specificity, accuracy, and area under the receiver operating characteristic curve (AUC). The visualization algorithm was implemented using Shapley's additive explanations.

Results: A total of 5,443 patients were enrolled during the study period. The mean EuroSCORE II score was 3.7%, and the actual in-hospital mortality rate was 2.7%. For predicting operative mortality after cardiac surgery, the AUC scores were 0.87, 0.79, 0.81, and 0.82 for random forest, neural network, support vector machine, and gradient boosting machine, compared with 0.70, 0.73, 0.71, and 0.74 for EuroSCORE I and II, STS, and logistic regression model. Shapley's additive explanations analysis of random forest yielded the top-20 predictors and individual-level explanations for each prediction.

Conclusions: Machine learning models based on available clinical data may be superior to clinical scoring tools in predicting postoperative mortality in patients following cardiac surgery. Explanatory models show the potential to provide personalized risk profiles for individuals by accounting for the contribution of influencing factors. Additional prospective multicenter studies are warranted to confirm the clinical benefit of these machine learning-driven models.

Keywords: Mortality prediction; machine learning; cardiac surgery; artificial intelligence; random forest

Submitted Oct 09, 2021. Accepted for publication Dec 28, 2021.
doi: 10.21037/cdt-21-648
View this article at: https://dx.doi.org/10.21037/cdt-21-648

Introduction

As patients undergoing cardiac surgery are at high risk for intra-operative and post-operative complications, a preoperative risk-benefit evaluation is of great importance. The risks of surgery are in some cases unpredictable and the decision to proceed with surgery based on individual circumstances may be complicated.

Clinical risk scoring systems can support surgeons in advising patients during the decision-making process and may also assist in managing surgical outcomes and cost-benefit analysis. A number of models for risk stratification have been developed, included the European EuroSCORE II (1) and the American Society of Thoracic Surgeons (STS) scores (2), both of which have been studied the most and are the most widely used among the many that have been proposed. However, several studies have shown limitations of the scores in certain procedures or patient subgroups, particularly in overestimating the risk in high-risk patient subgroups (3-6). This potentially enables surgeons and medical centers to make false assurances about efficacy performance and compromises decision making. Existing risk scores are developed from logistic regression analysis, which potentially compromises the complex interactions between features and the nonlinear relationships between features and outcomes, negatively impacting their predictive efficacy (7,8). In addition, traditional multivariate linear analysis methods limit the number of relevant variables that may be clinically meaningful, especially for some intraoperative events, such as operative time and blood transfusion (7). Alternatively, machine learning algorithms are not confined to linear relationships or to the number of variables included in the analysis, but one can automatically learn from the data and incorporate it into the analysis (9,10).

Artificial intelligence is a relatively broad concept that refers to the capability of a computer system to perform tasks in an intelligent fashion, as humans generally do. Machine learning is a branch of artificial intelligence, which is characterized by the ability of computer algorithms to learn from and understand the meaning of data, build recognition patterns, and automatically establish models for decision analysis (11). There are a number of real-world applications of machine learning, such as autonomous vehicles (12), picture and voice recognition (13), and web content recommendation algorithms (14). Though machine learning has been widely used in medical and health care, its specifically potential applications in cardiac surgery are not yet well developed. In this study, we attempted to develop a predictive model for the risk of postoperative hospital mortality in cardiac surgery patients by using machine learning algorithms. We present the following article in accordance with the TRIPOD reporting checklist (available at https://cdt.amegroups.com/article/view/10.21037/cdt-21-648/rc).

Methods

Study population

A flow chart depicting the study protocol is presented in *Figure 1*. A total of 5,443 consecutively enrolled adult inpatients, who underwent cardiac surgery with cardiopulmonary bypass at the First Medical Centre of Chinese PLA General Hospital between December 2008 and December 2017, were screened from the electronic medical record system. Exclusion criteria: (I) age <18 years; (II) data loss >10%. The study was conducted in accordance with the Declaration of Helsinki (as revised in 2013). The study was approved by the institutional review board of Chinese PLA General Hospital (S2020-121-02). Informed consent was waived due to the observational nature of the study.

Data collection

Clinically related data were recruited from institutional electronic medical records, cardiopulmonary bypass records, and the Anesthesia Information Management System (AIMS). Data on demographic characteristics, comorbidities, preoperative medical status, surgery-related information, STS scores and EuroSCORE scores were used as predictors. *Table 1* presents the list of factors used as predictors during modelling. Predicted risk of death was calculated for all patients, including STS scores, EuroSCORE I and EuroSCORE II scores, as all patients included in the study underwent the evaluation index of the procedure. In the absence of data records required to calculate EuroSCORE and STS, we assumed that this risk factor was not present (equivalent to the referenced level).

Primary outcome

The primary outcome of interest was post-operative mortality, as defined by the STS Adult Cardiac Surgery (version 2.81). It involved (I) all deaths, regardless of

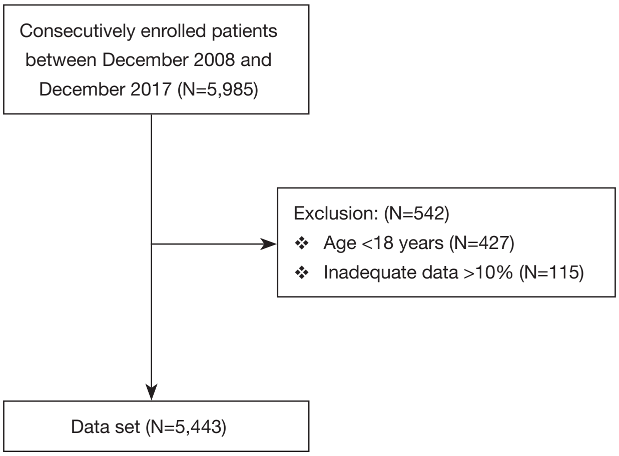
Consecutively enrolled patients between December 2008 and December 2017 (N=5,985)
Exclusion: (N=542)
❖ Age <18 years (N=427)
❖ Inadequate data >10% (N=115)
Data set (N=5,443)

Figure 1 Flow chart showing patient selection and exclusion criteria. A total of 5,985 consecutive patients who underwent cardiac surgery were included initially. 542 patients were excluded according to the exclusion criteria, and a total of 5,443 patients were eventually included in the study.

cause, that occurred during the hospital stay in which the procedure is performed, even after 30 days (including patients transferred to other critical care units); and (II) all deaths, irrespective of cause, that occurred after discharged but before the end of the 30th post-operative day. If a patient was discharged, they would be scheduled for a 30-day appointment or phone call to document the morbidity and mortality.

Models development

The whole database was randomly divided into two subsets: 80% for model training and 20% for model testing. The hyperparameters in each model were developed through a 10-fold cross-validation process. Specifically, the 80% training set was randomly split into 90% and 10% clusters for model training and testing, respectively. A 10-fold cross-validation process was to repeated this process 10 times, so as to obtain 10 individual probabilities. And the final tuning hyperparameters of each model were the average of these 10 individual probabilities. Finally, the models built from the training set were then fed into the 20% testing set for model validation.

Machine learning technique

We used univariate analysis to select variables for inclusion in the multivariate logistic regression models based on p values of less than 0.1. We applied the following machine learning approaches: (I) random forest. Random forest is an aggregate-learning approach for classification and regression analysis. The theory of random forest involves constructing multiple decision trees and returning the classification outcome given by the average predicted value of individual trees (15). Compared to conventional decision trees, random forest is more robust to over-fitting. (II) Neural network. Neural network is a functional network aimed to identify potential relationships in a set of data by mimicking the processes operating in the human brain. Neural networks, typically do not require programming with a particular rule to define what to expect from the input. Instead, neural network algorithms learn from dealing with labeled examples (i.e., data with an "answer") provided during modelling, and then use this key to learn what features are needed to construct the correct answer. By the time a sufficient number of samples have been handled, the algorithm can start processing new inputs and successfully output the correct answers. (III) Support vector machine. In a support vector machine, each patient's feature is a point in the space which is mapped into different classes of examples partitioned by a gap as wide as possible, thus parceling the input feature space into hyperplanes (the well-known kernel trick) (16). (IV) Gradient boosting machine. Gradient boosting machine is a member of supervised machine learning methods for classification problems that can be highly customized. A gradient boosting machine yields predictions based on assortment of weak prediction models (e.g., decision trees) (17). By using boosting approaches and bias-reducing meta-algorithms, gradient boosting machine can turn a set of weak predictors into strong ones. Besides, model performance was estimated using several metrics, including sensitivity, specificity, accuracy, and AUC.

Statistical analysis

Continuous variables were expressed as mean and standard deviation (SD). Categorical variables were presented as percentages. The Student's *t*-test was used to compare continuous variables, and the chi-squared test and Fischer's exact test were used for analysis of categorical variables. A significant difference was considered at P<0.05. Our open-source analysis was conducted in python version 3.6 (https://www.python.org) using the following libraries and packages: Pandas data analysis library, NumPy extension module, Matplotlib and SHapley Additive exPlanations

Table 1 Perioperative characteristics of patients

Characteristic	Overall	Survival	Mortality	P value
Patient population, n	5,443	5,296	147	
Demographic data				
Male, n (%)	3,678 (67.6)	3,569 (67.4)	109 (74.1)	0.011
Age (years), mean (SD)	63.1 (14.0)	63.0 (13.8)	67.8 (15.8)	<0.001*
BMI (kg/m^2), mean (SD)	25.8 (4.7)	25.8 (4.6)	26.0 (4.9)	0.019*
SBP (mmHg), mean (SD)	126 (7.3)	126 (7.3)	131 (7.8)	0.002*
DBP (mmHg), mean (SD)	72 (6.5)	72 (6.3)	73 (6.5)	0.436
MAP (mmHg), mean (SD)	91 (7.1)	90 (7.1)	93 (7.1)	0.014*
Smoking, n (%)	1,714 (31.5)	1,665 (31.4)	49 (33.3)	0.213
Alcohol, n (%)	1,143 (21.0)	1,110 (21.0)	33 (22.4)	0.121
EUROSCORE II score, mean (SD)	3.7 (4.6)	3.7 (4.6)	11.3 (8.5)	<0.001*
Medical history, n (%)				
Dyslipidemia	615 (11.3)	594 (11.2)	21 (14.3)	0.013*
Diabetes mellitus	1,214 (22.3)	1,174 (22.3)	40 (27.2)	<0.001*
Hypertension	2,231 (41.0)	2,145 (40.5)	86 (58.5)	<0.011*
Chronic kidney disease	228 (4.2)	214 (4.0)	14 (9.5)	<0.001*
Active endocarditis	125 (2.3)	117 (2.2)	8 (5.4)	0.016*
Neurological dysfunction	125 (2.3)	120 (2.3)	5 (3.4)	0.018*
Preoperative condition, n (%)				
MI within 90 days	348 (6.4)	333 (6.3)	15 (10.2)	0.026*
Critical preoperative state	773 (14.2)	745 (14.1)	28 (19.0)	<0.001*
Previous cardiac surgery	983 (18.1)	950 (17.9)	33 (22.4)	0.322
On dialysis	59 (1.1)	55 (1.0)	4 (2.7)	<0.001*
Atrial fibrillation	1,251 (23.0)	1,210 (22.8)	41 (27.9)	0.043*
Preoperative medications, n (%)				
β-block	2,667 (49.0)	2,572 (48.6)	95 (64.6)	<0.001*
ACEi	615 (11.3)	595 (11.2)	20 (13.6)	0.028*
ARB	718 (13.2)	695 (13.1)	23 (15.6)	0.013*
Aspirin	1,251 (23.0)	1,214 (22.9)	37 (25.2)	0.105
Insulin	816 (15.0)	783 (14.8)	33 (22.4)	<0.001*

Table 1 (*continued*)

Table 1 (*continued*)

Characteristic	Overall	Survival	Mortality	P value
Laboratory findings, mean (SD)				
Hgb	131.0 (14.3)	131.0 (14.4)	128.0 (18.9)	<0.001*
RBC	4.29 (0.58)	4.29 (0.59)	4.18 (0.73)	<0.001*
HCT	0.38 (0.04)	0.38 (0.04)	0.35 (0.05)	<0.001*
Scr	79.2 (15.4)	79.1 (15.4)	88.6 (18.3)	<0.001*
Creatinine clearance	80.1 (16.3)	80.1 (16.3)	72.3 (19.8)	<0.001*
Urea nitrogen	6.3 (1.3)	6.4 (1.3)	6.9 (1.8)	<0.001*
Type of surgery, n (%)				
Valve surgery only	1,994 (36.6)	1,978 (37.3)	16 (10.9)	0.001*
CABG only	2,533 (46.5)	2,473 (46.7)	60 (40.8)	
CABG + valve	609 (11.2)	570 (10.8)	39 (26.5)	
Surgery on thoracic aorta	307 (5.6)	275 (5.2)	32 (21.8)	
Minimally invasive	1,034 (19.0)	1,112 (21.0)	22 (15.0)	<0.001*
Intraoperative variables, mean (SD)				
Anesthesia time	5.4 (1.1)	5.5 (1.1)	6.0 (1.4)	<0.001*
Operation time	4.6 (1.0)	4.6 (1.0)	5.2 (1.3)	<0.001*
CPB time	119.9 (42.1)	120.0 (42.2)	139.5 (49.8)	<0.001*
Cross clamp time	91.0 (22.3)	90.9 (22.3)	99.8 (32.6)	<0.001*
Perioperative blood loss (mL/kg/h)	1.21 (0.5)	1.21 (0.5)	1.38 (0.6)	<0.001*
Urine output (mL/kg/h), mean (SD)	2.73 (1.50)	2.73 (1.50)	1.79 (1.10)	<0.001*
pRBC transfusion during surgery	2.5 (0.9)	2.5 (0.9)	4 (2.2)	<0.001*
FFP transfusion during surgery	4.6 (2.2)	4.6 (2.2)	5.5 (3.1)	<0.001*
PLT transfusion during surgery	0.8 (0.3)	0.8 (0.3)	1.5 (0.5)	<0.001*

Data are presented as mean (SD) or number (%) and statistical analysis was performed with Student's *t*-test, Pearson or Fisher Chi-square test. *, P value <0.05 when comparing patients with and without mortality following cardiac surgery. BMI, body mass index; SBP, systolic blood pressure; DBP, diastolic blood pressure; MAP, mean arterial pressure; MI, myocardial infarction; ACEi, angiotensin-converting-enzyme inhibitor; ARB, angiotensin II receptor blocker; Hgb, hemoglobin; RBC, red blood cell count; HCT, hematocrit; Scr, serum creatinine; CABG, coronary artery bypass grafting; CPB, cardiopulmonary bypass; pRBC, packed red blood cell; FFP, fresh frozen plasma; PLT, platelet.

(SHAP) package for statistical data visualization, and scikit-learn machine learning library. The Shapley Additive Explanations method (18) was used to achieve the explanatory modeling of the black box. The explanations of the top 20 factors in the random forest were visualized in SHAP summary plots. In addition, 2 individual patients with correct classification were displayed as force plots. These 2 patients were selected as showing high and low risk of mortality.

Results

Characteristics of patients

Data concerning 5,443 consecutive patients, who underwent cardiac surgery from December 2008 to December 2017, were collected from the electronic medical record system. Overall, the mean age was 63.1 years, and males comprised 67.6% of the study cohort. The most frequent procedures were isolated coronary artery bypass grafting (CABG)

(46.5%), followed by isolated valve surgery (36.6%). The overall proportions of missing data for the EuroSCORE covariates were very small (1.4%), yielding a mean (SD) EuroSCORE II of 3.7% (4.6%).

Postoperative mortality

Baseline characteristics of patients stratified by mortality were summarized in *Table 1*. In summary, there were 147 (2.7%) patients who died within 30 post-operative day, included 105 patients that died within 30 days during hospitalization, and 42 who died within 30 days but after hospital discharge. For these patients, age was significantly older [63.0 (13.8) *vs.* 67.8 (15.8) years, P<0.001], with prolonged duration of operative time [4.6 (1.0) *vs.* 5.2 (1.3) hours, P<0.001], cardiopulmonary bypass time (CPB) [120.0 (42.2) *vs.* 139.5 (49.8) minutes, P<0.001], and cross-clamp time [90.9 (22.3) *vs.* 99.8 (32.6) minutes, P<0.001], as compared to those who survival. Univariate analysis suggested a strong correlation between mortality and the factors included in EuroSCORE II, with the exception of previous cardiac surgery (*Table 2*).

Model discrimination

the performance of different models in predicting mortality, including EuroSCORE I and II, STS models, logistic regression models, and different machine learning models, was summarized in *Table 3*. All prediction models displayed favorable discrimination (AUC scores ≥0.7). By using 10-fold cross-validation to tune the hyperparameters, we observed a slight improvement in the predictive power of the models. Among all models tested, the random forest model with 10-fold cross-validation had the most robust predictive ability (AUC: 0.87), coupled with better sensitivity (0.83), specificity (0.92) and accuracy (0.89).

Notably, the logistic regression model based on this institutional data yielded an AUC score of 0.74. The AUC metrics for the STS model, EuroSCORE I and II, were 0.71, 0.70, and 0.73, respectively. In predicting mortality after cardiac surgery in all patients included in the study, the AUC metrics were significantly lower in the EuroSCORE I and II, STS model, and logistic regression model than in the random forest model (all, P<0.001).

Model visualization

The interpretive model of the best performer in the supervised machine learning algorithms, the random forest, was shown in *Figure 2*. The top 20 risk factors in the random forest classifier involved age, body mass index (BMI), left ventricular ejection fraction (LVEF), smoking, EuroSCORE II Score, chronic lung disease, diabetes mellitus, active endocarditis, myocardial infarction (MI) within 90 days, critical preoperative state, previous cardiac surgery, emergency, hemoglobin (Hgb), creatinine clearance, serum creatinine (Scr), surgery on thoracic aorta, operation time, CPB time, perioperative blood loss, packed red blood cell (P-RBC) transfusion during surgery. Many of those factors were covered in the EuroSCORE II model. Even so, there was a statistically significant difference between random forest and EuroSCORE II model in the predictive ability (e.g., AUC) of the risk of mortality for the whole study cohort, implying a real difference in predicting outcomes of interest within these models.

By comparison with the EuroSCORE II factors, the explanatory models for random forest showed similarities and some key discrepancies (*Table 2*, *Figure 2*). The variables that were agreed to be highly predictive in both models were age, LVEF, chronic lung disease, diabetes mellitus, active endocarditis, MI within 90 days, critical preoperative state, previous cardiac surgery, emergency, creatinine clearance, Scr, and surgery on thoracic aorta. Regarding disparities, the variables that were highly valued by EuroSCORE II rather than random forest were gender, extracardiac arteriopathy, poor mobility, New York Heart Association (NYHA) classification, and systolic pulmonary artery pressure. On the contrary, those covariates deemed to have an effect in random forest but not in EuroSCORE II were BMI, smoking, Hgb, operation time, CPB time, perioperative blood loss, and P-RBC transfusion during surgery. In addition, *Figure 3* displayed SHAP plot for individual predictions with correct classification in patients with high and low risk of mortality.

Discussion

The aim of this study was to perform the prediction of mortality in patients after cardiac surgery by commonly used machine learning models, incorporating preoperative and intraoperative factors during the modeling process. This study evaluated the risk of mortality after cardiac surgery and the main results were (I) machine learning models showed good predictive power in risk assessment compared to established risk scores, (II) random forest was a better predictor of mortality with higher discrimination (AUC),

Table 2 Distribution of features included in the EuroSCORE II stratified by mortality

Characteristic	Survival	Mortality	P value
Patient population, n	5,296	147	
Age, mean (SD)	63.0 (13.8)	67.8 (13.8)	<0.001*
Male, n (%)	3,569 (67.4)	109 (74.1)	0.011*
Creatinine clearance: 50–85 mL/min, n (%)	2,420 (45.7)	70 (47.6)	0.138
Creatinine clearance: <50 mL/min, n (%)	313 (5.9)	32 (21.8)	<0.001*
On dialysis, n (%)	55 (1.0)	4 (2.7)	<0.001*
Extracardiac arteriopathy, n (%)	477 (9.0)	25 (17.0)	<0.001*
Poor mobility, n (%)	120 (2.3)	5 (3.4)	0.018*
Previous cardiac surgery, n (%)	950 (17.9)	33 (22.4)	0.322
Chronic lung disease, n (%)	1,112 (21.0)	40 (27.2)	<0.001*
Active endocarditis, n (%)	117 (2.2)	8 (5.4)	0.016*
Critical preoperative state, n (%)	745 (14.1)	28 (19.0)	<0.001*
Diabetes on insulin, n (%)	435 (8.2)	22 (15.0)	<0.001*
NYHA class 2, n (%)	2,208 (41.7)	52 (35.4)	0.231
NYHA class 3	1,641 (31.0)	51 (34.7)	0.043*
NYHA class 4, n (%)	233 (4.4)	14 (9.5)	<0.001*
CCS class 4 angina, n (%)	169 (3.2)	19 (12.9)	<0.001*
LVEF 31–50%, n (%)	1,135 (21.4)	42 (28.6)	0.019*
LVEF 21–30%, n (%)	180 (3.4)	17 (11.6)	<0.001*
LVEF 20% or less, n (%)	90 (1.7)	8 (5.4)	<0.001*
MI within 90 days, n (%)	333 (6.3)	15 (10.2)	0.026*
PA systolic 31–55 mmHg, n (%)	386 (7.3)	22 (15.0)	<0.001*
PA systolic >55 mmHg, n (%)	111 (2.1)	7 (4.8)	<0.001*
Urgent, n (%)	954 (18.0)	34 (23.1)	0.057
Emergency, n (%)	239 (4.5)	42 (28.5)	<0.001*
Salvage, n (%)	317 (0.60)	32 (21.8)	<0.001*
Single non-CABG, n (%)	2012 (38.0)	20 (13.6)	<0.001*
2 procedures, n (%)	582 (11.0)	44 (29.9)	<0.001*
3 procedures, n (%)	151 (2.9)	22 (15.0)	<0.001*
Surgery on thoracic aorta, n (%)	275 (5.2)	32 (21.8)	<0.001*

*, P value <0.05 when comparing patients with and without mortality following cardiac surgery. NYHA, New York Heart Association; CCS, Canadian Cardiovascular Society; LVEF, left ventricular ejection fraction; MI, myocardial infarction; PA, pulmonary artery; CABG, coronary artery bypass grafting.

Table 3 The performance for each of the models

Models	Sensitivity	Specificity	Accuracy	AUC
EuroSCORE I	0.61	0.78	0.71	0.70*
EuroSCORE II	0.66	0.79	0.74	0.73*
STS model	0.62	0.79	0.73	0.71*
Logistic regression model	0.71	0.78	0.76	0.74*
Machine learning models without 10-fold cross validation				
Random forest	0.79	0.88	0.85	0.83
Neural network	0.61	0.83	0.78	0.75
Support vector machine	0.71	0.84	0.74	0.76
Gradient boosting machine	0.75	0.89	0.81	0.79
Machine learning models with 10-fold cross validation				
Random forest	0.83	0.92	0.89	0.87
Neural network	0.69	0.85	0.80	0.79
Support vector machine	0.75	0.84	0.82	0.81
Gradient boosting machine	0.78	0.85	0.84	0.82

*, P value <0.001 compared to the random forest with 10-fold cross validation. AUC, area under the curve; STS, Society of Thoracic Surgeons.

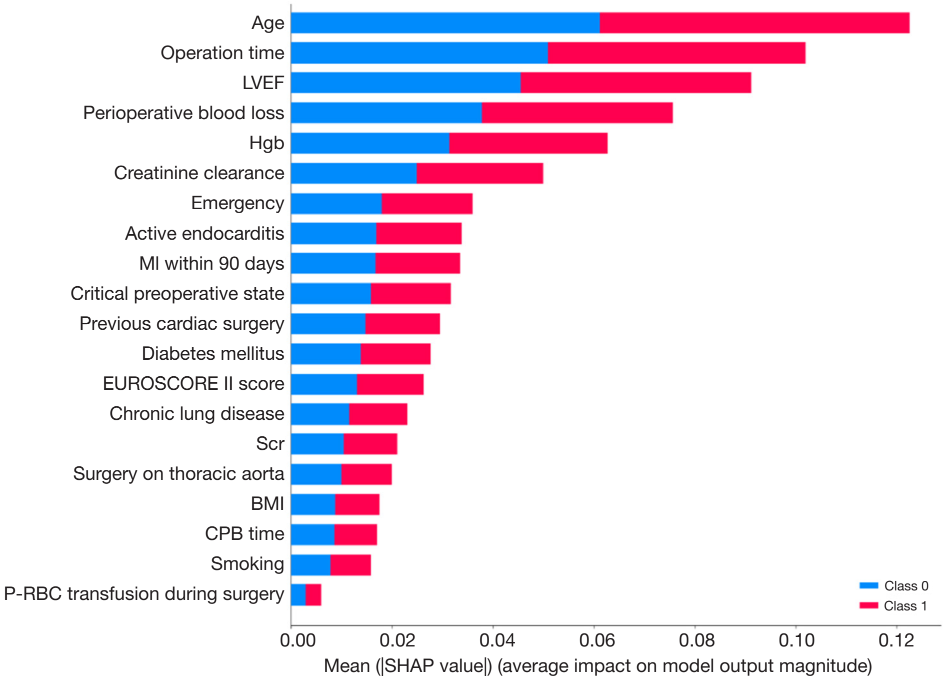

Figure 2 SHAP summary plot of the random forest model. This importance matrix plot depicts the importance of top-20 factors in the development of the random forest model. Feature importance is expressed as the sum of the absolute SHAP values. SHAP, SHapley Additive exPlanations; LVEF, left ventricular ejection fraction; Hgb, hemoglobin; MI, myocardial infarction; Scr, serum creatinine; BMI, body mass index; CPB, cardiopulmonary bypass; P-RBC, packed red blood cell.

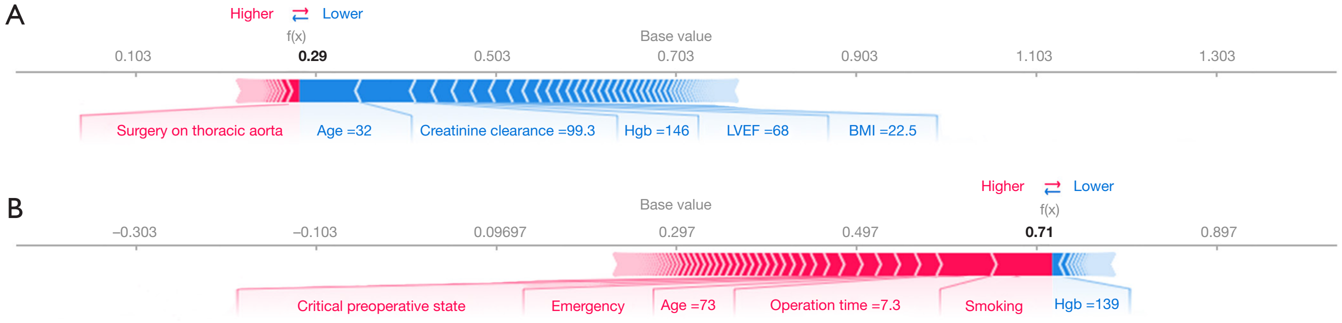

Figure 3 SHAP feature importance metrics for 2 patients who were correctly predicted as survival (A) or death (B). SHAP, SHapley Additive exPlanations; Hgb, hemoglobin; LVEF, left ventricular ejection fraction; BMI, body mass index.

accuracy, sensitivity and specificity than other machine learning models, and (III) by comparing with EuroSCORE II factors, the top 20 risk factors in the random forest screener displayed similarity and some critical disparity.

Since cardiac surgery has an effect on multiple systems throughout the body, the mortality rate after this type of surgery is high (19). Our reported mortality rate of 2.7% is slightly lower than that reported in the EuroSCORE II study (3.9%). This difference may derive from the much lower frequency of valve surgery and thoracic aortic surgery in our study cohort (1).

Investing interest in risk prediction models has blossomed clinically to facilitate multi-disciplinary shared decision making, which is also applied to the monitoring of innovation. In cardiac surgery, the two most commonly used risk stratification models are the EuroSCORE and STS risk scores (1,20,21). the derivation of which is anchored in logistic regression modelling. The risk models represented by them are particularly applicable in the era of expanding multimodality treatment for coronary and valve disease, where risk prediction plays an important role in determining which patients will benefit most from surgical or percutaneous treatment (22). Nevertheless, there are known inherent limitations within them. For instance, there may be influencing factors that are not collected by the EuroSCORE but have an impact on the risk of death, such as preoperative blood loss, operative time, and cross-clamp time. In addition, EuroSCORE I and II have been widely criticized, including poor performance in external validation, especially in high-risk subgroups (23-25). This may partially explain why the STS model and the EuroSCORE I and II showed only modest discrimination in our study.

Rather, machine learning models precisely stratified postoperative mortality in this study, which are in line with other risk classifiers based on machine learning techniques (26-28). Machine learning algorithms often behave as "data hungry" and tend to yield better performance with an increasing number of data points. We identified that machine learning models outperformed traditional logistic regression, as has been observed in many studies. Hence, our findings broadly provide an endorsement for previously published literature on machine learning and risk stratification (7,29,30). It is possible that the lower discriminatory power of EuroSCORE and STS scores, as well as logistic regression model, is partly driven by the fact that these linear regression-based models neglect the dynamic interactions and nonlinear effects between features. Alternatively, machine learning could automatically catch the interactions and nonlinear relationships, which might potentially lead to improvement in prediction.

Machine learning involves algorithms that make it possible to leverage and improve by itself, allowing prediction easier and more accurate. Recently, these models have been gradually applied in predicting post-operatively adverse complications, such as delirium (31) and atrial fibrillation (32,33). Unlike previous machine learning models that worked exclusively on preoperative factors, we integrated intra-operative variables, such as operative time, cross-clamp time, and peri-operative blood loss, all of which may pose an impact on later prognosis. It would thus be a reasonable approach to take preoperative and intraoperative factors into account, which may be advantageous over other machine learning-based models.

Random forest has previously been shown to be better classifier of risk stratification as compared with other models (30,34). Consistently, random forest classifier appeared to yield greater discrimination over other models in our study.

Discriminative power of the model is a measure of the trade-off between sensitivity and specificity, as measured by the area under the receiver operating characteristic curve, or AUC. For operative mortality, the AUC score in random forest was 0.83. After hyper-tuning with 10-fold cross validation, there was slight improvement in model discrimination, with and AUC level up to 0.87. The AUC score of the random forest model in our study seemed to be more favorable when compared with other published AUC scores for risk prediction after cardiac surgery. In a single-center study involving 11,190 individuals, the extreme gradient boosting model obtained an AUC score of 0.808 regarding the estimated risk of in-hospital fatality (35). Another study based on 28,761 patients undergoing cardiac surgery reported that the random forest yielded the highest AUC score of 0.80 (36).

Progress in visualization modeling algorithms, of the type used in this study, such as the Shapley additive explanations, allows for increased "black box" disclosure and thus enables clinically interpretable results (18). As shown in *Figure 3*, the contribution of variables to high- and low-risk cases makes the predictions more persuasive to clinician in decision making and accordingly, targeted prevention strategies are tailored for them.

Limitations

The present study is limited in some ways. The retrospective nature of the study exposes it to selection bias and makes it impossible to find causal relationships. In addition, our results may not be generalizable owing to the fact that the sample size is relatively small and the absence of external validation. The possibility also exists that there are confounding variables which were not taken into the analysis but did predict the outcome. Finally, the availability of predictive models for clinical use remains an important question mark. Therefore, future prospective multicenter studies with convincing evidences are warranted.

Conclusions

Machine learning models integrating intraoperative-related factors are likely to yield better discriminatory power for risk stratification in patients followed cardiac surgery. The results of the current study indicated that the deployment of machine learning algorithms could potentially offer significant gains in predicting in-hospital mortality after cardiac procedures, and it is possible to achieve accurate assessments at the individual level. In the age of precision medicine, the combination of machine learning models and big data may be of great use for clinical decision making.

Acknowledgments

We appreciate Prof. Jessica Cai for her support and help in polishing the article.
Funding: None.

Footnote

Reporting Checklist: The authors have completed the TRIPOD reporting checklist. Available at https://cdt.amegroups.com/article/view/10.21037/cdt-21-648/rc

Data Sharing Statement: Available at https://cdt.amegroups.com/article/view/10.21037/cdt-21-648/dss

Conflicts of Interest: All authors have completed the ICMJE uniform disclosure form (available at https://cdt.amegroups.com/article/view/10.21037/cdt-21-648/coif). The authors have no conflicts of interest to declare.

Ethical Statement: The authors are accountable for all aspects of the work in ensuring that questions related to the accuracy or integrity of any part of the work are appropriately investigated and resolved. The study was conducted in accordance with the Declaration of Helsinki (as revised in 2013). The study was approved by the institutional review board of Chinese PLA General Hospital (S2020-121-02). Informed consent was waived due to the observational nature of the study.

References

1. Nashef SA, Roques F, Sharples LD, et al. EuroSCORE II. Eur J Cardiothorac Surg 2012;41:734-44; discussion 744-5.

2. Edwards FH, Clark RE, Schwartz M. Coronary artery bypass grafting: the Society of Thoracic Surgeons National Database experience. Ann Thorac Surg 1994;57:12-9.
3. Chhor V, Merceron S, Ricome S, et al. Poor performances of EuroSCORE and CARE score for prediction of perioperative mortality in octogenarians undergoing aortic valve replacement for aortic stenosis. Eur J Anaesthesiol 2010;27:702-7.
4. Duchnowski P, Hryniewiecki T, Kuśmierczyk M, et al. Performance of the EuroSCORE II and the Society of Thoracic Surgeons score in patients undergoing aortic valve replacement for aortic stenosis. J Thorac Dis 2019;11:2076-81.
5. Guida P, Mastro F, Scrascia G, et al. Performance of the European System for Cardiac Operative Risk Evaluation II: a meta-analysis of 22 studies involving 145,592 cardiac surgery procedures. J Thorac Cardiovasc Surg 2014;148:3049-57.e1.
6. Provenchère S, Chevalier A, Ghodbane W, et al. Is the EuroSCORE II reliable to estimate operative mortality among octogenarians? PLoS One 2017;12:e0187056.
7. Churpek MM, Yuen TC, Winslow C, et al. Multicenter Comparison of Machine Learning Methods and Conventional Regression for Predicting Clinical Deterioration on the Wards. Crit Care Med 2016;44:368-74.
8. Weichenthal S, Ryswyk KV, Goldstein A, et al. A land use regression model for ambient ultrafine particles in Montreal, Canada: A comparison of linear regression and a machine learning approach. Environ Res 2016;146:65-72.
9. Kessler RC, van Loo HM, Wardenaar KJ, et al. Testing a machine-learning algorithm to predict the persistence and severity of major depressive disorder from baseline self-reports. Mol Psychiatry 2016;21:1366-71.
10. Taylor RA, Pare JR, Venkatesh AK, et al. Prediction of In-hospital Mortality in Emergency Department Patients With Sepsis: A Local Big Data-Driven, Machine Learning Approach. Acad Emerg Med 2016;23:269-78.
11. Meng F, Zhang Z, Hou X, et al. Machine learning for prediction of sudden cardiac death in heart failure patients with low left ventricular ejection fraction: study protocol for a retroprospective multicentre registry in China. BMJ Open 2019;9:e023724.
12. Shieh JL, Haq QMU, Haq MA, et al. Continual Learning Strategy in One-Stage Object Detection Framework Based on Experience Replay for Autonomous Driving Vehicle. Sensors (Basel) 2020;20:6777.
13. Sennott SC, Akagi L, Lee M, et al. AAC and Artificial Intelligence (AI). Top Lang Disord 2019;39:389-403.
14. Kwong M, Gardner HL, Dieterle N, et al. Optimization of Electronic Medical Records for Data Mining Using a Common Data Model. Top Companion Anim Med 2019;37:100364.
15. Sarica A, Cerasa A, Quattrone A. Random Forest Algorithm for the Classification of Neuroimaging Data in Alzheimer's Disease: A Systematic Review. Front Aging Neurosci 2017;9:329.
16. Camara C, Subramaniyam NP, Warwick K, et al. Non-Linear Dynamical Analysis of Resting Tremor for Demand-Driven Deep Brain Stimulation. Sensors (Basel) 2019;19:2507.
17. Nieuwenhuis M, Schnack HG, van Haren NE, et al. Multi-center MRI prediction models: Predicting sex and illness course in first episode psychosis patients. Neuroimage 2017;145:246-53.
18. Lundberg SM, Erion G, Chen H, et al. From Local Explanations to Global Understanding with Explainable AI for Trees. Nat Mach Intell 2020;2:56-67.
19. Efremov S, Lomivorotov V, Stoppe C, et al. Standard vs. Calorie-Dense Immune Nutrition in Haemodynamically Compromised Cardiac Patients: A Prospective Randomized Controlled Pilot Study. Nutrients 2017;9:1264.
20. Nashef SA, Roques F, Michel P, et al. European system for cardiac operative risk evaluation (EuroSCORE). Eur J Cardiothorac Surg 1999;16:9-13.
21. O'Brien SM, Feng L, He X, et al. The Society of Thoracic Surgeons 2018 Adult Cardiac Surgery Risk Models: Part 2-Statistical Methods and Results. Ann Thorac Surg 2018;105:1419-28.
22. Sullivan PG, Wallach JD, Ioannidis JP. Meta-Analysis Comparing Established Risk Prediction Models (EuroSCORE II, STS Score, and ACEF Score) for Perioperative Mortality During Cardiac Surgery. Am J Cardiol 2016;118:1574-82.
23. Ad N, Holmes SD, Patel J, et al. Comparison of EuroSCORE II, Original EuroSCORE, and The Society of Thoracic Surgeons Risk Score in Cardiac Surgery Patients. Ann Thorac Surg 2016;102:573-9.
24. Gummert JF, Funkat A, Osswald B, et al. EuroSCORE overestimates the risk of cardiac surgery: results from the national registry of the German Society of Thoracic and Cardiovascular Surgery. Clin Res Cardiol 2009;98:363-9.
25. Shahian DM, Blackstone EH, Edwards FH, et al. Cardiac surgery risk models: a position article. Ann Thorac Surg 2004;78:1868-77.

26. Harris AHS, Kuo AC, Weng Y, et al. Can Machine Learning Methods Produce Accurate and Easy-to-use Prediction Models of 30-day Complications and Mortality After Knee or Hip Arthroplasty? Clin Orthop Relat Res 2019;477:452-60.
27. Merath K, Hyer JM, Mehta R, et al. Use of Machine Learning for Prediction of Patient Risk of Postoperative Complications After Liver, Pancreatic, and Colorectal Surgery. J Gastrointest Surg 2020;24:1843-51.
28. Zhang G, Xu J, Yu M, et al. A machine learning approach for mortality prediction only using non-invasive parameters. Med Biol Eng Comput 2020;58:2195-238.
29. Shi HY, Lee KT, Lee HH, et al. Comparison of artificial neural network and logistic regression models for predicting in-hospital mortality after primary liver cancer surgery. PLoS One 2012;7:e35781.
30. Shouval R, Hadanny A, Shlomo N, et al. Machine learning for prediction of 30-day mortality after ST elevation myocardial infraction: An Acute Coronary Syndrome Israeli Survey data mining study. Int J Cardiol 2017;246:7-13.
31. Mufti HN, Hirsch GM, Abidi SR, et al. Exploiting Machine Learning Algorithms and Methods for the Prediction of Agitated Delirium After Cardiac Surgery: Models Development and Validation Study. JMIR Med Inform 2019;7:e14993.
32. Ambale-Venkatesh B, Yang X, Wu CO, et al. Cardiovascular Event Prediction by Machine Learning: The Multi-Ethnic Study of Atherosclerosis. Circ Res 2017;121:1092-101.
33. Karri R, Kawai A, Thong YJ, et al. Machine Learning Outperforms Existing Clinical Scoring Tools in the Prediction of Postoperative Atrial Fibrillation During Intensive Care Unit Admission After Cardiac Surgery. Heart Lung Circ 2021;30:1929-37.
34. 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.
35. 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;109:1811-9.
36. Benedetto U, Sinha S, Lyon M, et al. Can machine learning improve mortality prediction following cardiac surgery? Eur J Cardiothorac Surg 2020;58:1130-6.

Cite this article as: Fan Y, Dong J, Wu Y, Shen M, Zhu S, He X, Jiang S, Shao J, Song C. Development of machine learning models for mortality risk prediction after cardiac surgery. Cardiovasc Diagn Ther 2022;12(1):12-23. doi: 10.21037/cdt-21-648