Interpretable Machine Learning Model for Mortality Prediction in ICU: A Multicenter Study

Ka Man Fong
Intensive Care Unit, Queen Elizabeth Hospital

Shek Yin Au (h0145237@gmail.com)
Department of Intensive Care, Queen Elizabeth Hospital, 30 Gascoigne Road, Kowloon, Hong Kong SAR
https://orcid.org/0000-0001-8331-2250

George Wing Yiu Ng
Intensive Care Unit, Queen Elizabeth Hospital

Anne Kit Hung Leung
Intensive Care Unit, Queen Elizabeth Hospital

Research

Keywords: Mortality, Machine Learning, Critical Care

Posted Date: October 1st, 2020

DOI: https://doi.org/10.21203/rs.3.rs-83283/v1

Abstract

Background: Researchers have long been struggling to improve the disease severity score in mortality prediction in ICU. The digitalization of medical health records and advancement of computation power have promoted the use of machine learning in critical care. This study aimed to develop an interpretable machine learning model using datasets from multicenters, and to compare with the APACHE IV, in predicting hospital mortality of patients admitted to ICU.

Method: The datasets were assembled from the eICU database including 136145 patients across 208 hospitals throughout the U.S. and 5 ICUs in Hong Kong, including 10909 patients. The two datasets were first combined into one large dataset before 80:20 stratified split into the training set and the test set. The XGBoost machine algorithm was chosen to predict the hospital mortality. The variables in the model were the same as those included in the APACHE IV score. The discrimination and calibration of the model were assessed. The model would be interpreted using the Shapley Additive explanations values.

Results: Of the 147054 patients in the whole cohort, the hospital mortality was 9.3%. The area under the precision-recall curve for the XGBoost algorithm was 0.57, and 0.49 for APACHE IV. Similarly, the XGBoost reached an area under the receiving operating curve (AUROC) of 0.90, while APACHE IV had an AUROC of 0.87. Additionally, the XGBoost algorithm showed better calibration than the APACHE IV. The three most important variables were age, heart rate, and whether the patient was on ventilator.

Conclusions: The severity score developed by machine learning model using mutlicenter datasets outperformed the APACHE IV in predicting hospital mortality for patients admitted to ICU.

Introduction

Over the years there have been evolutions in the development of disease severity scores in Intensive Care Units (ICU). In 1985, the Acute Physiology and Chronic Health Evaluation (APACHE) II was designed by a group of experts led by Dr Knaus, who subjectively chose some variables and assigned weights to them based on their expert clinical judgements and some documented physiologic relationships from 5,815 ICU admissions. [1] Twenty years later, APACHE IV was published by Zimmerman et al., formulated using a non-linear logistic regression from data originated from 131,618 ICU admissions. These scores are now commonly used in many ICUs to quantify disease severity, to characterise organ dysfunction, to predict patient outcome and to facilitate resource allocation. [2] Although the use of statistical inference to characterise relationships between variables remains the cornerstone in medical research, statistical tests are not primarily built for making predictions. The difficulty lies in the complexity of critical illness which makes it unrealistic to form an accurate statistical model, while at the same time reasonably fulfilling the rigid assumptions behind these statistical tests. In contrast, machine learning models excel in their flexibility, which are particularly suitable for making prediction. The superior performance of machine learning has been well demonstrated in mortality prediction over SAPS II score, prediction of unplanned extubations and prediction of ICU readmissions.[3–5] However, the achievement of machine learning has

been slow to be recognized in the medical community. Many would see machine learning as a 'black-box' model, and question how the model derives such results. It is important to explain how a model works because if the users do not trust the working mechanism of a model or prediction, they will not use it no matter how accurate it could be. It has been shown that providing an explanation increased the acceptance of automation system. [6, 7]

In this study, to demonstrate the flexibility of machine learning model in making prediction, the very same APACHE IV variables would be used. As the prediction of SAPS II has been shown to be less accurate than the APACHE IV [8], it would be expected that the machine learning model would be more difficult to beat the record of the APACHE IV. The objective of this study was to develop an interpretable machine learning model and to compare with the APACHE IV, in predicting hospital mortality of patients admitted to ICU.

Methods

Study design and participants

To develop a machine learning prediction model that would be generalisable to ICUs across the US and Hong Kong, the dataset was built on the eICU dataset and Hong Kong datasets. The eICU Collaborative Research Database is a free, deidentified, multi-center ICU database including 200,859 ICU admissions in 2014–2015 accounted by 139,369 unique patients across 335 units at 208 hospitals throughout the U.S. [9] All tables were deidentified to meet the safe harbour provision of the US Health Insurance Portability and Accountability Act. The Hong Kong datasets comprised of data from 5 ICUs in Hong Kong, namely namely Queen Elizbeth Hospital, Kwong Wah Hospital, Pamela Youde Nethersole Eastern Hospital, Tuen Mun Hospital and Pok Oi Hospital, from 2017 to 2018. The study was approved by the Research Ethics Committee (Kowloon Central/Kowloon East) of Hospital Authority (Ref.: KC/KE-19-0215/ER-2). Variables used in APACHE IV calculation would be the input for the machine learning model in this study. Patients were excluded from the analysis according to the same exclusion criteria as in the original APACHE IV study: patients admitted for < 4 hours, patients with burns, patients < 16 years of age, patients admitted after transplant operations (except for hepatic and renal transplantation), patients remained in hospital for > 365 days, readmissions, and patients admitted from another ICU.[2] After exclusion, there were 136145 patients and 10909 patients from the eICU and the Hong Kong datasets respectively.

However, the 2 datasets represented patients of different disease spectrum. The Hong Kong dataset had higher Acute Physiology Score (APS) (55.63 ± 32.86 vs. 43.17 ± 23.40, $p < 0.001$) and APACHE IV score (67.09 ± 34.85 vs. 55.00 ± 25.34, $p < 0.001$) than the eICU dataset (Additional file 1: Table S1). Additional file 1: Figure S1 visualized the distribution of APACHE IV score, with the eICU dataset having a lower mean APACHE IV score, approximating to a normal distribution, whereas that distribution of the Hong Kong dataset was rather right skewed. Furthermore, additional file 1: Table S2 listed the top ten admission diagnoses in each dataset. The eICU dataset contained a significant proportion of patients admitted for medical diseases, such as acute myocardial infarction, cerebrovascular accidents,

congestive heart failure, and rhythm disturbances. Conversely, the Hong Kong dataset contained more patients admitted to ICU after neurosurgical operations, hepatobiliary operations, and emergency operations after gastrointestinal perforation. Therefore, to maintain the generalizability of the model across centers, the training dataset was built from both the eICU and the Hong Kong datasets. A stratified 80:20 split of the data was performed: 80% of the eICU datasets and 80% of the Hong Kong datasets were combined into the training set, while the rest would join to be the test sets. The datasets were pre-processed using the one-hot encoder and standard scaler.

XGBoost model

The primary outcome was hospital mortality. The predicted mortality by APACHE IV was extracted from the datasets. Then the same variables used in APACHE IV were inputted to compute the machine learning model using the extreme gradient boosting (XGBoost) algorithm.[10] The XGBoost model was trained with Stratified K Fold. The optimal combination of hyperparameters were sought by Grid Search with cross validation by looking for the highest area under the precision-recall curve (AUPRC).

Performance measures

While the most common measure for model discrimination has been the area under the receiver-operating characteristic curve (AUROC), it might mask the poor model performance in the case of imbalanced datasets.[11, 12] Therefore, the AUPRC instead of the AUROC would be used as the primary performance measure in this study. Meanwhile, the AUROC would also be shown for comparison.

Calibration of the model would be evaluated by the calibration curve and the Hosmer-Lemeshow goodness-of-fit test. Furthermore, with reference to the original APACHE IV study, the predicted mortality and the actual mortality in certain specific subgroups, would be compared to calculate the standardized mortality ratio (SMR). [2]

Model Interpretation

Lundberg and Lee proposed SHapley Additive exPlanations (SHAP) explain the output of machine learning models.[13] It would be calculated and visualised in SHAP summary plot. It was a concise plot that combined the feature importance with feature effect. Each point on the plot represented the Shapley value for the corresponding feature in a particular instance. The colour represented the value of the feature. The features were ordered in the y-axis in a descending order of importance. [14] Furthermore, relationship between individual variables and SHAP would be visualised using dependence plot. It was a scatter plot reflecting the effect one single variable had on the predictions made by the model.

Results

There were 147054 patients included in this study. Table 1 showed their baseline characteristics. The hospital mortality of the whole cohort was 9.3%.

Table 1
Baseline characteristics of the study population

	Overall population (n = 147054)	Alive at hospital discharge (n = 133328)	Dead at hospital discharge (n = 13726)
Age	62.8 ± 17.1	62.2 ± 17.2	69.2 ± 15.0
Source of admission[a]			
Floor	17670 (12.0)	14874 (11.2)	2796 (20.4)
OT/ Recovery	23718 (16.1)	22641 (17.0)	1077 (7.8)
AED	65316 (44.4)	59250 (44.4)	6066 (44.2)
Step-down units [b]	2030 (13.8)	1625 (1.2)	405 (3.0)
Other hospitals	2454 (1.6)	2095 (1.6)	359 (2.6)
Other sources	284 (0.2)	275 (0.2)	9 (0.7)
Pre-ICU LOS (days)	1.04 ± 4.64	0.93 ± 3.73	2.10 ± 9.70
Temperature (°C)	36.5 ± 1.0	36.5 ± 0.9	36.0 ± 1.7
Mean BP (mmHg)	86.4 ± 41.4	87.1 ± 40.8	79.3 ± 47.0
Heart rate (bpm)	100.7 ± 30.9	99.6 ± 30.1	111.2 ± 36.3
Respiratory rate (per minute)	25.0 ± 15.0	24.6 ± 14.9	28.8 ± 15.0
On ventilator	29884 (25.5)	23778 (22.4)	6062 (55.6)
FiO2 (%)	54.4 ± 26.6	51.4 ± 25.5	68.2 ± 27.5
PaO2 (mmHg)	122.1 ± 77.7	121.8 ± 75.2	123.3 ± 88.4
PaCO2 (mmHg)	42.7 ± 13.6	42.6 ± 13.0	43.0 ± 15.9
pH	7.36 ± 0.11	7.37 ± 0.1	7.30 ± 0.2

OT, operating theatre; AED, accident and emergency department; ICU, intensive care unit; LOS, length of stay; GCS, Glasgow coma scale; AIDS, acquired immunodeficiency syndrome

Data are represented as mean ± standard deviation, and n (%)

The number of patients with data available for each variable, p-value and confidence interval were shown in Table 7 in Appendix 6.4

[a]The percentages of admission source did not add up to 100% because of missing data

[b]High-dependency units were counted as step down units in the Hong Kong dataset.

	Overall population (n = 147054)	Alive at hospital discharge (n = 133328)	Dead at hospital discharge (n = 13726)
Serum sodium (mmol/L)	138.0 ± 5.7	137.9 ± 5.5	138.5 ± 7 .4
Urine output (mL / 24 hours)	1787.8 ± 1539.2	1838.7 ± 1523.6	1267.8 ± 1600.7
Serum creatinine (mg/dL)	1.59 ± 1.86	1.52 ± 1.84	2.19 ± 1.93
Urea nitrogen (mg/dL)	26.7 ± 22.7	25.4 ± 21.8	38.9 ± 27.4
Serum glucose (mg/dL)	165.6 ± 102.8	163.9 ± 101.6	181.6 ± 112.8
Albumin (g/dL)	2.89 ± 0.72	2.95 ± 0.70	2.46 ± 0.74
Bilirubin (mg/dL)	1.21 ± 2.28	1.09 ± 1.91	2.02 ± 3.89
Hemocrit (%)	32.7 ± 7.0	32.9 ± 6.9	31.0 ± 7.9
White blood cell (x10^9 cells/ L)	12.5 ± 8.6	12.1 ± 7.8	15.9 ± 13.6
Glasgow coma scale			
Eye	3.5 ± 1.0	3.5 ± 0.9	2.6 ± 1.4
Verbal	4.0 ± 1.6	4.2 ± 1.5	2.8 ± 1.8
Motor	5.5 ± 1.3	5.6 ± 1.1	4.2 ± 2.2
Could not assess GCS before medications	1535 (1.3)	1138 (1.1)	397 (3.6)
Comorbidities			
On dialysis	4213 (3.6)	3650 (3.4)	563 (5.2)
Cirrhosis	2035 (1.7)	1701 (1.6)	334 (3.1)
Hepatic failure	1735 (1.5)	1464 (1.4)	271 (2.5)
Metastatic cancer	2568 (2.2)	2113 (2.0)	455 (4.2)

OT, operating theatre; AED, accident and emergency department; ICU, intensive care unit; LOS, length of stay; GCS, Glasgow coma scale; AIDS, acquired immunodeficiency syndrome

Data are represented as mean ± standard deviation, and n (%)

The number of patients with data available for each variable, p-value and confidence interval were shown in Table 7 in Appendix 6.4

[a]The percentages of admission source did not add up to 100% because of missing data

[b]High-dependency units were counted as step down units in the Hong Kong dataset.

	Overall population (n = 147054)	Alive at hospital discharge (n = 133328)	Dead at hospital discharge (n = 13726)
Lymphoma	515 (0.4)	435 (0.4)	80 (0.7)
Leukemia	884 (0.8)	696 (0.7)	188 (1.7)
Immunosuppression	3454 (2.9)	2885 (2.7)	569 (5.2)
AIDS	137 (0.1)	113 (0.1)	24 (0.2)
Received thrombolysis	1834 (1.6)	1673 (1.6)	161 (1.5)
Emergency operation	4273 (3.6)	3733 (3.5)	540 (5.0)
OT, operating theatre; AED, accident and emergency department; ICU, intensive care unit; LOS, length of stay; GCS, Glasgow coma scale; AIDS, acquired immunodeficiency syndrome			
Data are represented as mean ± standard deviation, and n (%)			
The number of patients with data available for each variable, p-value and confidence interval were shown in Table 7 in Appendix 6.4			
[a]The percentages of admission source did not add up to 100% because of missing data			
[b]High-dependency units were counted as step down units in the Hong Kong dataset.			

Discrimination

The AUPRC was 0.57 for the XGBoost algorithm, and 0.49 for the APACHE IV in the whole cohort. (Fig. 1) Looking individually at the eICU and Hong Kong datasets, the XGBoost algorithm had higher AUPRC than the APACHE IV score (eICU: 0.55 vs. 0.45, Hong Kong dataset: 0.71 vs. 0.66). (Additional file 1: Table S2). The XGBoost algorithm reached an AUROC of 0.90, and APACHE IV had an AUROC of 0.87 in the combined data test set. (Additional file 1: Table S3)

Calibration

Figure 2 showed the calibration plot of the whole cohort. The closer curve to the diagonal reference line suggested a better calibration of the XGBoost algorithm (closer to the diagonal reference line) than the APACHE IV score. The calibration plots of the individual datasets were shown in Additional file 1: Table S4. While the Hosmer-Lemeshow chi-square was 22.31 ($p = 0.004$), caution should be made in interpretation of the p-value. A statistically significant Hosmer Lemeshow test does not mean that model fits poorly because the Hosmer-Lemeshow test, initially developed using a smaller dataset, would become overpowered when it is applied to a large sample. (19, 20) Rather, looking at the individual decile (Table 2), the absolute difference ranged from − 0.33 to 1.17% only. Table 3 tabulated the SMR comparing the predicted mortality over the actual mortality. The XGBoost algorithm demonstrated a SMR ranging from

0.70 to 1.28. The APACHE IV tended to overestimate the mortality, resulting in SMR varying from 0.44 to 0.82.

Table 2
Observed and predicted hospital mortality rates of the XGBoost model across risk deciles within the test set (n = 29957)

Risk decile[a]	Observed Deaths No. (%)	Predicted Deaths No. (%)	Difference %
1	2 (0.07)	8 (0.27)	-0.20
2	7 (0.23)	17 (0.57)	-0.33
3	21 (0.70)	28 (0.93)	-0.23
4	37 (1.23)	43 (1.44)	-0.20
5	68 (2.27)	65 (2.17)	0.10
6	78 (2.60)	100 (3.34)	-0.73
7	178 (5.94)	160 (5.64)	0.30
8	309 (10.31)	274 (9.15)	1.17
9	554 (18.49)	534 (17.82)	0.67
10	1571 (52.44)	1572 (52.47)	-0.03

[a]Risk decile: population sorted by increasing predicted risk and then split into deciles. Hosmer-Lemeshow Chi-square = 22.31, df = 8, p-value = 0.004

Table 3
Standardized mortality ratio for selected disease groups in the test set

	n			XGBoost		APACHE IV	
Disease Group	eICU	Hong Kong dataset	Observed mortality, %	Predicted mortality, %	SMR	Predicted mortality, %	SMR
Sepsis (non-urinary tract)	2717	373	19.9	19.5	1.02	25.0	0.79
Cardiac arrest	678	77	51.8	51.7	1.00	64.6	0.80
Emphysema/ bronchitis	629	16	5.4	7.7	0.70	12.3	0.44
Thoractomy of lung, neoplasm	138	17	2.6	2.5	1.05	5.3	0.49
Aortic aneurysm, elective repair	158	76	3.9	4.0	0.99	7.9	0.50
Stroke	1070	35	10.6	9.2	1.15	19.0	0.56
Hepatic failure	123	6	17.8	19.9	0.89	27.9	0.64
Respiratory arrest	364	2	21.9	17.1	1.28	26.6	0.82

Table 4 summarized the discrimination and calibration between the XGBoost model and the APACHE IV, showing the superior performance in the former.

Table 4
Comparison of discrimination and calibration of the XGBoost model and APACHE IV model when applied to the test set

	XGBoost	APACHE IV
Observed mortality rate (%)	9.44	9.44
Expected mortality rate (%)	9.37	13.83
SMR	1.01	0.68
AUPRC	0.57	0.49
AUROC	0.90	0.87
Hosmer-Lemeshow Chi-square	22.31 (p = 0.004)	711.73 (p < 0.001)

SMR, standardized mortality ratio; AUPRC, area under the precision-recall curve; AUROC, area under the receiver-operating characteristic curve

Variable Importance

Figure 3 showed the SHAP variable importance plot. It was made up of individual dots and each represented one training data. The feature importance was shown by the descending order of the variables. The x position of the dot reflected the impact of the prediction. A positive SHAP value was positively associated with higher mortality, and vice versa. The color of the dots represented the value of that variable for the prediction. For example, older age (as shown in red) was positively associated with mortality (as it was on the right side of the axis), whereas patients who were not on ventilators (as shown in blue as it was encoded as 0 in the data, compared with 1 representing patients who were on ventilator) were associated with a lower risk of mortality (left side of the axis). Therefore, age contributed most in predicting hospital mortality in the XGBoost model, followed by other factors like heart rate, whether the patient was on ventilator, bilirubin level, and whether the patient suffered from sepsis (non-urinary tract) etc.

Furthermore, the SHAP variable importance plot visualized the data in an intuitive way. For example, when looking at the right side of the plot (SHAP value > 0.0), both a high heart rate and a low heart rate were positively associated with mortality, but the effect of low heart rate was stronger than that of high heart rate. To investigate the effect of an individual variable like the heart rate, the dependence plot generated by the SHAP model had elegantly illustrated the U-shape relationship. (Fig. 4) The SHAP value was lowest with the heart rate ranging from about 50 to 100 beats per minute. The effect of severe bradycardia on mortality prediction was greater compared with that of tachycardia. The effect of interactions with other variables was also shown. For example, for patients who had heart rate of 150 bpm, younger patients had much lower SHAP than that of older patients. Different dependence plots between SHAP and independent variables, and interactions among these variables could also be plotted and studied from the system if needed.

Discussion

This study demonstrated the superior performance of model prediction based on big data using machine learning algorithm than traditional statistical inference, by utilizing the same variables in the original APACHE IV score. While the original APACHE IV model was built using solely the data from the U.S. population, the model in this study was built on a combination of data from the eICU and multiple centers in Hong Kong. This machine learning algorithm outperformed the original APACHE IV scoring in the whole cohort, as well as the individual eICU and the Hong Kong populations. Such improved generalisability of the model was an important quality of an outcome predictive scoring system which served as a tool to compare different patient populations in medical research, and to compare the quality of care across ICUs worldwide.

The use of SHAP has increased the interpretability of the model. It has been proven that machine learning model is no longer a ‘black-box’ model – that is, although the machine itself has no idea what these variables mean, compared with clinicians who could subjectively give weight to individual variables in the APACHE II era, the interpretation of the models by SHAP has proved that it is biologically sound and

clinically plausible. This message is paramount in boosting the confidence of clinicians who would consider utilizing this model.

The strength of this study was the enormous dataset built from the eICU dataset and multiple centers in Hong Kong. The XGBoost algorithm computed in this model was one of the state-of-the-art machine learning models with outstanding performance. This study also demonstrated the use of AUPRC in addition of AUROC to reveal the true performance of the model in face of an imbalanced dataset. Model discrimination was evaluated in multiple facets and calibration plot was used to avoid the pitfall of Hosmer-Lemeshow Chi-square test. The use of SHAP improved the interpretability of the model.

The limitation of this study was that most of the patients came from the eICU dataset representing the U.S. data. Based on the figures, the two groups of patients obviously differed in terms of disease severity, admission sources and admission diagnoses. Behind the scene there might be difference in case referral pattern and case management. The thresholds for ICU admission varied across centers and countries. The pathophysiology of disease might not be the same between Chinese and Caucasians. Therefore, it was also the reason why the eICU dataset and the Hong Kong dataset were combined before splitting into the training and test sets. Furthermore, the variables used in this study were limited to those used in the original APACHE IV, because the aim of this study was to show the superior performance of machine learning model versus traditional statistical methods. The mortality prediction model might be improved by using data from a more complex case-mix or recruiting more data from academic centers and rural hospitals worldwide. Particularly, there is a need to establish a global database recruiting ICUs across the continents in order to produce a generalisable prediction model. Other formats of data such as waveforms (electrocardiogram, photoplethysmography, arterial blood pressure), imaging data, and clinical texts might also be employed to boost the accuracy of mortality prediction model. On the other hand, one needs to strike a balance in determining the number of variables in a model because increasing the variables increases the difficulty in data collection. Lastly, other outcomes can also be predicted in the future studies, such as ICU and hospital length of stay, ventilator-free days, dialysis independence, and long-term mortality outcome.

Conclusion

Using the same variables in APACHE IV, the XGBoost algorithm outperformed the APACHE IV scoring in hospital mortality prediction for patients admitted to the ICU, based on the data from eICU dataset and Hong Kong datasets. There is an emerging need to establish a global database of patients in critical care.

List Of Abbreviations

ICU, intensive care unit; APACHE, Acute Physiology and Chronic Health Evaluation; APS, Acute Physiology Score; XGBoost, eXtreme Gradient Boosting; AUPRC, area under the precision-recall curve; AUROC, area under the receiver-operating characteristic curve; SMR, standardized mortality ratio; SHAP, SHapley Additive exPlanations

Declarations

Acknowledgement

This study contained materials from the thesis from KF for the fellowship of Hong Kong College of Physicians. The authors would like to express their gratitude to the COC (ICU) Committee members, ICU colleagues in Kwong Wah Hospital, Pamela Youde Nethersole Eastern Hospital, Tuen Mun Hospital, Pok Oi Hospital and Queen Elizbeth Hospital for their support of this study and the approval to access the local APACHE IV data.

Funding

There is no funding for this review.

Availability of data and materials

All data generated or analyzed during the present study are included in this published article and its supplementary information files.

Authors' contribution

KF conceived the study, performed data cleaning and analysis, and drafted the manuscript. SA, GN and AL did the revision of the manuscript. All authors read and approved the final manuscript.

Ethics approval and consent to participate

Not applicable

Consent for publication

Not applicable

Competing interests

The authors declare that they have no competing interests.

References

1. Knaus WA, Draper EA, Wagner DP, Zimmerman JE. APACHE II: a severity of disease classification system. Crit Care Med. 1985;13(10):818-29.
2. Zimmerman JE, Kramer AA, McNair DS, Malila FM. Acute Physiology and Chronic Health Evaluation (APACHE) IV: hospital mortality assessment for today's critically ill patients. Crit Care Med. 2006;34(5):1297-310.

3. Pirracchio R, Petersen ML, Carone M, Rigon MR, Chevret S, van der Laan MJ. Mortality prediction in intensive care units with the Super ICU Learner Algorithm (SICULA): a population-based study. Lancet Respir Med. 2015;3(1):42-52.
4. Hsieh MH, Hsieh MJ, Chen CM, Hsieh CC, Chao CM, Lai CC. Comparison of machine learning models for the prediction of mortality of patients with unplanned extubation in intensive care units. Sci Rep. 2018;8(1):17116.
5. Desautels T, Das R, Calvert J, Trivedi M, Summers C, Wales DJ, et al. Prediction of early unplanned intensive care unit readmission in a UK tertiary care hospital: a cross-sectional machine learning approach. BMJ Open. 2017;7(9):e017199.
6. Ribeiro MT, Singh S, Guestrin C. "Why Should I Trust You?": Explaining the Predictions of Any Classifier. arXiv preprint arXiv.1602.04938. 2016 Feb.
7. Dzindolet MT, Peterson SA, Pomranky RA, Pierce LG, Beckc HP. The role of trust in automation reliance. International Journal of Human-Computer Studies.Volume 58, Issue 6, June 2003, Pages 697-718.
8. Kuzniewicz MW, Vasilevskis EE, Lane R, Dean ML, Trivedi NG, Rennie DJ, et al. Variation in ICU risk-adjusted mortality: impact of methods of assessment and potential confounders. Chest. 2008;133(6):1319-27.
9. Pollard TJ, Johnson AEW, Raffa JD, Celi LA, Mark RG, Badawi O. The eICU Collaborative Research Database, a freely available multi-center database for critical care research. Sci Data. 2018;5:180178.
10. Chen T, Guestrin CE. XGBoost: A Scalable Tree Boosting System. Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. 2016:Pages 785–94.
11. Saito T, Rehmsmeier M. The precision-recall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced datasets. PLoS One. 2015;10(3):e0118432.
12. Jeni LAJ, Cohn JF, Torre FDL. Facing Imbalanced Data: Recommendations for the Use of Performance Metrics. 2013 Humaine Association Conference on Affective Computing and Intelligent Interaction. 2013:pp. 245-51.
13. Lundberg SM, Lee SI. A Unified Approach to Interpreting Model Predictions. Advances in neural information processing systems. 2017:pp. 4765-74.
14. Molnar C. Interpretable Machine Learning 2019 [Available from: https://christophm.github.io/interpretable-ml-book/shap.html.

Figures

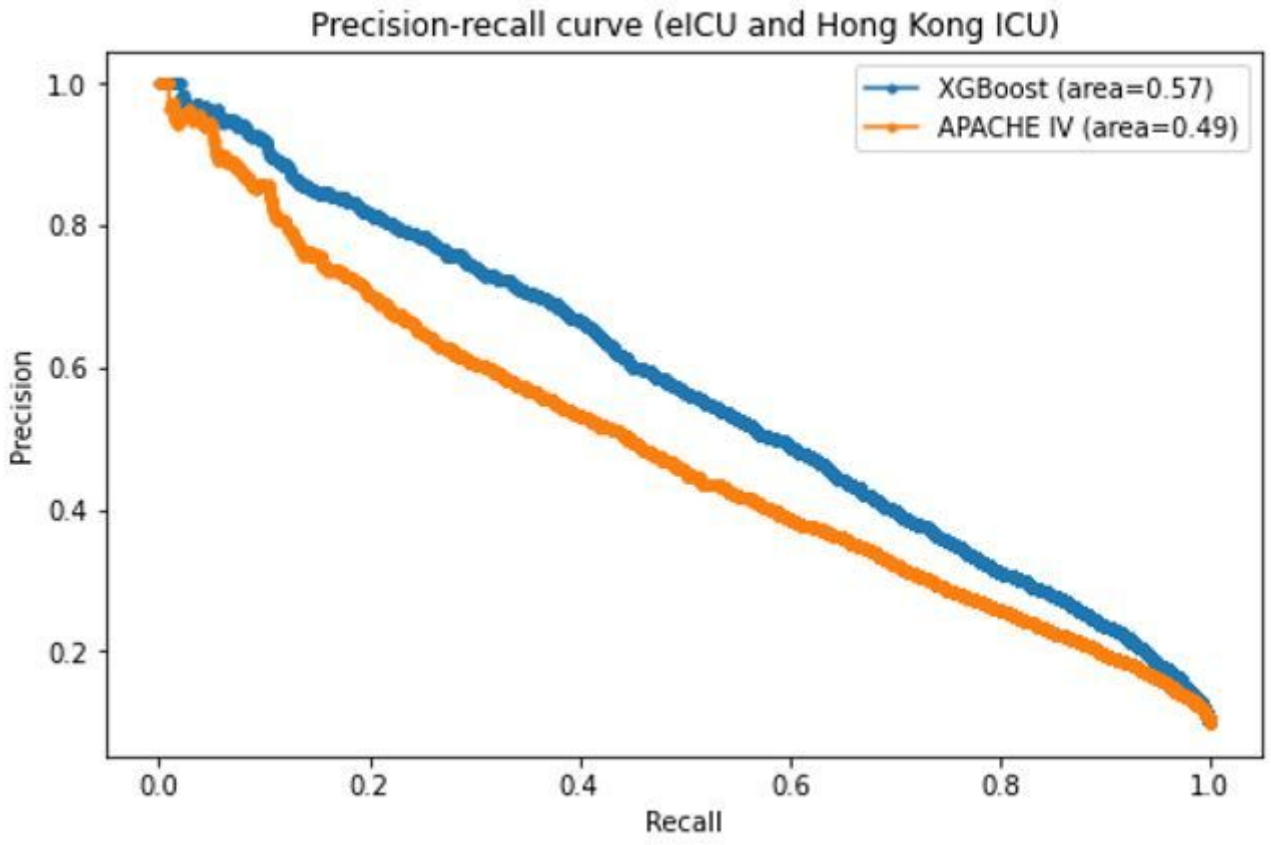


Figure 1

Precision-recall curves of the whole cohort

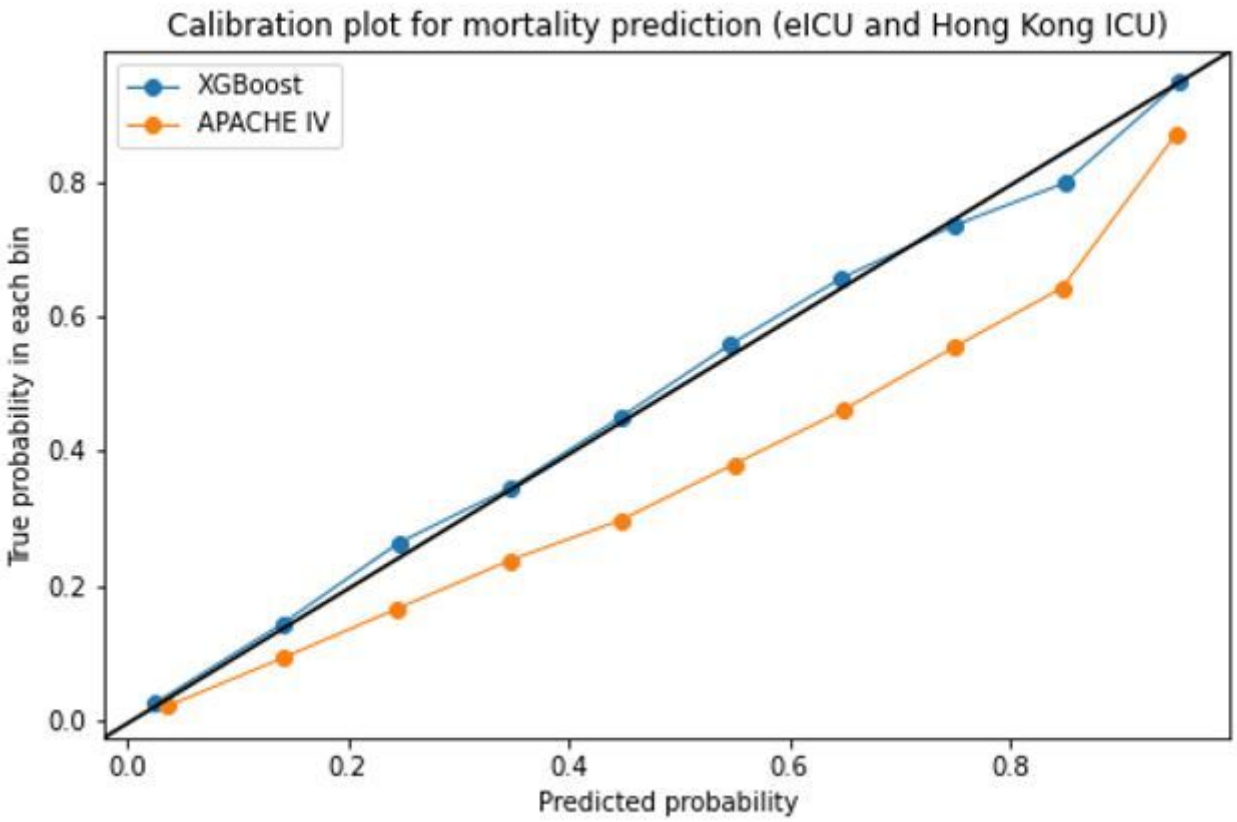
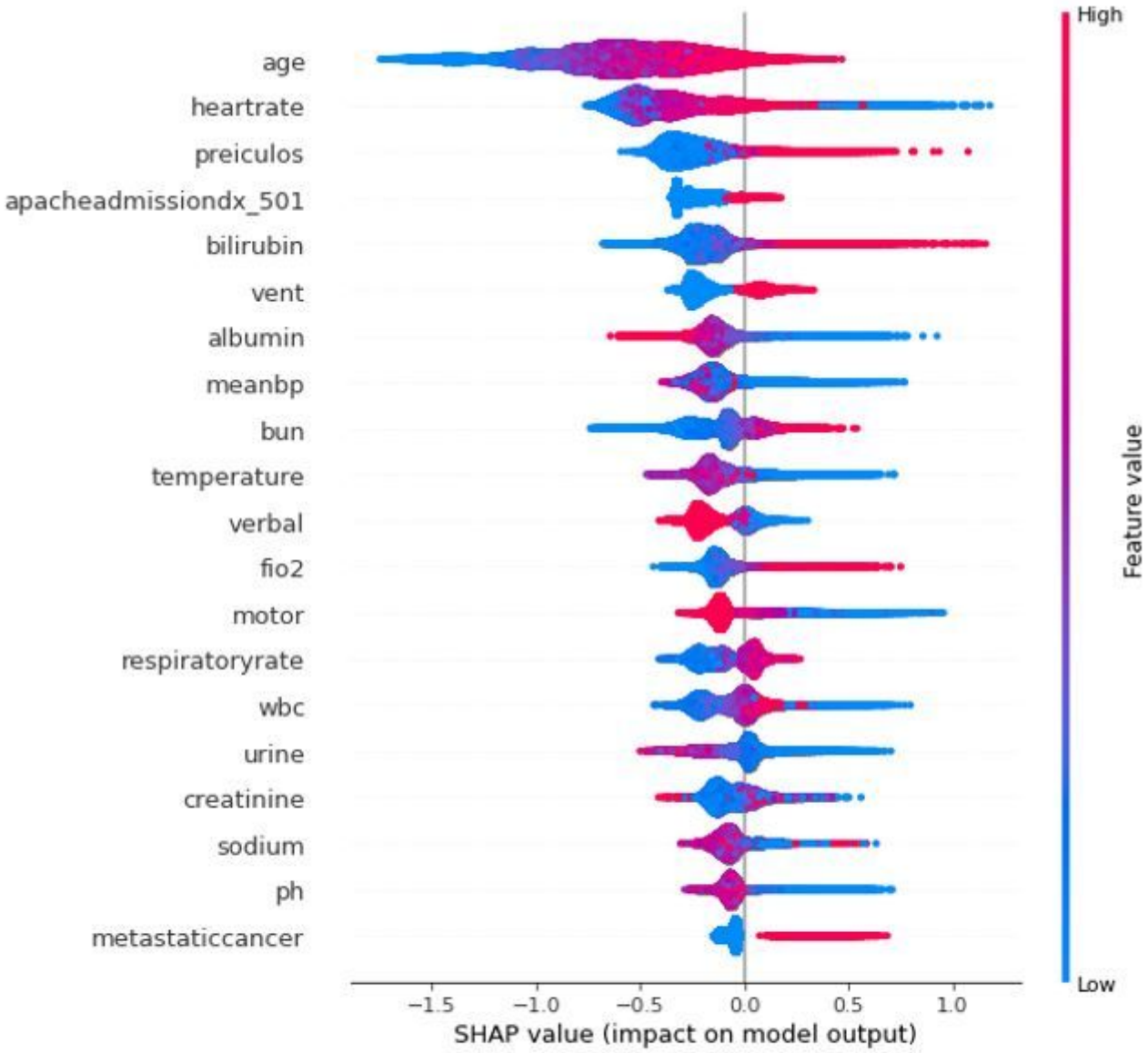


Figure 2

Calibration plots of the whole cohort



Note: Apacheadmissiondx_501 represented the diagnosis of sepsis (non-urinary tract). Eyes was one of the components in the Glasgow coma scale.

Figure 3

The SHAP Variable Importance Plot

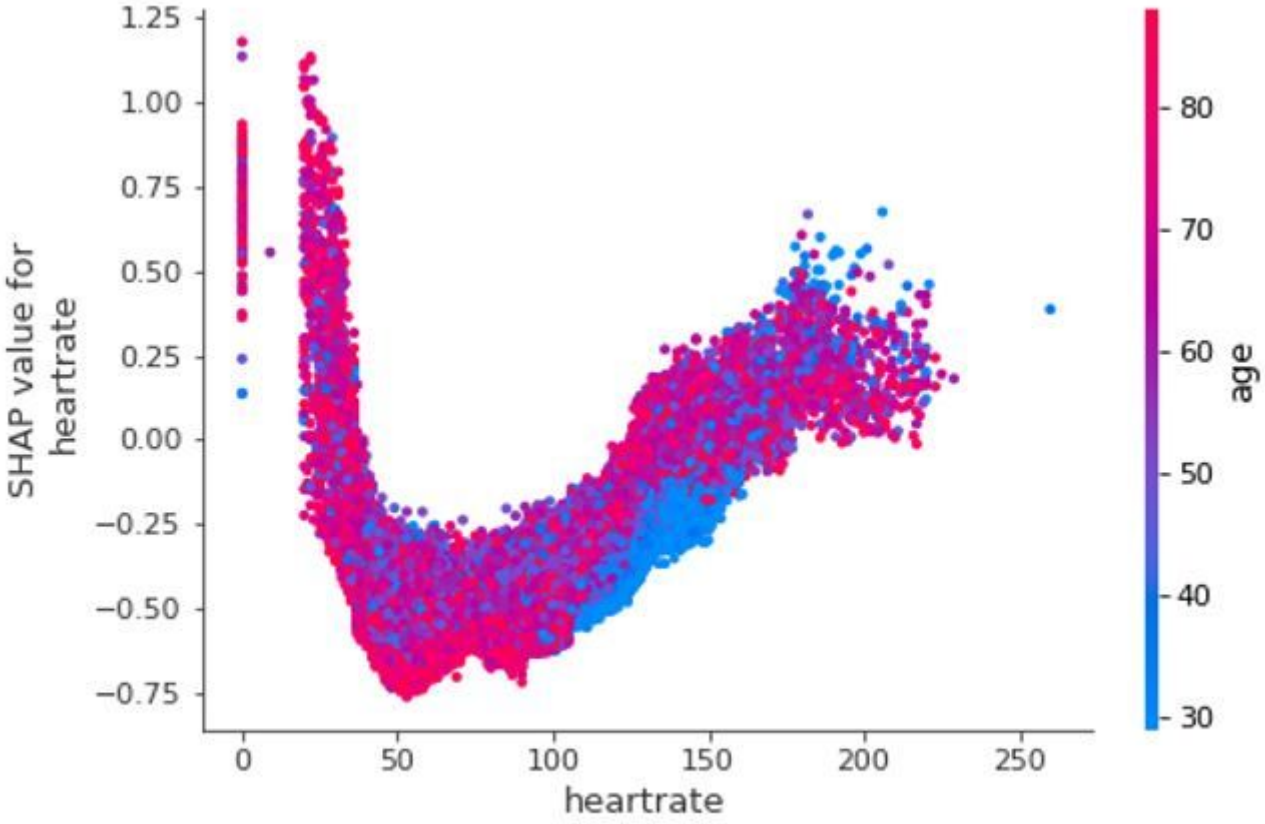


Figure 4

SHAP Dependence plot showing the interaction between heart rate and age

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- MLjournalcriticalcareadditionalfile1.docx