International Journal of General Medicine

Dovepress
open access to scientific and medical research

Open Access Full Text Article

ORIGINAL RESEARCH

Improving Mortality Risk Prediction with Routine Clinical Data: A Practical Machine Learning Model Based on eICU Patients

Shangping Zhao[1], Guanxiu Tang[2], Pan Liu[1], Qingyong Wang[3], Guohui Li[1], Zhaoyun Ding[1]

[1]Laboratory for Big Data and Decision, National University of Defense Technology, ChangSha, Hunan, People's Republic of China; [2]The Nursing Department, The Third Xiangya Hospital of Central South University, ChangSha, Hunan, People's Republic of China; [3]School of Information and Computer, Anhui Agricultural University, Hefei, Anhui, People's Republic of China

Correspondence: Zhaoyun Ding, Laboratory for Big Data and Decision, National University of Defense Technology, No. 109 Deya Road, Kaifu District, ChangSha, Hunan, 410003, People's Republic of China, Tel +86 17607310865, Fax +86 731 88618837, Email dingzhaoyun1983@163.com

Purpose: Mortality risk prediction helps clinicians make better decisions in patient healthcare. However, existing severity scoring systems or algorithms used in intensive care units (ICUs) often rely on laborious manual collection of complex variables and lack sufficient validation in diverse clinical environments, thus limiting their practical applicability. This study aims to evaluate the performance of machine learning models that utilize routinely collected clinical data for short-term mortality risk prediction.

Patients and Methods: Using the eICU Collaborative Research Database, we identified a cohort of 12,393 ICU patients, who were randomly divided into a training group and a validation group at a ratio of 9:1. The models utilized routine variables obtained from regular medical workflows, including age, gender, physiological measurements, and usage of vasoactive medications within a 24-hour period prior to patient discharge. Four different machine learning algorithms, namely logistic regression, random forest, extreme gradient boosting (XGboost), and artificial neural network were employed to develop the mortality risk prediction model. We compared the discrimination and calibration performance of these models in assessing mortality risk within 1-week time window.

Results: Among the tested models, the XGBoost algorithm demonstrated the highest performance, with an area under the receiver operating characteristic curve (AUROC) of 0.9702, an area under precision and recall curves (AUPRC) of 0.8517, and a favorable Brier score of 0.0259 for 24-hour mortality risk prediction. Although the model's performance decreased when considering larger time windows, it still achieved a comparable AUROC of 0.9184 and AUPRC of 0.5519 for 3-day mortality risk prediction.

Conclusion: The findings demonstrate the feasibility of developing a highly accurate and well-calibrated model based on the XGBoost algorithm for short-term mortality risk prediction with easily accessible and interpretative data. These results enhance confidence in the application of the machine learning model to clinical practice.

Keywords: intensive care unit, XGBoost, routinely collected data, short-term mortality risk

Plain Language Summary

Mortality risk prediction helps clinicians make better decisions in patient healthcare. Machine learning methods are flexible algorithms that offer potential advantages over conventional scoring systems and demonstrate promising performance in medical data analysis. In this study, we developed a highly accurate and well-calibrated short-term mortality risk prediction model based on XGBoost. The model utilized routine clinical data collected within a 24-hour window prior to patient discharge. This timeframe allowed us to capture physiological data that accurately reflected the differences in health status between the dead and the surviving patients. Furthermore, the model was applied in short-term mortality risk prediction at different time points during a patient's hospital stay, which enhances its reliability in assessing dynamic mortality risk for patients at any given time. With its excellent prediction performance across various time windows and easy availability of features, our model has the potential to accurately identify high-risk patients at an earlier stage. It can be efficiently used even in low-resource healthcare environments, thereby assisting healthcare professionals in making better therapeutic decisions, optimizing resource allocation, and addressing other challenges in patient healthcare.

Received: 8 April 2023
Accepted: 16 July 2023
Published: 26 July 2023

Introduction

Reliable mortality predictions are essential for assessing the severity of illness and determining the effectiveness of new interventions for intensive care unit (ICU) patients, which may help to promote the quality of care and clinical outcomes.[1] For the past decades, significant efforts have been invested in predicting the risk of death for ICU patients, and several severity scoring systems have been developed. Some widely documented tools for mortality risk prediction include the Acute Physiology and Chronic Health Evaluation (APACHE), the Mortality Probability Model (MPM), and the Simplified Acute Physiology Score (SAPS).[2,3] However, studies indicate that only 12% of ICUs actually use these severity scoring systems.[4] This low adoption rate may mainly be attributed to two main factors. At first, the deployment of these mortality risk assessment tools requires clinicians to engage in the labor-consuming collection of patient data that is not captured in routine workflows. For instance, the well-known APACHE model heavily relies on manual comprehensive evaluations of a patient's chronic health status, which may increase the workload on clinicians. Secondly, these prediction models seem to show rapid deteriorates in performance over time. In a study by Kramer, the SAPS II was reported to be out of calibration by 2005.[5] Some subsequent research has also highlighted the calibration issues with APACHE and SAPS.[6,7] Additionally, the diverse patient cohorts and variations in medical treatments have been associated with an overestimation of mortality rates.

The limitations of existing scoring systems have led to the emergence of novel machine learning (ML) algorithms for mortality prediction. It is inspiring that ML algorithms can be continuously trained using newly obtained data in clinical settings, allowing for updates and recalibration over time. This dynamic nature improves the accuracy of the model as populations and treatments evolve. Over the past few decades, researchers have demonstrated that ML techniques outperform traditional prediction methods in predicting outcomes such as mortality, complications, and length of stay.[8–12] In a study by Delehanty,[8] automated risk adjustment algorithms based on extreme gradient boosting (XGBoost) trees were developed for adult ICU patients, achieving an impressive area under the receiver-operator curve (AUROC) 0f 0.94. However, this model heavily relies on manual diagnoses made by clinicians, which may not be available early in hospitalization, limiting its applicability in hospitals that do not use the All Patients Refined Diagnosis Related Groups code. Recent advancements have focused on using more accessible features in ML models for mortality prediction.[13–15] For instance, the XGBoost algorithm achieved an excellent AUROC of 0.97 in neonatal mortality prediction using routinely collected data in 2021.[14] Another study by Alghatani et al[15] compared various ML algorithms in mortality prediction for adult ICU patients using only baseline demographic and vital sign features. The Random Forest (RF) model demonstrated the highest accuracy with an accuracy of 0.8861.

Though the promising results highlight the use of routinely collected clinical data and state-of-The-art ML algorithms in mortality risk prediction, there are still several limitations that need to be addressed. Firstly, most models in the literature have primarily focused on predicting mortality at the time of admission, typically based on data collected within 24 hours pre or post-ICU admission.[16–18] While this approach is valuable for the early detection of mortality risk, it lacks flexibility and fails to consider how patients respond to treatments after admission. This limitation can potentially reduce the accuracy of predictions when used at later time points. It would be beneficial to develop models that can adapt and update predictions based on patients' responses to treatments during their hospital stay. Secondly, many published papers[19,20] mainly report the discrimination performance of the models, typically evaluated by metrics like the AUROC. However, they may neglect the crucial aspect of calibration evaluation, which becomes particularly important when applying models to new population groups. In a system review of applications of ML to routinely collected ICU data by Shillan et al,[21] the researcher noted that only 21(13.6%) papers reported the calibration. In addition, it has also been suggested that some studies validate their predictions using random subsets of the development data, which can lead to an overestimate of the model's performance. To ensure a more robust evaluation, it is preferable to validate the model using an external dataset, separate from the one used for development. In cases where external validation is not feasible, proper procedures such as bootstrapping or cross-validation should be applied to compensate for statistical over-optimism and mitigate statistical over-optimism.

The ICU patients are always under intensive monitoring and thus produce multiple physiological data and treatment records in regular care workflows, which provide valuable data resources for mortality prediction. Mortality risk

assessments integrating the available clinical data and outstanding ML algorithms with rigorous evaluation are potential to facilitate the feasibility and confidence of the model in clinical practice. In this study, we aimed to develop machine learning models for in-hospital mortality risk prediction models by using routinely collected clinical data, and further evaluate the model' performance in short-term mortality risk prediction.

Materials and Methods

Study Design and Population

This study was conducted as a retrospective cohort study, utilizing data extracted from the eICU- CRD, a large, multi-center, and publicly available critical care database.[22] Access to this database is freely granted to researchers under the data usage agreement established by the review board of PhysioNet. We have obtained permission after the application and completion of the Protecting Human Research Participants course and test (37796533). All procedures undertaken in this study adhered to the ethical standards set by the responsible committee on human experimentation at The Third Xiangya Hospital of Central South University (Number 22296) and with the Helsinki Declaration of 1975, as revised in 2008. Furthermore, the study design and use of de-identified data exempted the requirement for informed consent from the study participants, as determined by the hospital review board. The ethical guidelines outlined in the Helsinki Declaration of 1975, revised in 2008, were also strictly followed throughout the study.

The eICU-CRD integrates de-identified clinical data of 139,367 patients who came from more than 200 US hospitals between 2014 and 2015. For this study, all inpatients aged 14 years or over were considered. Subjects with documented length-of-stay and survival for less than 24h following admission were excluded to ensure sufficient data for analysis. Ultimately, a total of 123,929 patients were included in the final analysis. Furthermore, these subjects were randomly divided into two groups at a ratio of 9:1, consisting of a training dataset (n=111,536) and a test dataset (n=12,393). The prediction model was developed based on the training dataset, and the test dataset was reserved for performance measurement. The detailed process of cohort selection is illustrated in Figure 1.

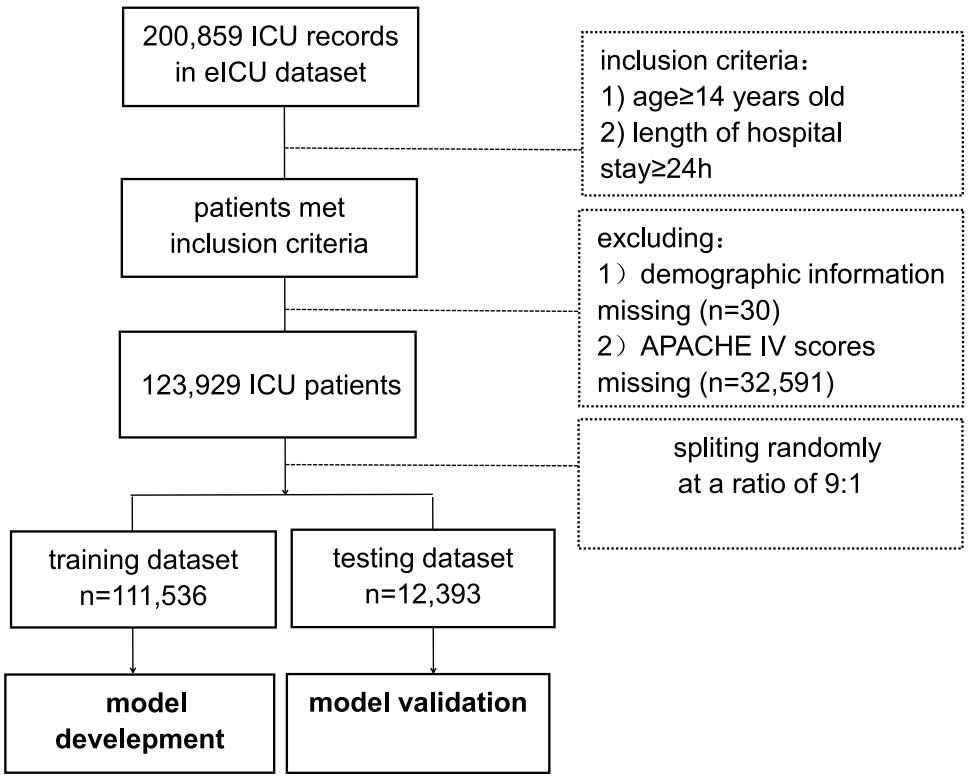

Figure 1 The process of cohort selection.

Data Extraction and Feature Selection

Actual hospital mortality (labeled as 0 or 1) and predicted hospital mortality based on APACHE score were obtained from the database for all patients. The predicted hospital mortality served as an evaluation benchmark since the widespread use of APACHE scores in critical care settings. Additionally, administration information, such as hospital stay length, ICU stay length, and patient discharge time, was extracted during the data selection process to further analyze population characteristics.

As for features, a 7-day window of data was obtained for each patient to facilitate short-term mortality risk assessments. For patients whose stays exceeded 7 days in length, the 7 days of data prior to their discharge or death time were obtained; otherwise, patients' data from admission to discharge were obtained. Variables routinely measured in medical practice were selected for model development, based on previous literature and consultations with critical care physicians.

The feature extraction and selection process proceeded as follows. Firstly, age and gender were extracted and selected for prediction as they provide basic information about the patient. Then, all vital signs and several laboratory tests screened by clinicians were obtained from the interested time window. These physiological variables were chosen as features due to their fundamental importance as indicators of health status, and their universal understanding among healthcare professionals. They are continuously collected through automatic or regular manual measurements during patient care, allowing for better quantification of health status variability and facilitating mortality risk prediction. To capture trends in vital signs and laboratory parameters, we calculated the minimum, maximum, and mean values over the 24-hour window as inputs to quantify the variability of each physiological variable. By limiting the features to a 24-hour window, we ensure that the prediction model considers the patient's current health status. Furthermore, the usage of vasoactive medications (such as epinephrine, noradrenaline, and dopamine) was examined and used as predictors. Medication usage was recorded as a binary value, with 1 indicating the presence of a medication record in the infusion drug table of the eICU database within the 24-hour window, and 0 indicating its absence. Vasoactive medications were chosen as predictors due to their fundamental role in improving patient circulatory function and maintaining homeostasis.

To ensure the accuracy of the prediction, we removed variables with more than 20% missing values. We did not employ any statistical methods or ML algorithms for further feature selection to avoid excluding potentially influential factors. The feature selection process resulted in a total of 138 features, which were listed in Table S1 (see Additional File 1). To fill in missing data and obtain the most realistic patient status possible, we applied the nearest neighbors approach[23] to impute missing values. For patients without any record of a certain variable, a pre-specified normal value was used for imputation.[24]

Model Development

The features were derived from a 24-hour window prior to the patient's discharge or death time and were used for training the model. The specific time period was chosen because it is more likely to capture the intrinsic characteristics of a patient experiencing deterioration or stabilization. Patients who died typically exhibit highly unstable patterns across various features towards the end of their hospital stay.

Before constructing the ML model, numerical variables were normalized to a range of 0 to 1 based on prior experience. Four ML models, namely logistic regression (LR), RF, XGBoost, and artificial neural network (ANN), were trained using the selected variables for mortality prediction in the training dataset.

1. LR is a widely utilized statistical method in medical data analysis and serves as a fundamental algorithm for ML development. We used binary logistic regression to predict the relationship between the outcome and the predictors. The LR model employed an L_2 penalty with a stopping criterion of 0.0001.
2. RF is a bagging ensemble learning method based on multiple independent decision trees. The classification result is determined by the voting of decision trees. It is known to be relatively stable in high-dimensional data analysis compared to other ML algorithms.[25] In this study, an RF model consisting of 200 trees was constructed, and the Gini measure was used as a criterion. There were no restrictions on tree depth, and at least 2 samples were required to split an internal node. To address overfitting, bootstrapping was allowed during the training process.

3. XGBoost[26] is an implementation of the gradient-boosted decision trees ensemble algorithm. It consists of numerous simple decision trees that subsequent trees based on the errors made by previous trees, thereby reducing variance and bias. It is not only robust to over-fitting issues but is quite fast as it utilizes parallel and distributed computing. Also, it has been generally considered a better classifier than RF for imbalanced datasets. A model with 100 estimators was built using the Python XGBoost library, with a learning rate set at 0.3. The maximum depth of the tree was limited to 6.
4. ANN[27,28] is a well-known supervised ML algorithm based on the structure of a human neuron. It has been widely applied in data mining and has shown promise in risk stratification and early warning of critical complications. In this study, a deep neural network (DNN) with 6 hidden layers was constructed, where each hidden layer consisted of 256 nodes and a dropout rate of 10% was applied. To handle imbalanced samples, the focal loss function was used. The ANN model employed the Adam optimizer with a learning rate of 0.001 and the SeLU activation function. The ANN model was built using the Python PyTorch framework and the scikit-learn library.

Model Evaluation

Overall, we used a combination of discrimination and calibration metrics to comprehensively evaluate the performance of ML-based models in comparison to the APACHE-IV scores, which served as the reference standard for outcome prediction. The predicted mortality based on APACHE-IV scores was obtained from the eICU-CRD database. Given the imbalanced nature of the data, where the mortality prevalence for the entire population was 7.9%, AUROC alone may not be a reliable evaluation indicator. Therefore, we considered the Precision and Recall (PR) curve. The PR curve is a graph that plots precision values on the y-axis against recall values on the x-axis. We calculated the area under the PR curve (AUPRC) as it is a sensitive discrimination indicator for imbalanced datasets.[29] This allows us to compare the models effectively. To assess calibration, traditional methods such as the Hosmer-Lemeshow test may become unreliable when dealing with large sample sizes.[30,31] Instead, we utilized the Brier score as a quantitative indicator for calibration assessment. Brier score measures the mean squared difference between the predicted probabilities and the observed outcomes, with lower values indicating better calibration. Additionally, calibration curves were plotted to visually represent the calibration performance. The slope of these curves indicates the relationship between the predicted and observed outcomes, providing an intuitive understanding of the calibration performance.

Optimal Model Application in Short-Term Mortality Risk Prediction

In order to validate the ML-based models' performance in short-term mortality risk prediction within one week, the optimal model that demonstrated the highest overall performance in in-hospital mortality prediction was further validated. By using cohorts of patients with a hospital stay length ranging from 1 day to 7 days. For each validation cohort, we defined specific time points based on the patient's proximity to death or discharge. The last 24-hour window to either event was defined as day 1, day 2, and so on until day 7. On each of these time points, we extracted relevant features for mortality prediction, resulting in seven distinct mortality prediction tasks: 1-day mortality risk prediction, 2-day mortality risk prediction, and so forth up to 7-day mortality risk prediction. Similar to the previous evaluation, we assessed the performance of the model using AUROC, AUPRC, and Brier scores for each of the mortality prediction tasks.

Statistical Analysis

To compare the characteristics of patients between the death and survival groups, we utilized appropriate statistical tests based on the type and distribution of data. Categorical variables were presented as proportions, while continuous variables were reported as mean with standard deviation (SD), or median with interquartile range (IQR). Comparative analysis was conducted using Student's *t*-tests, chi-squared tests, or the Mann–Whitney *U*-tests. A two-tailed *P* value less than 0.05 was considered statistically significant. Statistical analyses and ML algorithms were conducted using SPSS 18.0 and Python version 3.6, along with the scikit-learn version 0.22.1.

Results

Baseline Characteristics

A total of 123,929 ICU patients were included in this study. Among them, 9844 patients died during hospitalization, while 114,085 survived and were discharged. The overall mortality rate was only 7.9%, indicating that the dataset was highly imbalanced. The average age of all patients included in the study was 64 years, and 53.8% of them were men. The majority of the subjects were Caucasian individuals, accounting for 76.7% of the sample, The median length of hospitalization for these patients was 5 days. The basic characteristics of the dead and surviving patients on admission were shown in Table 1. Compared to the survival patients, the dead patients were significantly older, had higher APACHE IV sores, and experienced longer stays in both ICU and the hospital (P<0.05). The detailed comparison of these characteristics between the two groups were described in Table 1.

Comparison of Models' Performance

We used the APACHE IV scores model as a benchmark for performance evaluation. As demonstrated in Table 2, the APACHE IV scores exhibited an AUROC of 0.8598 in predicting in-hospital mortality in the independent testing set. However, the ML models showcased superior performance compared to the APACHE IV scores. Notably, the XGBoost model achieved the highest AUROC value(0.9702) and AUPRC value (0.8517). As for calibration, all ML models demonstrated good calibration (Figure 2), with the XGBoost model attaining the best Brier score of 0.0259. Overall it appears that XGBoost performed exceptionally well in both discrimination and calibration evaluations.

Optimal Model Analysis

The XGBoost model was selected for further application due to its superior performance in terms of AUROC and AUPRC. We examined the discrimination and calibration performance of the XGBoost model in predicting mortality risk within different time frames, specifically 1 to 7 days. The results are presented in Table 3 and Figure 3. The validation cohorts for short-term mortality risk prediction ranged from 4667 to 12,393 samples. When the XGBoost was used for mortality prediction within longer periods. The values of AUROC and AUPRC decreased. Even in the 7-day mortality risk prediction task, the XGBoost model achieved a respectable AUROC of 0.8406, albeit with a relatively lower

Table 1 Basic Characteristics for the Survival and Death Groups

Variable Name	All Patients n=123,929	Alive n=114,085	Death n=9844	*P* value
Age (years), mean (SD)	63.5±17.7	62.9±17.7	70.2±15.9	<0.01
Male	66,704(53.8%)	61,423(53.8%)	5281(54.0%)	0.72
APACHE IV scores, mean (SD)	55.2±24.9	52.4±22.1	87.6±31.4	<0.01
Ethnicity (%)				<0.01
Asian	1783(1.4%)	1626(1.4%)	157(1.6%)	
Hispanic	4853(3.9%)	4475(3.9%)	378(3.8%)	
Caucasian	95,102(76.7%)	87,399(76.6%)	7703(78.3%)	
American	15,264(12.3%)	14,152(12.4%)	1113(11.2%)	
Other	5137(5.7%)	6434(5.7%)	493(5.0%)	
Length of stay in ICU (days), median(IQR)	1.9 (1.1–3.4)	1.8 (1.0–3.2)	3.1 (1.5–6.2)	<0.01
Length of stay in hospital (days), median (IQR)	5.3 (3.1–9.1)	5.3 (1.0–9.0)	5.2 (2.6–10.1)	<0.01

Abbreviations: SD, standard deviation; IQR, interquartile range (25–75%).

Table 2 Model Performance in the Independent Testing Dataset for Different Machine Learning Algorithms

Algorithms	AUROC	AUPRC	Brier Score
APACHE IV scores	0.8598	0.4341	0.0624
Logistic Regression	0.9357	0.7775	0.0306
Random Forest	0.9559	0.8172	0.0285
Artificial Neural Network	0.9620	0.8256	0.0501
Extreme Gradient Boosting	**0.9702**	**0.8517**	**0.0259**

Abbreviations: AUROC, area under the receiver-operator curve; AUPRC, area under the Precision and Recall curve.

AUPRC of 0.3533. Regarding calibration, the Briers scores ranged from 0.0295 to 0.0592, which were within an acceptable range.

Discussion

Overall, we aimed to assess the feasibility of using ML algorithms to predict mortality risk among ICU patients, utilizing routinely collected physical and medication data in clinical care. The XGBoost algorithm was demonstrated outstanding performance in identifying high-risk patients. The model's predictive capability validated on an independent test dataset and compared with the traditional severity scoring system, APACHE IV scores, to enhance reliability. Further, we validate the performance of the XGBoost model in short-term mortality risk prediction.

In our study, the ML-based prediction models outperformed APACHE IV scores in predicting in-hospital mortality for ICU patients, with the XGBoost algorithm yielding better performance compared to other methods. XGBoost is an ensemble method based on decision trees, known for its excellent performance in various prediction tasks. Compared to RF, the XGBoost algorithm assigns higher learning weights to the samples with lower accuracy in prior rounds of decision tree training, thereby improving overall algorithm accuracy.[26] The results indicated that the XGBoost algorithm exhibited exceptional performance in both AUROC and AUPRC. It achieved an impressive AUROC score of 0.9702 in the 24-hour mortality risk prediction task, indicating its ability to effectively classify high-risk and low-risk patients. However, since AUROC may not fully reflect model performance on imbalanced data[32,33], AUPRC was also employed as an evaluation metric. The XGBoost algorithm yielded an AUPRC of 0.8517, further bolstering the credibility and

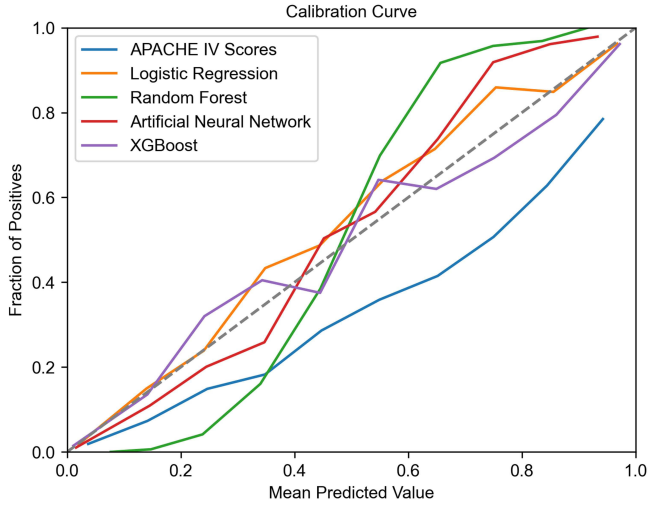

Figure 2 Calibration curves of the machine learning models in the test dataset. X-axis indicates the predicted mean mortality risk, Y-axis indicates the actual mean mortality risk, the slope indicates the relation between the predicted and observed outcomes.

Table 3 The XGBoost Model Performance in Different Short-Term Mortality Risk Prediction Tasks

Prediction Tasks	AUROC	AUPRC	Brier Score
1-day mortality prediction (n=12,393)	0.9702	0.8517	0.0259
2-day mortality prediction (n=10,907)	0.9310	0.6437	0.0427
3-day mortality prediction (n=9403)	0.9148	0.5519	0.0477
4-day mortality prediction (n=7935)	0.8971	0.4835	0.0517
5-day mortality prediction (n=6697)	0.8627	0.3720	0.0592
6-day mortality prediction (n=5148)	0.8544	0.3731	0.0647
7-day mortality prediction (n=4667)	0.8406	0.3533	0.0668

Abbreviations: AUROC, area under the receiver-operator curve; AUPRC, area under the Precision and Recall curve.

robustness of the model. A higher AUPRC indicates that the model achieved better precision-recall trade-off, addressing the limitations associated with imbalanced data.

The remarkable prediction capabilities of the XGBoost model can be attributed to the well-defined features used in this study. These features were extracted from a specific 24-hour time window prior to patient death or discharge, capturing crucial physiological data reflecting the differences in health status between the dead and surviving patients. Generally, within this time period, patients either recover from a severe illness or unfortunately died after this time moment. Therefore, the physiological data obtained during this timeframe accurately reflects the disparities in health status between the dead and the survival patients. This enables the classifiers to effectively capture variations in patients who will be died, leading to more precise predictions. Moreover, we incorporated multi-dimensional features, including fundamental indicators of health status that are easily understood by healthcare professionals. Our model not only included objectively measured variables during critical care but also considered pharmacological therapy as predictors. This integration reflects the direct influence of human intelligence[34] and enhances the richness of features, consequently increasing the accuracy and reliability of the prediction model.[35]

All proposed ML models in our study achieved favorable results (AUROC=0.9357–0.9702) when compared to published in-hospital mortality risk prediction tools.[36] This highlights the potential and promise of ML techniques in

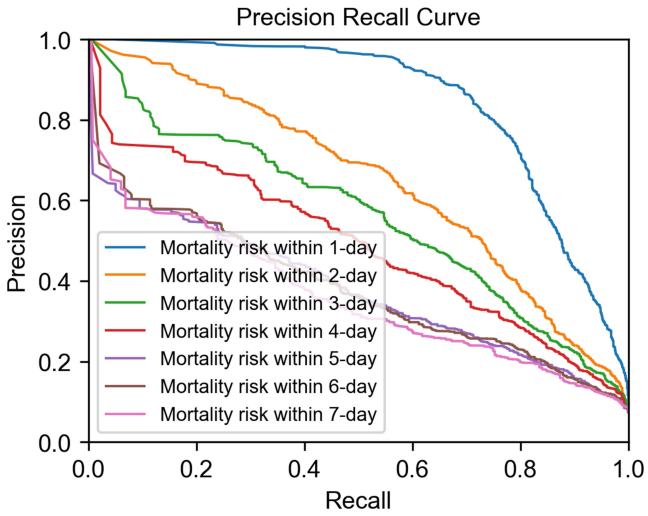

Figure 3 Precision-Recall curve of the XGBoost model in different tasks. Lines with different color indicate mortality risk prediction task within different time periods. Task with bigger area under Precision-Recall curves shows better discrimination performance.

accurately predicting mortality risk. Furthermore, the data elements employed in our model can be seamlessly accessed through the hospital information system without requiring additional data entry by clinicians. This ensures that our predictions are up-to-date and less susceptible to missing data. It is worth noting that the cost of manual data collection poses a significant impediment to the clinical application of predictive models.[4] However, all vital signs, laboratory tests, and infusion drug information used in our study were collected automatically or through regular manual measurements during routine patient care workflows. Given the widespread adoption of hospital information systems and the accessibility of ML algorithms, our models can be broadly and efficiently used even in low-resource healthcare environments.

To our knowledge, this research represents one of the limited efforts to apply a mortality prediction model from a clinician's perspective. A mortality risk model that holds clinical significance is capable of providing dynamic and reliable predictions throughout a patient's hospital stay. We employed the XGBoost model for predicting mortality risks within a 7-day time frame. This served as an assessment of the model's performance in short-term mortality prediction. As shown in Table 3, the proposed XBGoost model exhibited superior AUROC values compared to severity scoring systems currently in use and most ML models reported in previous studies. Although the AUROC and AUPRC scores were least optimal when predicting 7-day mortality, they progressively increased over the course of the hospital stay. This suggests that in our model, using the most up-to-date data for predictions consistently yields more accurate forecasts, thereby enhancing clinicians' trust in the model. Baker et al also reported a similar trend in AUROC, where they developed hybrid neural network models to predict mortality risk for the 3, 7, and 14-day windows.[37] Our XGBoost model achieved comparable AUROC(0.9148) and AUPRC(0.5519) values in 3-day mortality risk prediction when compared to the model presented by Baker. However, it is worth noting that few researchers have reported AUPRC scores and calibration for their models, making comprehensive comparisons challenging. Considering the XGBoost model's excellent discrimination performance and reasonable calibration, we can conclude that our model possesses stable and reliable capabilities in short-term mortality prediction within a 3-day window. These findings instill confidence in the practical application of our model in clinical settings, as it has the potential to accurately identify high-risk patients at an earlier stage. Consequently, health professionals can make better decisions regarding treatment, resource allocation, and other aspects of patient care.[15]

Limitations

Admittedly, there are some limitations that should be addressed. At first, although we used the multi-center eICU database for training and validation of the model, external validation focused on Asians and Hispanics is necessary before applying the model to these populations, as the proportions of these ethnic groups in the eICU database are small. Additionally, the XGBoost algorithm was applied in a retrospective cohort to predict short-term mortality risk within 7 days and showed promising AUROC performance. However, its ability to predict mortality risk over a longer period of time remains unclear. In our analysis, 37% of patients who met our selection criteria in the eICU database stayed in the hospital for more than 7 days, indicating that the performance of the proposed model in predicting mortality risk beyond 7 days has not been verified. Moreover, the relatively poor AUPRC in the 7-day mortality risk prediction task should not be overlooked. Therefore, the clinical significance of the model still requires further examination through well-designed prospective studies. In future work, we plan to deploy the model on a small scale in clinical practice to assess its effectiveness in real-world scenarios.

Conclusion

In conclusion, highly accurate and well-calibrated mortality risk prediction based on the XGboost model with routinely recorded and interpretative features derived from a 24-hour window prior to patient discharge is feasible and has the potential to augment the clinician's decision-making process. The excellent performance of the XGBoost model in short-term mortality risk prediction tasks strengthens its reliability for dynamic mortality risk assessments in patient care.

Disclosure

The authors report no conflicts of interest in this work.

References

1. Keegan MT, Gajic O, Afessa B. Severity of illness scoring systems in the intensive care unit. *Crit Care Med*. 2011;39(1):163–169.
2. Salluh JI, Soares M. ICU severity of illness scores: APACHE, SAPS and MPM. *Curr Opin Crit Care*. 2014;20(5):557–565. doi:10.1097/MCC.0000000000000135
3. Knaus WA, Draper EA, Wagner DP, et al. APACHE II: a severity of disease classification system. *Crit Care Med*. 1985;13(10):818–829. doi:10.1097/00003246-198510000-00009
4. Breslow MJ, Badawi O. Severity scoring in the critically ill: part 2: maximizing value from outcome prediction scoring systems. *Chest*. 2012;141(2):518–527. doi:10.1378/chest.11-0331
5. Kramer AA. Predictive mortality models are not like fine wine. *Crit Care*. 2005;9(6):636–637. doi:10.1186/cc3899
6. Sakr Y, Krauss C, Amaral AC, et al. Comparison of the performance of SAPS II, SAPS 3, APACHE II, and their customized prognostic models in a surgical intensive care unit. *Br J Anaesth*. 2008;101:798–803. doi:10.1093/bja/aen291
7. Falcão ALE, Barros AGA, Bezerra AAM, et al. The prognostic accuracy evaluation of SAPS 3, SOFA and APACHE II scores for mortality prediction in the surgical ICU: an external validation study and decision-making analysis. *Ann Intensive Care*. 2019;9(1):18. doi:10.1186/s13613-019-0488-9
8. Delahanty RJ, Kaufman D, Jones SS. Development and evaluation of an automated machine learning algorithm for in-hospital mortality risk adjustment among critical care patients. *Crit Care Med*. 2018;46(6):e481–e488. doi:10.1097/CCM.0000000000003011
9. Ruyssinck J, van der Herten J, Houthooft R, et al. Random survival forests for predicting the bed occupancy in the intensive care unit. *Comput*. 2016;2016:7087053.
10. Ngufor C, Murphree D, Upadhyaya S, et al. Predicting prolonged stay in the ICU attributable to bleeding in patients offered plasma transfusion. *AMIA Annu Symp Proc*. 2017;2016:954–963.
11. Zhang L, Huang T, Xu F, et al. Prediction of prognosis in elderly patients with sepsis based on machine learning (random survival forest). *BMC Emerg Med*. 2022;22(1):1–10.
12. Raita Y, Goto T, Faridi MK, et al. Emergency department triage prediction of clinical outcomes using machine learning models. *Crit Care*. 2019;23(1):64. doi:10.1186/s13054-019-2351-7
13. Hu C, Li L, Li Y, et al. Explainable machine-learning model for prediction of in-hospital mortality in septic patients requiring intensive care unit readmission. *Infect Dis Ther*. 2022;11(4):1695–1713.
14. Batista A, Diniz C, Bonilha EA, et al. Neonatal mortality prediction with routinely collected data: a machine learning approach. *BMC Pediatr*. 2021;21(1):322. doi:10.1186/s12887-021-02788-9
15. Alghatani K, Ammar N, Rezgui A, et al. Predicting intensive care unit length of stay and mortality using patient vital signs: machine learning model development and validation. *JMIR Med Inform*. 2021;9(5):e21347. doi:10.2196/21347
16. Subudhi S, Verma A, Patel AB, et al. Comparing machine learning algorithms for predicting ICU admission and mortality in COVID-19. *NPJ Digit Med*. 2021;4(1):87.
17. Deliberato RO, Escudero GG, Bulgarelli L, et al. SEVERITAS: an externally validated mortality prediction for critically ill patients in low and middle-income countries. *Int J Med Inform*. 2019;131:103959. doi:10.1016/j.ijmedinf.2019.103959
18. Yu R, Zheng Y, Zhang R, et al. Using a multi-task recurrent neural network with attention mechanisms to predict hospital mortality of patients. *IEEE J Biomed Health Inform*. 2020;24(2):486–492. doi:10.1109/JBHI.2019.2916667
19. Hou N, Li M, He L, et al. Predicting 30-days mortality for MIMIC-III patients with sepsis-3: a machine learning approach using XGboost. *J Transl Med*. 2020;18(1):462.
20. Liu J, Wu J, Liu S, et al. Predicting mortality of patients with acute kidney injury in the ICU using XGBoost model. *PLoS One*. 2021;16(2):e0246306. doi:10.1371/journal.pone.0246306
21. Shillan D, Sterne JAC, Champneys A, et al. Use of machine learning to analyse routinely collected intensive care unit data: a systematic review. *Crit Care*. 2019;23(1):284.
22. Pollard TJ, Johnson AEW, Raffa JD, et al. The eICU collaborative research database, a freely available multi-center database for critical care research. *Scientific Data*. 2018;5:180178. doi:10.1038/sdata.2018.178
23. Hameed M, Alamgir Z. Improving mortality prediction in Acute Pancreatitis by machine learning and data augmentation. *Comput Biol Med*. 2022;150:106077. doi:10.1016/j.compbiomed.2022.106077
24. Zhao SP, Liu P, Tang GX, et al. External validation of a deep learning prediction model for in-hospital mortality among ICU patients. 2022 IEEE 2nd International Conference on Power, Electronics and Computer Applications (ICPECA); 2022 January 21–23; Shenyang, China: IEEE; 2022.
25. Sarica A, Cerasa A, Quattrone A, et al. Random forest algorithm for the classification of neuroimaging data in Alzheimer's disease: a systematic review. *Front Aging Neurosci*. 2017;9:329.
26. Chen T, Guestrin C. XGBoost: a scalable tree boosting system. Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining-KDD. 2016 August 13–17; San Francisco, CA, USA; 2016.
27. Renganathan V. Overview of artificial neural network models in the biomedical domain. *Bratisl Lek Listy*. 2019;120(7):536–540. doi:10.4149/BLL_2019_087
28. Miotto R, Wang F, Wang S, et al. Deep learning for healthcare: review, opportunities and challenges. *Brief Bioinform*. 2018;19(6):1236–1246. doi:10.1093/bib/bbx044
29. Yuan Y, Su W, Zhu M. Threshold-free measures for assessing the performance of medical screening tests. *Front Public Health*. 2015;3:57. doi:10.3389/fpubh.2015.00057
30. Paul P, Pennell ML, Lemeshow S. Standardizing the power of the Hosmer-Lemeshow goodness of fit test in large data sets. *Stat Med*. 2013;32:67–80. doi:10.1002/sim.5525
31. Kramer AA, Zimmerman JE. Assessing the calibration of mortality benchmarks in critical care: the Hosmer-Lemeshow test revisited. *Crit Care Med*. 2007;35:2052–2056. doi:10.1097/01.CCM.0000275267.64078.B0
32. Schrynemackers M, Küffner R, Geurts P. On protocols and measures for the validation of supervised methods for the inference of biological networks. *Front Genet*. 2013;4:262. doi:10.3389/fgene.2013.00262

33. Boyd K, Santos Costa V, Davis J, et al. Unachievable region in precision-recall space and its effect on empirical evaluation. *Proc Int Conf Mach Learn*. 2012;2012:349.
34. Meyer A, Zverinski D, Pfahringer B, et al. Machine learning for real-time prediction of complications in critical care: a retrospective study. *Lancet Respir Med*. 2018;6(12):905–914. doi:10.1016/S2213-2600(18)30300-X
35. Davenport TH, Barth P, Bean R. Why detailed data is as important as big data. *MIT Sloan Manage Rev*. 2012;53(4):1–3.
36. Kramer AA, Higgins TL, Zimmerman JE. Comparison of the mortality probability admission model iii, national quality forum, and acute physiology and chronic health evaluation IV hospital mortality models: implications for national benchmarking. *Crit Care Med*. 2014;42:544–553. doi:10.1097/CCM.0b013e3182a66a49
37. Baker S, Xiang W, Atkinson I. Continuous and automatic mortality risk prediction using vital signs in the intensive care unit: a hybrid neural network approach. *Sci Rep*. 2020;10(1):21282. doi:10.1038/s41598-020-78184-7