frontiers | Research Topics

# Clinical application of artificial intelligence in emergency and critical care medicine, volume III

**Edited by**
Zhongheng Zhang, Rahul Kashyap, Longxiang Su, Nan Liu and Qinghe Meng

**Published in**
Frontiers in Medicine

ISSN 1664-8714
ISBN 978-2-83251-256-2
DOI 10.3389/978-2-83251-256-2

## About Frontiers

Frontiers is more than just an open access publisher of scholarly articles: it is a pioneering approach to the world of academia, radically improving the way scholarly research is managed. The grand vision of Frontiers is a world where all people have an equal opportunity to seek, share and generate knowledge. Frontiers provides immediate and permanent online open access to all its publications, but this alone is not enough to realize our grand goals.

## Frontiers journal series

The Frontiers journal series is a multi-tier and interdisciplinary set of open-access, online journals, promising a paradigm shift from the current review, selection and dissemination processes in academic publishing. All Frontiers journals are driven by researchers for researchers; therefore, they constitute a service to the scholarly community. At the same time, the *Frontiers journal series* operates on a revolutionary invention, the tiered publishing system, initially addressing specific communities of scholars, and gradually climbing up to broader public understanding, thus serving the interests of the lay society, too.

## Dedication to quality

Each Frontiers article is a landmark of the highest quality, thanks to genuinely collaborative interactions between authors and review editors, who include some of the world's best academicians. Research must be certified by peers before entering a stream of knowledge that may eventually reach the public - and shape society; therefore, Frontiers only applies the most rigorous and unbiased reviews. Frontiers revolutionizes research publishing by freely delivering the most outstanding research, evaluated with no bias from both the academic and social point of view. By applying the most advanced information technologies, Frontiers is catapulting scholarly publishing into a new generation.

## What are Frontiers Research Topics?

Frontiers Research Topics are very popular trademarks of the *Frontiers journals series*: they are collections of at least ten articles, all centered on a particular subject. With their unique mix of varied contributions from Original Research to Review Articles, Frontiers Research Topics unify the most influential researchers, the latest key findings and historical advances in a hot research area.

Find out more on how to host your own Frontiers Research Topic or contribute to one as an author by contacting the Frontiers editorial office: frontiersin.org/about/contact

# Clinical application of artificial intelligence in emergency and critical care medicine, volume III

**Topic editors**

Zhongheng Zhang — Department of Emergency Medicine, Sir Run Run Shaw Hospital, China

Rahul Kashyap — WellSpan Health, United States

Longxiang Su — Peking Union Medical College Hospital (CAMS), China

Nan Liu — National University of Singapore, Singapore

Qinghe Meng — Upstate Medical University, United States

**Citation**

Zhang, Z., Kashyap, R., Su, L., Liu, N., Meng, Q., eds. (2023). *Clinical application of artificial intelligence in emergency and critical care medicine, volume III*. Lausanne: Frontiers Media SA. doi: 10.3389/978-2-83251-256-2

# Table of contents

frontiers | Frontiers in Medicine

TYPE Editorial
PUBLISHED 19 December 2022
DOI 10.3389/fmed.2022.1075023



OPEN ACCESS

EDITED AND REVIEWED BY
Alfredo Vellido,
Universitat Politecnica de Catalunya, Spain

*CORRESPONDENCE
Zhongheng Zhang
zh_zhang1984@zju.edu.cn

SPECIALTY SECTION
This article was submitted to Intensive Care Medicine and Anesthesiology, a section of the journal Frontiers in Medicine

RECEIVED 20 October 2022
ACCEPTED 04 November 2022
PUBLISHED 19 December 2022

CITATION
Zhang Z, Kashyap R, Su L and Meng Q (2022) Editorial: Clinical application of artificial intelligence in emergency and critical care medicine, volume III. *Front. Med.* 9:1075023. doi: 10.3389/fmed.2022.1075023



# Editorial: Clinical application of artificial intelligence in emergency and critical care medicine, volume III

Zhongheng Zhang[1,2]*, Rahul Kashyap[3,4], Longxiang Su[5] and Qinghe Meng[6]

[1]Department of Emergency Medicine, Key Laboratory of Precision Medicine in Diagnosis and Monitoring Research of Zhejiang Province, Sir Run Run Shaw Hospital, Zhejiang University School of Medicine, Hangzhou, China, [2]Key Laboratory of Digital Technology in Medical Diagnostics of Zhejiang Province, Hangzhou, China, [3]Critical Care Independent Multidisciplinary Program, Mayo Clinic, Rochester, MN, United States, [4]Department of Anesthesiology and Perioperative Medicine, Mayo Clinic, Rochester, MN, United States, [5]State Key Laboratory of Complex Severe and Rare Diseases, Department of Critical Care Medicine, Peking Union Medical College Hospital, Chinese Academy of Medical Science and Peking Union Medical College, Beijing, China, [6]Department of Surgery, State University of New York Upstate Medical University, Syracuse, NY, United States

KEYWORDS

**artificial intelligence, critical care, heterogeneity, prediction, diagnosis**

Editorial on the Research Topic
Clinical application of artificial intelligence in emergency and critical care medicine, volume III

Two years have passed since the first launch of the Research Topic on the application of artificial intelligence (AI) in emergency and critical care settings (1). We have witnessed increasing submissions to this topic over these years, indicating continued research interest among the critical care community. AI is a data analysis approach that has revolutionized many industry areas (2, 3), as well as clinical medicine (4). With more data being captured and stored during routine clinical practice, the large volumes of data have the potential to reveal more knowledge to better inform clinical decision makings (5). In general, clinical questions involving all stages of clinical practice including diagnosis, treatment, and prognosis can be well-investigated by the employment of AI technology. Figure 1 illustrates how AI can help to make better patient care in all stages of clinical practice. Diagnosis is usually the first step in the management of patients. Prompt and accurate diagnosis can help better patient treatment in the critical care setting. For instance, there has been a large body of evidence showing that early initiation of a sepsis care bundle can help to improve survival outcomes, and thus strenuous efforts have been made to provide early warning for sepsis. The automated early warning system has been widely applied in the clinical setting, and preliminary studies show promising results (6). With the help of AI, the identification of sepsis can be done earlier with increased accuracy (7). The second aspect relating to the diagnosis refers to the sub-classification of a heterogeneous syndrome. Many diseases or syndromes in the critical care

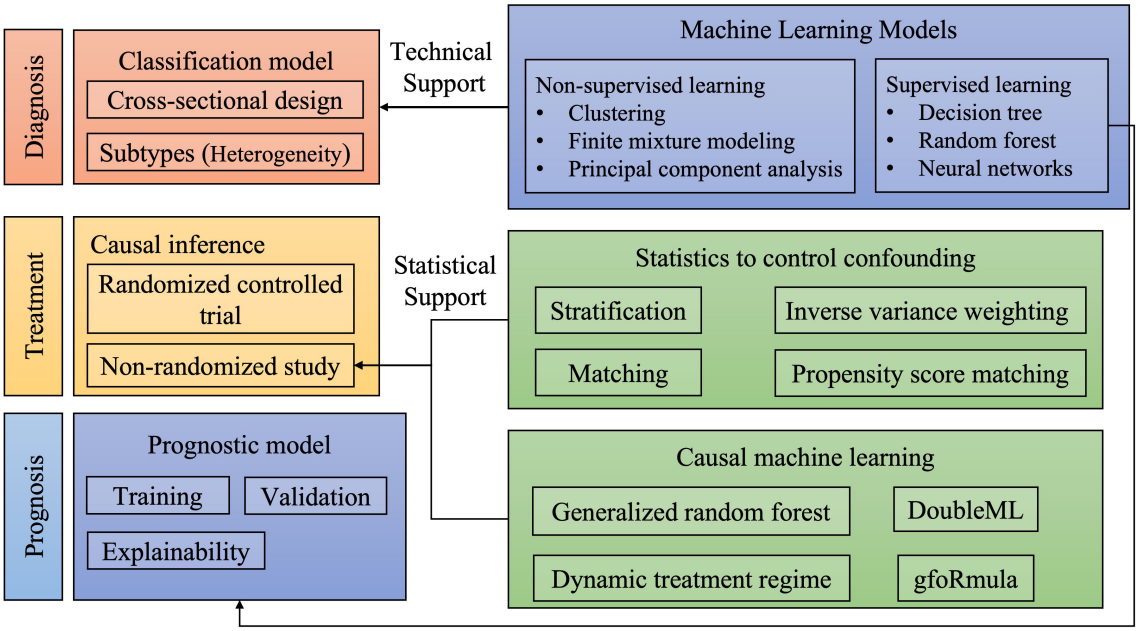



FIGURE 1
Illustrations of the application of artificial intelligence to various aspects of clinical practice.

setting encompass a heterogenous population and the identification of subtypes can help tailor treatment strategies (8). In volume III of the topic series, Wu et al. trained a classification model on facial expressions video clips, and their deep learning method is shown to accurately classify patients with or without pain. This important study implies that pain assessment can be achieved by an automated computer system, thereby providing high granularity time-varying facial expression data for patient management. Sepsis-Associated Thrombocytopenia (SAT) is an important complication in sepsis patients and early risk stratification can help to tailor individualized treatment. Jiang et al. trained multiple machine learning (ML) models for the prediction of SAT in a Chinese cohort, and then these models were validated in an open-access critical care database.

The second step in patient management involves the treatment strategy. Since critically ill patients are usually treated with multi-module strategies, the effectiveness of treatment strategies is time sensitive, and varies across the individual subject. Thus, an individualized treatment strategy is needed, in line with the idea of precision medicine. Mechanical ventilator (MV) weaning is an important medical decision-making process for the management of patients on MV. Liu et al. developed an AI algorithm to dictate MV weaning. This study paves the way for the realization of personalized medicine in the management of MV patients.

Finally, the prognosis is also important for the management of critically ill patients. Risk stratification for intensive care unit (ICU) patients is useful for clinicians to make better decisions and for consulting with family members. The diagnosis and prognosis can be studied with the supervised ML algorithm. The difference lies in the study design. While the studies on diagnostic performance require cross-sectional data to train the model, those involving prognostic performance require a follow-up period allowing the outcome (label) to occur. ICU readmission is an important indicator of the quality of care and is an important outcome measurement. In this topic issue, Hegselmann et al. developed an explainable boosting machine to predict ICU re-admission using a German dataset.

In conclusion, the successful launch of the special issue on the application of AI in critical care medicine indicates that researchers continue to be interested in this particular field. The power of big data and AI are revolutionizing clinical practice in the near future. The ICU is a highly technological environment where each patient generates a large volume of data per day, such special characteristics make it the best place for AI applications.

## Author contributions

ZZ conceived the idea and drafted the manuscript. LS and QM revised the paper. RK worked to organize the topic issue

and made contributions to the editorial contents. All authors contributed to the article and approved the submitted version.

## Funding

ZZ received funding from Open Foundation of Key Laboratory of Digital Technology in Medical Diagnostics of Zhejiang Province (SZZD202206).

## Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

## Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

## References

1. Zhang Z, Liu N, Meng Q, Su L. Editorial: clinical application of artificial intelligence in emergency and critical care medicine, volume I. *Front Med.* (2021) 8:809478. doi: 10.3389/fmed.2021.809478

2. Zhang Z, Navarese EP, Zheng B, Meng Q, Liu N, Ge H, et al. Analytics with artificial intelligence to advance the treatment of acute respiratory distress syndrome. *J Evid Based Med.* (2020) 13:301–312. doi: 10.1111/jebm.12418

3. Hashimoto DA, Witkowski E, Gao L, Meireles O, Rosman G. Artificial intelligence in anesthesiology: current techniques, clinical applications, and limitations. *Anesthesiology.* (2020) 132:379–394. doi: 10.1097/ALN.0000000000002960

4. Zhang Z, Chen L, Xu P, Hong Y. Predictive analytics with ensemble modeling in laparoscopic surgery: a technical note. *Laparosc Endosc Robot Surg.* (2022) 5:25–34. doi: 10.1016/j.lers.2021.12.003

5. Greener JG, Kandathil SM, Moffat L, Jones DT. A guide to machine learning for biologists. *Nat Rev Mol Cell Biol.* (2021) 23:40–55. doi: 10.1038/s41580-021-00407-0

6. Zhang Z, Chen L, Xu P, Wang Q, Zhang J, Chen K, et al. Effectiveness of automated alerting system compared to usual care for the management of sepsis. *NPJ Digit Med.* (2022) 5:101. doi: 10.1038/s41746-022-00650-5

7. Umscheid CA, Betesh J, VanZandbergen C, Hanish A, Tait G, Mikkelsen ME, et al. Development, implementation, and impact of an automated early warning and response system for sepsis. *J Hosp Med.* (2015) 10:26–31. doi: 10.1002/jhm.2259

8. Wiersema R, Jukarainen S, Vaara ST, Poukkanen M, Lakkisto P, Wong H, et al. Two subphenotypes of septic acute kidney injury are associated with different 90-day mortality and renal recovery. *Crit Care.* (2020) 24:150. doi: 10.1186/s13054-020-02866-x

frontiers in Medicine

ORIGINAL RESEARCH
published: 27 January 2022
doi: 10.3389/fmed.2022.837382

# Prediction Models for Sepsis-Associated Thrombocytopenia Risk in Intensive Care Units Based on a Machine Learning Algorithm

*Xuandong Jiang, Yun Wang, Yuting Pan and Weimin Zhang**

*Intensive Care Unit, Dongyang Hospital of Wenzhou Medical University, Jinhua, China*

Sepsis-associated thrombocytopenia (SAT) is a common complication in the intensive care unit (ICU), which significantly increases the mortality rate and leads to poor prognosis of diseases. Machine learning (ML) is widely used in disease prediction in critically ill patients. Here, we aimed to establish prediction models for platelet decrease and severe platelet decrease in ICU patients with sepsis based on four common ML algorithms and identify the best prediction model. The research subjects were 1,455 ICU sepsis patients admitted to Dongyang People's Hospital affiliated with Wenzhou Medical University from January 1, 2015, to October 31, 2019. Basic clinical demographic information, biochemical indicators, and clinical outcomes were recorded. The prediction models were based on four ML algorithms: random forest, neural network, gradient boosting machine, and Bayesian algorithms. Thrombocytopenia was found to occur in 732 patients (49.7%). The mechanical ventilation time and length of ICU stay were longer, and the mortality rate was higher for the thrombocytopenia group than for the non-thrombocytopenia group. The models were validated on an online international database (Medical Information Mart for Intensive Care III). The areas under the receiver operating characteristic curves (AUCs) of the four models for the prediction of thrombocytopenia were between 0.54 and 0.72. The AUCs of the models for the prediction of severe thrombocytopenia were between 0.70 and 0.77. The neural network and gradient boosting machine models effectively predicted the occurrence of SAT, and the Bayesian models had the best performance in predicting severe thrombocytopenia. Therefore, these models can be used to identify such high-risk patients at an early stage and guide individualized clinical treatment, to improve the prognosis of diseases.

**Keywords: sepsis-associated thrombocytopenia, intensive care unit, machine learning, artificial intelligence, prediction**

OPEN ACCESS

***Edited by:***
*Zhongheng Zhang,*
*Sir Run Run Shaw Hospital, China*

***Reviewed by:***
*Qi Guo,*
*Sun Yat-sen University, China*
*Shaobin Duan,*
*Central South University, China*

****Correspondence:***
*Weimin Zhang*
*jalzhan@163.com*

***Specialty section:***
*This article was submitted to Intensive Care Medicine and Anesthesiology, a section of the journal Frontiers in Medicine*

***Received:*** *16 December 2021*
***Accepted:*** *07 January 2022*
***Published:*** *27 January 2022*

***Citation:***
*Jiang X, Wang Y, Pan Y and Zhang W (2022) Prediction Models for Sepsis-Associated Thrombocytopenia Risk in Intensive Care Units Based on a Machine Learning Algorithm. Front. Med. 9:837382. doi: 10.3389/fmed.2022.837382*

## INTRODUCTION

Artificial intelligence (AI) has enabled many cutting-edge scientific research achievements in the field of medical care, especially for acute and severe diseases. In fields such as disease risk assessment, early warning of disease deterioration, and early warning of death, AI can alert officials regarding potential risks earlier and more accurately. Machine learning (ML) is a branch of AI, and

it has been used for predicting disease outcomes. Using the Medical Information Mart for Intensive Care (MIMIC) database, Garcia Gallo et al. (1) established a model to predict the mortality of patients with severe sepsis based on the ML algorithm, which achieved better evaluation results than traditional scoring systems such as Sequential (sepsis-related) Organ Failure Assessment (SOFA) Score and Simplified Acute Physiology Score II. Thorsen-Meyer et al. (2) applied the ML algorithm and further employed intensive care unit (ICU) time series data analysis to predict the 90-day mortality in real-time, thus improving the prognosis of diseases in ICU patients.

Sepsis-related thrombocytopenia (SAT) is a common complication in the ICU; in particular, the incidence of thrombocytopenia in patients with septic shock can be as high as 55% (3). SAT involves many mechanisms (4), which might include inflammation-mediated platelet production changes, endothelial dysfunction, abnormal blood coagulation function, and hemodilution. Thrombocytopenia can significantly increase the incidence of complications and mortality in patients with sepsis (5). A study by Azkárate et al. (6) showed that thrombocytopenia was associated with a 1.7-fold increased risk of mortality in severe sepsis patients. Thrombocytopenia may cause severe hemorrhage; a multicenter observational study (7) in UK ICU found that a total of 169 patients (9% of the study population) received platelet transfusion, and the prevalence of severe thrombocytopenia ($<50 \times 10(9)$ /L) was 12.4, and 35.4% of the patients finally died in the ICU. In actual clinical work, when a decrease in platelet count is observed for a patient, especially a severe decrease, platelets should be infused in time to reduce the risk of bleeding because platelets cannot be stored for a long time. However, patients may have to wait for 2–3 days from the beginning of platelet reservation to the actual infusion of platelets. In this process, the patients are at a high risk of bleeding and may even experience hemorrhagic shock, which is life-threatening. Early detection of platelet decrease is crucial for critically ill patients.

Presently, there are many related models for predicting sepsis using artificial intelligence (8, 9), which can enhance doctors' medical decision-making ability for patients with sepsis. However, research on predicting SAT and severe thrombocytopenia in the ICU is lacking, and effective models for predicting SAT using ML algorithms have not yet been established. Therefore, we used a large amount of real-time data from the ICU to establish prediction models for thrombocytopenia in ICU sepsis patients for the early identification of patients with a high risk of thrombocytopenia, which would help reduce the occurrence of bleeding events and improve the prognosis of diseases in patients.

# MATERIALS AND METHODS

## Study Design and Research Subjects

Our study was reported according to the guidelines of the TRIPOD (10) statement (Checklist in **Additional File 1**). A retrospective study was conducted with 1,455 sepsis patients who were admitted to the ICU of Dongyang People's Hospital between January 1, 2015, and October 31, 2019. External validation was performed using the MIMIC III dataset (11), a freely accessible online critical care database. The inclusion criteria were age $\geq$18 years and admission to the ICU with sepsis. The exclusion criteria were patients who had hematological malignancy, cirrhosis patients who had underlying thrombocytopenia before ICU admission, and patients who had undergone splenectomy.

This study was approved by the Ethics Committee of Dongyang People's Hospital (DRY-2021-YX-178). The need for informed consent was waived because of the retrospective, observational study design. The data were anonymously analyzed after the removal of personal information from the data. One author (XJ) obtained permission for accessing the MIMIC database after the completion of "Protecting Human Research Participants," an online training course launched by the National Institutes of Health (certification number: 7632299).

## Data Collection and Grouping

### Data Collection

Data were collected using the medical record information mining software provided by Shanghai Le9 Healthcare Technology Co., Ltd. The collected information included the following: (1) basic clinicodemographic information [age, sex, disease severity (Acute Physiology and Chronic Health Evaluation, APACHE II score, SOFA score), smoking history, alcohol abuse history, and complications]; (2) blood gas, blood routine, biochemistry, and liver function indicators on the first day of ICU admission; and (3) clinical outcomes (mortality, time on ventilator, length of ICU stay, length of hospital stay, and hospitalization cost).

### Diagnostic Criteria

Definition of SAT: Sepsis patients with thrombocytopenia.

Thrombocytopenia (12, 13):Platelet count $<100 \times 109$/L or a 30% relative decrease of the baseline platelet count during ICU stay; the baseline platelet count was defined as the highest value over the past seven days before ICU admission. We used the initial platelet count in ICU admission as baseline platelet count for patients without platelet count measurement before ICU admission.

Severe thrombocytopenia (14, 15): Platelet count $<50 \times 109$/L during ICU stay.

Sepsis 3.0 (16): Organ dysfunction triggered by an infection that endangers the patient's life and rapid increase in the SOFA score, with a total score of two points.

Sepsis shock (16): The patient with sepsis requiring vasopressors to maintain mean blood pressure at 65 mmHg or higher and having a serum lactate level higher than 2 mmol/L (18 mg/dL) after fluid resuscitation.

## Data Processing

### Selection of Independent Variables

Sixty-five potentially related variables were preliminarily screened. After excluding three variables with more than 15% of missing values, the remaining 62 variables were subjected to data preprocessing using CARET in R language. Thirteen variables showing a strong correlation (correlation coefficient $>0.9$) with other independent variables were eliminated. The remaining 57 variables were then subjected to feature

**TABLE 1 |** Comparison of the additional evaluation metrics of four machine learning models in external validation.

| | Models for predicting thrombocytopenia | | | |
|---|---|---|---|---|
| | **RF** | **Bayesian** | **NNET** | **GBM** |
| Accuracy | 0.61 | 0.55 | 0.68 | 0.67 |
| Precision | 0.61 | 0.59 | 0.71 | 0.67 |
| Recall | 0.72 | 0.40 | 0.65 | 0.74 |
| Specificity | 0.50 | 0.70 | 0.71 | 0.61 |
| | **Models for predicting severe thrombocytopenia** | | | |
| | **RF** | **Bayesian** | **NNET** | **GBM** |
| Accuracy | 0.71 | 0.68 | 0.72 | 0.72 |
| Precision | 0.47 | 0.45 | 0.48 | 0.48 |
| Recall | 0.55 | 0.84 | 0.59 | 0.49 |
| Specificity | 0.77 | 0.62 | 0.77 | 0.81 |

*RF, random forest; NNET, neural network; GBM, gradient boosting machine.*

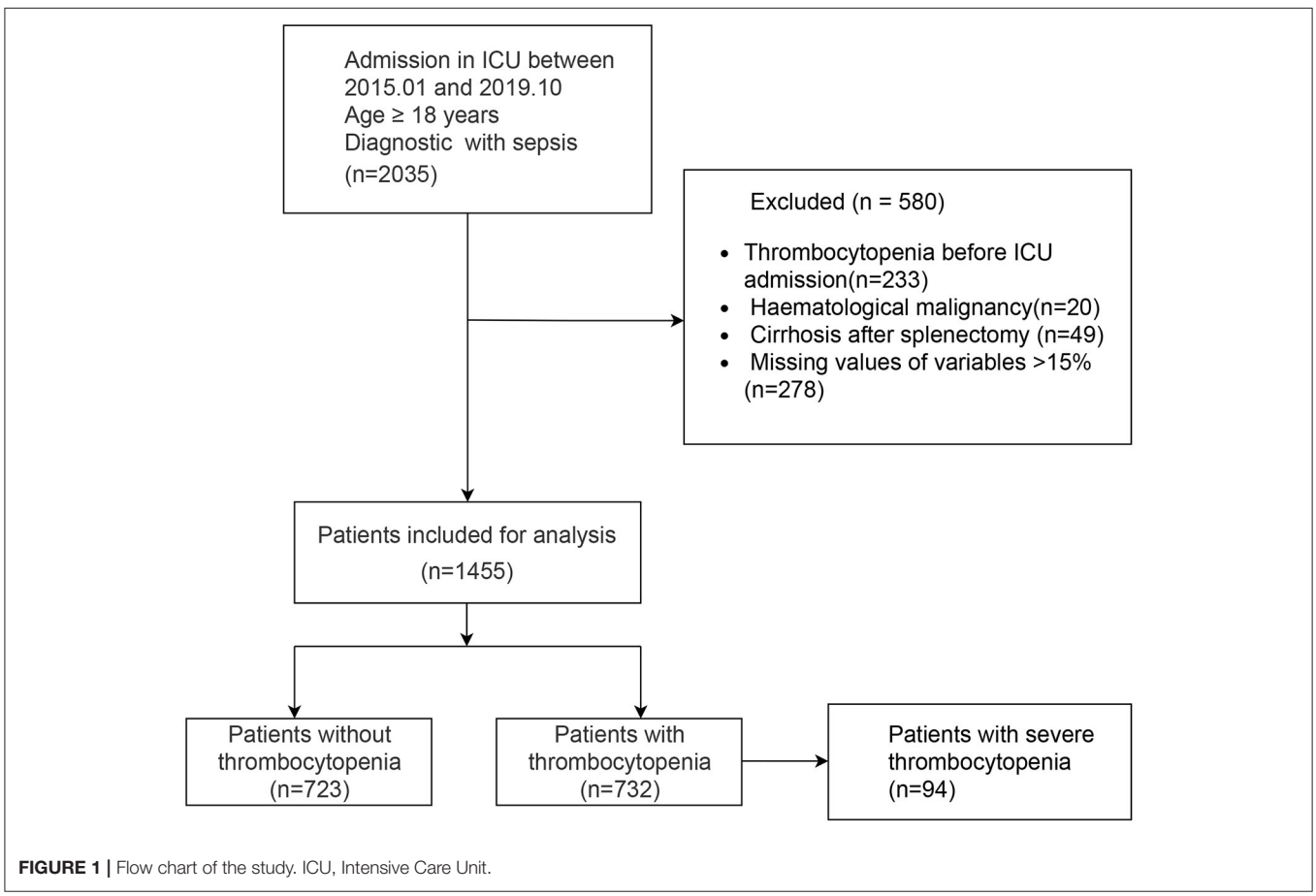



**FIGURE 1 |** Flow chart of the study. ICU, Intensive Care Unit.

selection using the backward selection method, random forest (RF) sampling, and 10% cross-checking. Then, the efficiency (precision, recall, accuracy, and specificity, the cutoff point was 0.5) was calculated, and the variables were ranked according to their importance. The 10 most important variables were retained.

## Handling of Missing Values

Variables with >15% missing values were deleted. If the incidence of missing values was <2%, the mean value of the variable was used to replace the missing values. The missing values of variables with loss rates of >2 and <15% were replaced using multiple imputations.

**TABLE 2 |** Comparisons of baseline characteristics between with thrombocytopenia and without thrombocytopenia.

| | No-SAT ($n$ = 723) | SAT ($n$ = 732) | Total ($n$ = 1,455) | $P$ |
|---|---|---|---|---|
| Age (years) | 65.6 ± 16.1 | 65.6 ± 17.1 | 65.6 ± 16.6 | 0.97 |
| Male [*n* (%)] | 455 (62.9) | 471 (64.3) | 926 (63.6) | 0.613 |
| Alcohol drinking [*n* (%)] | 277 (38.3) | 262 (35.8) | 539 (37) | 0.347 |
| Smoking [*n* (%)] | 275 (38) | 289 (39.5) | 564 (38.8) | 0.609 |
| CKD [*n* (%)] | 17 (2.4) | 19 (2.6) | 36 (2.5) | 0.896 |
| Cancer [*n* (%)] | 84 (11.6) | 82 (11.2) | 166 (11.4) | 0.867 |
| Diabetes [*n* (%)] | 122 (16.9) | 108 (14.8) | 230 (15.8) | 0.3 |
| Hypertension [*n* (%)] | 366 (50.6) | 311 (42.5) | 677 (46.5) | 0.002 |
| APACHE-II | 17.1 ± 6.1 | 19.7 ± 7 | 18.4 ± 6.7 | < 0.001 |
| SOFA | 6 ± 2.7 | 7.8 ± 3.3 | 6.9 ± 3.2 | < 0.001 |
| Sepsis_shock [*n* (%)] | 44 (6.1) | 145 (19.8) | 189 (13) | < 0.001 |
| Antiplatelet drug used [*n* (%)] | 185 (25.6) | 101 (13.8) | 286 (19.7) | < 0.001 |
| **Biochemical indexes on ICU admission** | | | | |
| Red blood cell (x$10^9$/L) | 3.8 ± 0.7 | 3.7 ± 0.8 | 3.8 ± 0.7 | 0.038 |
| Hematokrit (L/L) | 0.3 ± 0.1 | 0.3 ± 0.1 | 0.3 ± 0.1 | 0.045 |
| White blood cell (x$10^9$/L) | 11.3 (8.3, 14.77) | 12.03 (8.04, 16.71) | 11.54 (8.13, 15.39) | 0.052 |
| Neutrophil count (x$10^9$/L) | 9.82 (6.88, 13.19) | 10.47 (6.92, 14.93) | 10.06 (6.89, 13.93) | 0.013 |
| Lymphocyte count (x$10^9$/L) | 0.83 (0.5, 1.22) | 0.69 (0.43, 1.06) | 0.75 (0.47, 1.14) | < 0.001 |
| Platelet count (x$10^9$/L) | 193 (153, 243.5) | 211 (154, 274) | 201 (153, 256) | 0.002 |
| Platelet distribution width (%) | 16 (15.5, 16.4) | 16.2 (15.8, 16.5) | 16.1 (15.7, 16.5) | < 0.001 |
| Mean platelet volume (fl) | 9.8 ± 1.3 | 9.9 ± 1.3 | 9.8 ± 1.3 | 0.031 |
| pH | 7.42 (7.37, 7.47) | 7.38 (7.3, 7.43) | 7.4 (7.34, 7.45) | < 0.001 |
| Serum sodium (mmol/L) | 140.2 (137.5, 142.8) | 141.4 (138.6, 144.1) | 140.9 (138, 143.5) | < 0.001 |
| Serum calcium (mmol/L) | 2.1 ± 0.2 | 2 ± 0.2 | 2 ± 0.2 | < 0.001 |
| Serum lactic acid (mmol/L) | 1.7 (1.2, 2.6) | 3.1 (1.8, 5.2) | 2.2 (1.4, 3.85) | < 0.001 |
| Serum bicarbonate (mmol/L) | 96 ± 7.3 | 94.7 ± 8.7 | 95.4 ± 8 | 0.002 |
| Prothrombin time(s) | 14.4 (13.6, 15.3) | 15.4 (14.2, 17.03) | 14.8 (13.9, 16.1) | < 0.001 |
| Activated partial thromboplastin time(s) | 39.1 (35.4, 44.35) | 40.55 (36.07, 47.73) | 39.8 (35.7, 46) | < 0.001 |
| International normalized ratio | 1.12 (1.05, 1.23) | 1.23 (1.12, 1.41) | 1.17 (1.08, 1.3) | < 0.001 |
| D-dimer (μg/L) | 2.61 (1.28, 5.43) | 4.88 (2.21, 12.03) | 3.5 (1.58, 8.09) | < 0.001 |
| Alanine aminotransferase (U/L) | 20 (13, 37) | 24 (15, 55.25) | 23 (13, 44) | < 0.001 |
| Aspartate aminotransferase (U/L) | 29 (22, 54.5) | 45 (26, 99) | 36 (23, 70) | < 0.001 |
| Serum albumin (g/L) | 32.2 ± 5.1 | 30.5 ± 5.6 | 31.3 ± 5.4 | < 0.001 |
| C-reactive protein (mg/L) | 40 (9.95, 99.85) | 62.1 (21.27, 144.92) | 55.87 (14.61, 125.15) | < 0.001 |
| Serum urea (mmol/L) | 6.92 (5.08, 9.49) | 8.08 (5.74, 12.09) | 7.53 (5.43, 10.76) | < 0.001 |
| Serum creatinine (mmol/L) | 68 (53, 89) | 82 (59, 123.25) | 74 (56, 105.5) | < 0.001 |
| Procalcitonin (ug/L) | 0.41 (0.12, 1.5) | 1.04 (0.3, 5.74) | 0.67 (0.17, 2.92) | < 0.001 |

*Continuous variables are described by means and quarterbacks. Categories variables are analyzed by $\chi^2$ test and continuous variables are analyzed by Wilcoxon rank sum test. SAT, sepsis-associated thrombocytopenia; APACHE, acute physiology and chronic health evaluation; ICU, Intensive Care Unit; CKD, Chronic kidney disease; SOFA, Sepsis-related Organ Failure Assessment.*

## Handling of Outliers

Outliers were detected using the interquartile range (IQR), i.e., the difference between the upper and lower quartiles of the boxplot. We used 1.5 times of IQR as the standard, and points exceeding this criterion (the upper quartile + 1.5 times of IQR, or the lower quartile – 1.5 times of IQR) were defined as outliers. The excluded outliers were handled as missing values.

## Model Establishment

The following R packages for the ML method were used: caret, ipred, ranger, arm, nnet, and gbm. Samples were randomly divided into training set and test set in a 7:3 ratio. All ML models were evaluated using 10× cross-validation.

The hyperparameters were adjusted by grid search as follows. For the RF model, the number of trees and mtry parameters were adjusted. For the neural network (NNET) model, size and decay

**TABLE 3 |** Comparison of infection site and clinical outcomes between groups.

| | No-SAT ($n$ = 723) | SAT ($n$ = 732) | Total ($n$ = 1,455) | $P$ |
|---|---|---|---|---|
| Ventilation duration (days) | 0.96 (0.28, 5) | 3.91 (0.8, 8.8) | 2.12 (0.47, 7.38) | <0.001 |
| ICU length of stay (days) | 3.88 (1.88, 8.47) | 6.97 (3.62, 12.02) | 5.22 (2.6, 10.65) | <0.001 |
| Hosp. LOS (days) | 19 (13, 29) | 18 (11, 28) | 19 (12, 28) | 0.022 |
| Hospital mortality [*n* (%)] | 94 (13) | 221 (30.2) | 315 (21.6) | <0.001 |
| Cost (x$10^3$, yuan) | 51.2 (33.5, 79.0) | 55.54 (36.3, 87.6) | 53.6 (34.5, 82.7) | 0.002 |
| **Infection site [*n* (%)]** | | | | |
| Pulmonary | 510 (70.5) | 509 (69.5) | 1019 (70) | 0.718 |
| Urinary | 54 (7.5) | 78 (10.7) | 132 (9.1) | 0.043 |
| Blood stream | 67 (9.3) | 150 (20.5) | 217 (14.9) | < 0.001 |

*Continuous variables are described by means and quarterbacks. Categories variables are analyzed by $\chi^2$ test and continuous variables are analyzed by Wilcoxon rank sum test. SAT, sepsis-associated thrombocytopenia; ICU, Intensive Care Unit; Hosp. LOS, length of hospital stay.*

parameters were adjusted. For the gradient boosting machine (GBM) model, n.trees, interaction.depth, and shrinkage were adjusted. Finally, the importance of variables was sorted using the function "varImpPlot" within the "caret" package in *R*.

### *Model Validation and Evaluation*

The area under the receiver operating characteristic curve (AUC), sensitivity, specificity, and 95% CI of each model were calculated. The confusion matrix was evaluated using accuracy, precision, and recall as parameters presented in **Table 1**. Local Interpretable Model-Agnostic Explanations (LIME) provides another method for model interpretation (17).

### Statistical Analysis

Descriptive statistics were analyzed conventionally using the CBCgrps package in R (18). Normally distributed measurement data were expressed as x ± s and compared between groups using the two-independent-samples *t*-test. Meanwhile, non-normally distributed data were expressed as M (P25, P75) and compared using the Mann–Whitney *U* test. Enumeration data were expressed in terms of the rate and percentage and compared between the groups using the $\chi^2$ test. All statistical analyses were performed using R (software version 3.6.3). A *P*-value of 0.05 was considered significant.

## RESULTS

### Comparison of Basic Information and Clinical Outcomes

A total of 1,455 patients with sepsis were included in this study. The flow chart of the study is shown in **Figure 1**, including 732 SAT patients (49.7%). Regarding the sources of infection, pulmonary infection accounted for the highest proportion, with 1,019 cases (70%), followed by blood-borne infection, with 217 cases (14.9%), and urinary tract infection, with 132 cases (9.1%). There were 189 patients with septic shock, and 76.7% of them had SAT.

**Table 2** shows a comparison of general clinical information and clinical outcomes between the thrombocytopenia and non-thrombocytopenia groups. There was no significant difference in age and gender between the two groups, with an average age of 65.6 ± 16.6 years and 63.6% of subjects being male. The disease conditions in the thrombocytopenia group were more serious, and the APACHE and SOFA scores were significantly higher than those in the non-thrombocytopenia group, with statistically significant differences ($P < 0.001$). There were significant differences in terms of mechanical ventilation time, length of ICU stays, length of hospital stays, and mortality between the two groups ($P < 0.001$), and the clinical outcome of the thrombocytopenia group was worse. **Table 3** shows a comparison of infection site and clinical outcomes between the groups. We compared the baseline characteristics and clinical outcomes of the external validation set, shown in **Supplementary Table S1**. The comparison of feature distribution between the training, internal validation, and external validation is shown in in **Supplementary Table S2**. The incidence rate of SAT in the three groups of patients was similar, and there was no significant difference in age, SOFA score, and initial platelet count in ICU admission.

### Evaluation of Machine Learning Algorithm Models

**Figure 2** shows the ROC comparison of four ML models for thrombocytopenia prediction, with internal validation showing AUCs between 0.74 and 0.79 and external validation showing AUCs between 0.54 and 0.72. **Table 3** shows the pairwise comparison in external validation. Results of external validation show that NNET and GBM had the best prediction, with no significant difference between the two models, while the prediction accuracy of RF and Bayesian models was slightly worse. Additional evaluation metrics for the four machine learning models in external validation are presented in **Table 4**. We established the model for predicting severe thrombocytopenia using the same method. **Figure 3** shows the ROC comparison of ML models for the prediction of severe thrombocytopenia, with internal validation showing AUCs between 0.84 and 0.89 and external validation showing AUCs between 0.70 and 0.77. The prediction was better than for thrombocytopenia, with the Bayesian model showing the best results. The calibration curve analysis of models is

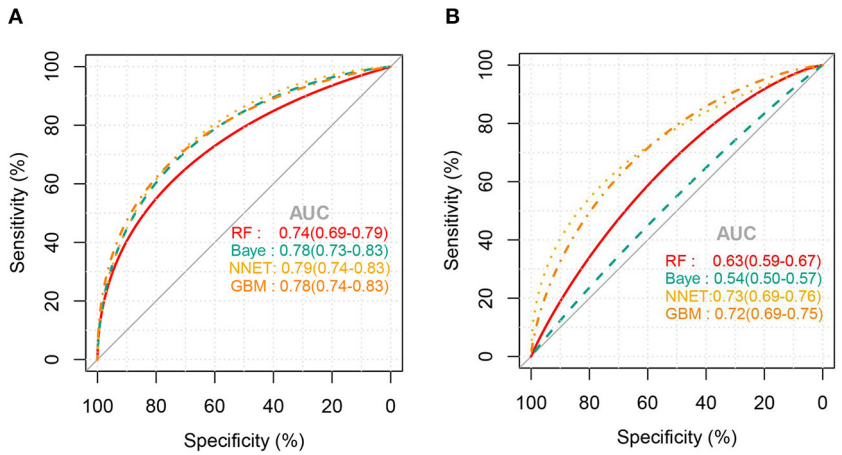



**FIGURE 2 |** ROC curves of the four machine learning models for predicting thrombocytopenia. **(A)**, Internal validation; **(B)**, external validation; RF, random forest; NNET, neural network; GBM, gradient boosting machine; Baye, bayesian.

**TABLE 4 |** Comparison of the area under the roc curve of four machine learning models in external validation.

| **Models for predicting thrombocytopenia** | | | | |
|---|---|---|---|---|
| | **RF** | **Bayesian** | **NNET** | **GBM** |
| RF | / | 0.001 | 0.001 | 0.001 |
| Bayesian | 0.001 | / | 0.001 | 0.001 |
| NNET | 0.001 | | / | 0.94 |
| GBM | 0.001 | 0.001 | 0.94 | / |
| **Models for predicting severe thrombocytopenia** | | | | |
| | **RF** | **Bayesian** | **NNET** | **GBM** |
| RF | / | 0.001 | 0.913 | 0.127 |
| Bayesian | 0.001 | / | 0.001 | 0.001 |
| NNET | 0.913 | 0.001 | / | 0.662 |
| GBM | 0.127 | 0.001 | 0.662 | / |

*ROC, Receiver operating characteristic; RF, random forest; NNET, neural network; GBM, gradient boosting machine.*

shown in **Supplementary Figure S1**. **Figures 4**, **5** showed the top 10 variables of the four models ordered by importance. LIME provides explanations for any individual patient, and the contribution of a given variable may change depending on other features of the patient in **Supplementary Figures S2**, **S3** shows contributions by the variables for two patients (#2, #3). The red (blue) color indicates that the variable contradicts (supports) a given class.

## DISCUSSION

Our study found that SAT had high morbidity and mortality, as well as poor clinical outcomes in ICU, and RF, Bayesian, NNET, and GBM prediction models achieved good predictions.

Thrombocytopenia is very common in ICU patients, with sepsis being its main cause (12). Previous studies on SAT have shown that the incidence rate in critically ill patients (3, 19) was approximately 50%—similar to our findings. Platelets play crucial roles in inflammatory response (20), such as promoting immune response and blood coagulation activation. Presently, many published articles have shown that thrombocytopenia is significantly related to the poor prognosis of patients and is closely related to the degree of thrombocytopenia (21).

Thrombocytopenia is a common reason of platelet transfusion in the ICU. When the platelet count is $<50 \times 10^9$/ L, clinicians often transfuse platelets (22, 23) to reduce bleeding events. A British prospective multicenter observation study (7) showed that, in ICU patients with severe thrombocytopenia, the mortality rate was as high as 35.4%. Therefore, we also predicted severe thrombocytopenia in patients with sepsis. The models had higher accuracy and better prediction effect. For such patients, early discontinuation of antiplatelet drugs, use of platelet-increasing drugs, and early reservation of platelets might help prevent bleeding events and improve the prognosis of patients.

In this study, among the four ML models, the top variables in terms of importance scores were SOFA score, serum lactic acid, serum sodium bicarbonate, and dimer, which suggested that these factors had a significant correlation with SAT.

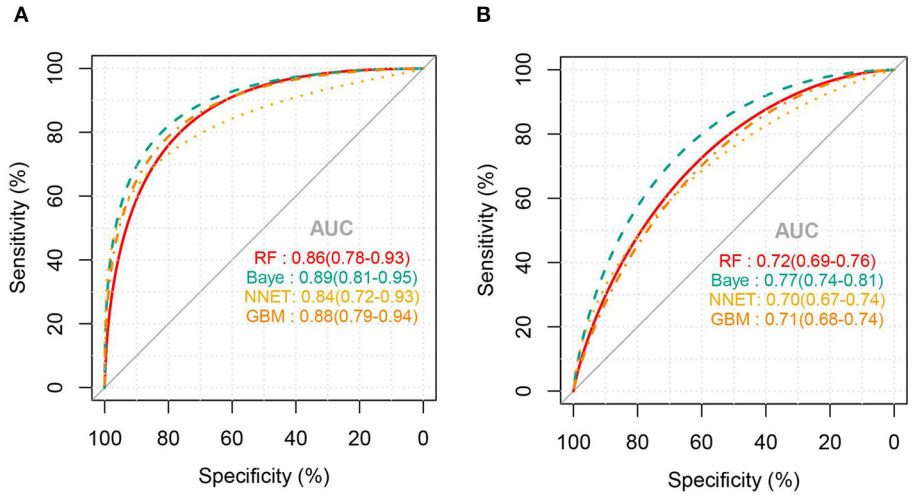



**FIGURE 3 |** ROC curves of the four machine learning models for predicting severe thrombocytopenia. **(A)**, Internal validation; **(B)**, external validation; RF, random forest; NNET, neural network; GBM, gradient boosting machine; Baye, bayesian.

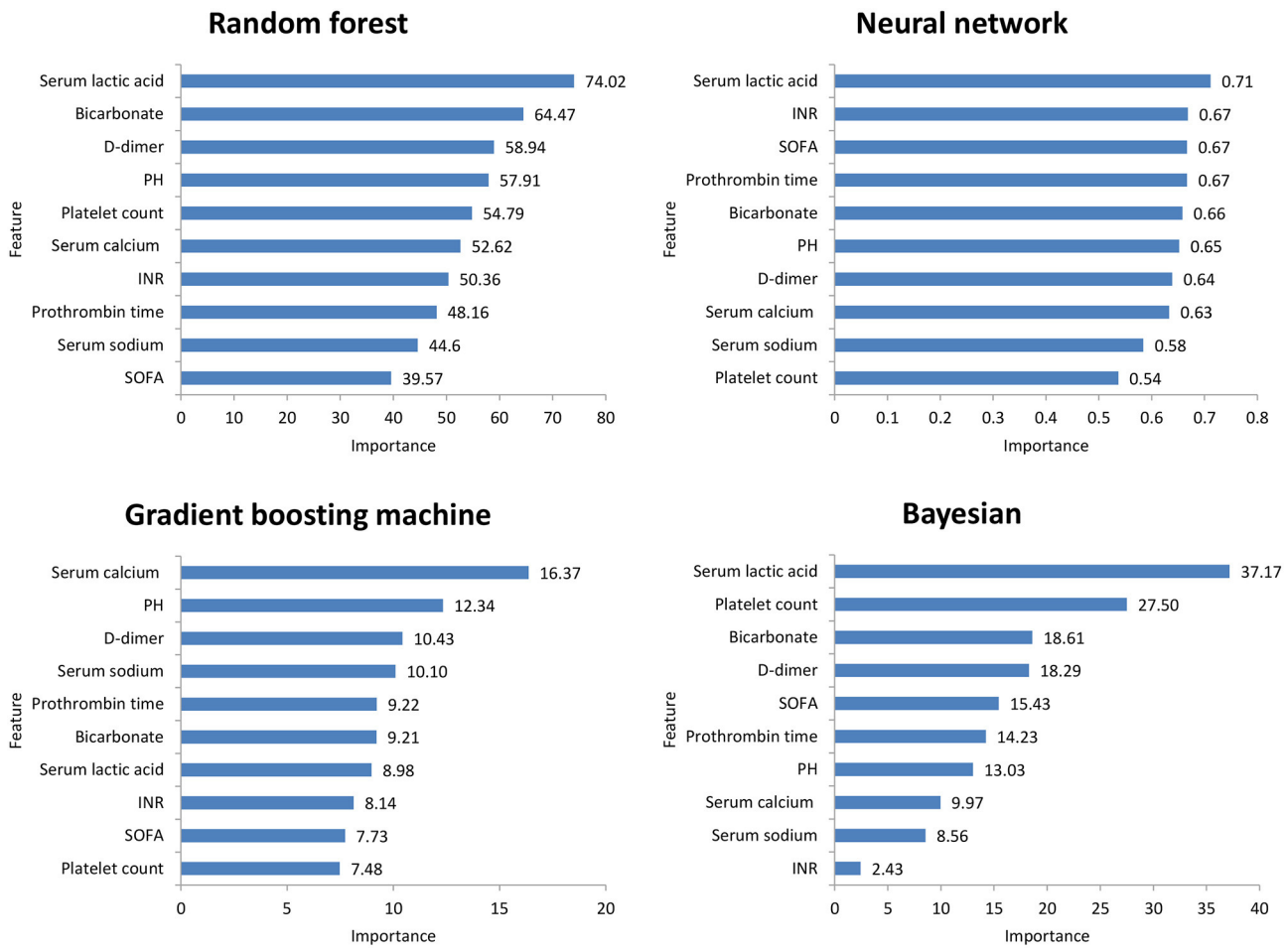



**FIGURE 4 |** Top 10 variables of the four machine learning models for predicting thrombocytopenia ordered by importance. SOFA, Sepsis-related Organ Failure Assessment; INR, International normalized ratio.

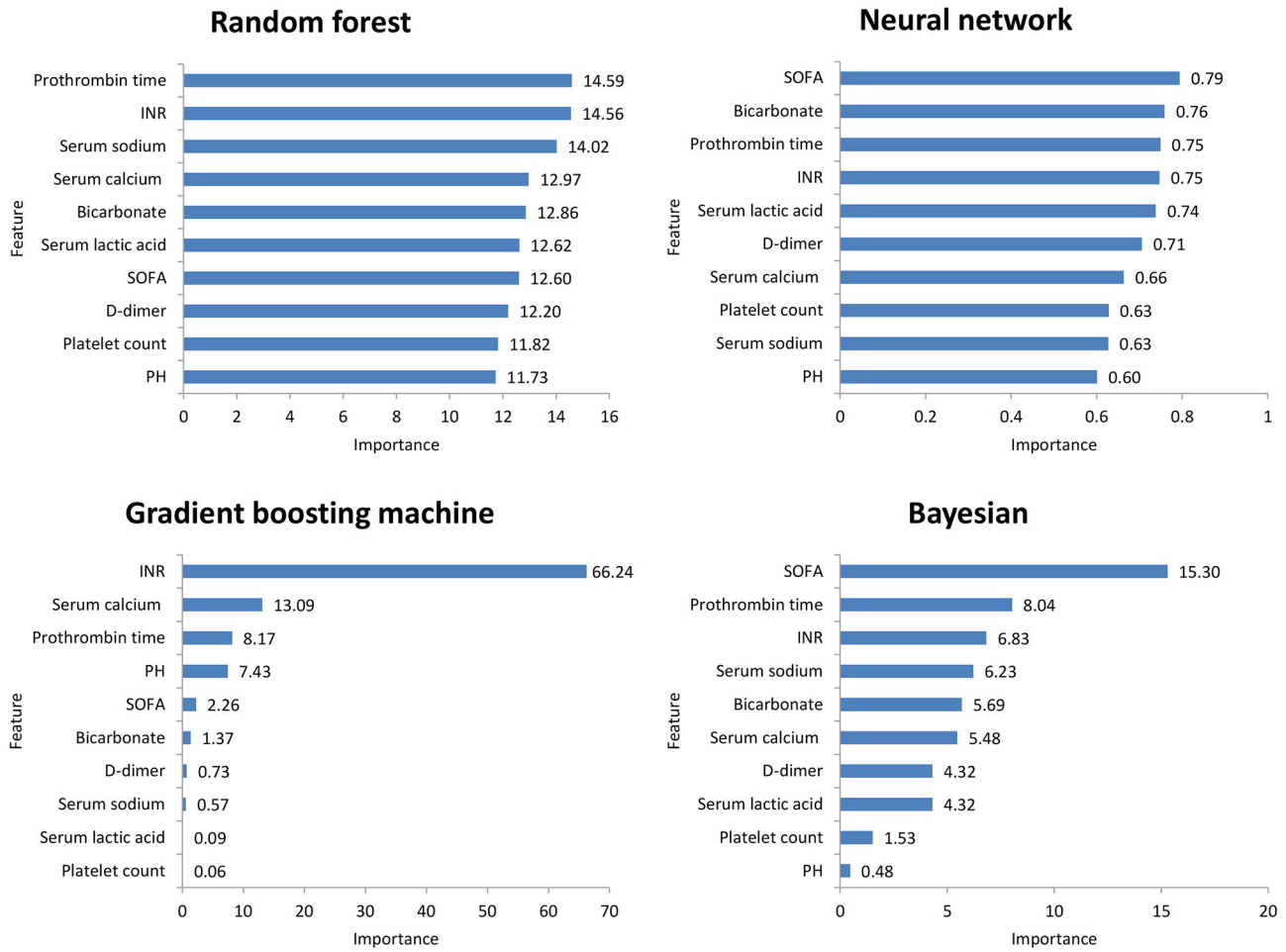



**FIGURE 5 |** Top 10 variables of the four machine learning models for predicting severe thrombocytopenia ordered by importance. SOFA, Sepsis-related Organ Failure Assessment; INR, International normalized ratio.

A retrospective study including 267 patients with abdominal infection showed that a high SOFA score was an important risk factor for hospital-acquired thrombocytopenia. A systematic evaluation (24) found that disease severity was an influencing factor of thrombocytopenia, while serum lactic acid and serum sodium bicarbonate were classic indicators reflecting the severity of the patient's disease. Plasma D-dimer is an important marker of thrombosis activity. In sepsis patients, fibrinolysis activation and D-dimer level have been independently correlated with mortality (25). Therefore, monitoring the SOFA score, serum lactic acid, serum sodium bicarbonate, and dimer levels is helpful for the early detection of thrombocytopenia patients.

This study has some limitations. First, this was a single-center, retrospective study, and some data were missing. We supplemented the data through multiple imputation functions of statistical software to reduce the bias of research results. Second, there are many reasons for thrombocytopenia. For example, some patients with sepsis were treated with hemodialysis, and heparin-induced thrombocytopenia was reported after using heparin. These patients were not excluded, which influenced the results. Third, due to the algorithm characteristics of ML, the models could not clarify the specific relationship between variables and thrombocytopenia, and they were not suitable for all people, which limited the performance of the models. Therefore, based on the algorithms, we showed the measurement of variable importance in the four models and LIME feature plot explained the relationship between variables in the models and thrombocytopenia to a certain extent. Finally, our ML models to predict SAT between ICU stays, the models to predict SAT each day of the ICU stays will be more clinically meaningful. In the future, we will develop software and join the electronic information system to predict SAT each day of the ICU stays.

## CONCLUSION

We established four ML models to predict SAT and severe thrombocytopenia. The models were validated in MIMIC III and can be used to identify such high-risk patients at an early stage and guide individualized clinical treatment. In the future, we will conduct a prospective cohort study and apply these models to clinical practice.

## DATA AVAILABILITY STATEMENT

The raw data supporting the conclusions of this article will be made available by the authors, without undue reservation.

## ETHICS STATEMENT

The studies involving human participants were reviewed and approved by the Ethics Committee of Dongyang People's Hospital. Written informed consent for participation was not required for this study in accordance with the national legislation and the institutional requirements.

## AUTHOR CONTRIBUTIONS

YW and YP carried out the design and contributed to manuscript revision. XJ participated in data analysis and drafted the manuscript. WZ provided overall supervision and undertook the responsibility of submitting the manuscript for publication. All authors contributed to the article and approved the submitted version.

## FUNDING

This study was supported by the Clinical Research Fund Project of the Zhejiang Medical Association (2020ZYC-B44 and 2018KY866) and the Conba Hospital Management Project of the Zhejiang Hospital Association (2021ZHA-KEB335).

## ACKNOWLEDGMENTS

We would like to thank Editage (www.editage.cn) for English language editing.

## SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fmed.2022.837382/full#supplementary-material

**Supplementary Figure S1 |** Calibration of four machine learning models. **(A)**: Models for predicting thrombocytopenia in Internal validation set; **(B)**: Models for predicting thrombocytopenia in external validation set. **(C)**: Models for predicting severe thrombocytopenia in Internal validation set; **(D)**: Models for predicting severe thrombocytopenia in external validation set; RF, random forest; NNET, neural network; GBM, gradient boosting machine; Baye, bayesian.

**Supplementary Figure S2 |** Heatmap plot showing the contribution of each variable to the classification of sample patients. The relative contribution of each variable was calculated using the LIME algorithm. Patients #2, #3 are shown as examples. Red (blue) color indicates that the relevant variable contradicts (supports) a given label. SOFA, Sepsis-related Organ Failure Assessment; INR, International normalized ratio; LIME, Local Interpretable Model-Agnostic Explanations.

**Supplementary Figure S3 |** LIME feature plot shows the contribution of each variable to the classification of sample patients. Red (blue) color indicates that the relevant variable contradicts (supports) a given label. SOFA, Sepsis-related Organ Failure Assessment; INR, International normalized ratio; LIME, Local Interpretable Model-Agnostic Explanations.

**Supplementary Table S1 |** Baseline characteristics and clinical outcomes between SAT and No-SAT groups in the MIMIC III cohort. Continuous variables are described by means and quarterbacks. Categories variables are analyzed by $\chi^2$ test and continuous variables are analyzed by Wilcoxon rank sum test. SAT, sepsis-associated thrombocytopenia; SOFA, Sepsis-related Organ Failure Assessment; DM, diabetes mellitus; COPD, Chronic Obstructive Pulmonary Disease; AST, aspartate aminotransferase; INR, International normalized ratio; PT, prothrombin time; Hosp. hospital, LOS length of stay; ICU LOS, ICU length of stay.

**Supplementary Table S2 |** Comparison of feature distribution between the training, internal validation, and external validation. Continuous variables are described by means and quarterbacks. Categories variables are analyzed by $\chi^2$ test and continuous variables are analyzed by Wilcoxon rank sum test. SAT, sepsis-associated thrombocytopenia; ICU, Intensive Care Unit; SOFA, Sepsis-related Organ Failure Assessment; Hosp. LOS, length of hospital stay.

## REFERENCES

1. García-Gallo JE, Fonseca-Ruiz NJ, Celi LA, Duitama-Muñoz JF. A machine learning-based model for 1-year mortality prediction in patients admitted to an Intensive Care Unit with a diagnosis of sepsis. *Med intensiva.* (2020) 44:160–70. doi: 10.1016/j.medin.2018.07.016
2. Thorsen-Meyer HC, Nielsen AB, Nielsen AP, Kaas-Hansen BS, Toft P, Schierbeck J, et al. Dynamic and explainable machine learning prediction of mortality in patients in the intensive care unit: a retrospective study of high-frequency data in electronic patient records. *Lancet Digit Health.* (2020) 2:e179–91. doi: 10.1016/S2589-7500(20)30018-2
3. Sharma B, Sharma M, Majumder M, Steier W, Sangal A, Kalawar M, et al. Thrombocytopenia in septic shock patients – a prospective observational study of incidence, risk factors and correlation with clinical outcome. *Anaesth Intensive Care.* (2007) 35:874–80. doi: 10.1177/0310057X0703500604
4. Bedet A, Razazi K, Boissier F, Surenaud M, Hue S, Giraudier S. Mechanisms of thrombocytopenia during septic shock: a multiplex cluster analysis of endogenous sepsis mediators. *Shock.* (2018) 49:641–8. doi: 10.1097/SHK.0000000000001015
5. Xie Y, Tian R, Xie H, Jin W, Du J, Huang P. The clinical significance of thrombocytopenia complicating sepsis: a meta-analysis. *J Infect.* (2019) 78:323–37. doi: 10.1016/j.jinf.2018.12.002
6. Azkárate I, Choperena G, Salas E, Sebastián R, Lara G, Elósegui I, et al. Epidemiology and prognostic factors in severe sepsis/septic shock. Evolution over six years. *Med Intensiva.* (2016) 40:18–25. doi: 10.1016/j.medine.2015.01.002
7. Stanworth SJ, Walsh TS, Prescott RJ, Lee RJ, Watson DM, Wyncoll DL, et al. Thrombocytopenia and platelet transfusion in UK critical care: a multicenter observational study. *Transfusion.* (2013) 53:1050-8. doi: 10.1111/j.1537-2995.2012.03866.x
8. Giannini HM, Ginestra JC, Chivers C, Draugelis M, Hanish A, Schweickert WD, et al. A machine learning algorithm to predict severe sepsis and septic shock: development, implementation, and impact on clinical practice. *Crit Care Med.* (2019) 47:1485–92. doi: 10.1097/CCM.0000000000003891
9. Giacobbe DR, Signori A, Del Puente F, Mora S, Carmisciano L, Briano F, et al. Early detection of sepsis with machine learning techniques: a brief clinical perspective. *Front Med.* (2021) 8:617486. doi: 10.3389/fmed.2021.617486
10. Collins, GS, Reitsma JB, Altman DG, Moons KG. Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD): the TRIPOD statement. *BMJ.* (2015) 350:g7594. doi: 10.1161/CIRCULATIONAHA.114.014508
11. Johnson A, Pollard T, Shen L, Lehman L, Feng M, Ghassemi M, et al. MIMIC-III, a freely accessible critical care database. *Sci Data.* (2016) 3:160035. doi: 10.1038/sdata.2016.35
12. Thiolliere F, Serre-Sapin AF, Reignier J, Benedit M, Constantin JM, Lebert C, et al. Epidemiology and outcome of thrombocytopenic patients in the intensive care unit: results of a prospective multicenter study. *Intensive Care Med.* (2013) 39:1460–8. doi: 10.1007/s00134-013-2963-3

13. Ben HC, Lauzet JY, Rézaiguia DS, Duvoux C, Cherqui D, Duvaldestin P, et al. Effect of severe thrombocytopenia on patient outcome after liver transplantation. *Intensive Care Med.* (2003) 29:756–62. doi: 10.1007/s00134-003-1727-x
14. Hui P, Cook DJ, Lim W, Fraser GA, Arnold DM. The frequency and clinical significance of thrombocytopenia complicating critical illness: a systematic review. *Chest.* (2011) 139:271–8. doi: 10.1378/chest.10-2243
15. Greinacher A, Selleng K. Thrombocytopenia in the intensive care unit patient. *Hematology Am Soc Hematol Educ Program.* (2010) 2010:135–43. doi: 10.1182/asheducation-2010.1.135
16. Singer M, Deutschman CS, Seymour CW, Shankar-Hari M, Annane D, Bauer M, et al. The third international consensus definitions for sepsis and septic shock (Sepsis-3). *JAMA.* (2016) 315:801–10. doi: 10.1001/jama.2016.0287
17. Zhang Zh, Chen L, Xu P, Hong YC. Predictive analytics with ensemble modeling in laparoscopic surgery: A technical note. *Laparosc Endosc Robot Surg.* (2022). doi: 10.1016/j.lers.2021.12.003
18. Zhang Z, Gayle AA, Wang J, Zhang H, Cardinal FP. Comparing baseline characteristics between groups: an introduction to the CBCgrps package. *Ann Transl Med.* (2017) 5:484. doi: 10.21037/atm.2017.09.39
19. Venkata C, Kashyap R, Farmer JC, Afessa B. Thrombocytopenia in adult patients with sepsis: incidence, risk factors, and its association with clinical outcome. *J Intensive Care.* (2013) 1:9. doi: 10.1186/2052-0492-1-9
20. Vardon BF, Ruiz S, Gratacap MP, Garcia C, Payrastre B, Minville V. Platelets are critical key players in sepsis. *Int J Mol Sci.* (2019) 20:3494. doi: 10.3390/ijms20143494
21. Vandijck DM, Blot SI, De WJ, Hoste EA, Vandewoude KH, Decruyenaere JM. Thrombocytopenia and outcome in critically ill patients with bloodstream infection. *Heart Lung.* (2010) 39:21–6. doi: 10.1016/j.hrtlng.2009.07.005
22. Ning S, Barty R, Liu Y, Heddle NM, Rochwerg B, Arnold DM. Platelet transfusion practices in the ICU: data from a large transfusion registry. *Chest.* (2016) 150:516–23. doi: 10.1016/j.chest.2016.04.004
23. Arnold DM, Crowther MA, Cook RJ, Sigouin C, Heddle NM, Molnar L, et al. Utilization of platelet transfusions in the intensive care unit: indications, transfusion triggers, and platelet count responses. *Transfusion.* (2006) 46:1286–91. doi: 10.1111/j.1537-2995.2006.00892.x
24. Jonsson AB, Rygård SL, Hildebrandt T, Perner A, Møller MH, Russell L. Thrombocytopenia in intensive care unit patients: a scoping review. *Acta Anaesthesiol Scand.* (2021) 65:2–14. doi: 10.1111/aas.13699
25. Semeraro F, Colucci M, Caironi P, Masson S, Ammollo CT, Teli R, et al. Platelet drop and fibrinolytic shutdown in patients with sepsis. *Crit Care Med.* (2018) 46:e221–8. doi: 10.1097/CCM.0000000000000291

**Conflict of Interest:** The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

**Publisher's Note:** All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

frontiers in Medicine

ORIGINAL RESEARCH
published: 17 March 2022
doi: 10.3389/fmed.2022.851690

# Deep Learning-Based Pain Classifier Based on the Facial Expression in Critically Ill Patients

*Chieh-Liang Wu*[1,2,3,4†], *Shu-Fang Liu*[5†], *Tian-Li Yu*[6], *Sou-Jen Shih*[5], *Chih-Hung Chang*[6], *Shih-Fang Yang Mao*[7], *Yueh-Se Li*[7], *Hui-Jiun Chen*[5], *Chia-Chen Chen*[7*] and *Wen-Cheng Chao*[1,4,8,9*]

[1] Department of Critical Care Medicine, Taichung Veterans General Hospital, Taichung, Taiwan, [2] Department of Industrial Engineering and Enterprise Information, Tunghai University, Taichung, Taiwan, [3] Artificial Intelligence Studio, Taichung Veterans General Hospital, Taichung, Taiwan, [4] Colledge of Medicine, National Chung Hsing University, Taichung, Taiwan, [5] Department of Nursing, Taichung Veterans General Hospital, Taichung, Taiwan, [6] Department of Electrical Engineering, National Taiwan University, Taipei, Taiwan, [7] Electronic and Optoelectronic System Research Laboratories, Industrial Technology Research Institute, Hsinchu, Taiwan, [8] Department of Automatic Control Engineering, Feng Chia University, Taichung, Taiwan, [9] Big Data Center, National Chung Hsing University, Taichung, Taiwan

**Objective:** Pain assessment based on facial expressions is an essential issue in critically ill patients, but an automated assessment tool is still lacking. We conducted this prospective study to establish the deep learning-based pain classifier based on facial expressions.

**Methods:** We enrolled critically ill patients during 2020–2021 at a tertiary hospital in central Taiwan and recorded video clips with labeled pain scores based on facial expressions, such as relaxed (0), tense (1), and grimacing (2). We established both image- and video-based pain classifiers through using convolutional neural network (CNN) models, such as Resnet34, VGG16, and InceptionV1 and bidirectional long short-term memory networks (BiLSTM). The performance of classifiers in the test dataset was determined by accuracy, sensitivity, and F1-score.

**Results:** A total of 63 participants with 746 video clips were eligible for analysis. The accuracy of using Resnet34 in the polychromous image-based classifier for pain scores 0, 1, 2 was merely 0.5589, and the accuracy of dichotomous pain classifiers between 0 vs. 1/2 and 0 vs. 2 were 0.7668 and 0.8593, respectively. Similar accuracy of image-based pain classifier was found using VGG16 and InceptionV1. The accuracy of the video-based pain classifier to classify 0 vs. 1/2 and 0 vs. 2 was approximately 0.81 and 0.88, respectively. We further tested the performance of established classifiers without reference, mimicking clinical scenarios with a new patient, and found the performance remained high.

**Conclusions:** The present study demonstrates the practical application of deep learning-based automated pain assessment in critically ill patients, and more studies are warranted to validate our findings.

**Keywords: pain, critically ill patients, facial expression, artificial intelligence, classifier**

OPEN ACCESS

***Edited by:***
*Nan Liu,*
*National University of Singapore, Singapore*

***Reviewed by:***
*Kenneth Craig,*
*University of British Columbia, Canada*
*Narendra Londhe,*
*National Institute of Technology, India*

***Correspondence:**
*Chia-Chen Chen*
*ChiaChen@itri.org.tw*
*Wen-Cheng Chao*
*cwc081@hotmail.com*

*[†]These authors have contributed equally to this work*

***Specialty section:***
*This article was submitted to Intensive Care Medicine and Anesthesiology, a section of the journal Frontiers in Medicine*

***Received:*** *10 January 2022*
***Accepted:*** *17 February 2022*
***Published:*** *17 March 2022*

***Citation:***
*Wu C-L, Liu S-F, Yu T-L, Shih S-J, Chang C-H, Yang Mao S-F, Li Y-S, Chen H-J, Chen C-C and Chao W-C (2022) Deep Learning-Based Pain Classifier Based on the Facial Expression in Critically Ill Patients. Front. Med. 9:851690. doi: 10.3389/fmed.2022.851690*

## BACKGROUND

Pain is an essential medical issue but somehow difficult to assess in critically ill patients who cannot report their pain (1). Therefore, the Critical-Care Pain Observation Tool (CPOT) has been developed to grade the pain through assessing behavior alternations, such as facial expressions, among critically ill patients in the past two decades (2). The facial expression is the fundamental behavior alternation in CPOT and consists of relaxed, tense, and grimacing (pain score 0, 1, and 2) (3). Currently, facial expression-based pain assessment is graded by the nurse, and there is an unmet need to develop an automated pain assessment tool based on facial expression to relieve the medical staff from the aforementioned workload (4).

A number of automated recognition of facial expressions of pain and emotion has been developed through using distinct approaches (5–9). Pedersen et al. used Support Vector Machine (SVM) as a facial expression-based pain classifier in UNBC-McMaster Shoulder Pain Expression Archive Database, consisting of 200 video sequences obtained from 25 patients with shoulder pain, and reported that the accuracy of the leave-one-subject-out 25-fold cross was 0.861 (7). Given that video sequences contain temporal information with respect to pain, two studies were used Recurrent Neural Network (RNN) and hybrid network to extract the time-frame feature among images and reported an improved performance (8, 9). Furthermore, recent studies have employed fusion network architectures and further improved the F1 score to ∼0.94 (10, 11). Therefore, the recent advancements in deep learning might enable us to establish a facial expressed-based pain assessment tool in critically ill patients.

Notably, the application of the aforementioned methods in critically ill patients might not be straightforward due to real-world difficulties to obtain standardized and whole unmasked facial images of patients admitted to the intensive care unit (ICU) (12). Unlike the high-quality whole facial image in the UNBC-McMaster Shoulder Pain Expression Archive Database, critically ill patients may have masks on the face due to needed medical devices, such as endotracheal tube, nasoesophageal tube, and oxygen mask. Furthermore, pain-associated facial muscle movements might hence be subtle due to sedation and tissue oedema in critically ill patients. Therefore, there is a substantial need for using facial images obtained in sub-optimal real-world conditions at ICUs to establish an automated facial expression-based assessment tool for pain in critically ill patients. In the present prospective study, we recorded facial video clips in critically ill patients at the ICUs of Taichung Veterans General Hospital (TCVGH) and employed an ensemble of three Convolutional Neural Network (CNN) models as well as RNN to establish the pain classifier based on facial expressions.

## MATERIALS AND METHODS

### Ethical Approval

This study was approved by the Institutional Review Board approval of the Taichung Veterans General Hospital (CE20325A). Informed consent was obtained from all of the participants prior to the enrollment in the study and collection of data.

### Study Population

We conducted this prospective study by enrolling patients who were admitted to medical and surgical ICUs at TCVGH, a referral hospital with 1,560 beds in central Taiwan, between 2020-Nov and 2021-Nov. The CPOT is a standard of care in the study hospital, and grading of the facial expression-based pain score is in accordance with the guideline (3). In detail, a score of 0 is given if there is no observed muscle tension in the face, and the score of 1 is composed of a tensed muscle contraction, such as the presence of frowning, brow lowering, orbit tightening as well as levator muscle contraction. The score of 2 consists of grimacing, which is a contraction of facial muscles, particularly muscles nearby the eyebrow area, plus eyelid tightly closed.

### Video Sequences With Labeled Pain Grade Based on Facial Expression

**Figure 1** depicts the protocol of video record, labeling, and image preprocessing of the present study (**Figure 1**). Video record and labeling were performed by three experienced nurses after training for inter-rater concordance, and the labeling was further validated by two senior registered nurses. To mitigate information bias and synchronize the recording and labeling, we designed a user interface that enables the study nurse to observe the patient for 10 s, to record a video for 20 s, and then label the pain score at the end of the video. To further reduce the potential sampling errors, we recorded three labeled videos in each observation; therefore, each 90-s video sequence has three 20-s clips (**Figure 1A**). Given the nature of observation of this study, we conducted the recording per day during the ICU admission of participants, particularly before and after suction, dressing change as well as invasive procedures, to obtain the videos with distinct pain grades in individual critically ill patients. With regards to the hardware, the frame per second of the applied camera was 30, and the total frames of a 20-s video clip were nearly 400–600 frames per clip. To standardize the video clips, we used 50 frames in each 20-s clip; therefore, there were 2.5 representative frames per second for the following experiments. To avoid any interference with critical care, we designed a portable camera rack that enables us to take high-quality video ∼1–2 m from the patient.

### Image Preprocessing

We used a facial landmark tracker to locate the facial area (13). Due to the face that was masked by the aforementioned medical devices might not be detected by the facial landmark tracker, we further used multi-task CNN to locate the facial area if the face was not located by the facial landmark tracker (14). Given that the area nearby the eyebrow is the key area to interpret pain score, we hence cropped the face between hairline and nose not only to focus on the eyebrow area, but also to avoid the confounding of the aforementioned medical devices. We further cropped the central part of the eyebrow area with a fixed ratio of height/width (3/4) for the following experiments. Given that facial images with extreme angles may lead to the facial landmark misalignment

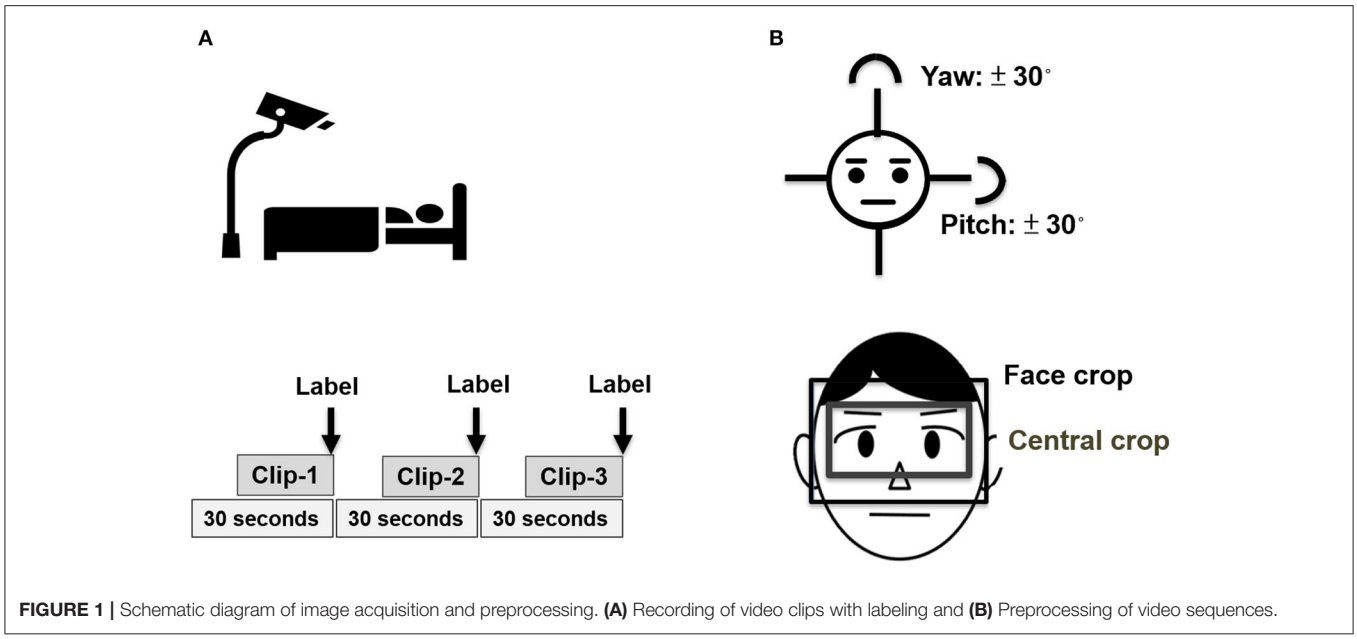



**FIGURE 1 |** Schematic diagram of image acquisition and preprocessing. **(A)** Recording of video clips with labeling and **(B)** Preprocessing of video sequences.

and affect the following experiments, we hence excluded the faces with yaw or pitch angle over 30 degrees (**Figure 1B**).

## Image-Based Pain Classifiers

**Figure 2** illustrates the deep learning-based Siamese network architectures for image- and video-based pain classifiers in this study (15) (**Figure 2**). To reduce the need for an extremely high number of labeled but unrelated images for learning, we employed a relation network architecture for the image-based pain classifier (16). In brief, the aforementioned relation network is designed for learning to compare the differences among labeled images of each individual patient; therefore, the essential need is the images with distinct grades among individual patients, instead of a high number of unrelated images from patients with high heterogeneity. Therefore, we used the data of the 63 participants who had images of all of 0, 1, 2 labeled images. In detail, by feeding grade-0 facial expression image and grade 1/2 images into CNN encoder, two vectors were obtained to represent the subtle difference between the image of grade-0 and grade-1/2, instead of calculating the complex distance metric of two images in high dimensions. Indeed, the application of relation network should be in line with clinical grading of pain by the nurse, who had to recognize the baseline facial appearance of an individual patient prior to grade pain-score based on the facial expression. In this study, we used three CNN models that have fewer vanishing gradient issues, such as Resnet34, VGG16, and inceptionV1, as well as two types of the fully connected layer set up with one and two layers (17–19). Therefore, there were a total of six combinations for the image-based pain classifier, and we applied the voting to optimize the classifier performance through averaging outputs of different models. With regards to the main hyperparameters, we used the cross-entropy loss as the loss function in the image-based pain classifier, and the learning rate, optimizer, and trained epochs were 1e-4, Adam, and 60 epochs, respectively.

## Video-Based Pain Classifiers

With regard to the video-based pain classifier, we employed a many-to-one sequence model given that the output of this study is a one pain grade. Similar to the image-based pain classifier, we used a Siamese network architecture as feature extractors. Given that multiple CNN encoders were used in the present study, we hence processed the image through three CNN encoders to get three vectors and concatenate these vectors to a relatively low-dimensional space. The concatenated output vectors of each frame were then fed into the bidirectional long short-term memory networks (BiLSTM) for the classification of pain (20). Given that the CNN encoder had been trained in the image-based pain classifier, we hence reduced the learning rate to 1e-5 on the video-based pain classifier and froze the weights of the CNN encoder in the first 10 epochs, and this approach may facilitate to focus on training BiLSTM in the first 10 epochs. The other parameters, such as loss function, optimizer, and trained epochs, were in line with those used in the image-based pain classifier.

## Statistical Analyses

Data were expressed in frequency of occurrence (percentages) for categorical variables and as means ± *SD* for continuous variables. Differences between the survivor and non-survivor groups were analyzed using Student's *t*-test for continuous variables and Fisher's exact test for categorical variables. The proportion of train, validation, and test datasets were 60, 20, and 20%, respectively. The performance of the pain classifier in the test dataset was determined by accuracy, sensitivity, and F1-score. Python version 3.8, PyTorch 1.9.1, and CUDA 11.1 were used in this study.

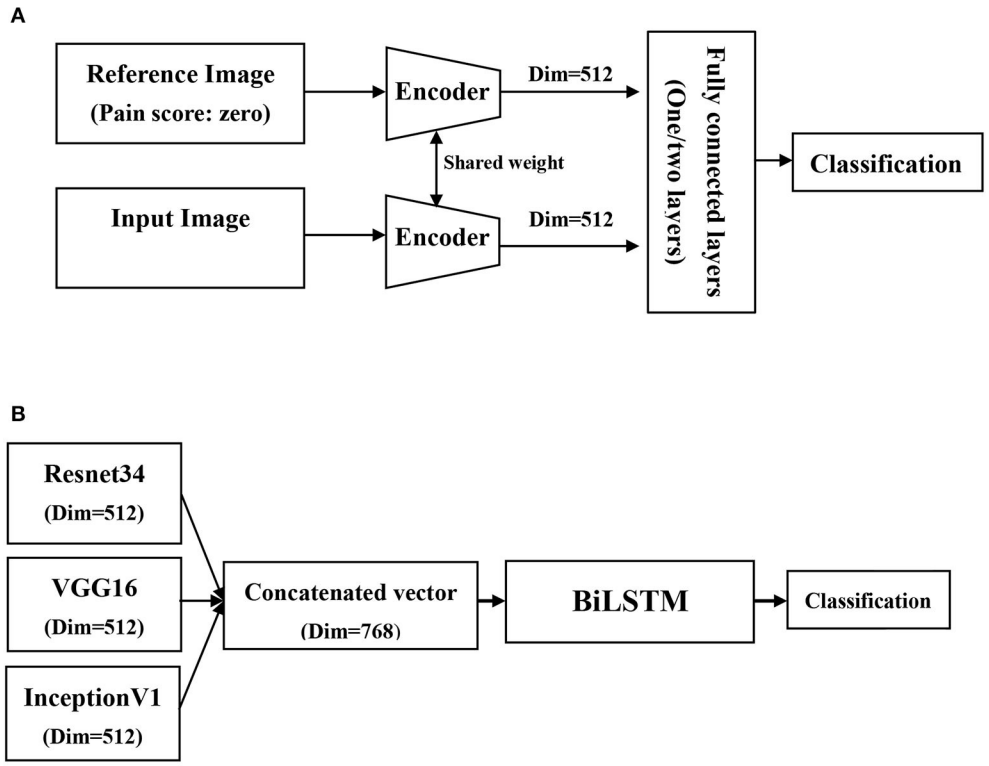



**FIGURE 2 |** Schematic diagram of network architectures in the present study. **(A)** Image-based pain classifiers using relation and siamese network architecture, **(B)** Video-base pain classifier using bidirectional long short-term memory networks (BiLSTM).

# RESULTS

## Patients' Characteristics

A total of 341 participants were enrolled, and there were 7,813 qualified videos, of which the number of scores 0, 1, and 2 were 5,717, 1,714, and 382, respectively. Given that we employed relation network architecture in this study, we hence used images among 63 participants who had all of the pain-score 0, 1, and 2 labeled video clips, and the number of videos with 0, 1, and 2 were 351, 253, and 142, respectively. The mean age of included patients for analyses was 69.3 ± 14.6 years, and 55.6 (35/63) of them was male (**Table 1**). The majority (81.0%, 51/63) of enrolled participants were critically ill patients who were admitted to medical ICUs. The ICU severity scores of acute physiology and chronic health evaluation II (APACHE II), sequential organ failure assessment (SOFA) day-1, SOFA day-3, and SOFA day-7 were 25.3 ± 5.7, 9.0 ± 3.7, 8.5 ± 4.1, and 8.2 ± 3.8, respectively.

## Performance of Image-Based Pain Classifiers

In image-based pain classifiers, we attempted to classify with three pain categories (0, 1, and 2) and dichotomous pain classifiers (0 vs. 1/2 and 0 vs. 2) given pain score = 2 reflects a clinical warning signaling requiring immediate clinical evaluation and management (**Table 2**). In Resnet34 with one fully connected layer (1024, 3), the performance of the polychromous classifier for 0, 1, and 2 appeared to be suboptimal, with the accuracy, sensitivity, and F1 score were merely 0.5589, 0.5589, and 0.5495, respectively. The performance of the two dichotomous image-based pain classifiers was much higher than that in polychromous pain classifier. The accuracy, sensitivity, and F1 score were 0.7668, 0.8422, and 0.8593 to classify 0 vs. 1/2 and were 0.8593, 0.8925, and 0.8638 to classify 0 vs. 2. We further tested the performance of using VGG16, InceptionV1, and two fully connected layers. The performances of Resnet34 and VGG16 were slightly higher than that of InceptionV1. For example, the accuracy of dichotomous pain classifier between 0 vs. 1/2 in Resnet34, VGG16, and Inception were 0.7668, 0.7578, and 0.7055, respectively. With regard to the efficacy of using two fully connected layers ([1024, 256] followed by [256, 3]), the performance tended to improve in a few models, such as dichotomous pain classifier between 0 vs. 1/2 in InceptionV1 (accuracy increased from 0.7055 to 0.7587).

## Performance of Video-Based Pain Classifiers and the Pain Classifier Without Reference

We then examined the performance of a video-based pain classifier through concatenating vectors of the aforementioned three CNN encoders and BiLSTM with distinct hidden layers (**Table 3**). We found that the performance of video-based

**TABLE 1 |** Characteristics of the enrolled 63 participants who had videos with all of three pain-score categories.

| **Basic data** | |
|---|---|
| Age, years | 69.3 ± 14.6 |
| Sex (male) | 35 (55.6%) |
| Height (cm) | 160.1 ± 8.0 |
| Body weight (kgs) | 57.3 ± 10.0 |
| **ICU types** | |
| Medical ICUs | 51 (81.0%) |
| Surgical ICUs | 12 (19.0%) |
| **Laboratory data (Day-1)** | |
| White blood cell counts (/ml) | 13,670.7 ± 11,259.5 |
| Hematocrit (%) | 28.8 ± 8.4 |
| Creatinine (mg/dl) | 1.9 ± 1.4 |
| Sodium (mg/dl) | 140.3 ± 5.5 |
| Potassium (mg/dl) | 4.0 ± 0.7 |
| **Severity scores** | |
| APACHE II score | 25.3 ± 5.7 |
| SOFA score, day-1 | 9.0 ± 3.7 |
| SOFA score, day-3 | 8.5 ± 4.1 |
| SOFA score, day-7 | 8.2 ± 3.8 |

*Data were presented as mean ± standard deviation and number (percentage). ICU, intensive care unit; APACHE II, acute physiology and chronic health evaluation II; SOFA, sequential organ failure assessment.*

**TABLE 2 |** Performance image-based pain classifiers with pain score zero as the reference in different settings.

| | **CNN model** | **Fully connected layers** | **Pain score 0 vs. 1 vs. 2** | **Pain score 0 vs. 1/2** | **Pain score 0 vs. 2** |
|---|---|---|---|---|---|
| Accuracy | Resnet34 | 1 layer (1024, 3) | 0.5589 | 0.7668 | 0.8593 |
| Sensitivity | | | 0.5589 | 0.8422 | 0.8925 |
| F1-score | | | 0.5495 | 0.7832 | 0.8638 |
| Accuracy | | 2 layers (1,024, 256) (256, 3) | 0.6032 | 0.7711 | 0.8568 |
| Sensitivity | | | 0.6032 | 0.8380 | 0.8514 |
| F1-score | | | 0.5969 | 0.7855 | 0.8561 |
| Accuracy | VGG16 | 1 layer (1024, 3) | 0.5914 | 0.7578 | 0.8557 |
| Sensitivity | | | 0.5914 | 0.6665 | 0.8499 |
| F1-score | | | 0.5867 | 0.7141 | 0.8548 |
| Accuracy | | 2 layers (1,024, 256) (256, 3) | 0.5871 | 0.7578 | 0.8276 |
| Sensitivity | | | 0.5871 | 0.6908 | 0.8064 |
| F1-score | | | 0.5811 | 0.7405 | 0.8239 |
| Accuracy | InceptionV1 | 1 layer (1024, 3) | 0.5872 | 0.7055 | 0.8302 |
| Sensitivity | | | 0.5872 | 0.8216 | 0.8782 |
| F1-score | | | 0.5788 | 0.7362 | 0.8380 |
| Accuracy | | 2 layers (1,024, 256) (256, 3) | 0.5567 | 0.7587 | 0.8035 |
| Sensitivity | | | 0.5567 | 0.8159 | 0.8338 |
| F1-score | | | 0.5556 | 0.7718 | 0.8093 |

*CNN, convolutional neural network.*

**TABLE 3 |** Performance of video-based pain classifiers with different numbers of hidden layers in bidirectional long short-term memory (BiLSTM) networks.

| | **Hidden layers** | **Pain score 0 vs. 1 vs. 2** | **Pain score 0 vs. 1/2** | **Pain score 0 vs. 2** |
|---|---|---|---|---|
| Accuracy | 64 | 0.6144 | 0.8145 | 0.8810 |
| Sensitivity | | 0.6144 | 0.7947 | 0.8755 |
| F1-score | | 0.6123 | 0.8107 | 0.8803 |
| Accuracy | 128 | 0.5941 | 0.8054 | 0.8461 |
| Sensitivity | | 0.5942 | 0.7858 | 0.7589 |
| F1-score | | 0.5902 | 0.8015 | 0.8314 |
| Accuracy | 256 | 0.6006 | 0.8268 | 0.8367 |
| Sensitivity | | 0.6006 | 0.8244 | 0.7500 |
| F1-score | | 0.5948 | 0.8264 | 0.8212 |

*BiLSTM, bidirectional long short-term memory.*

pain classifiers among the polychromous classifier and two dichotomous classifiers was higher than those in the image-based pain classifier. The accuracy in classifying 0 vs. 1/2 was nearly 0.8 and reached ∼0.88 to classify 0 vs. 2. Additionally, we further tested the performance of the established classifier without reference, mimicking the clinical scenario in a new patient without an image score of 0 as the reference (**Table 4**). We found that the performance of both image- and video-based classifiers slightly decreased in classifiers without reference. Notably, the performance of a video-based classifier without reference to differentiate 2 from 0 was up to 0.8906, indicating the established classifier had learned the difference between 0 and 2. Collectively, we established the image and video facial expression-based pain classifier in critically ill patients, with the accuracy to classify 0 vs. 1/2 and 0 vs. 2 were ∼0.8 and 0.9, respectively.

## DISCUSSION

In this prospective study, we developed a protocol to obtain video clips of facial expressions in critically ill patients and employed the deep learning-based approach to establish the facial expression-based pain classifier. We focused on the area nearby eyebrow that is less likely to be masked by medical devices and employed an ensemble of three CNN models, such as Resnet34, VGG16, and InceptionV1, to learn pain-associated facial features and BiLSTM for temporal relation between video frames. The accuracy of the dichotomous classifier to differentiate tense/grimacing (1/2) from relaxed (0) facial expression was ∼80%, and the accuracy to detect grimacing (2) was nearly 90%. The present study demonstrates the practical application of deep learning-based automated pain assessment in ICU, and the findings shed light on the application of medical artificial intelligence (AI) not only to improve patient care, but also to relieve healthcare workers from the routine workload.

Pain is the fifth vital sign in hospitalized patients but is somehow difficult to assess in critically ill patients who cannot self-report the pain (21, 22). Facial expressions of pain consist of coordinated pain-indicative muscle movements, particularly the contraction of muscles surrounding the eyes, i.e., orbicularis oculi muscle (23). Notably, facial pain responses appear to be

**TABLE 4 |** Accuracy of proposed image- and video-based pain classifiers with and without reference.

| | Reference | Pain score 0 vs. 1 vs. 2 | Pain score 0 vs. 1/2 | Pain score 0 vs. 2 |
|---|---|---|---|---|
| **Image-based pain classifiers** | | | | |
| Accuracy | Pain score 0 | 0.6347 | 0.8000 | 0.8937 |
| Sensitivity | | 0.6347 | 0.8022 | 0.8826 |
| F1-score | | 0.6321 | 0.8004 | 0.8953 |
| Accuracy | No reference | 0.6421 | 0.7954 | 0.8771 |
| Sensitivity | | 0.6421 | 0.7974 | 0.9074 |
| F1-score | | 0.6371 | 0.7947 | 0.8724 |
| **Video-based pain classifiers** | | | | |
| Accuracy | Pain score 0 | 0.6144 | 0.8268 | 0.8810 |
| Sensitivity | | 0.6144 | 0.8244 | 0.8755 |
| F1-score | | 0.6123 | 0.8264 | 0.8803 |
| Accuracy | No reference | 0.6130 | 0.7858 | 0.8906 |
| Sensitivity | | 0.6130 | 0.8016 | 0.8344 |
| F1-score | | 0.6102 | 0.7892 | 0.8841 |

consistent across distinct types of pain stimulation, such as pressure, temperature, electrical current, and ischemia (23, 24). A number of studies have explored the physiological basis of how pain signaling leads to pain-indicative muscle movement. Kuramoto et al. recently used facial myogenic potential topography in 18 healthy adult participants to investigate the facial myogenic potential and subsequent facial expressions (25). Furthermore, Kunz used functional MRI (fMRI) to address the association between brain responses in areas that processed the sensory dimension of pain and activation of the orbicularis oculi muscle (26). Although promising, monitoring of facial myogenic potential might be infeasible in critically ill patients given that contact device-associated issues regarding infection control and the potential interference with critical care (27). The possibility of application of fMRI in ICU appears to be low; therefore, using a portable camera to take high-quality video ∼1–2 m from the patient as well as AI-based image analyses focusing on eyebrow area as we have shown in the present study has high applicative value in critically ill patients.

It is estimated that more than 50% of patients in ICU experienced experience moderate to severe pain at rest, and 80% of critically ill patients experience pain during procedures (28, 29). Therefore, CPOT, as well as Behavioral Pain Scale (BPS), has been introduced for pain assessment in patients at ICUs in the past two decades, and facial expression is the fundamental domain in both BPS and CPOT given that muscle tension in facial areas, particularly facial area nearby eyebrow, can be directly observed by the caring staff without contact (3, 30). Notably, contactless monitoring in ICU is of increasing importance in the post-coronavirus disease (COVID) era (27). A number of AI-based tools, such as the dynamic relationship of facial landmarks or CNN-learned facial features, have been developed to assess pain in non-ICU patients (7, 23, 31). Nevertheless, the subtle pain-associated movement of facial muscles/landmarks in the non-ICU patient is largely distinct from those in critically ill patients under sedation. Given that patients in ICU often received mechanical ventilation, experienced fear were deprived of normal sleep, felt isolation; therefore, appropriate sedation, at least light sedation, is recommended as a standard of care in critically ill patients and hence leads to difficulties to identify pain based on facial expressions (32). In addition to the impact of sedation on pain assessment, subtle facial muscle movements might also be confounded by facial oedema resulting from fluid overload, which is highly prevalent in critically ill patients who underwent fluid resuscitation, as we have shown in our previous studies (33, 34). Collectively, automated pain assessment based on facial expressions in critically ill patients is currently an unmet need in the research field of medical AI due to the aforementioned difficulties.

Intriguingly, we found a suboptimal performance in the polychromous classifier, whereas the performance in dichotomous classifiers was high. We postulated that the relatively little difference between pain grades 1 and 2 may lead to the reduced performance to differentiate between 1 and 2, and the performance of dichotomous classifiers was high due to the apparent difference between 0 and 1/2. We found a higher performance in video classifiers than those in image classifiers, and this finding indicates that the temporal relation among image frames is crucial to classify pain by facial expressions. A similar finding has been found in pain classifiers using the UNBC-McMaster shoulder pain database (7–9). The accuracy of the leave-one-subject-out 25-fold cross in facial expression-based pain classifier by machine learning approach was ∼0.861 using the UNBC-McMaster database (7). Similar to our approach, Rodriguez et al. used VGG to learn basic facial features as well as LSTM to exploit the temporal relation between video frames and reported a further increased accuracy (0.933) in the aforementioned UNBC-McMaster database (8). Similarly, Huang *et al.* proposed an end-to-end hybrid network to extract multidimensional features including time-frame features from images of the UNBC-McMaster database and also found an improved performance (9). Recently, Semwal and Londhe further used distinct fusion network architectures, including CNN-based fusion network to learn both the spatial appearance and shape-based descriptors, as well as decision-level fusion network to learn the domain-specific spatial appearance and complementary features, to improve the performance of pain intensity assessment, with the F1 score, was ∼0.94 (10, 11). This evidence highlights the potential application of automated pain assessment based on facial expressions in hospital.

The inevitable medical devices and high heterogeneity in critically ill patients have led to technical difficulties as we have shown in this study. We choose to crop the facial area nearby the eyebrow area, and this approach not only keeps the essential area to detect painful facial expressions but also is essential to extend the established model to clinical scenarios with distinct facial masks, such as the increasing prevalence of wearing a facial mask in the post-COVID era. Moreover, we used a pain score of 0 to train the pain classifiers in this study and further tested the performance of established classifiers without reference (**Table 4**). Notably, the performance of dichotomous

classifiers, particularly the 0 vs. 2 classifier, remains high without reference, indicating that the established model has learned the pain-associated facial expression in critically ill patients.

Timely detection of severe pain, such as pain score 2, is crucial in critical care. Frequent pain assessment is substantial for the identification of the existence of pain and the adjustment dosage of pharmacological analgesic agents or the intensity of non-pharmacological management (1). The previous studies have shown that regular pain assessment is associated with a better outcome, such as ventilator-day, in critically ill patients (35, 36). Severe pain may reflect not only inadequate pain control, but also the potential deterioration of critical illness. For example, increasing pain has been implicated with anxiety, delirium, and poor both short-term and long-term outcomes in critically ill patients (37). Therefore, the automated AI-based pain assessment, particularly timely identification of severe pain/pain score 2, should serve as an actionable AI target, i.e., the detection of pain score 2 indicates the need for immediate evaluation and management by the healthcare worker. Additionally, we have established the user interface to guide the user with regard to quality of the image and the real-time classification of pain based on facial expressions, and the application of the established model should hence reach level 5 of technology readiness level (TRL) (**Supplementary Demonstration Video 1**) (38, 39).

There are limitations in this study. First, this study is a single center study. However, the pain relevant management in the study hospital is in accordance with the guideline; therefore, the generalization issue should be at least partly mitigated. Second, we recorded the video for 90 s in each record, and a longer duration could further improve the accuracy. Third, we focused on the facial expression in the present study, and more sensors for the other domains of CPOT/BPS are warranted in the future.

## CONCLUSION

Autonomous facial expression-based pain assessment is an essential issue in critical care but is somehow difficult in critically ill patients due to inevitable masked areas by medical devices and relatively subtle muscle movement resulting from sedation/oedema. In the present prospective study, we established the deep learning-based pain classifier based on facial expression focusing on the area nearby eyebrow, with the accuracy to detect tense/grimacing and grimacing were ∼80 and 90%, respectively. These findings indicate a real-world application of AI-based pain assessment based on the facial expression in ICU, and more studies are warranted to validate the performance of the automated pain assessment tool.

## DATA AVAILABILITY STATEMENT

The original contributions presented in the study are included in the article/**Supplementary Material**, further inquiries can be directed to the corresponding author/s.

## ETHICS STATEMENT

The study was approved by the Institutional Review Board of Taichung Veterans General Hospital (TCVGH: CE20325A). The patients/participants provided their written informed consent to participate in this study.

## AUTHOR CONTRIBUTIONS

C-LW, S-JS, S-FY, C-CC, and W-CC: study concept and design. S-FL, S-JS, and H-JC: acquisition of data. T-LY, C-HC, S-FY, Y-SL, C-CC, and W-CC: analysis and interpretation of data. C-LW and W-CC: drafting the manuscript.

## FUNDING

This study was supported by and Ministry of Science and Technology Taiwan (MOST 109-2321-B-075A-002). The funders had no role in the study design, data collection and analysis, decision to publish, or preparation of the manuscript.

## SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fmed.2022.851690/full#supplementary-material

## REFERENCES

1. Devlin JW, Skrobik Y, Gelinas C, Needham DM, Slooter AJC, Pandharipande PP, et al. Clinical practice guidelines for the prevention and management of pain, agitation/sedation, delirium, immobility, and sleep disruption in adult patients in the ICU. *Crit Care Med.* (2018) 46:e825–e73. doi: 10.1097/CCM.0000000000003259
2. Payen JF, Bru O, Bosson JL, Lagrasta A, Novel E, Deschaux I, et al. Assessing pain in critically ill sedated patients by using a behavioral pain scale. *Crit Care Med.* (2001) 29:2258–63. doi: 10.1097/00003246-200112000-00004
3. Gelinas C, Fillion L, Puntillo KA, Viens C, Fortier M. Validation of the critical-care pain observation tool in adult patients. *Am J Crit Care.* (2006) 15:420–7. doi: 10.4037/ajcc2006.15.4.420
4. Buchanan C, Howitt ML, Wilson R, Booth RG, Risling T, Bamford M. Predicted influences of artificial intelligence on the domains of nursing: scoping review. *JMIR Nurs.* (2020) 3:e23939. doi: 10.2196/23939
5. Bartlett MS, Littlewort GC, Frank MG, Lee K. Automatic decoding of facial movements reveals deceptive pain expressions. *Curr Biol.* (2014) 24:738–43. doi: 10.1016/j.cub.2014.02.009
6. Sikka K, Ahmed AA, Diaz D, Goodwin MS, Craig KD, Bartlett MS, et al. Automated assessment of children's postoperative pain using computer vision. *Pediatrics.* (2015) 136:e124–31. doi: 10.1542/peds.2015-0029
7. Pedersen H. Learning appearance features for pain detection using the Unbc-mcmaster shoulder pain expression archive database. *Computer Vision Syst.* (2015) 15:12. doi: 10.1007/978-3-319-20904-3_12
8. Rodriguez P, Cucurull G, Gonalez J, Gonfaus JM, Nasrollahi K, Moeslund TB, et al. Deep pain: exploiting long short-term memory networks for facial expression classification. *IEEE Trans Cybern.* (2017) 17:2199. doi: 10.1109/TCYB.2017.2662199
9. Huang Y, Qing L, Xu S, Wang L, Peng Y. HybNet: a hybrid network structure for pain intensity estimation. *Visual Comput.* (2021) 21:56. doi: 10.1007/s00371-021-02056-y

10. Semwal A, Londhe ND. Computer aided pain detection and intensity estimation using compact CNN based fusion network. *Appl Soft Comput.* (2021) 112:107780. doi: 10.1016/j.asoc.2021.107780
11. Semwal A, Londhe ND. MVFNet: A multi-view fusion network for pain intensity assessment in unconstrained environment. *Biomed Signal Process Control.* (2021) 67:102537. doi: 10.1016/j.bspc.2021.102537
12. Davoudi A, Malhotra KR, Shickel B, Siegel S, Williams S, Ruppert M, et al. Intelligent ICU for autonomous patient monitoring using pervasive sensing and deep learning. *Sci Rep.* (2019) 9:8020. doi: 10.1038/s41598-019-44004-w
13. Sanchez-Lozano E, Tzimiropoulos G, Martinez B, Torre F, Valstar M. A functional regression approach to facial landmark tracking. *IEEE Trans Pattern Anal Mach Intell.* (2018) 40:2037–50. doi: 10.1109/TPAMI.2017.2745568
14. Ge H, Dai Y, Zhu Z, Wang B. Robust face recognition based on multi-task convolutional neural network. *Math Biosci Eng.* (2021) 18:6638–51. doi: 10.3934/mbe.2021329
15. Chicco D. Siamese neural networks: an overview. *Methods Mol Biol.* (2021) 2190:73–94. doi: 10.1007/978-1-0716-0826-5_3
16. Sung F, Yang Y, Zhang L, Xiang T, Torr PHS, Hospedales TM. "Learning to compare: relation network for few-shot learning," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition.* (2018), p. 1199–208.
17. Kaiming He, Xiangyu Zhang, Shaoqing Ren, Sun. J. "Deep residual learning for image recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition.* (2016), p. 770-8.
18. Karen S, Zisserman A. "Very deep convolutional networks for large-scale image recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (2014).
19. Szegedy C, Wei L, Yangqing J, Sermanet P, Reed S, Anguelov D, et al. "Going deeper with convolutions," in *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR).* (2015), pp. 7–12.
20. Sharfuddin AA, Tihami MN, Islam MS, "A deep recurrent neural network with BiLSTM model for sentiment classification," in *2018 International Conference on Bangla Speech and Language Processing (ICBSLP).* (2018), p. 21–22.
21. Quality improvement guidelines for the treatment of acute pain and cancer pain. American Pain Society Quality of Care Committee. *JAMA.* (1995) 274:1874–80. doi: 10.1001/jama.274.23.1874
22. Arif-Rahu M, Grap MJ. Facial expression and pain in the critically ill non-communicative patient: state of science review. *Intensive Crit Care Nurs.* (2010) 26:343–52. doi: 10.1016/j.iccn.2010.08.007
23. Kunz M, Meixner D, Lautenbacher S. Facial muscle movements encoding pain-a systematic review. *Pain.* (2019) 160:535–49. doi: 10.1097/j.pain.0000000000001424
24. Prkachin KM. The consistency of facial expressions of pain: a comparison across modalities. *Pain.* (1992) 51:297–306. doi: 10.1016/0304-3959(92)90213-U
25. Kuramoto E, Yoshinaga S, Nakao H, Nemoto S, Ishida Y. Characteristics of facial muscle activity during voluntary facial expressions: Imaging analysis of facial expressions based on myogenic potential data. *Neuropsychopharmacol Rep.* (2019) 39:183–93. doi: 10.1002/npr2.12059
26. Kunz M, Chen JI, Rainville P. Keeping an eye on pain expression in primary somatosensory cortex. *Neuroimage.* (2020) 217:116885. doi: 10.1016/j.neuroimage.2020.116885
27. Lyra S, Mayer L, Ou L, Chen D, Timms P, Tay A, et al. A deep learning-based camera approach for vital sign monitoring using thermography images for ICU patients. *Sensors (Basel).* (2021) 21:1495. doi: 10.3390/s21041495
28. Chanques G, Sebbane M, Barbotte E, Viel E, Eledjam JJ, Jaber S. A prospective study of pain at rest: incidence and characteristics of an unrecognised symptom in surgical and trauma versus medical intensive care unit patients. *Anesthesiology.* (2007) 107:858–60. doi: 10.1097/01.anes.0000287211.98642.51
29. Puntillo KA, Max A, Timsit JF, Vignoud L, Chanques G, Robleda G, et al. Determinants of procedural pain intensity in the intensive care unit. *The Europain(R) study. Am J Respir Crit Care Med.* (2014) 189:39–47. doi: 10.1164/rccm.201306-1174OC
30. Chanques G, Payen JF, Mercier G, de Lattre S, Viel E, Jung B, et al. Assessing pain in non-intubated critically ill patients unable to self report: an adaptation of the Behavioral Pain Scale. *Intensive Care Med.* (2009) 35:2060–7. doi: 10.1007/s00134-009-1590-5
31. Neshov N, Manolova A. "Pain detection from facial characteristics using supervised descent method," in *IEEE 8th International Conference on Intelligent Data Acquisition and Advanced Computing Systems: Technology and Applications (IDAACS).* (2015), p. 251–6.
32. Chanques G, Constantin JM, Devlin JW, Ely EW, Fraser GL, Gelinas C, et al. Analgesia and sedation in patients with ARDS. *Intensive Care Med.* (2020) 46:2342–56. doi: 10.1007/s00134-020-06307-9
33. Chen YC, Zheng ZR, Wang CY, Chao WC. Impact of early fluid balance on 1-year mortality in critically ill patients with cancer: a retrospective study in Central Taiwan. *Cancer Control.* (2020) 27:1073274820920733. doi: 10.1177/1073274820920733
34. Wu CL, Pai KC, Wong LT, Wang MS, Chao WC. Impact of early fluid balance on long-term mortality in critically ill surgical patients: a retrospective cohort study in Central Taiwan. *J Clin Med.* (2021) 10:73. doi: 10.3390/jcm10214873
35. Payen JF, Bosson JL, Chanques G, Mantz J, Labarere J, Investigators D. Pain assessment is associated with decreased duration of mechanical ventilation in the intensive care unit: a post Hoc analysis of the DOLOREA study. *Anesthesiology.* (2009) 111:1308–16. doi: 10.1097/ALN.0b013e3181c0d4f0
36. Georgiou E, Hadjibalassi M, Lambrinou E, Andreou P, Papathanassoglou ED. The impact of pain assessment on critically ill patients' outcomes: a systematic review. *Biomed Res Int.* (2015) 2015:503830. doi: 10.1155/2015/503830
37. Reade MC, Finfer S. Sedation and delirium in the intensive care unit. *N Engl J Med.* (2014) 370:444–54. doi: 10.1056/NEJMra1208705
38. Martínez-Plumed F, Gómez E, Hernández-Orallo J. Futures of artificial intelligence through technology readiness levels. *Telematics and Informatics.* (2021) 58:101525. doi: 10.1016/j.tele.2020.101525
39. Komorowski M. Artificial intelligence in intensive care: are we there yet? *Intensive Care Med.* (2019) 45:1298–300. doi: 10.1007/s00134-019-05662-6


**Conflict of Interest:** The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

frontiers | Frontiers in Medicine

ORIGINAL RESEARCH
published: 15 June 2022
doi: 10.3389/fmed.2022.853102

# Development and Validation of Machine Learning Models for Real-Time Mortality Prediction in Critically Ill Patients With Sepsis-Associated Acute Kidney Injury

*Xiao-Qin Luo, Ping Yan, Shao-Bin Duan*, Yi-Xin Kang, Ying-Hao Deng, Qian Liu, Ting Wu and Xi Wu*

*Department of Nephrology, Hunan Key Laboratory of Kidney Disease and Blood Purification, The Second Xiangya Hospital of Central South University, Changsha, China*

OPEN ACCESS

***Edited by:***
*Longxiang Su,*
*Peking Union Medical College Hospital (CAMS), China*

***Reviewed by:***
*Yi Yang,*
*Southeast University, China*
*Jun Lyu,*
*First Affiliated Hospital of Jinan University, China*

****Correspondence:***
*Shao-Bin Duan*
*duansb528@csu.edu.cn*

***Specialty section:***
*This article was submitted to Intensive Care Medicine and Anesthesiology, a section of the journal Frontiers in Medicine*

***Received:*** *12 January 2022*
***Accepted:*** *19 May 2022*
***Published:*** *15 June 2022*

***Citation:***
*Luo X-Q, Yan P, Duan S-B, Kang Y-X, Deng Y-H, Liu Q, Wu T and Wu X (2022) Development and Validation of Machine Learning Models for Real-Time Mortality Prediction in Critically Ill Patients With Sepsis-Associated Acute Kidney Injury. Front. Med. 9:853102. doi: 10.3389/fmed.2022.853102*

**Background:** Sepsis-associated acute kidney injury (SA-AKI) is common in critically ill patients, which is associated with significantly increased mortality. Existing mortality prediction tools showed insufficient predictive power or failed to reflect patients' dynamic clinical evolution. Therefore, the study aimed to develop and validate machine learning-based models for real-time mortality prediction in critically ill patients with SA-AKI.

**Methods:** The multi-center retrospective study included patients from two distinct databases. A total of 12,132 SA-AKI patients from the Medical Information Mart for Intensive Care IV (MIMIC-IV) were randomly allocated to the training, validation, and internal test sets. An additional 3,741 patients from the eICU Collaborative Research Database (eICU-CRD) served as an external test set. For every 12 h during the ICU stays, the state-of-the-art eXtreme Gradient Boosting (XGBoost) algorithm was used to predict the risk of in-hospital death in the following 48, 72, and 120 h and in the first 28 days after ICU admission. Area under the receiver operating characteristic curves (AUCs) were calculated to evaluate the models' performance.

**Results:** The XGBoost models, based on routine clinical variables updated every 12 h, showed better performance in mortality prediction than the SOFA score and SAPS-II. The AUCs of the XGBoost models for mortality over different time periods ranged from 0.848 to 0.804 in the internal test set and from 0.818 to 0.748 in the external test set. The shapley additive explanation method provided interpretability for the XGBoost models, which improved the understanding of the association between the predictor variables and future mortality.

**Conclusions:** The interpretable machine learning XGBoost models showed promising performance in real-time mortality prediction in critically ill patients with SA-AKI, which are useful tools for early identification of high-risk patients and timely clinical interventions.

**Keywords: sepsis, acute kidney injury, mortality, machine learning, critical care**

# INTRODUCTION

Sepsis is life-threatening organ dysfunction due to a dysregulated host response to infection. It is a major cause of health loss worldwide (1, 2). Acute kidney injury (AKI), characterized by an abrupt increase in serum creatinine (SCr) or decrease in urine output, is a common complication of critical illness (3–5). AKI has been shown to be more frequent, less likely to resolve, and associated with higher mortality in critically ill patients with sepsis than in those without (6). Considering the critical condition of patients with sepsis-associated AKI (SA-AKI), the accurate prediction of their outcomes is a topic of interest.

Studies have shown that widely-used severity scores, such as the Simplified Acute Physiology Score II (SAPS-II) and the Sequential Organ Failure Assessment (SOFA) score, exhibit insufficient power for outcome prediction in SA-AKI patients (7, 8). A few prediction models for mortality in patients with SA-AKI have been established (7, 8). However, they were limited to small sample size or inadequate predictive performance. In addition, the models incorporated static measurements at single time points, typically in the early period after intensive care unit (ICU) admission, and failed to reflect patients' dynamic clinical evolution. There is still a lack of feasible ways to assess the real-time risk of death and guide individualized treatment decisions in critically ill patients with SA-AKI.

The rapid development in big data analytics and machine learning techniques, along with the data-rich environment in ICU settings, provide unprecedented opportunities to establish novel mortality prediction tools in SA-AKI patients (9–11). Advanced machine learning methods are adept at handling high-order interactions and fitting complex non-linear relationships, which can be used to integrate large amounts of data from electronic health records (EHRs). The application of data-driven analytics by machine learning has shown promise to improve predictive performance in medical fields (12–15).

The study aimed to develop and validate machine learning-based models for real-time mortality prediction in critically ill patients with SA-AKI, in an attempt to provide useful tools for early prognostic assessment and clinical decision-making.

# METHODS

## Source of Data

Data were obtained from the Medical Information Mart for Intensive Care IV (MIMIC-IV) v1.0 and the eICU Collaborative Research Database (eICU-CRD) v2.0 (16–19). The MIMIC-IV is a large and publicly available database containing records from patients admitted to the ICUs of the Beth Israel Deaconess Medical Center from 2008 to 2019. The eICU-CRD is a multi-center telehealth database including data from more than 200,000 admissions to 335 ICUs at 208 hospitals across the United States between 2014 and 2015. The study was an analysis of the third-party databases with pre-existing institutional review board approval and all protected patient information de-identified. One of the authors has completed the Collaborative Institutional Training Initiative course and can access the databases (certification number 40010711).

## Study Population

The study included adult patients with sepsis who developed AKI within 48 h after ICU admission. In the MIMIC-IV, sepsis was diagnosed based on the Sepsis-3 criteria, including suspected infection and a SOFA score $\geq 2$ (1). We identified patients with suspected infection (antibiotics administration concomitant with body fluid cultures) during the first 24 h after ICU admission and calculated SOFA scores using data from the same period (20). In the eICU-CRD, sepsis was identified according to the admission diagnosis recorded on the Acute Physiology and Chronic Health Evaluation IV dataset (21). AKI was defined based on the 2012 Kidney Disease: Improving Global Outcomes Clinical Practice Guideline, using both SCr and urine output criteria (3). Baseline SCr was defined as the minimum SCr value in the 7 days prior to ICU admission, or the first SCr value after ICU admission if no pre-admission SCr was available (22, 23). If the patient had multiple ICU admissions during a hospital stay, only the first ICU stay was included in the analysis to ensure the independence of the data. Patients with age $< 18$ years old, end-stage renal disease (identified by diagnosis codes), and ICU stay $< 48$ hours were excluded.

## Outcomes and Predictor Variables

The primary outcome was in-hospital mortality within 28 days after ICU admission, censored at hospital discharge or 28 days, whichever occurred first. Each patient's ICU stay within 28 days was separated into 12-hour windows, which were labeled as "death" or "survival". Specifically, to predict mortality in the next 48, 72, and 120 h, the time windows in the corresponding hours before death were labeled as "death" and the remaining as "survival". To predict mortality in the first 28 days after ICU admission, all time windows were labeled as "death" in patients who died and "survival" in patients who survived. The final objective of the model was to predict the correct label for each time window. Additionally, the secondary outcomes were ICU length of stay, hospital length of stay and use of renal replacement therapy (RRT) within the first 28 days.

The predictor variables within each time window contained four static features (age, sex, ethnicity, and baseline SCr) and sets of dynamic features including hours from ICU admission, vital signs, laboratory values, and interventions. The list of all predictor variables included for modeling is provided in **Table 1**. For dynamic features, their values were time-varying and updated on a 12-hour basis. We used the mean value of variables measured multiple times and the lowest Glasgow Coma Scale (GCS) score in each time window. For variables with no recorded measurements during the 12-hour windows, their values were carried forward from the most recent measurements.

## Statistical Analysis

Statistical analyses were performed using R 4.1.2 (https://cran.r-project.org). Continuous variables were presented as medians with interquartile ranges and categorical variables were presented as numbers with percentages. The schematic diagram of methods is shown in **Supplementary Figure S1**. We divided the study population in the MIMIC-IV into the training (50%), validation (30%), and internal test (20%) sets, randomized at the patient

level to ensure that each patient was allocated to only a subset. We used the cohort of SA-AKI patients in the eICU-CRD as an external test set. In the training set, the eXtreme Gradient Boosting (XGBoost) algorithm was used to establish mortality prediction models with all predictor variables input. XGBoost, a scalable end-to-end tree boosting system, is an optimized implementation of the gradient boosting framework designed to be highly efficient, flexible, and portable (24). During the training process, it generates a series of decision trees, each of which is generated based on the previous one to decrease the gradient of the loss function. After that, a prediction model composed of multiple decision trees is obtained. The XGBoost algorithm can handle missing values by adding a default direction for them in each tree node and learning the optimal direction from the data. Therefore, missing values were directly input into the XGBoost models as not available values. **Supplementary Table S1** provides the percentages of missing values in the predictor variables. For machine learning approaches, hyperparameter tuning is required to fit the complex relationship in the data and avoid overfitting. The hyperparameters in the XGBoost models (learning rate, minimum sum of instance weight, maximum tree depth, and minimum loss reduction) and max number of boosting iterations were optimized on the validation set to achieve the maximum area under the receiver operating characteristic curves (AUCs). The *xgboost* package was used for XGBoost modeling. Details on the functions and tuning parameters used for the XGBoost algorithm can be found in **Supplementary Table S2**. More

**TABLE 1 |** List of the predictor variables.

| | Variables | Type |
|---|---|---|
| Demographics | age, sex, ethnicity | static |
| Length of stay | hours from admission | dynamic |
| Vital signs | systolic blood pressure, diastolic blood pressure, heart rate, respiratory rate, body temperature, oxygen saturation, glasgow coma scale score, urine output | dynamic |
| Laboratory data | baseline serum creatinine | static |
| | hemoglobin, white blood cells, platelets, serum total bilirubin, serum albumin, serum creatinine, blood urea nitrogen, arterial pH, partial pressure of arterial oxygen, partial pressure of arterial carbon dioxide, serum sodium, serum potassium, serum chloride, serum bicarbonate, lactate, international normalized ratio, partial thromboplastin time | dynamic |
| Interventions | mechanical ventilation, vasopressors, renal replacement therapy, loop diuretics | dynamic |

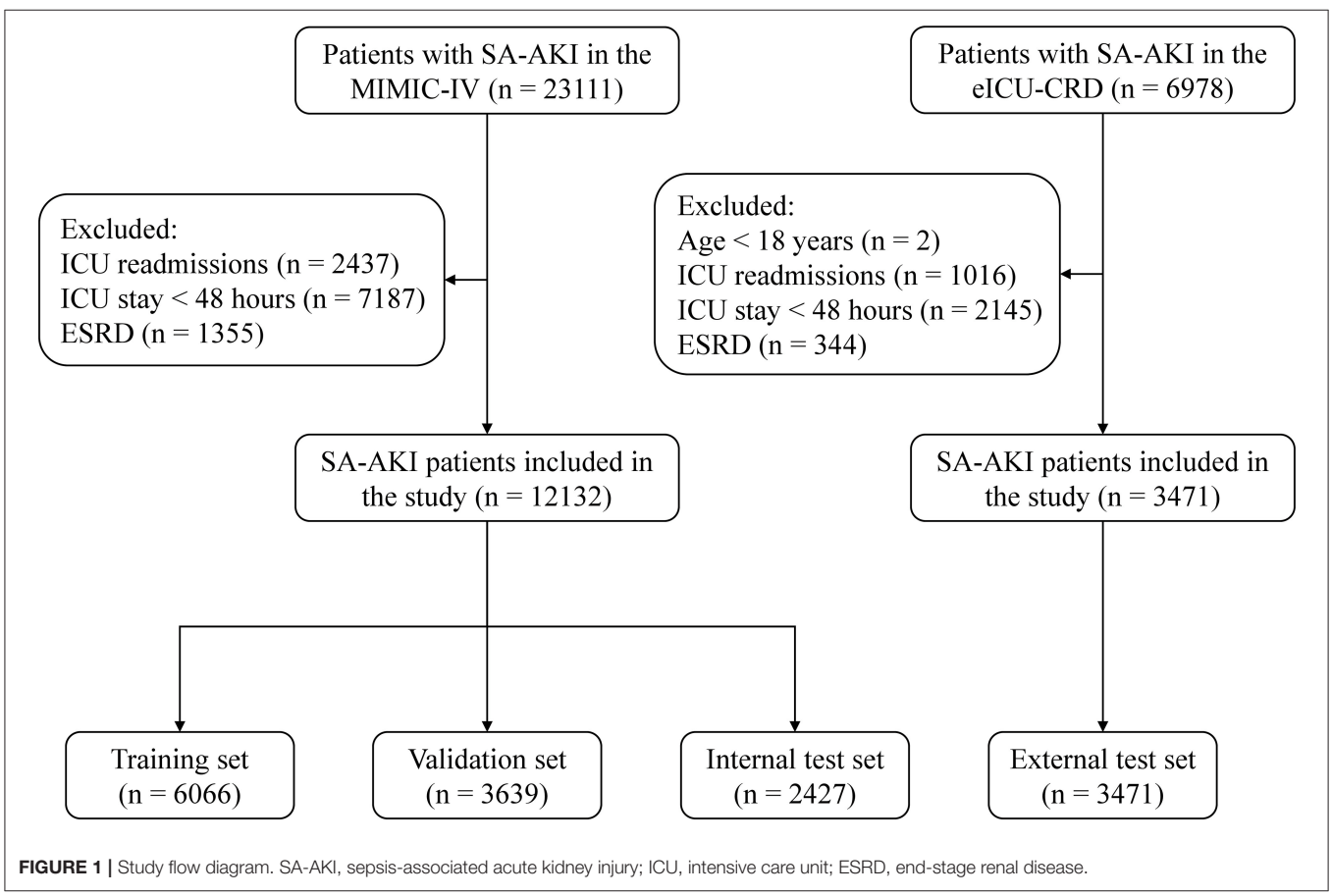



**FIGURE 1 |** Study flow diagram. SA-AKI, sepsis-associated acute kidney injury; ICU, intensive care unit; ESRD, end-stage renal disease.

**TABLE 2 |** Baseline characteristics and outcomes of SA-AKI patients in the training, validation and internal test sets.

| Variables | Training set ($n$ = 6,066) | Validation set ($n$ = 3,639) | Internal test set ($n$ = 2,427) |
|---|---|---|---|
| Age (year) | 69 (58–79) | 70 (59–80) | 69 (59–80) |
| Sex, male, *n* (%) | 3,501 (57.7) | 2,089 (57.4) | 1,330 (54.8) |
| **Ethnicity, *n* (%)** | | | |
| White | 4,182 (68.9) | 2,505 (68.8) | 1,730 (71.3) |
| Black | 512 (8.4) | 332 (9.1) | 219 (9.0) |
| Hispanic | 195 (3.2) | 109 (3.0) | 71 (2.9) |
| Asian | 139 (2.3) | 94 (2.6) | 34 (1.4) |
| Other/Unknown | 1,038 (17.1) | 599 (16.5) | 373 (15.4) |
| Baseline serum creatinine | 1.1 (0.8–1.5) | 1.0 (0.8–1.5) | 1.1 (0.8–1.5) |
| **Positive cultures*, *n* (%)** | | | |
| Respiratory culture | 960 (15.8) | 603 (16.6) | 347 (14.3) |
| Blood culture | 648 (10.7) | 391 (10.7) | 229 (9.4) |
| Urine culture | 1,044 (17.2) | 616 (16.9) | 375 (15.5) |
| Wound culture | 213 (3.5) | 134 (3.7) | 75 (3.1) |
| Fluid culture | 199 (3.3) | 116 (3.2) | 90 (3.7) |
| MRSA screen | 310 (5.1) | 189 (5.2) | 109 (4.5) |
| Tissue | 97 (1.6) | 67 (1.8) | 38 (1.6) |
| Anaerobic culture | 108 (1.8) | 64 (1.8) | 41 (1.7) |
| Fungal culture | 154 (2.5) | 86 (2.4) | 65 (2.7) |
| **KDIGO diagnostic criteria, *n* (%)** | | | |
| Serum creatinine | 540 (8.9) | 345 (9.5) | 229 (9.4) |
| Urine output | 3,467 (57.2) | 2,035 (55.9) | 1,368 (56.4) |
| Both | 2,059 (33.9) | 1,259 (34.6) | 830 (34.2) |
| **Outcomes** | | | |
| In-hospital mortality#, *n* (%) | 1,127 (18.6) | 620 (17.0) | 444 (18.3) |
| ICU length of stay | 4 (3–8) | 4 (3–8) | 4 (3–8) |
| Hospital length of stay | 10 (6–16) | 9 (6–16) | 9 (6–16) |
| Use of RRT#, *n* (%) | 562 (9.3) | 325 (8.9) | 224 (9.2) |

*MRSA, methicillin-resistant Staphylococcus aureus; KDIGO, kidney disease: improving global outcomes; ICU, intensive care unit; RRT, renal replacement therapy. Continuous variables were presented as median (interquartile range) and categorical variables were presented as n (%).*

** Positive cultures taken during the suspected infection time.*

*# In the first 28 days after ICU admission.*

details about the XGBoost algorithm can be found at XGBoost Documentation (https://xgboost.readthedocs.io/).

The performance of the prediction models was assessed on the internal and the external test sets. AUC was selected as the primary evaluation metric. Other metrics included sensitivity, specificity, and accuracy. We reported the metrics under multiple cutoff values, based on the local maximas of the receiver operating characteristic curves. We compared the performance of the XGBoost models with traditional risk scores, including the SOFA score (25) and SAPS-II (26). We did not calculate the risk scores in each 12-hour window for patients in the eICU-CRD because some required variables were unavailable.

The XGBoost algorithm provides the importance of features in predicting the outcome. We used the gain as the measure, representing the fractional contribution of each feature to the model output based on the total gain of this feature's splits. To explore the interpretability of the XGBoost models, we used the Shapley Additive exPlanations (SHAP) method (27), which provides consistent and locally accurate attribution values for each feature. The influence of the predictor variables on the outcome can be explained by the summing effects of variable attributions in calculating the output risk for each observation.

In sensitivity analysis, we applied other frequently used machine learning algorithms such as random forest and support vector machine to our dataset for comparison (28, 29). Additionally, we assessed the performance of the SOFA score, SAPS-II and XGBoost model using data gathered in the early period after ICU admission, i.e., the first 12 h, in predicting in-hospital mortality in the first 28 days.

# RESULTS

## Baseline Characteristics and Outcomes

A total of 15,603 critically ill patients with SA-AKI were included in our study, with 6,066 in the training set, 3,639 in the validation set, 2,427 in the internal test set, and 3,471 in the external test set (**Figure 1**). Baseline characteristics and outcomes of the study population in each dataset are shown in **Table 2** and **Supplementary Table S3**. In the MIMIC-IV, 56.6% of SA-AKI patients were diagnosed by urine output

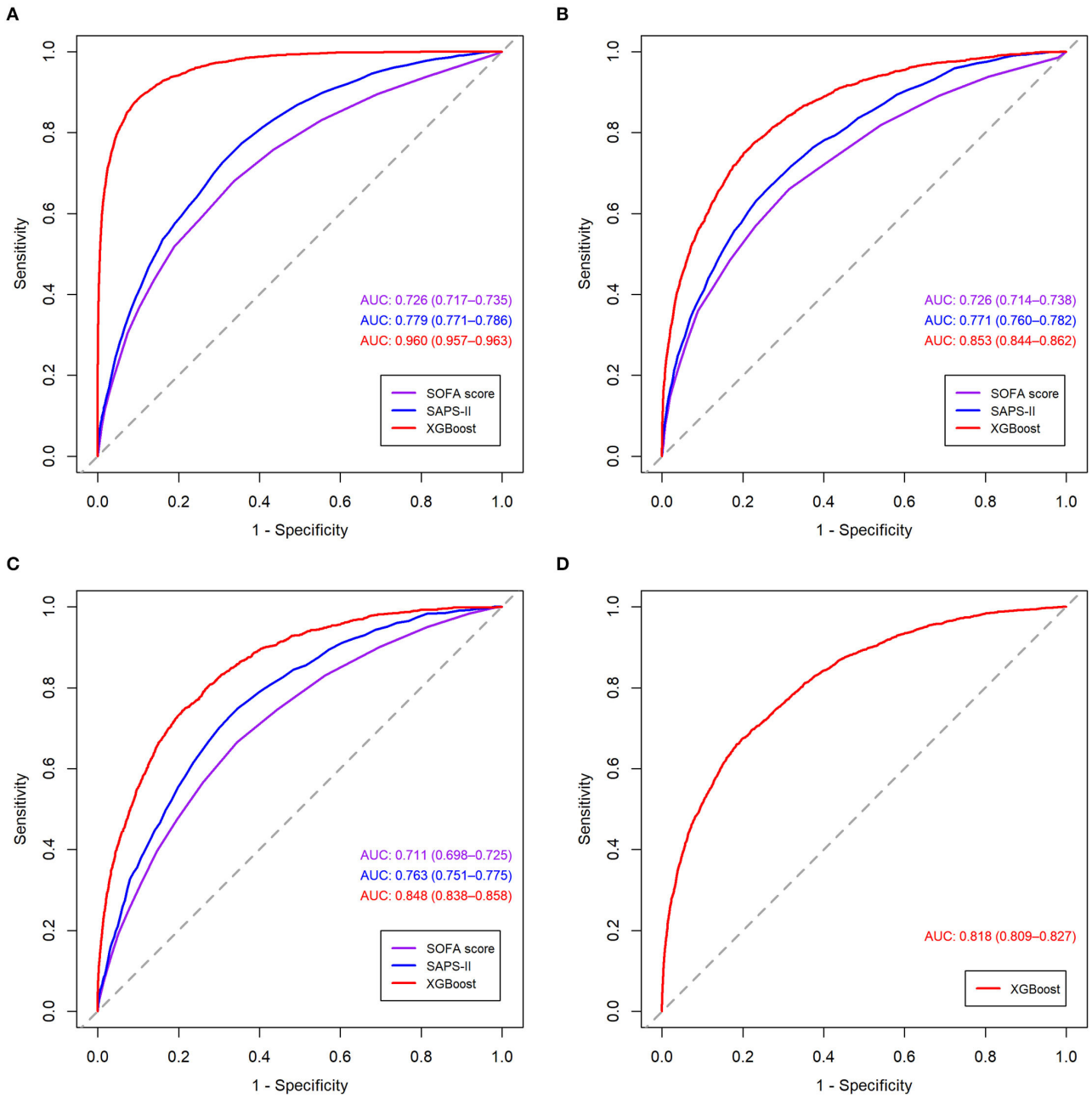



**FIGURE 2 |** Receiver operating characteristic curves of the models for mortality in the following 48 h in the training set. **(A)** validation set; **(B)** internal test set; **(C)** external test set; **(D)** SOFA, sequential organ failure assessment; SAPS-II, simplified acute physiology score II; XGBoost, extreme gradient boosting.

criteria, 9.2% by SCr criteria, and 34.2% by both criteria. In the eICU-CRD, the proportions of SA-AKI patients meeting urine output criteria, SCr criteria, and both criteria were 38.5, 40.9, and 20.5%, respectively. The overall in-hospital mortality within 28 days was 18.6% in the training set, 17.0% in the validation set, 18.3% in the internal test set, and 22.7% in the external test set. For each 12 h window of the ICU stays, the number of in-hospital deaths in the first 28 days is shown in **Supplementary Table S4**. Distribution of the predictor variables within each 12-hour window of the ICU stays is shown in **Supplementary Table S5**.

## Model Performance

The receiver operating characteristic curves of the models for mortality in the following 48, 72, and 120 h and in the first 28 days after ICU admission are shown in **Figure 2** and **Supplementary Figures S2–S4**. The XGBoost models showed better discrimination than the SOFA score and SAPS-II,

with the AUCs ranging from 0.848 to 0.804 in the internal test set and from 0.818 to 0.748 in the external test set. The sensitivity, specificity, and accuracy of the XGBoost models at different cutoffs for mortality prediction in the internal and the external test sets are provided in **Table 3** and **Supplementary Tables S6–S8**. In the internal test set, the XGBoost model achieved a sensitivity of 80.1% and specificity of 72.9% at the cutoff of 0.0349 for mortality in the following 48 h. The sensitivity was slightly higher, and the specificity was lower in the external test set than in the internal test set across different cutoffs. The calibration curves of the XGBoost models comparing the predicted and observed probability across deciles in the internal and the external test sets are shown in **Figure 3** and **Supplementary Figures S5–S7**. The XGBoost models were well-calibrated, except that they might underestimate or overestimate the probability at the higher risk deciles.

**TABLE 3 |** Performance of the XGBoost model for mortality in the following 48 h at different cutoffs.

| Cutoffs | Sensitivity (%) | Specificity (%) | Accuracy (%) |
|---|---|---|---|
| **Internal test set** | | | |
| 0.0214 | 90.0 | 58.5 | 60.0 |
| 0.0280 | 85.0 | 66.6 | 67.4 |
| 0.0349 | 80.1 | 72.9 | 73.2 |
| 0.0431 | 75.0 | 78.3 | 78.1 |
| 0.0445* | 74.3 | 79.1 | 78.9 |
| 0.0515 | 70.0 | 82.5 | 81.9 |
| 0.0600 | 65.1 | 85.6 | 84.7 |
| 0.0676 | 60.0 | 87.9 | 86.6 |
| **External test set** | | | |
| 0.0214 | 93.2 | 40.9 | 43.8 |
| 0.0280 | 89.2 | 50.5 | 52.7 |
| 0.0349 | 85.5 | 57.8 | 59.4 |
| 0.0431 | 81.1 | 64.7 | 65.7 |
| 0.0515 | 75.8 | 70.4 | 70.7 |
| 0.0600 | 71.6 | 75.0 | 74.8 |
| 0.0676 | 69.1 | 78.1 | 77.6 |
| 0.0735* | 67.4 | 80.3 | 79.5 |

**The cutoff value corresponding to the maximum Youden index (sensitivity + specificity - 1).*

## Model Interpretability

**Figure 4** and **Supplementary Figures S8–S10** illustrate the feature importance derived from the XGBoost models. The top five most important predictor variables in the XGBoost model for mortality in the following 48 h were urine output, GCS score, hours from admission, serum lactate level, and age. **Figure 5** and **Supplementary Figures S11–S13** provide the SHAP summary plots of the XGBoost models, revealing the impact of the predictor variables on model output. Lower GCS score, decreased urine output, prolonged ICU length of stay, older age, and higher blood urea nitrogen (BUN) level were the top five factors associated with increased risk of death in the following 48 h.

## Sensitivity Analysis

In sensitivity analysis, the XGBoost models showed higher AUCs than the random forest and the support vector machine models in the internal and the external test sets (**Supplementary Table S9**). In addition, the XGBoost model using data gathered during the first 12 h after ICU admission showed poor predictive

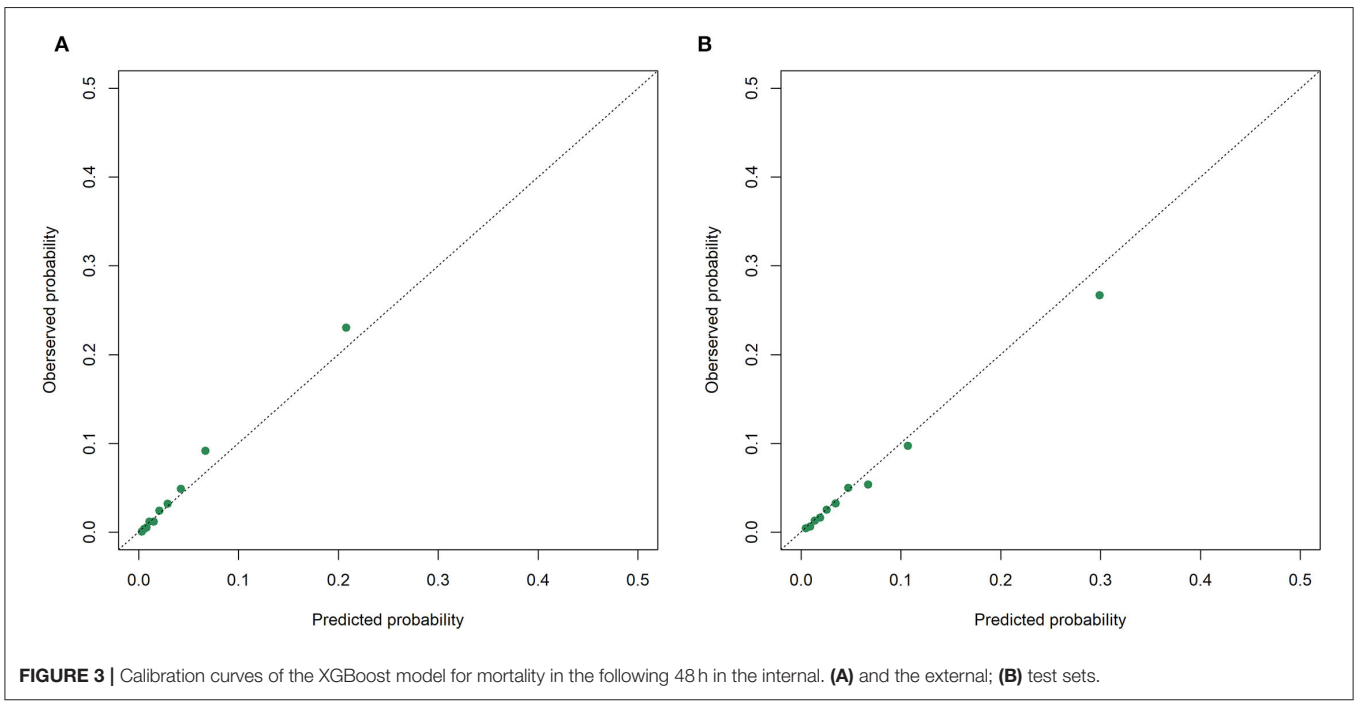



**FIGURE 3 |** Calibration curves of the XGBoost model for mortality in the following 48 h in the internal. **(A)** and the external; **(B)** test sets.

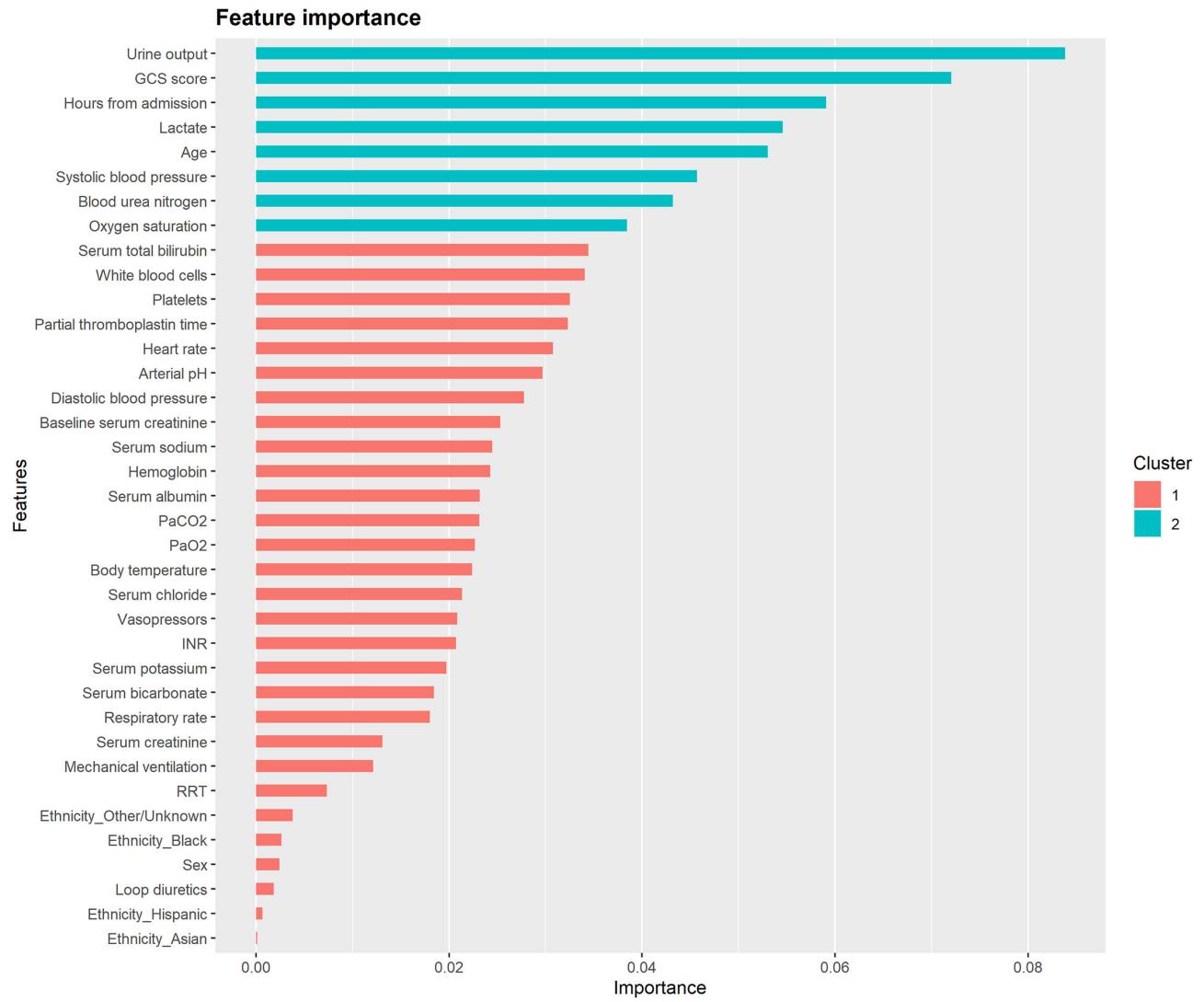



**FIGURE 4 |** Feature importance derived from the XGBoost model for mortality in the following 48 h. The importance value represents the fractional contribution of each feature to the XGBoost model based on the total gain of this feature's splits. Higher percentage means a more important feature. GCS, glasgow coma scale; $PaCO_2$, partial pressure of arterial carbon dioxide; $PaO_2$, partial pressure of arterial oxygen; INR, international normalized ratio; RRT, renal replacement therapy.

performance for in-hospital mortality in the first 28 days, with the AUC being 0.770 (95% CI 0.747–0.794) in the internal test set and 0.676 (95% CI 0.655–0.697) in the external test set (**Supplementary Figure S14**).

## DISCUSSION

In this multi-center retrospective study, we developed and validated interpretable machine learning-based models using the XGBoost algorithm for real-time mortality prediction in critically ill patients with SA-AKI. The XGBoost models exhibited better performance than traditional risk scores (including the SOFA score and SAPS-II) or other machine learning models (including the random forest and support vector machine models) in predicting death in the following 48, 72, and 120 h and in the first 28 days after ICU admission. The XGBoost models could help identify high-risk patients in real time for early clinical interventions.

SA-AKI is common in critically ill patients with rapid clinical evolution and significantly higher mortality than those without AKI or with AKI attributed to other causes (6). Reliable prediction models are essential for clinicians to assess the risk of death and make proper clinical decisions in critically ill patients with SA-AKI. Generic scores, such as the SOFA score and SAPS-II, are widely used for outcome prediction in critical care settings. However, they have shown controversial results on predictive performance for mortality in AKI patients (7, 8, 30–32). Recently, several models have been proposed to predict AKI mortality in unselected ICU patients (31, 32), but few have been validated in patients with SA-AKI. Da Hora Passos et al. (7) proposed a

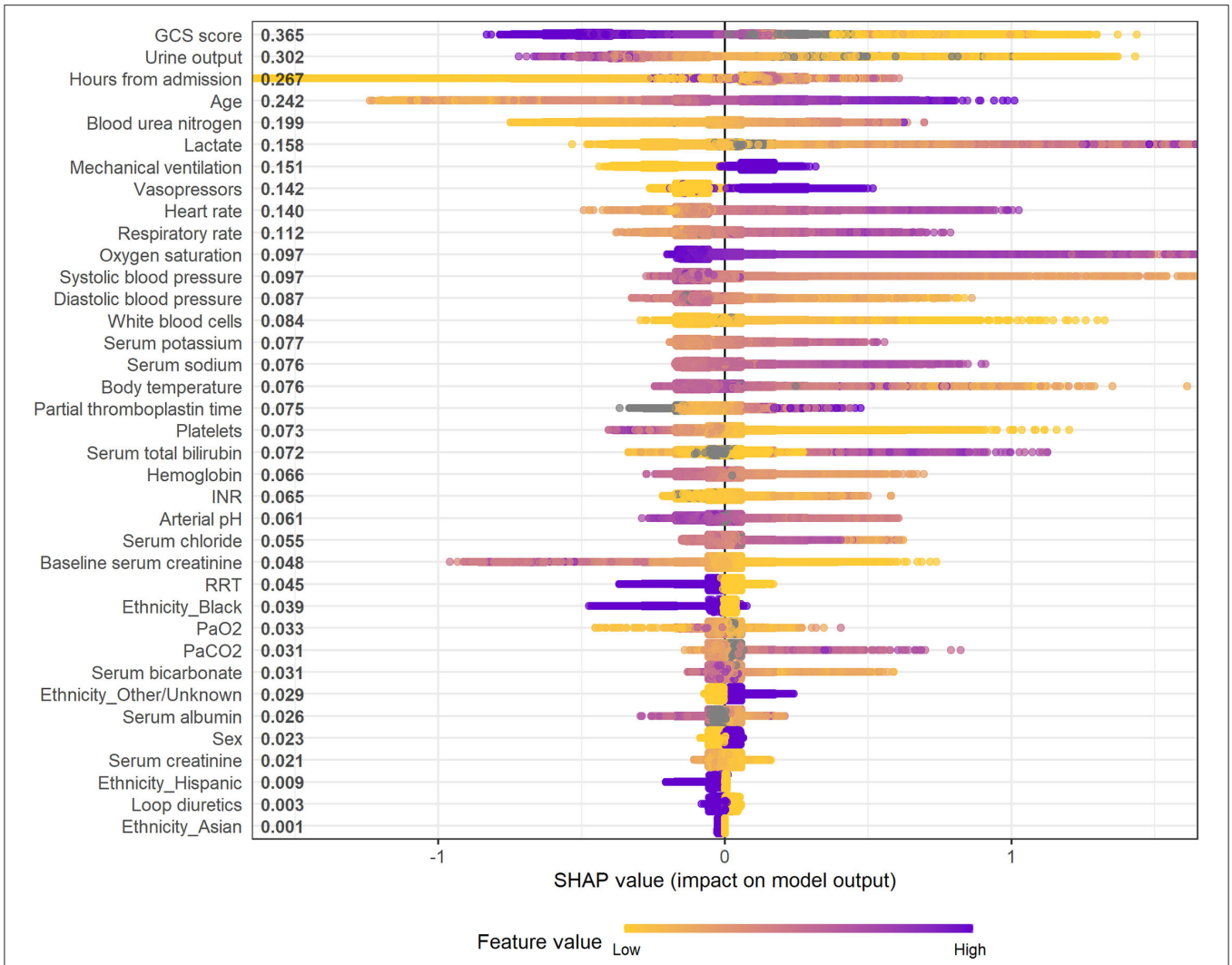



**FIGURE 5 |** SHAP summary plot of the XGBoost model for mortality in the following 48 h. Higher SHAP value means a higher probability of death within the next 48 h. Purple represents higher feature values and yellow represents lower feature values. A dot is created for each feature attribution in calculating the output risk for each observation. GCS, glasgow coma scale; INR, international normalized ratio; RRT, renal replacement therapy; $PaO_2$, partial pressure of arterial oxygen; $PaCO_2$, partial pressure of arterial carbon dioxide.

clinical score to predict 7 days mortality in a cohort of 186 SA-AKI patients who required continuous RRT. The five-variable score showed better performance than the generic models, with a C-statistic of 0.82, but was limited to a single center and small sample size. In addition, Hu et al. (8) established a prediction model for in-hospital mortality in critically ill patients with SA-AKI. However, the model included only static clinical variables and showed insufficient predictive power.

Compared with the other risk prediction tools, our models have several strengths. First, the study demonstrated the applicability of the XGboost algorithm in mortality prediction in critically ill patients with SA-AKI. The XGBoost models had stronger predictive power than the traditional risk scores. Sensitivity analysis further showed that the XGBoost models were superior to the random forest and the support vector machine models. XGBoost-based models have shown exciting performance in various situations, such as volume responsiveness in patients with oliguric AKI (14), long-term kidney outcomes in patients with IgA nephrology (33), and mortality in ICU patients with rhabdomyolysis (34). The reasons for the improvement in predictive abilities observed in the XGBoost models may be multifactorial. The XGBoost algorithm, based on the gradient tree boosting framework, is adept at fitting non-linearities, discontinuities and complex high-order interactions. It is also robust to outliers in and multicollinearity among predictor variables. Besides, the XGBoost algorithm can handle missing values automatically, allowing the input of only available predictor variables in its clinical application.

Second, the real-time mortality prediction models can provide dynamic risk assessment and guide clinical decision-making. Patients in the ICU environment are clinically unstable, change rapidly between states of deterioration and improvement, and

require continuous monitoring and interventions (35). It has promoted the establishment of real-time prediction models in critical care, such as models for mortality in critically ill children (35), the development of AKI (36), and sepsis onset (37, 38). Previously published models for mortality prediction in SA-AKI patients included static physiological parameters gathered during the early stages of the ICU stays. However, SA-AKI patients with similar disease severity at the early stage of ICU admission may exhibit different clinical outcomes due to distinct disease trajectories and treatment responses. The real-time prediction models can provide the risk of death updated on a 12-hour basis, which is more accurate and allows clinicians to make predictions dynamically.

Third, our models achieved promising predictive performance in both the internal and the external test sets, which demonstrated their robustness and generalizability. The predictor variables included in our model are routinely collected and usually available in the EHRs, and their values are rarely influenced by the examiner. Using only the most basic and commonly measured clinical data can facilitate the generalizability of the prediction model in other ICUs. Our models were further validated in an external test set, including 3,471 SA-AKI patients from a large multi-center critical care database with significantly different distributed features. Furthermore, automated data extraction from EHRs and data input can save additional labor and cost and reduce the possibility of incorrect entry in future clinical applications of the models (35).

Fourth, the interpretability of the models was explored to reveal the predictors for death over different time periods. Most recently, the relationship between the evolution of SA-AKI and mortality has been revealed. Uhel et al. (39) found that persistent AKI, but not transient AKI, was associated with increased mortality in critically ill septic patients. Ozrazgat-Baslanti et al. (40) also showed that persistent AKI and the absence of renal recovery were associated with worse clinical outcomes. Our results further demonstrated that decreased urine output and higher BUN level were important factors for increased real-time risk of death, suggesting the necessity for continuous renal function monitoring in SA-AKI patients. Additionally, the discovery of other potentially modifiable extra-renal risk factors, such as lower GCS score, higher lactate level, higher heart rate, and higher respiratory rate, may help improve patient care and outcomes.

Our study was subject to some limitations. Firstly, it was a retrospective analysis based on the publicly accessible databases. The diagnosis of sepsis in the eICU-CRD may not meet the updated Sepsis-3 criteria. It remains unclear whether the prediction model performs well for individual prognostication and whether its clinical application can improve patient outcomes. Secondly, although the XGBoost algorithm can handle missing values automatically, the presence of missing data may lead to bias. Thirdly, clinical data beyond the ICU stays were unavailable, limiting the continuous assessment of the risk of death for SA-AKI patients who were transferred to the general wards or other locations. Finally, the visualization and application of the models are still limited. In our subsequent study, we will prospectively investigate the effectiveness of our models and develop a web-based risk calculator that automatically extracts data from EHRs and performs risk calculations.

## CONCLUSIONS

This study developed and externally validated interpretable machine learning XGBoost models for real-time mortality prediction in critically ill patients with SA-AKI. The XGBoost models, based on routine clinical variables updated every 12 h, showed promising performance in predicting death in the following 48, 72, and 120 h and in the first 28 days after ICU admission. The real-time prediction models are useful tools for early identification of high-risk patients and timely clinical interventions. Future studies are required to determine the robustness and effectiveness of the prediction models in a prospective way.

## DATA AVAILABILITY STATEMENT

The datasets analyzed for this study can be found in the MIMIC-IV (https://mimic.mit.edu/) and eICU-CRD (https://eicu-crd.mit.edu/).

## ETHICS STATEMENT

The studies involving human participants were reviewed and approved by the Institutional Review Boards of the Beth Israel Deaconess Medical Center and Massachusetts Institute of Technology. Written informed consent for participation was not required for this study in accordance with the National Legislation and the Institutional Requirements.

## AUTHOR CONTRIBUTIONS

S-BD designed, supervised the study, and drafted the manuscript. X-QL performed the data extraction, analysed, interpreted the data, and drafted the manuscript. PY and Y-XK analyzed and interpreted the data and critically revised the manuscript. Y-HD, TW, and XW analyzed the data and revised the manuscript critically for important intellectual content. All authors have read and approved the final manuscript.

## FUNDING

This study was supported by National Natural Science Foundation of China (Grant No. 81873607).

## SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fmed.2022.853102/full#supplementary-material

## REFERENCES


1. Singer M, Deutschman CS, Seymour CW, Shankar-Hari M, Annane D, Bauer M, et al. The third international consensus definitions for sepsis and septic shock (Sepsis-3). *JAMA*. (2016) 315:801–10. doi: 10.1001/jama.2016.0287
2. Rudd KE, Johnson SC, Agesa KM, Shackelford KA, Tsoi D, Kievlan DR, et al. Global, regional, and national sepsis incidence and mortality, 1990–2017: analysis for the Global Burden of Disease Study. *Lancet*. (2020) 395:200–11. doi: 10.1016/s0140-6736(19)32989-7
3. KDIGOKAKIW GROUP. KDIGO clinical practice guideline for acute kidney injury. *Kidney Int Suppl.* (2012) 2:1–138. doi: 10.1038/kisup.2012.1
4. Hoste EA, Bagshaw SM, Bellomo R, Cely CM, Colman R, Cruz DN, et al. Epidemiology of acute kidney injury in critically ill patients: the multinational AKI-EPI study. *Intensive Care Med.* (2015) 41:1411–23. doi: 10.1007/s00134-015-3934-7
5. Hoste EAJ, Kellum JA, Selby NM, Zarbock A, Palevsky PM, Bagshaw SM, et al. Global epidemiology and outcomes of acute kidney injury. *Nat Rev Nephrol.* (2018) 14:607–25. doi: 10.1038/s41581-018-0052-0
6. Peters E, Antonelli M, Wittebole X, Nanchal R, Francois B, Sakr Y, et al. A worldwide multicentre evaluation of the influence of deterioration or improvement of acute kidney injury on clinical outcome in critically ill patients with and without sepsis at ICU admission: results from the intensive care over nations audit. *Crit Care.* (2018) 22:188. doi: 10.1186/s13054-018-2112-z
7. da Hora Passos R, Ramos JG, Mendonca EJ, Miranda EA, Dutra FR, Coelho MF, et al. A clinical score to predict mortality in septic acute kidney injury patients requiring continuous renal replacement therapy: the helenicc score. *BMC Anesthesiol.* (2017) 17:e21. doi: 10.1186/s12871-017-0312-8
8. Hu H, Li L, Zhang Y, Sha T, Huang Q, Guo X, et al. A Prediction model for assessing prognosis in critically ill patients with sepsis-associated acute kidney injury. *Shock.* (2021) 56:564–72. doi: 10.1097/SHK.0000000000001768
9. Bailly S, Meyfroidt G, Timsit JF. What's new in ICU in 2050: big data and machine learning. *Intensive Care Med.* (2018) 44:1524–7. doi: 10.1007/s00134-017-5034-3
10. Sanchez-Pinto LN, Luo Y, Churpek MM. Big data and data science in critical care. *Chest*. (2018) 154:1239–48. doi: 10.1016/j.chest.2018.04.037
11. Gutierrez G. Artificial intelligence in the intensive care unit. *Crit Care.* (2020) 24:101. doi: 10.1186/s13054-020-2785-y
12. Kang MW, Kim J, Kim DK, Oh KH, Joo KW, Kim YS, et al. Machine learning algorithm to predict mortality in patients undergoing continuous renal replacement therapy. *Crit Care.* (2020) 24:42. doi: 10.1186/s13054-020-2752-7
13. Meyer A, Zverinski D, Pfahringer B, Kempfert J, Kuehne T, Sündermann SH, et al. Machine learning for real-time prediction of complications in critical care: a retrospective study. *Lancet Respir Med.* (2018) 6:905–14. doi: 10.1016/s2213-2600(18)30300-x
14. Zhang Z, Ho KM, Hong Y. Machine learning for the prediction of volume responsiveness in patients with oliguric acute kidney injury in critical care. *Crit Care.* (2019) 23:112. doi: 10.1186/s13054-019-2411-z
15. Luo X-Q, Yan P, Zhang N-Y, Luo B, Wang M, Deng Y-H, et al. Machine learning for early discrimination between transient and persistent acute kidney injury in critically ill patients with sepsis. *Sci Rep.* (2021) 11:20269. doi: 10.1038/s41598-021-99840-6
16. 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. doi: 10.1038/sdata.2018.178
17. Johnson A, Bulgarelli L, Pollard T, Horng S, Celi LA, Mark R. MIMIC-IV (version 1.0). *PhysioNet*. (2020). doi: 10.13026/s6n6-xd98
18. Yang J, Li Y, Liu Q, Li L, Feng A, Wang T, et al. Brief introduction of medical database and data mining technology in big data era. *J Evid Based Med.* (2020) 13:57–69. doi: 10.1111/jebm.12373
19. Wu WT, Li YJ, Feng AZ, Li L, Huang T, Xu AD, et al. Data mining in clinical big data: the frequently used databases, steps, and methodological models. *Mil Med Res.* (2021) 8:44. doi: 10.1186/s40779-021-00338-z
20. Johnson AEW, Aboab J, Raffa JD, Pollard TJ, Deliberato RO, Celi LA, et al. A comparative analysis of sepsis identification methods in an electronic database. *Crit Care Med.* (2018) 46:494–9. doi: 10.1097/CCM.0000000000002965
21. 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) 34:1297–310. doi: 10.1097/01.CCM.0000215112.84523.F0
22. Zhao GJ, Xu C, Ying JC, Lu WB, Hong GL, Li MF, et al. Association between furosemide administration and outcomes in critically ill patients with acute kidney injury. *Crit Care.* (2020) 24:75. doi: 10.1186/s13054-020-2798-6
23. Chaudhary K, Vaid A, Duffy A, Paranjpe I, Jaladanki S, Paranjpe M, et al. Utilization of deep learning for subphenotype identification in sepsis-associated acute kidney injury. *Clin J Am Soc Nephrol.* (2020) 15:1557–65. doi: 10.2215/CJN.09330819
24. Chen T, Guestrin C. XGBoost: a scalable tree boosting system. In: *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. San Francisco, Ca (2016). p. 785–94.
25. Vincent JL, Moreno R, Takala J, Willatts S, De Mendonça A, Bruining H, et al. The SOFA (Sepsis-related Organ Failure Assessment) score to describe organ dysfunction/failure. On behalf of the Working Group on Sepsis-Related Problems of the European Society of Intensive Care Medicine. *Intensive Care Med.* (1996) 22:707–10. doi: 10.1007/bf01709751
26. Le Gall JR, Lemeshow S, Saulnier F. A new simplified acute physiology score (SAPS II) based on a European/North American multicenter study. *JAMA*. (1993) 270:2957–63. doi: 10.1001/jama.270.24.2957
27. Lundberg SM, Lee S-I. A unified approach to interpreting model predictions. *Adv Neural Inf Process Syst.* (2017) 30:4765–74.
28. Breiman L. Random forests. *Mach Learn.* (2001) 45:5–32. doi: 10.1023/A:1010933404324
29. Cortes C, Vapnik V. Support-vector networks. *Mach Learn.* (1995) 20:273–97. doi: 10.1023/A:1022627411411
30. Demirjian S, Chertow GM, Zhang JH, O'Connor TZ, Vitale J, Paganini EP, et al. Model to predict mortality in critically ill adults with acute kidney injury. *Clin J Am Soc Nephrol.* (2011) 6:2114–20. doi: 10.2215/CJN.02900311
31. Lin K, Hu Y, Kong G. Predicting in-hospital mortality of patients with acute kidney injury in the ICU using random forest model. *Int J Med Inform.* (2019) 125:55–61. doi: 10.1016/j.ijmedinf.2019.02.002
32. Huang H, Liu Y, Wu M, Gao Y, Yu X. Development and validation of a risk stratification model for predicting the mortality of acute kidney injury in critical care patients. *Ann Transl Med.* (2021) 9:323. doi: 10.21037/atm-20-5723
33. Chen T, Li X, Li Y, Xia E, Qin Y, Liang S, et al. Prediction and risk stratification of kidney outcomes in IgA nephropathy. *Am J Kidney Dis.* (2019) 74:300–9. doi: 10.1053/j.ajkd.2019.02.016
34. Liu C, Liu X, Mao Z, Hu P, Li X, Hu J, et al. Interpretable machine learning model for early prediction of mortality in ICU patients with rhabdomyolysis. *Med Sci Sports Exerc.* (2021) 53:1826–34. doi: 10.1249/mss.0000000000002674
35. Kim SY, Kim S, Cho J, Kim YS, Sol IS, Sung Y, et al. A deep learning model for real-time mortality prediction in critically ill children. *Crit Care.* (2019) 23:279. doi: 10.1186/s13054-019-2561-z
36. Le S, Allen A, Calvert J, Palevsky PM, Braden G, Patel S, et al. Convolutional Neural Network Model for Intensive Care Unit Acute Kidney Injury Prediction. *Kidney Int Rep.* (2021) 6:1289–98. doi: 10.1016/j.ekir.2021.02.031
37. Nemati S, Holder A, Razmi F, Stanley MD, Clifford GD, Buchman TG. An Interpretable machine learning model for accurate prediction of sepsis in the ICU. *Crit Care Med.* (2018) 46:547–53. doi: 10.1097/CCM.0000000000002936
38. Li X, Xu X, Xie F, Xu X, Sun Y, Liu X, et al. A time-phased machine learning model for real-time prediction of sepsis in critical care. *Crit Care Med.* (2020) 48:e884–8. doi: 10.1097/CCM.0000000000004494
39. Uhel F, Peters-Sengers H, Falahi F, Scicluna BP, van Vught LA, Bonten MJ, et al. Mortality and host response aberrations associated with transient and persistent acute kidney injury in critically ill patients with sepsis: a prospective cohort study. *Intensive Care Med.* (2020) 46:1576–89. doi: 10.1007/s00134-020-06119-x

40. Ozrazgat-Baslanti T, Loftus TJ, Mohandas R, Wu Q, Brakenridge S, Brumback B, et al. Clinical trajectories of acute kidney injury in surgical sepsis: a prospective observational study. *Ann Surg.* [Epub ahead of print]. (2020). doi: 10.1097/SLA.0000000000004360

**Conflict of Interest:** The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

**Publisher's Note:** All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

frontiers | Frontiers in Medicine

ORIGINAL RESEARCH
published: 04 July 2022
doi: 10.3389/fmed.2022.880896

# Identifying Novel Clusters of Patients With Prolonged Mechanical Ventilation Using Trajectories of Rapid Shallow Breathing Index

*Tsung-Ming Yang[1,2,3†], Lin Chen[4,5†], Chieh-Mo Lin[1,6,7], Hui-Ling Lin[8,9], Tien-Pei Fang[3,9], Huiqing Ge[10], Huabo Cai[11], Yucai Hong[11] and Zhongheng Zhang[11*]*

[1] Division of Pulmonary and Critical Care Medicine, Chiayi Chang Gung Memorial Hospital, Chiayi, Taiwan, [2] School of Traditional Chinese Medicine, Chang Gung University, Taoyuan, Taiwan, [3] Department of Respiratory Care, Chang Gung University of Science and Technology, Chiayi, Taiwan, [4] Department of Critical Care Medicine, Affiliated Jinhua Hospital, Zhejiang University School of Medicine, Jinhua, China, [5] Key Laboratory of Emergency and Trauma, Ministry of Education, College of Emergency and Trauma, Hainan Medical University, Haikou, China, [6] College of Medicine, Graduate Institute of Clinical Medical Sciences, Chang Gung University, Taoyuan, Taiwan, [7] Department of Nursing, Chang Gung University of Science and Technology, Chiayi, Taiwan, [8] Department of Respiratory Therapy, Chang Gung University, Taoyuan, Taiwan, [9] Department of Respiratory Therapy, Chiayi Chang Gung Memorial Hospital, Chiayi, Taiwan, [10] Department of Respiratory Care, Sir Run Run Shaw Hospital, Zhejiang University School of Medicine, Hangzhou, China, [11] Key Laboratory of Precision Medicine in Diagnosis and Monitoring Research of Zhejiang Province, Department of Emergency Medicine, Sir Run Run Shaw Hospital, Zhejiang University School of Medicine, Hangzhou, China

OPEN ACCESS

**Edited by:**
*Eizo Watanabe,*
*Aichi Medical University, Japan*

**Reviewed by:**
*Narongkorn Saiphoklang,*
*Thammasat University, Thailand*
*Raffaele Campisi,*
*Azienda Ospedaliera Universitaria Policlinico G. Rodolico-San Marco, Italy*
*Hideto Yasuda,*
*Jichi Medical University Saitama Medical Center, Japan*

***Correspondence:**
*Zhongheng Zhang*
*zh_zhang1984@zju.edu.cn*

*[†] These authors have contributed equally to this work and share first authorship*

**Specialty section:**
*This article was submitted to Intensive Care Medicine and Anesthesiology, a section of the journal Frontiers in Medicine*

**Received:** *22 February 2022*
**Accepted:** *08 April 2022*
**Published:** *04 July 2022*

**Citation:**
*Yang T-M, Chen L, Lin C-M, Lin H-L, Fang T-P, Ge H, Cai H, Hong Y and Zhang Z (2022) Identifying Novel Clusters of Patients With Prolonged Mechanical Ventilation Using Trajectories of Rapid Shallow Breathing Index. Front. Med. 9:880896. doi: 10.3389/fmed.2022.880896*

**Objective:** Patients with prolonged mechanical ventilation (PMV) are comprised of a heterogeneous population, creating great challenges for clinical management and study design. The study aimed to identify subclusters of PMV patients based on trajectories of rapid shallow breathing index (RSBI), and to develop a machine learning model to predict the cluster membership based on baseline variables.

**Methods:** This was a retrospective cohort study conducted in respiratory care center (RCC) at a tertiary academic medical center. The RCC referral criteria were patients with mechanical ventilation for at least 21 days with stable hemodynamic and oxygenation status. Patients admitted to the RCC from April 2009 to December 2020 were screened. Two-step clustering through linear regression modeling and k-means was employed to find clusters of the trajectories of RSBI. The number of clusters was chosen by statistical metrics and domain expertise. A gradient boosting machine (GBM) was trained, exploiting variables on RCC admission, to predict cluster membership.

**Results:** A total of 1371 subjects were included in the study. Four clusters were identified: cluster A showed persistently high RSBI; cluster B was characterized by a constant low RSBI over time; Cluster C was characterized by increasing RSBI; and cluster D showed a declining RSBI. Cluster A showed the highest mortality rate (72%), followed by cluster D (63%), C (62%) and B (61%; $p = 0.005$ for comparison between 4 clusters). GBM was able to predict cluster membership with an accuracy of > 0.95 in ten-fold cross validation. Highly ranked variables for the prediction of clusters included thyroid-stimulating hormone (TSH), cortisol, platelet, free thyroxine (T4) and serum magnesium.

**Conclusions:** Patients with PMV are composed of a heterogeneous population that can be classified into four clusters by using trajectories of RSBI. These clusters can be easily predicted with baseline clinical variables.

**Keywords: prolonged mechanical ventilation, rapid shallow breathing index, gradient boosting machine, mortality, ICU**

## BACKGROUND

Prolonged mechanical ventilation (PMV) after critical illness has long been noticed as an emerging public health challenge. It is reported that patients with PMV have a 1-year mortality rate of 50–70% (1). This group of patients is typically characterized by old age, high comorbidity burden, high frailty score and increased likelihood of in-hospital complications (2). Great efforts have been made to improve the clinical outcomes of these patients. For example, many hospitals established specialized ventilator weaning unit such as respiratory care center (RCC) to manage these patients (3). In the literature, there have been many studies reporting the epidemiological characteristics of PMV patients, including risk factors for PMV, prediction of weaning probability, short and long-term mortality (4–6). The results are inconsistent across studies due to the heterogeneity of the PMV patients.

While PMV is well described in the literature, it has been noted that PMV patients are heterogeneous, comprising subclusters with distinct clinical characteristics and clinical outcomes. The heterogeneity creates great challenges for the clinical management and study designs. To the best of our knowledge, there has been no study to address the heterogeneity of PMV patients in the literature. Since MV liberation is the primary aim in the management of these patients, many studies have developed models and/or scores for the prediction of ventilator weaning (7–10). Rapid shallow breathing index (RSBI), defined as the ratio of respiratory frequency to tidal volume, is a canonical index to predict weaning success (11, 12). People on a ventilator who cannot tolerate independent breathing tend to breathe rapidly and shallowly and will therefore have a high RSBI. It is reasonable to characterized patients into subclusters based on longitudinal changes of RSBI. The present study aimed to explore the latent subclusters of PMV patients based on the trajectories of RSBI. A machine learning (ML) model based on variables collected upon RCC arrival was trained to predict cluster membership of PMV patients. Important variables associated with cluster assignment were explored in the ML model. We hypothesized that PMV patients could be well separated into several subtypes. The subtypes would have prognostic value for weaning and mortality outcomes. More importantly, these subtypes can be predicted early by using machine learning method trained on routinely collected variables.

**Abbreviations:** RCC, respiratory care center; RSBI, Rapid shallow breathing index; IMV, invasive mechanical ventilation; WBC, white blood cell count; RCC, respiratory care center; Q1, the first quartile; Q3, the third quartile; BUN, blood urea nitrogen; Cr, creatinine; RDW, red cell distribution width; MCV, mean corpuscular volume; NLR, neutrophil to lymphocyte ratio; SD, standard deviation; LIME, local interpretable model-agnostic explanations.

## METHODS

### Source of Data

This is a retrospective study conducted in the RCC of the Chang Gung Memorial Hospital from April 2009 to December 2020. All patients admitted to the RCC was screened for potential eligibility. The study was approved by the institutional review board (IRB) of the Chang Gung Memorial Hospital (Approval number: 202101862B0). The written informed consent was waived by the IRB because the study did not involve any interventions. Data were deidentified and stored in an encrypted computer. One patient with positive for HIV was excluded for confidential issues. The study was conducted according to the Helsinki declaration and was reported in accordance to the transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD) checklist (13).

### Participants

All patients admitted to the RCC was screened for potential eligibility. The indications for RCC admission must fulfill all the following criteria: (1) patients with mechanical ventilation for at least 21 days; (2) stable hemodynamic status (mean blood pressure $> 70$ mmHg with normal serum lactate) without vasopressors to maintain blood pressure; (3) stable oxygenation status with FiO2 $< 40\%$ and positive end expiratory pressure (PEEP) $< 10$ cm $H_2O$. Patients met one of the following criteria were excluded: (1) duplicated admissions to the RCC of the same patient; (2) patients who declined weaning attempts; (3) withdrawal of life support; (4) Transfer to other facility before weaning attempt started and (5) no spontaneous breathing.

### Patient Characteristics

Demographics, clinical and laboratory variables on RCC entry were extracted from the medical records. Demographic and clinical variables included age, sex, etiology of mechanical ventilation, hospital days upon RCC arrival, ventilation days upon RCC arrival, use of non-invasive ventilation (NIV) upon RCC arrival, Glasgow coma scale (GCS) upon RCC arrival, and comorbidities. Laboratory variables included blood gas, white blood cell count (WBC), hemoglobin (Hb), hematocrit (Hct), mean corpuscular volume (MCV), red cell distribution width (RDW), platelet, segment, lymphocyte, monocyte, eosinophil, basophil, neutrophil to lymphocyte ratio (NLR), blood urea nitrogen (BUN), creatinine (Cr), ionized calcium (Ca), phosphorus (P), magnesium (Mg), albumin, cortisol (AM), cortisol (PM), thyroid-stimulating hormone (TSH),

Free thyroxine (T4), pH, blood gas, dead-space fraction, and prealbumin were extracted.

Weaning indices were measured upon RCC arrival and then once a week as part of the routine practices to assess the patient's readiness for weaning, unless the patient was in respiratory distress requiring $FiO_2$ of 50% or higher, or in unstable hemodynamic status requiring vasopressor. Before measurement, the patient was disconnected from the mechanical ventilator. A handheld haloscale respirometer (Ferraris Medical, London, UK) was attached to the endotracheal tube to measure the minute ventilation (L/min). The average tidal volume (ml) was obtained by dividing the minute ventilation by the respiratory rate. Rapid shallow breaths index (RSBI) was calculated by dividing the respiratory rate (breaths/min) by average tidal volume in liter. Maximal negative inspiratory pressure (Pimax) was measured by inspiratory force meter (Boehringer Laboratories, Norristown, PA) when the patient was instructed to inhale forcefully and maximally. Finally, we obtained ventilatory parameters including tidal volume, respiratory rate, minute ventilation, maximal negative inspiratory pressure, and RSBI (14).

## Outcome Measurements

The following clinical outcomes were recorded for the study: long-term mortality outcome followed until December 2021, successful weaning from mechanical ventilation on RCC discharge, post-weaning respiratory failure after RCC discharge, days of duration from RCC discharge to respiratory failure, post-weaning respiratory failure before hospital discharge, days of duration from RCC discharge to respiratory failure in hospital, non-invasive mechanical ventilation (NIV) for post-weaning respiratory failure, invasive mechanical ventilation (IMV) for post-weaning respiratory failure, hospital length of stay, weaning and mortality outcome on hospital discharge, and long-term outcome at most recent follow up.

## Two-Step Clustering Through Linear Regression Modeling and K-Means

Two-step clustering through linear regression modeling and k-means was employed to identified clusters of the RSBI trajectories. Each trajectory was represented by the coefficients of an individually fitted linear regression model. The trajectories are then clustered based on the coefficients using k-means clustering (15, 16). The best number of clusters was determined by multiple metrics including log likelihood value, Bayesian information criterion (BIC), and Akaike's information criterion (AIC). We also considered to merge the cluster with fewer than 20 subjects. The trajectories of weaning indices were visualized for each latent cluster.

## Statistical Analysis

Baseline characteristics and laboratory variables were compared across the identified latent clusters. Categorical variables were reported as number (percentage) and were compared across latent clusters with $\chi^2$ test. Numeric variables were firstly tested for normality distribution and then compared across latent clusters using analysis of variance or Kruskal-Wallis rank sum test as appropriate (17). A $P < 0.05$ was considered as statistical significance.

## Model Development and Cross Validation

To predict RSBI trajectory clusters on RCC admission, we trained a GBM to predict cluster membership. Since the response was multiclass variable, cross entropy was employed as the loss function. The metric accuracy was used to evaluate the model performance in ten-fold cross validation procedure. GBMs build an ensemble of shallow and weak successive trees with each tree learning and improving on the previous (18, 19). The advantage of GBM includes its flexibility in allowing optimization on different loss functions and providing several hyperparameter tuning options that make the function fit very flexible. No data pre-processing is required that GBM often works great with categorical and numerical values as is. The hyperparameters in our GBM include the number of trees (from 1 to 15 at step 1), learning rate (0.1), and the interaction depth (depth of trees: 10, 15, 20, and 25). The minimum number of observations in terminal nodes was set to 30. A grid search strategy was employed to tune the hyperparameters. The accuracy was used in the 10-fold cross validation process for the hyperparameter tuning.

To understand the potential association between risk factors and latent clusters, we reported model specific variable importance for the GBM model. Variable importance is determined by calculating the relative influence of each variable: whether that variable was selected to split on during the tree building process, and how much the squared error (over all trees) improved (decreased) as a result. A greater value of variable importance indicates its higher association with latent clusters. Model interpretation was also performed by using local interpretable model-agnostic explanations (LIME) and iBreakdown algorithms (20). The intuition behind LIME is to learn the behavior of the underlying model (model-agnostic) by perturbing the predictors to see how the predictions change (21, 22). However, the explanation in LIME is additive while some complex relationships between predictors and clusters are non-additive. To address this limitation of LIME, we employed iBreakdown algorithm to detect interactions for instance-level explanations (23). All statistical analyses were performed with R (version 4.1.1).

# RESULTS

## Participants

A total number of 1,720 RCC admissions were screened from April 2009 to December 2020. 349 admissions were excluded due to reasons such as duplicated RCC admission, decline weaning attempt, transfer to other facility before weaning attempt started, no spontaneous breathing, withdrawal of life support and missing data on ventilator parameters (**Figure 1**). A number of 1,371 RCC

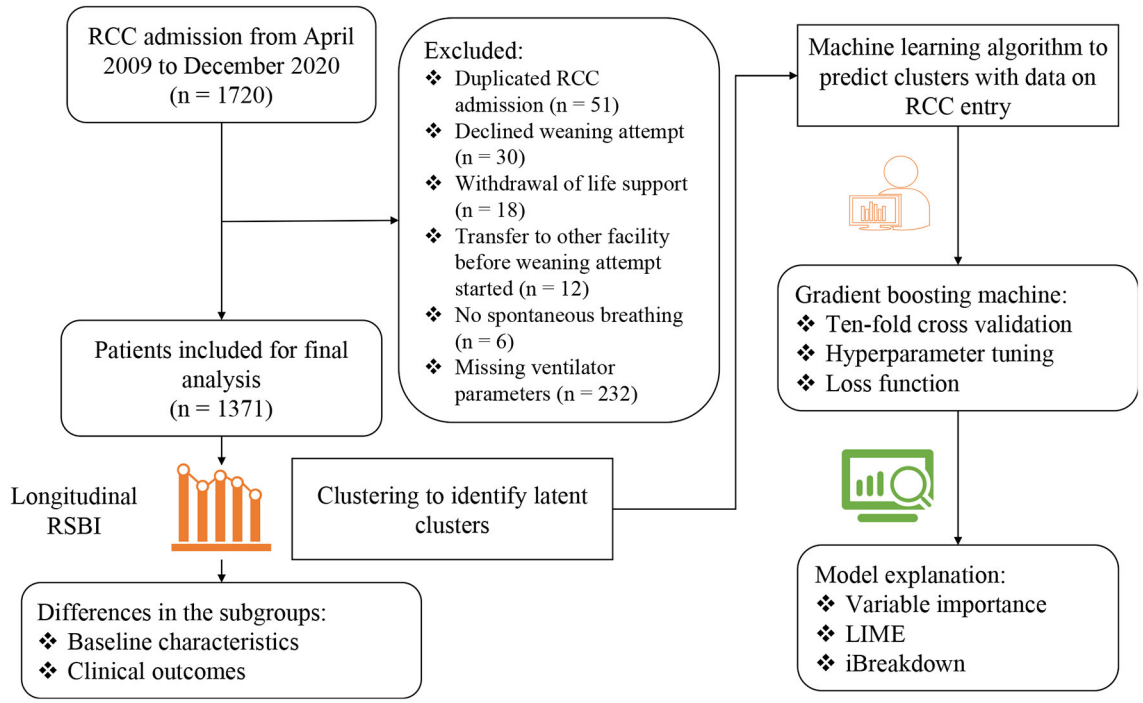



**FIGURE 1 |** Flowchart of patient enrollment and schematic illustration of the analysis workflow. RSBI, Rapid shallow breathing index; RCC, respiratory care center; LIME, local interpretable model-agnostic explanations.

admissions were included for the analysis. The median age of the study population was 76 (65–83) years. The median Charlson comorbidity index was 4 (3–7) and the most commonly reason for MV was acute lung injury (37%). The median follow-up days after RCC arrival was 105 (42–512) days (**Table 1**).

## Clusters of RSBI Trajectory

The 4-cluster model was considered as the best model it showed low BIC and AIC values, and high Log likelihood value (**Figure 2A**). Cluster B accounted for the largest proportion of patients and showed a constantly low RSBI during RCC stay. Cluster C was characterized by increasing RSBI (**Figures 2B,C**).

The clinical characteristics were compared across the clusters. Cluster B showed the highest proportion of male, while cluster D showed the lowest proportion of male patients (63% vs. 51%; p = 0.008). The APACHE II upon RCC arrival was the highest in cluster A and was the lowest in cluster D [median [Q1, Q3]: 24 (20, 28) vs. 23 (19, 26); $p = 0.021$, **Table 1**]. Interestingly, patients in cluster C showed lower plasma magnesium on RCC entry than that in cluster A (1.87 (1.63, 2.17) vs. 1.99 (1.72, 2.27) mg/dl; $p = 0.007$). The serum cortisone level on RCC entry was also associated with subsequent trajectory clusters (**Table 2**).

There were significant differences in clinical outcomes between the four clusters (**Table 1**). For the mortality outcome, cluster B showed the lowest mortality rate and cluster A showed the highest mortality (72 vs. 61%; $p = 0.005$). The weaning probability was highest in cluster B and the lowest in cluster A on hospital discharge (52 vs. 40%; $p = 0.007$). However, there was no significant difference on respiratory failure rate across clusters after successful weaning (p = 0.231).

## Predicting Trajectory Clusters on RCC Admission

The model hyperparameters of the GBM model were chosen by grid search to achieve the highest accuracy (> 0.95; **Figure 3A**). The top variables that are predictive of trajectory clusters included age, serum cortisol, BUN, platelet, and serum magnesium upon RCC arrival (**Figure 3B**). Four representative samples (sample ID = 1, 2, 4 and 5) were explored by LIME algorithm, which showed variables supporting or contradicting the assignment to a specific cluster (**Figures 3C,D**). The result indicated that TSH, cortisol, platelet, free T4 and serum magnesium were important predictors of clusters in many instances.

We also trained random forest (RF) and LASSO regression models, against which the GBM model was compared. The results showed that the GBM model outperformed LASSO and RF models with resampling method (**Figure 4**).

# DISCUSSION

The study for the first time explored the latent trajectories of patients with PMV (IMV duration > 21 days with stable hemodynamic and respiratory conditions) using RSBI. Four clusters were identified for the study population, namely, cluster

**TABLE 1 |** Baseline characteristics in the total population and across clusters.

| Variables | Total ($n$ = 1,371) | A ($n$ = 349) | B ($n$ = 461) | C ($n$ = 323) | D ($n$ = 238) | $p$ |
|---|---|---|---|---|---|---|
| Gender, Male (%) | 799 (58) | 178 (51) | 289 (63) | 194 (60) | 138 (58) | 0.008 |
| Age (years), Median (Q1,Q3) | 75.99 (64.89, 82.55) | 76.84 (66.39, 82.61) | 75.1 (65.08, 82.83) | 75.71 (62.3, 82.57) | 76.36 (64.41, 82.04) | 0.403 |
| APACHE II upon RCC arrival, Median (Q1,Q3) | 23 (20, 27) | 24 (20, 28) | 23 (20, 28) | 23 (19.5, 27) | 23 (19, 26) | 0.021 |
| Tracheostomy, n (%) | 371 (27) | 94 (27) | 127 (28) | 92 (28) | 58 (24) | 0.738 |
| Pre-Albumin (mg/dl, RCC Day 1), Median (Q1,Q3) | 16.3 (11.4, 21.4) | 16.21 (11.12, 20.65) | 16.9 (11.9, 22) | 16.2 (11.4, 22.15) | 15 (11.52, 21.78) | 0.292 |
| Charlson comorbidity index, Median (Q1,Q3) | 4 (3, 7) | 5 (3, 7) | 4 (3, 7) | 4 (3, 7) | 4 (3, 6) | 0.251 |
| GCS upon RCC arrival, Median (Q1,Q3) | 9 (7, 11) | 9 (7, 11) | 9 (7, 11) | 10 (8, 11) | 10 (9, 11) | 0.085 |
| **Etiology of mechanical ventilation, n (%)** | | | | | | 0.107 |
| Acute lung injury | 505 (37) | 120 (34) | 161 (35) | 140 (43) | 84 (35) | |
| Neurologic disease | 331 (24) | 69 (20) | 120 (26) | 76 (24) | 66 (28) | |
| Miscellaneous | 210 (15) | 60 (17) | 69 (15) | 47 (15) | 34 (14) | |
| Cardiac disease | 156 (11) | 51 (15) | 56 (12) | 26 (8) | 23 (10) | |
| Post-thoracic or abdominal surgery | 100 (7) | 27 (8) | 30 (7) | 22 (7) | 21 (9) | |
| Chronic lung injury | 69 (5) | 22 (6) | 25 (5) | 12 (4) | 10 (4) | |
| Equivalent hydrocortisone steroid dose (mg), Median (Q1,Q3) | 60 (40, 100) | 60 (40, 80) | 60 (40, 100) | 80 (40, 100) | 60 (40, 100) | 0.572 |
| Hospital days upon RCC arrival, Median (Q1,Q3) | 24 (21, 33) | 24 (21, 34) | 24 (21, 33) | 25 (21, 34) | 23 (20, 31) | 0.162 |
| Ventilation days upon RCC arrival, Median (Q1,Q3) | 21 (20, 25) | 21 (20, 25) | 21 (20, 26) | 22 (20, 25) | 21 (20, 24) | 0.191 |
| Ventialtor days upon extubation, Median (Q1,Q3) | 38 (32, 47) | 38 (32, 49) | 39 (34, 47) | 39 (32, 49) | 35 (31, 42) | < 0.001 |
| Post-weaning respiratory failure after RCC discharge, n (%) | 456 (33) | 105 (30) | 164 (36) | 97 (30) | 90 (38) | < 0.001 |
| Follow up days after RCC arrival, Median (Q1,Q3) | 105 (42, 512) | 119 (45, 513) | 111 (44, 524) | 96 (40, 428) | 84 (36, 612.5) | 0.542 |
| **Last follow up condition, n (%)** | | | | | | 0.005 |
| Dead | 885 (65) | 250 (72) | 283 (61) | 201 (62) | 151 (63) | |
| No ventilator | 451 (33) | 90 (26) | 168 (36) | 108 (33) | 85 (36) | |
| On ventilator | 35 (3) | 9 (3) | 10 (2) | 14 (4) | 2 (1) | |
| In-hospital mortality, n (%) | 363 (26) | 86 (25) | 125 (27) | 84 (26) | 68 (29) | 0.735 |
| Hospital length of stay, Median (Q1,Q3) | 65 (53, 82) | 65 (54, 83) | 65 (55, 81) | 65 (56, 86) | 61 (49, 76.75) | 0.019 |
| Weaning from MV in hospital or RCC, n (%) | 654 (48) | 141 (40) | 239 (52) | 151 (47) | 123 (52) | 0.007 |
| IMV for post-weaning respiratory failure, n (%) | 283 (21) | 67 (19) | 99 (21) | 57 (18) | 60 (25) | 0.231 |

*IMV, invasive mechanical ventilation; Q1, the first quartile; Q3, the third quartile; RCC, respiratory care center; GCS, Glasgow coma scale; APACHE II, The Acute Physiology and Chronic Health Evaluation II.*

A, B, C and D. Cluster B was characterized by a constant low RSBI over time; Cluster C was characterized by increasing RSBI; cluster D showed a declining RSBI, and cluster A showed persistently high RSBI. Many variables on RCC entry were associated with cluster membership including TSH, cortisol, platelet, free T4 and serum magnesium. These variables were also confirmed to be top ranked variables in GBM to classify trajectory clusters. It is feasible to predict the trajectories of RSBI upon RCC arrival using machine learning methods. Further external validation of the GBM is mandatory before this model can be used in clinical practice.

The identification of clusters for PMV patients has several implications. First, the heterogeneity of the population is addressed by classifying patients into clinically meaningful subgroups. These subgroups showed distinct clinical characteristics and outcomes, which is helpful for risk

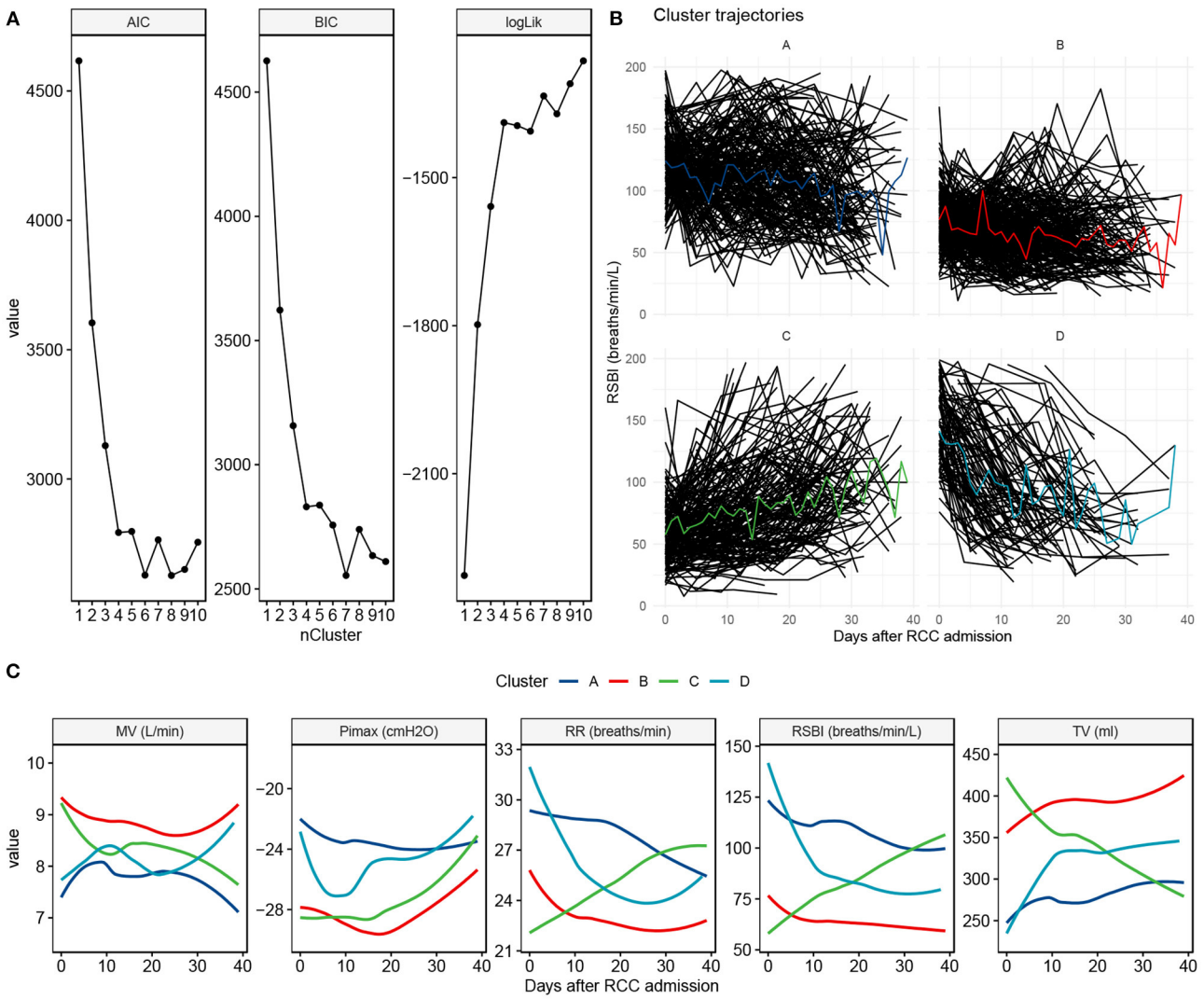



**FIGURE 2 |** Clustering to identify clusters of patients with prolonged mechanical ventilation. **(A)** The best number of clusters was chosen by using statistical metrics. Greater values of log likelihood indicate better model fit, whereas lower values of BIC and AIC indicate better model fit. **(B)** Trajectory characteristics of each cluster. Individual trajectories are represented by black lines and the cluster trajectory is colored. The cluster label and percentage are shown on the top of each panel. **(C)** Trajectory and 90% confidence interval for each of the ventilator parameters, stratified by the cluster membership. BIC, Bayesian information criterion; AIC, Akaike's information criterion; MV, minute ventilation; Pimax, maximum inspiratory pressure; RR, respiratory rate; RSBI, Rapid shallow breathing index; TV, tidal volume.

stratification and clinical decision making (24). Cluster A showed the lowest survival probability as compared to other clusters. Since it is feasible to predict patient trajectory on RCC admission, such early risk stratification can help resource allocation and family consultation. Second, individualized treatment strategy can be implemented for different subgroups. For example, we observed that low serum magnesium was associated with increased risk of cluster C trajectory with worsening RBSI during RCC treatment. This unfavorable outcome might be addressed by supplementing magnesium for this group of patients. Third, the identification of subtypes of patients can help to design clinical trials. Some interventions may have beneficial effects in a subgroup of patients, and trials investigating such interventions should target this subgroup. Such implementation of trial design has been explored in sepsis, showing that the probability of obtaining statistically significant beneficial/harmful effects vary by the proportion of subtypes (25).

The associations of several variables with cluster membership are supported by the literature. Serum magnesium has long been noticed to be associated with prolonged mechanical ventilation (26, 27). Hypomagnesemia is common in mechanically ventilated patients, and there is strong, consistent observational evidence that hypomagnesemia is significantly associated with increased need for prolonged mechanical ventilation and

**TABLE 2 |** Laboratory findings on RCC entry.

| Variables | Total ($n$ = 1,371) | A ($n$ = 349) | B ($n$ = 461) | C ($n$ = 323) | D ($n$ = 238) | $p$ |
|---|---|---|---|---|---|---|
| WBC ($\times 10^9/L$), median (Q1,Q3) | 9.3 (7.05, 12.4) | 9.2 (7.1, 12.1) | 9.4 (7, 12.5) | 9.5 (7.1, 12.55) | 9.35 (7.12, 12.2) | 0.943 |
| Hb (mg/dl), median (Q1,Q3) | 9.7 (8.9, 10.5) | 9.6 (8.9, 10.3) | 9.7 (9, 10.6) | 9.7 (8.9, 10.45) | 9.75 (8.9, 10.7) | 0.396 |
| Hct, median (Q1,Q3) | 0.3 (0.28, 0.32) | 0.3 (0.28, 0.32) | 0.3 (0.28, 0.33) | 0.3 (0.27, 0.32) | 0.3 (0.28, 0.33) | 0.349 |
| MCV, median (Q1,Q3) | 90.6 (87, 94.3) | 90.9 (87.1, 95) | 90.5 (87.1, 93.9) | 90.5 (86.65, 93.7) | 90.5 (87.23, 94.27) | 0.339 |
| RDW, median (Q1,Q3) | 0.16 (0.15, 0.18) | 0.16 (0.15, 0.18) | 0.16 (0.15, 0.18) | 0.16 (0.15, 0.18) | 0.16 (0.15, 0.17) | 0.024 |
| Platelet ($\times 10^9/L$), median (Q1,Q3) | 218 (147, 307.5) | 210 (138, 299) | 224 (154, 314) | 213 (142, 295) | 223.5 (160.5, 316.5) | 0.057 |
| Segment ($\times 10^9/L$), median (Q1,Q3) | 0.79 (0.72, 0.86) | 0.79 (0.71, 0.86) | 0.79 (0.72, 0.85) | 0.8 (0.73, 0.86) | 0.79 (0.73, 0.86) | 0.9 |
| Lymohocyte ($\times 10^9/L$), median (Q1,Q3) | 0.1 (0.06, 0.16) | 0.11 (0.06, 0.15) | 0.1 (0.06, 0.16) | 0.1 (0.07, 0.15) | 0.11 (0.06, 0.16) | 0.946 |
| Monocyte ($\times 10^9/L$), median (Q1,Q3) | 0.06 (0.04, 0.08) | 0.06 (0.04, 0.08) | 0.06 (0.04, 0.08) | 0.06 (0.04, 0.08) | 0.06 (0.04, 0.08) | 0.893 |
| Eosinophil ($\times 10^9/L$), median (Q1,Q3) | 0.01 (0, 0.03) | 0.01 (0, 0.03) | 0.01 (0, 0.03) | 0.01 (0, 0.03) | 0.01 (0, 0.03) | 0.753 |
| Basophil ($\times 10^9/L$), median (Q1,Q3) | 0 (0, 0) | 0 (0, 0) | 0 (0, 0) | 0 (0, 0) | 0 (0, 0) | 0.571 |
| NLR, median (Q1,Q3) | 7.5 (4.66, 13) | 7.41 (4.79, 13.23) | 7.55 (4.62, 12.43) | 7.8 (4.66, 12.91) | 7.45 (4.52, 13.42) | 0.936 |
| BUN (mg/dl), median (Q1,Q3) | 27.8 (16.3, 54) | 32.1 (17.5, 59.4) | 27.7 (15.9, 54.9) | 26.2 (16.55, 50.55) | 25.3 (15.93, 49.3) | 0.174 |
| Cr (mg/dl), median (Q1,Q3) | 0.75 (0.48, 1.71) | 0.78 (0.47, 1.69) | 0.75 (0.49, 1.93) | 0.73 (0.46, 1.79) | 0.74 (0.47, 1.46) | 0.652 |
| Ca (mg/dl), median (Q1,Q3) | 8.2 (7.9, 8.7) | 8.3 (7.9, 8.8) | 8.3 (7.9, 8.7) | 8.2 (7.8, 8.6) | 8.2 (7.9, 8.6) | 0.23 |
| P (mg/dl), median (Q1,Q3) | 3.5 (2.9, 4.2) | 3.6 (2.9, 4.2) | 3.5 (2.9, 4.2) | 3.5 (2.9, 4.4) | 3.4 (2.8, 4.1) | 0.59 |
| Mg (mg/dl), median (Q1,Q3) | 1.91 (1.68, 2.2) | 1.99 (1.72, 2.27) | 1.88 (1.67, 2.18) | 1.87 (1.63, 2.17) | 1.92 (1.72, 2.17) | 0.007 |
| Albumin (mg/dl), median (Q1,Q3) | 2.5 (2, 2.9) | 2.4 (2, 2.8) | 2.5 (2.1, 2.9) | 2.5 (2.02, 2.9) | 2.5 (2, 2.8) | 0.089 |
| Cortisol (mcg/dl, AM), median (Q1,Q3) | 14.32 (10.39, 18.15) | 14.51 (10.62, 19.12) | 14.97 (10.89, 18.2) | 13.91 (10.66, 17.91) | 13.04 (9.33, 16.86) | 0.013 |
| Cortisol (mcg/dl, PM), median (Q1,Q3) | 15.07 (10.6, 20.03) | 15.45 (10.52, 20.52) | 14.78 (10.72, 20.01) | 15.46 (11.31, 20.03) | 14.5 (10.19, 18.98) | 0.313 |
| TSH (mIU/L), median (Q1,Q3) | 2.19 (1.18, 4.24) | 2.51 (1.24, 4.46) | 2.05 (1.09, 4.32) | 2.11 (1.17, 4.18) | 2.13 (1.19, 3.9) | 0.26 |
| Free T4 (Free T4), median (Q1,Q3) | 0.97 (0.8, 1.16) | 0.95 (0.79, 1.13) | 0.98 (0.8, 1.16) | 0.98 (0.82, 1.17) | 0.99 (0.8, 1.14) | 0.737 |
| pH (Upon RCC arrival), median (Q1,Q3) | 7.49 (7.46, 7.52) | 7.49 (7.45, 7.51) | 7.49 (7.46, 7.52) | 7.49 (7.46, 7.52) | 7.49 (7.46, 7.52) | 0.178 |
| PaCO2 (mmHg, Upon RCC arrival), median (Q1,Q3) | 38 (32.92, 43.18) | 38.45 (33.7, 44.42) | 37.5 (32.4, 42.6) | 38 (32.75, 42.8) | 37.8 (33.12, 43.1) | 0.072 |
| PaO2 (mmHg, Upon RCC arrival), median (Q1,Q3) | 101.8 (84.53, 121.92) | 101 (85.6, 120.12) | 101.3 (83.1, 123.5) | 103.8 (86.45, 124.45) | 102 (87.82, 119.65) | 0.607 |
| HCO3 (mmol/L, Upon RCC arrival), median (Q1,Q3) | 29.3 (25.4, 32.9) | 29.7 (25.5, 33.6) | 29 (25.4, 32.6) | 29.1 (25.35, 32.65) | 29.65 (25.92, 33.2) | 0.263 |
| SaO2 (Upon RCC arrival), median (Q1,Q3) | 0.98 (0.97, 0.99) | 0.98 (0.97, 0.99) | 0.98 (0.97, 0.99) | 0.98 (0.97, 0.99) | 0.98 (0.97, 0.99) | 0.64 |
| FiO2 (Upon RCC arrival), Median (Q1,Q3) | 0.35 (0.3, 0.35) | 0.35 (0.35, 0.35) | 0.35 (0.3, 0.35) | 0.35 (0.3, 0.35) | 0.35 (0.35, 0.35) | 0.268 |
| End-tidal CO2 (mmHg, Upon RCC arrival), median (Q1,Q3) | 34 (30, 38) | 33 (31, 41) | 34 (31, 38) | 33.5 (29, 37) | 34.5 (30.75, 38) | 0.578 |
| Dead space fraction (Upon RCC arrival), mean ± SD | 0.08 ± 0.18 | 0.1 ± 0.17 | 0.08 ± 0.19 | 0.08 ± 0.17 | 0.08 ± 0.18 | 0.858 |
| Pre-Alb (mg/dl, RCC Day 1), median (Q1,Q3) | 16.3 (11.4, 21.4) | 16.21 (11.12, 20.65) | 16.9 (11.9, 22) | 16.2 (11.4, 22.15) | 15 (11.52, 21.78) | 0.292 |
| Pre-Alb (mg/dl, RCC Day 14), median (Q1,Q3) | 17.85 (13.1, 23.7) | 17.9 (13.5, 23.85) | 17.75 (13.03, 23.62) | 16.65 (12.23, 22.2) | 20 (15.5, 25.2) | 0.022 |

*WBC, white blood cell count; Hb, hemoglobin; Hct, hematocrit; RCC, respiratory care center; Q1, the first quartile; Q3, the third quartile; BUN, blood urea nitrogen; Cr, creatinine; RDW, red distribution width; MCV, mean corpuscular volume; NLR, neutrophil to lymphocyte ratio; P, phosphorus; Mg, magnesium; TSH, thyroid-stimulating hormone; T4, thyroxine; PaCO2, arterial partial pressure of carbon dioxide; PaO2, arterial partial pressure of oxygen; HCO3, Bicarbonate; SaO2, arterial oxygen saturation; FiO2, inspired oxygen fraction; SD, standard deviation.*

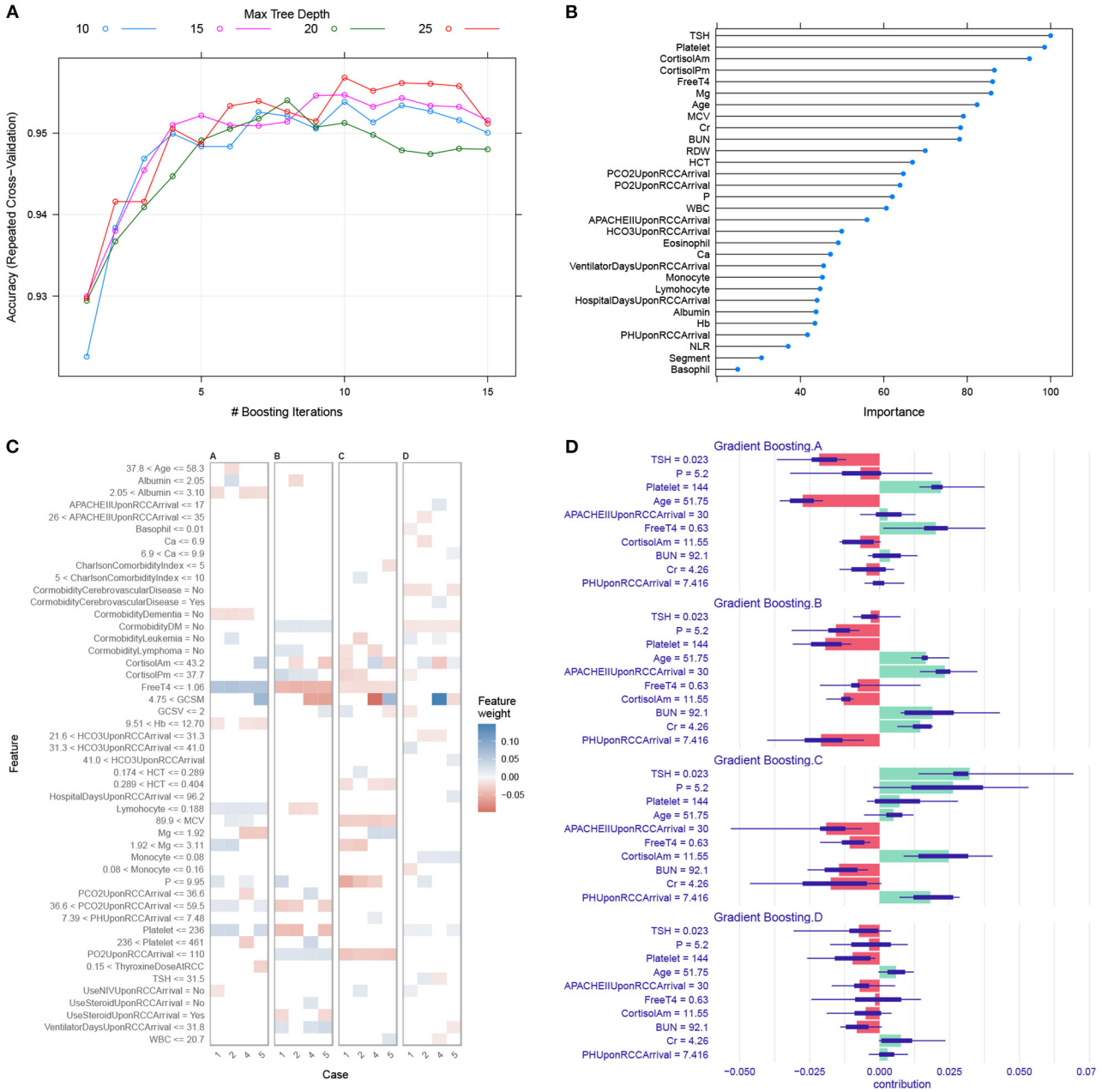



**FIGURE 3 |** Gradient boosting machine training and interpretation. **(A)** Hyperparameter tuning for the gradient boosting machine model. We used grid search strategy to select hyperparameters with the highest accuracy. **(B)** Variable importance in the GBM model. Higher importance value indicates greater influence of the variable in differentiating the clusters. **(C)** LIME interpretation for four sample subjects. The horizontal axis is labeled by the sample ID. The observed cluster membership for patients 1, 2, 4 and 5 were A, B, C and D respectively. Blue (red) color indicates the variable is supporting for (contradicting against) a given cluster. For example, the subject 4 has magnesium $< 1.92$ supporting for cluster C. **(D)** The iBreakdown explainer for patient #4 showed that there was more support for allocation to cluster C than to other clusters. The feature TSH $= 0.023$ strongly supports its assignment to cluster C, whereas the APACHE II $= 30$ on RCC arrival contradicts its assignment to cluster C. The short bar indicates the confidence interval for uncertainty. LIME, local interpretable model-agnostic explanations; HCT, hematocrit; WBC, white blood cell count; BUN, blood urea nitrogen; Cr, creatinine; RDW, red distribution width; MCV, mean corpuscular volume; GCS, Glasgow coma scale; GCSM, motion component of GCS.

increased mortality (28). The causality of hypomagnesemia and PMV has not been firmly established in the critical care literature. In a randomized controlled trial involving liver transplantation, Gucyetmez B and colleagues reported that intravenous magnesium sulfate administration was associated with shortened duration of mechanical ventilation

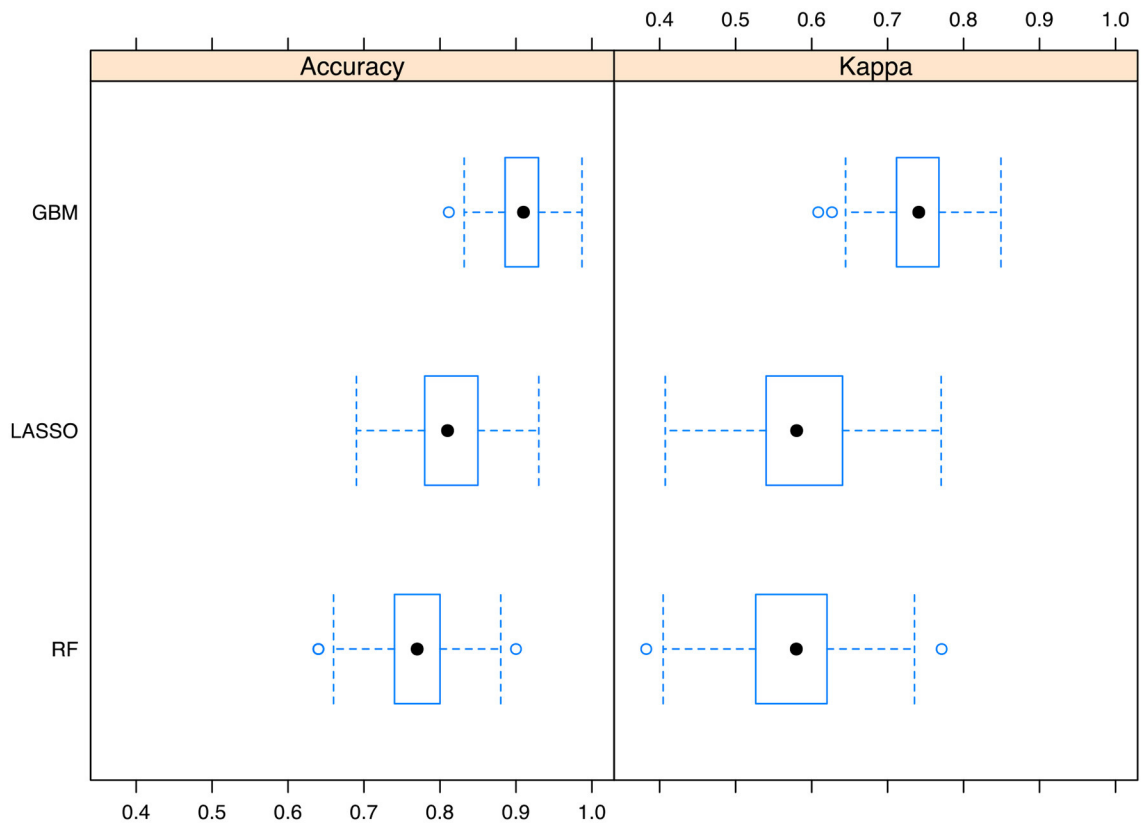



**FIGURE 4 |** Comparisons of the Gradient boosting machine with other models. The performance metrics of accuracy and kappa was reported. The boxplot shows the median and range of the performance metrics in resampled datasets. GBM, gradient boosting machine; LASSO, Least Absolute Shrinkage and Selection Operator; RF, random forest.

(29). However, the association of magnesium and trajectory clusters in RCC has not been explored and this is a novelty in our study. Although our latent clusters were identified by using longitudinal RSBI, changes of other ventilator parameters also have important clinical implications. For example, the cluster A shows constant RR with slightly increasing Pimax (i.e., less negative value indicates less inspiratory efforts) over RCC treatment, indicating less demand of ventilation with recovered critical illness. It is reasonable to deduce that oxygen consumption will decline after resolution of critical illness, which is reflected by reduced minute ventilation. Collectively, these changes in ventilator parameters indicate recovered overall condition and improved lung function.

Several limitations must be acknowledged in the study. First, the study was retrospective in design and there are many missing values in ventilator parameters. We had to exclude these patients due to missingness. It is largely unknown whether this exclusion will compromise the representativeness of our sample for the study population. Second, although we trained a GBM model for the prediction of subsequent trajectory clusters, the model was not validated in external dataset. We used 10-fold cross validation for training the model, but this cannot preclude the possibility of poor performance in other datasets. Third, the causality of baseline characteristic variables and cluster assignment cannot be fully confirmed in the present study design due to potential unmeasured confounding factors. Further randomized controlled trials are mandatory to confirm potential causal associations. P1.0 is another important parameter to predict weaning failure. It was not included in the analysis because this variable was not routinely measured. Finally, RSBI was the primary index used for trajectory clustering, which had its inherent strengths and limitations. RSBI is widely used to predict the weaning success and its measurement is easy at bedside. However, RSBI can be affected by pressure augmentation, PEEP, and a bias flow (30, 31).

## CONCLUSIONS

The study identified four clusters of patients requiring PMV based on longitudinal RSBI. These clusters have distinct clinical characteristics and outcomes, which is implicative for the implementation of precise medicine for this study population. It is also feasible to predict cluster assignment with variables collected upon RCC arrival with machine learning algorithms.

## DATA AVAILABILITY STATEMENT

The raw data supporting the conclusions of this article will be made available by the authors, without undue reservation.

## ETHICS STATEMENT

The studies involving human participants were reviewed and approved by the Institutional Review Board (IRB) of the Chang Gung Memorial Hospital (Approval number: 202101862B0). Written informed consent for participation was not required for this study in accordance with the national legislation and the institutional requirements.

## AUTHOR CONTRIBUTIONS

T-MY and ZZ designed the study and drafted the manuscript. H-LL helped interpret the results and write some discussions. ZZ and LC performed statistical analysis and result interpretation. YH and HC prepared the figures and interpret the results. T-PF, C-ML, and HG provided critical review of the manuscript. T-MY, T-PF, and C-ML were responsible for patient enrolment and data entry. ZZ is identified as the guarantor of the paper and taking responsibility for the integrity of the work as a whole, from inception to published article. All authors contributed to the article and approved the submitted version.

## FUNDING

ZZ had received funding from Yilu Gexin–Fluid Therapy Research Fund Project (YLGX-ZZ-2020005) and the Health Science and Technology Plan of Zhejiang Province (2021KY745). LC had received funding from Zhejiang Provincial Project of Medical and Health Technology (2022PY099), Youth Foundation of Jinhua Municipal Central Hospital (JY2020-2-10), Key Project of Jinhua City (2021-3-037), Key Laboratory of Emergency and Trauma (Hainan Medical University), and Ministry of Education (KLET-202118). H-LL had received funding from Chang Gung Memorial Foundation (BMRPE83).

## REFERENCES

1. Damuth E, Mitchell JA, Bartock JL, Roberts BW, Trzeciak S. Long-term survival of critically ill patients treated with prolonged mechanical ventilation: a systematic review and meta-analysis. *Lancet Respir Med.* (2015) 3:544–53. doi: 10.1016/S2213-2600(15)00150-2
2. Dettmer MR, Damuth E, Zarbiv S, Mitchell JA, Bartock JL, Trzeciak S. Prognostic factors for long-term mortality in critically Ill patients treated with prolonged mechanical ventilation: a systematic review. *Crit Care Med.* (2017) 45:69–74. doi: 10.1097/CCM.0000000000002022
3. Cox CE, Carson SS. Medical and economic implications of prolonged mechanical ventilation and expedited post-acute care. *Semin Respir Crit Care Med.* (2012) 33:357–61. doi: 10.1055/s-0032-1321985
4. Sierros V, Fleming R, Cascioli M, Brady T. The prognostic value of C-reactive protein in long-term care patients requiring prolonged mechanical ventilation. *Chron Respir Dis.* (2009) 6:149–55. doi: 10.1177/1479972309104660
5. Hill AD, Fowler RA, Burns KEA, Rose L, Pinto RL, Scales DC. Long-term outcomes and health care utilization after prolonged mechanical ventilation. *Ann Am Thorac Soc.* (2017) 14:355–62. doi: 10.1513/AnnalsATS.201610-792OC
6. Unroe M, Kahn JM, Carson SS, et al. One-year trajectories of care and resource utilization for recipients of prolonged mechanical ventilation: a cohort study. *Ann Intern Med.* (2010) 153:167–75. doi: 10.7326/0003-4819-153-3-201008030-00007
7. Rittayamai N, Ratchaneewong N, Tanomsina P, Kongla W. Validation of rapid shallow breathing index displayed by the ventilator compared to the standard technique in patients with readiness for weaning. *BMC Pulm Med.* (2021) 21:310. doi: 10.1186/s12890-021-01680-7
8. Torrini F, Gendreau S, Morel J, et al. Prediction of extubation outcome in critically ill patients: a systematic review and meta-analysis. *Crit Care.* (2021) 25:391. doi: 10.1186/s13054-021-03802-3
9. Burns KEA, Lellouche F, Nisenbaum R, Lessard MR, Friedrich JO. Automated weaning and SBT systems versus non-automated weaning strategies for weaning time in invasively ventilated critically ill adults. *Cochrane Database Syst Rev.* (2014) (9):CD008638. doi: 10.1002/14651858.CD008638.pub2
10. Wu TJ, Shiao JSC Yu HL, Lai RS. An integrative index for predicting extubation outcomes after successful completion of a spontaneous breathing trial in an adult medical intensive care unit. *J Intensive Care Med.* (2019) 34:640–5. doi: 10.1177/0885066617706688
11. Karthika M, Al Enezi FA, Pillai LV, Arabi YM. Rapid shallow breathing index. *Ann Thorac Med.* (2016) 11:167–76. doi: 10.4103/1817-1737.176876
12. Yang KL, Tobin MJ, A. prospective study of indexes predicting the outcome of trials of weaning from mechanical ventilation. *N Engl J Med.* (1991) 324:1445–50. doi: 10.1056/NEJM199105233242101
13. Collins GS, Reitsma JB, Altman DG, Moons KGM. Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD): the TRIPOD Statement. *Eur J Clin Invest.* (2015) 45:204–14. doi: 10.1111/eci.12376
14. Huaringa AJ, Wang A, Haro MH, Leyva FJ. The weaning index as predictor of weaning success. *J Intensive Care Med.* (2013) 28:369–74. doi: 10.1177/0885066612463681
15. Den Teuling NGP, Pauws SC, van den Heuvel ER. A comparison of methods for clustering longitudinal data with slowly changing trends. *Commun Stat Simul Comput.* (2021) 1–28. doi: 10.1080/03610918.2020.1861464. [Epub ahead of print].
16. Teuling ND, Pauws S, Heuvel E. van den. Clustering of longitudinal data: a tutorial on a variety of approaches. *arXiv [Preprint].* (2021). doi: 10.48550/arXiv.2111.05469
17. Zhang Z, Gayle AA, Wang J, Zhang H, Cardinal-Fernández P. Comparing baseline characteristics between groups: an introduction to the CBCgrps package. *Ann Transl Med.* (2017) 5:484. doi: 10.21037/atm.2017.09.39
18. Friedman JH. Greedy function approximation: a gradient boosting machine. *Ann. Stat.* (2001) 29:1189–232. doi: 10.1214/aos/1013203451
19. Zhang Z, Zhao Y, Canes A, Steinberg D, Lyashevska O. Written on behalf of AME Big-Data Clinical Trial Collaborative Group. Predictive analytics with gradient boosting in clinical medicine. *Ann Transl Med.* (2019) 7:152. doi: 10.21037/atm.2019.03.29
20. Zhang Z, Chen L, Xu P, Hong Y. Predictive analytics with ensemble modeling in laparoscopic surgery: a technical note. *Surg Laparosc Endosc Percutan Tech.* (2022) 5:25–34. doi: 10.1016/j.lers.2021.12.003
21. Zhang Z, Beck MW, Winkler DA, et al. Opening the black box of neural networks: methods for interpreting neural network models in clinical applications. *Ann Transl Med.* (2018) 6:216. doi: 10.21037/atm.2018.05.32
22. Ribeiro MT, Singh S, Guestrin C. "Why should i trust you?": Explaining the predictions of any classifier. In: *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining.* San Francisco, CA; New York, NY: Association for Computing Machinery (2016). p. 1135–44. doi: 10.1145/2939672.2939778
23. Gosiewska A, Biecek P. Do not trust additive explanations. *arXiv [Preprint].* (2020). doi: 10.48550/arXiv.1903.11420

24. Zhang Z, Zhang G, Goyal H, Mo L, Hong Y. Identification of subclasses of sepsis that showed different clinical outcomes and responses to amount of fluid resuscitation: a latent profile analysis. *Crit Care.* (2018) 22:347. doi: 10.1186/s13054-018-2279-3
25. Seymour CW, Kennedy JN, Wang S, Chang C-CH, Elliott CF, Xu Z, et al. Derivation, validation, and potential treatment implications of novel clinical phenotypes for sepsis. *JAMA*. (2019) 321:2003–17. doi: 10.1001/jama.2019.5791
26. Jiang P, Lv Q, Lai T, Xu F. Does hypomagnesemia impact on the outcome of patients admitted to the intensive care unit? A systematic review and meta-analysis. *Shock.* (2017) 47:288–95. doi: 10.1097/SHK.0000000000000769
27. Limaye CS, Londhey VA, Nadkart MY, Borges NE. Hypomagnesemia in critically ill medical patients. *J Assoc Physicians India.* (2011) 59:19–22.
28. Velissaris D, Karamouzos V, Pierrakos C, Aretha D, Karanikolas M. Hypomagnesemia in critically ill sepsis patients. *J Clin Med Res.* (2015) 7:911–8. doi: 10.14740/jocmr2351w
29. Gucyetmez B, Atalan HK, Aslan S, Yazar S, Polat KY. Effects of intraoperative magnesium sulfate administration on postoperative tramadol requirement in liver transplantation: a prospective, double-blind study. *Transplant Proc.* (2016) 48:2742–6. doi: 10.1016/j.transproceed.2016.08.033
30. El-Khatib MF, Jamaleddine GW, Khoury AR, Obeid MY. Effect of continuous positive airway pressure on the rapid shallow breathing index in patients following cardiac surgery. *Chest.* (2002) 121:475–9. doi: 10.1378/chest.121.2.475
31. Bien MY, Shui Lin Y, Shih CH, et al. Comparisons of predictive performance of breathing pattern variability measured during T-piece, automatic tube compensation, and pressure support ventilation for weaning intensive care unit patients from mechanical ventilation. *Crit Care Med.* (2011) 39:2253–62. doi: 10.1097/CCM.0b013e31822279ed

**Conflict of Interest:** The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

**Publisher's Note:** All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

frontiers | Frontiers in Medicine

TYPE Original Research
PUBLISHED 22 July 2022
DOI 10.3389/fmed.2022.920040



OPEN ACCESS

EDITED BY
Zhongheng Zhang,
Sir Run Run Shaw Hospital, China

REVIEWED BY
Zhi Mao,
People's Liberation Army General Hospital, China
Xiaohua Qian,
Shanghai Jiao Tong University, China

*CORRESPONDENCE
Haiyan Yin
yinhaiyan1867@126.com
Jun Lyu
lyujun2020@jnu.edu.cn

†These authors have contributed equally to this work

SPECIALTY SECTION
This article was submitted to Intensive Care Medicine and Anesthesiology, a section of the journal Frontiers in Medicine

RECEIVED 14 April 2022
ACCEPTED 01 July 2022
PUBLISHED 22 July 2022

CITATION
Huang X, Li B, Huang T, Yuan S, Wu W, Yin H and Lyu J (2022) External validation based on transfer learning for diagnosing atelectasis using portable chest X-rays. *Front. Med.* 9:920040. doi: 10.3389/fmed.2022.920040



# External validation based on transfer learning for diagnosing atelectasis using portable chest X-rays

Xiaxuan Huang[1,2†], Baige Li[3†], Tao Huang[2], Shiqi Yuan[1,2], Wentao Wu[4], Haiyan Yin[5*] and Jun Lyu[2,6*]

[1]Department of Neurology, The First Affiliated Hospital of Jinan University, Guangzhou, China, [2]Department of Clinical Research, The First Affiliated Hospital of Jinan University, Guangzhou, China, [3]Department of Radiology, The First Affiliated Hospital of Jinan University, Guangzhou, China, [4]School of Public Health, Xi'an Jiaotong University Health Science Center, Xi'an, China, [5]Department of Intensive Care Unit, The First Affiliated Hospital of Jinan University, Guangzhou, China, [6]Guangdong Provincial Key Laboratory of Traditional Chinese Medicine Informatization, Guangzhou, China

**Background:** Although there has been a large amount of research focusing on medical image classification, few studies have focused specifically on the portable chest X-ray. To determine the feasibility of transfer learning method for detecting atelectasis with portable chest X-ray and its application to external validation, based on the analysis of a large dataset.

**Methods:** From the intensive care chest X-ray medical information market (MIMIC-CXR) database, 14 categories were obtained using natural language processing tags, among which 45,808 frontal chest radiographs were labeled as "atelectasis," and 75,455 chest radiographs labeled "no finding." A total of 60,000 images were extracted, including positive images labeled "atelectasis" and positive X-ray images labeled "no finding." The data were categorized into "normal" and "atelectasis," which were evenly distributed and randomly divided into three cohorts (training, validation, and testing) at a ratio of about 8:1:1. This retrospective study extracted 300 X-ray images labeled "atelectasis" and "normal" from patients in ICUs of The First Affiliated Hospital of Jinan University, which was labeled as an external dataset for verification in this experiment. Data set performance was evaluated using the area under the receiver operating characteristic curve (AUC), sensitivity, specificity, and positive predictive values derived from transfer learning training.

**Results:** It took 105 min and 6 s to train the internal training set. The AUC, sensitivity, specificity, and accuracy were 88.57, 75.10, 88.30, and 81.70%. Compared with the external validation set, the obtained AUC, sensitivity, specificity, and accuracy were 98.39, 70.70, 100, and 86.90%.

**Conclusion:** This study found that when detecting atelectasis, the model obtained by transfer training with sufficiently large data sets has excellent external verification and acculturate localization of lesions.

KEYWORDS

**atelectasis, transfer learning, ResNet, artificial intelligence (AI), ICUs**

# Introduction

Atelectasis, as the most common postoperative pulmonary complication (PPC) (1), is also the most common disease in intensive care units (ICUs), and is often accompanied by pneumothorax, pleural effusion, pulmonary edema, and other pulmonary diseases, and requires reintubation within 48 h after a complication. Portable chest radiography (2) is one of the most common non-invasive radiological tests for rapid and straightforward atelectasis detection in ICUs. The main direct signs (3) of atelectasis on chest radiographs include defect migration, parenchymal opacity with unbroken linear boundaries, and vascular displacement. Indirect signs (4) have ipsilateral diaphragmatic elevation, hilar removal, heart involvement, and mediastinum and trachea dysfunction; however, it is difficult to rapidly distinguish the characteristics of early chest radiographs of patients with atelectasis from pleural effusion and lung consolidation with increased density. With the rapid development of deep-learning technology, the convolutional neural network (5) extracts inherent characteristics from medical image data for classification and recognition based on images much more effectively than do traditional recognition algorithms. The problem of inaccurate diagnoses caused by the continuous increase in the number of chest X-rays that exceeds the increase in the number of radiologists has somewhat been solved. Because previous studies have developed chest X-ray diagnostic algorithms based on deep neural networks, 14 basic lung diseases can be diagnosed.

The present study was the first to extract large-scale positive atelectasis data from the MIMIC-CXR-JPG database (6–8), an intensive care medical information database, to obtain a model with high accuracy and obtain reliable external validation. The purpose of this study was to determine the feasibility of transfer learning methods for detecting atelectasis using portable chest X-rays based on large data sets. In addition, this study used multi-center data set for the first time to realize the early diagnosis and prediction of bedside portable chest radiographs and obtained good external validation.

# Materials and methods

## Data source

### Training and testing cohorts

In this study, 14 categories that were clearly diagnosed as "positive" or "negative" were obtained from the MIMIC-CXR-JPG database using the open-source tagging tools NegBio9 (9) and CheXpert10 (10) (473,057 chest radiographs and 206,563 text reports). There were 45,808 chest radiographs labeled "atelectasis" and 75,455 chest radiographs labeled "no finding." A total of 60,000 images were extracted, including positive images labeled "atelectasis" and positive X-ray images labeled "no finding," and blank text was excluded. The data were defined as "normal" and "atelectasis" with an even distribution. At the same time, the data set was classified and randomly divided into three sets (training, validation, and testing) at a ratio of about 8:1:1. The MIMIC-CXR-JPG common database was used as the internal testing data of this experiment.

### External validation cohort

An external testing data set was developed in this study, with primary image data from ICUs patients at The First Affiliated Hospital of Jinan University, from which data during 2017–2021 were randomly selected by three senior attending physicians for a definitive diagnosis of the patients with atelectasis, who found no apparent abnormalities in the chest radiological image data, and professional radiologists carried out a review of the random tag data set. After excluding the data of poor posture placement and unclear diagnoses, 300 images with complete labels were finally extracted: 150 with atelectasis and 150 without abnormalities. The flow-chart of the training data creation is shown in **Figure 1**.

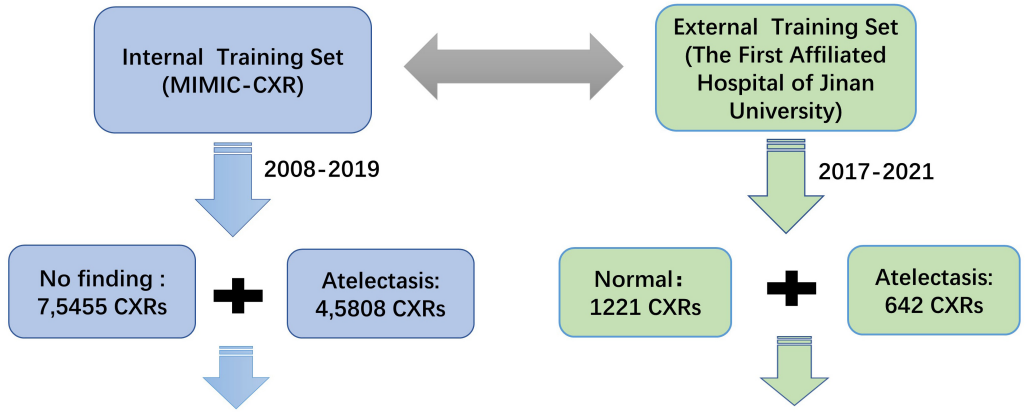
## Experimental environment

All experiments were performed using the Ubuntu 20.04 64-bit operating system. For the training process of the CNN model, MATLAB language was adopted as the programming environment. The specific software and hardware configuration are listed in **Table 1**.

## Data preprocessing

This study processed both internal and external testing cohorts. The frontal chest images from the MIMIC-CXR-JPG database had a resolution of 256 × 256 × 1 pixels, while those from The First Affiliated Hospital of Jinan University were 512 × 512 × 3 pixels. Since the image resolutions of the two data sets had different heights and widths, the image data from the two other sources had to be standardized. We achieved this by automatically cropping the chest region and adjusting it to 224 × 224 × 3 pixels in order to fit the model input resolution more conveniently. At the same time, the image was randomly scaled horizontally and vertically, and processing methods were cut and shifted to enhance the data processing (11) in order to further optimize the verification and evaluation of the testing results.

**Abbreviations:** MIMIC-CXR, medical information mart for intensive care chest X-ray; AUC, area under the curve; ROC, curves receiver operating characteristic curves; ICUs, intensive care units; AI, artificial intelligence.




| | Training | Validation | Testing |
|---|---|---|---|
| Normal | 1,9200 | 4800 | 6000 |
| Atelectasis | 1,9200 | 4800 | 6000 |

| | Training | Testing |
|---|---|---|
| Normal | 150 | 150 |
| Atelectasis | 150 | 150 |

FIGURE 1
Data flow design of the internal and external datasets.

## Model

The selected study model was the ResNet50 (12) network model, which is based on the residual error learning method from the VGG19 improved classification model. Based on the existing training network depth, a more-optimized residual learning framework was put forward, not only to solve the problem of the gradient disappearance and explosion (13), but also to further deepen the network depth. The problem of network performance degradation is also avoided. ResNet50 covers 49 convolutional layers and one full connection layer, retains the convolutional layer with a core size of 7 × 7 in VGG19 (14) to learn more features, and uses the maximum pooling layer for downsampling. In addition, there are five stages and two significant boards in the ResNet50 network. Stage 0 has a simple structure and is mainly used for the preprocessing of input images. It has gone through the convolutional layer, batch normalization, and ReLU activation function (12). The next four stages are composed of a bottleneck and have similar structures. After continuous convolution operation of residual blocks, the number of channels in the image pixel matrix becomes deeper and deeper, then passes through the flatten layer, and finally input into the full connection layer and output the corresponding category probability through SoftMax layer, the typical features of the image are automatically extracted, and the last constitute a classifier to divide the image into "normal" and "atelectasis." The specific network architecture is shown in **Figure 2**.

TABLE 1 Hardware and software configuration.

| Experimental environment | Configuration instructions | |
|---|---|---|
| Hardware environment | CPU | Intel (R) Xeon 5218 16C 2.3 GHz |
| | GPU | NVIDIA TESLA V100, 32 GB |
| | Memory | 32 GB |
| Software environment | Operating system | Ubuntu 20.04 |
| | Programming environment | MATLAB 2021a |

## Training strategy

Considering the extensive data set analyzed in this study, a general machine-learning method called transfer learning was used to improve the rate and performance of model learning. Transfer learning (15, 16) can transfer the knowledge learned by the model from the source domain to another target domain, so that the model can better acquire the understanding of the target domain, and improve the speed and simplicity of the learning, compared with the initial training network that uses randomly initialized weights. This study used a pretrained ResNet50 network model to randomly divide the data into the training (80%) and validation (20%) sets, and the same pretreatment operation was adopted. Since the chest radiography images were asymmetric (16), we adopted random undersampling technology and adaptive moment estimation (Adam) using vector momentum, which was adapted to increase

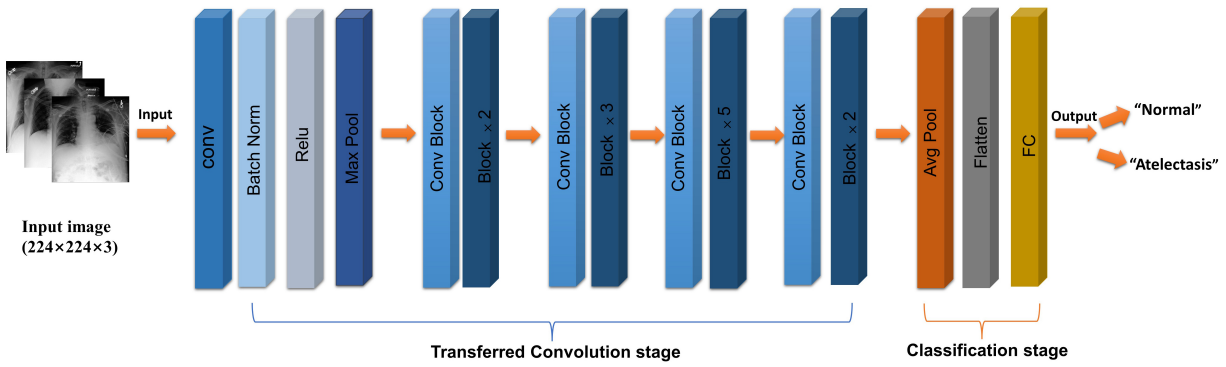



FIGURE 2
Flow chart of ResNet50 (CoNV is convolution operation, Batch Norm is Batch regularization processing, ReLU is activation function, MAXP00L and AvgPOOL are two pooling operations, including convolution transformation stage, and classification stage, respectively).

the convergence speed. Our model was trained with a 0.0001 learning rate, minimum batch size of 64, and maximum of 8 epochs, in order to achieve the maximum number of 7,200 iterations. The training lasted 105 min and 6 s. By fine-tuning the experimental parameters, the best experimental results were obtained. The training effect is shown in Figure 2.

## Performance analysis

In the experiments, the area under the receiver operating characteristic curve (AUC) (17) and the accuracy, specificity, and sensitivity (Eqs. 1, 2, and 3 below) were used as evaluation indicators (18). A larger AUC indicates that the prediction result is closer to the actual situation, and hence better model performance. Through the following formula, we can get the following indicators, respectively: false positives (FP), true negatives (TN), true positives (TP), and false negatives (FN). The confusion matrices were calculated from the following indexes, which can further help to analyze the model performance and calculate the above evaluation values.

$$\text{Accuracy} = (\text{TP} + \text{TN})/(\text{TP} + \text{FN} + \text{TN} + \text{FP}) \quad (1)$$

$$\text{Specificity} = \text{TN}/(\text{TN} + \text{FP}) \quad (2)$$

$$\text{Sensitivity} = \text{TP}/(\text{TP} + \text{FN}) \quad (3)$$

## Visualization

Based on the model obtained by the experimental training, Grad-CAMs (19, 20)can be used to generate a gradient class activation map. Grad-CAMs generated the class activation map and highlighted the areas that are important for the classification of the image, which not only provides insight into the nature of the black box of the model (16), but is also helpful for obtaining the key feature extraction of images and enables predictions of the classification decision interpretation model.

# Results

The whole training process took 105 min and 6 s. The learning curve of the model is shown in Figure 3, and the verification accuracy reached 79.91%. The partial accuracy curve reached 80% when the training was completed, and the loss rate decreased significantly to below 40%.

The confusion matrix results of the model on the testing and validation cohorts were calculated and shown in Figure 4. The testing cohort was distributed in a 2 × 2 matrix according to the labeled labels and the predicted results. Each square represents the ratio of predicted positives to actual positives. Total data volume and prediction are shown for each level. The trained model can classify the testing cohort at an accuracy of 81.70%, and the specificity and sensitivity were 88.30 and 75.10%, respectively. An accuracy of 86.90% was obtained by classifying the external validation set, and the corresponding specificity and sensitivity in the calculations were 100 and 70.70%, respectively. Table 2 lists all the evaluation indexes obtained in the internal and external testing cohort of this experiment. By calculating the above scores, the confusion matrixes of the internal and external testing cohorts were obtained, which could help to get TPs, TNs, FPs, and FNs. Meanwhile, the AUC was used as the evaluation index, with the vertical axis representing the true category rate. The horizontal and vertical axes represented the FP and TP rates, respectively, and the ROC curve was drawn. Larger AUC values indicate that the model prediction result is closer to the actual situation. The AUCs of the internal and external testing

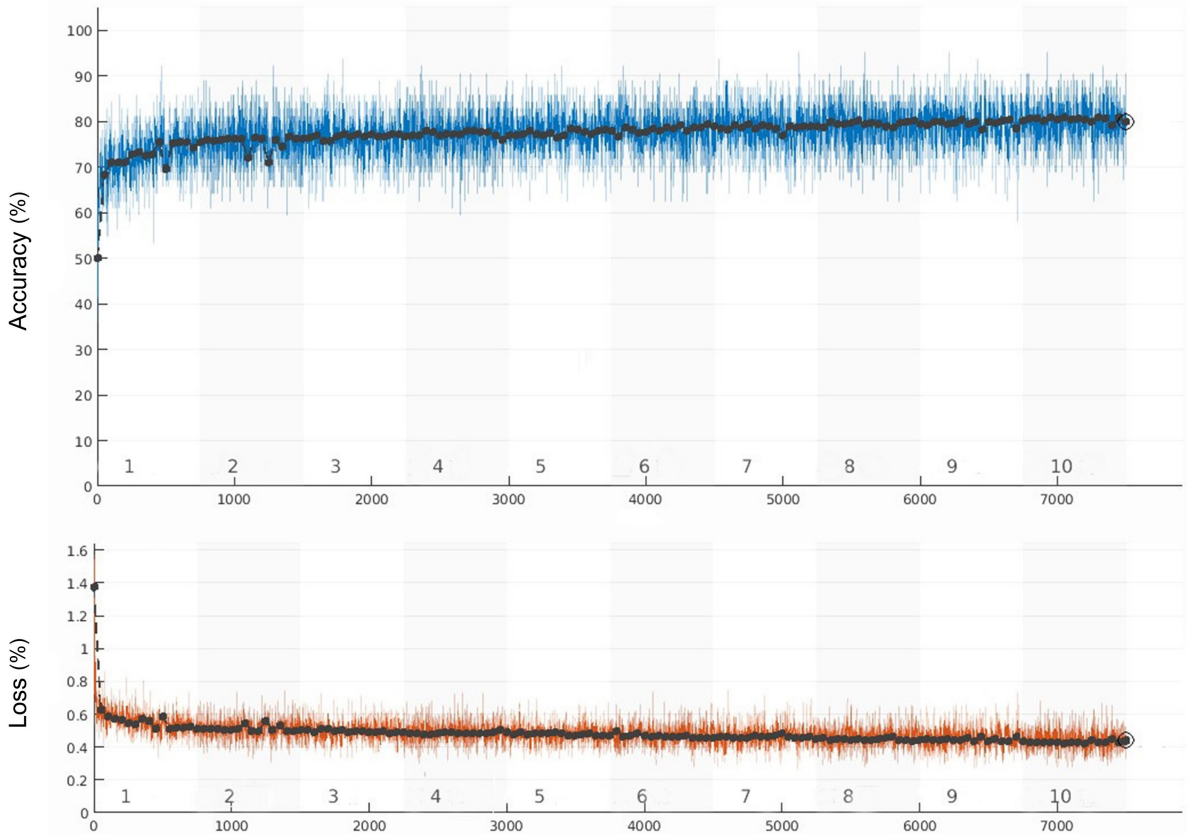



FIGURE 3
Training accuracy and training loss curves (the black curve and blue curve represent the accuracy of training set and verification set, respectively, the orange curve represents the loss).

cohorts obtained in the model training were 88.57 and 98.39%, respectively, as shown in **Figure 5**.

On this basis, the Grad-CAMs map was drawn to predict the deep model. **Figure 6** shows that the proposed model can accurately distinguish between normal chest radiographs and those with atelectasis. The red area on the chest radiographic heat map offers the critical area for the machine to determine its classification (21). Randomly generated heat maps focused on the lungs and heart.

## Discussion

Atelectasis remains a significant challenge for physicians in general anesthesia and ICUs treatment and diagnosis. Undiagnosed or late-diagnosed atelectasis can have a significant mortality risk (22). From a pathological point of view (3), atelectasis mainly manifests as reversible alveolar or lobe collapse, which is generally caused by obstruction of the affected alveoli in the airways, resulting in damage to the exchange of carbon dioxide and oxygen. According to preliminary studies, almost all patients undergoing major surgery will present with some degree of atelectasis (23, 24). Typically 2–4% of elective thoracic surgeries and 20% of emergency surgeries are related to PPCs, among which atelectasis is the most common respiratory complication. Without timely early diagnosis and intervention, a series of serious and often fatal complications will occur as the disease progresses. Eventually, due to decreased lung compliance, hypoxemia, decreased pulmonary vascular resistance, hypoxemia, postoperative infection, diffuse alveolar injury, respiratory failure, or even death (in extreme cases) may occur. So far, X-ray imaging has always been an essential means of atelectasis diagnosis, and portable chest X-ray in ICUs (25) is a rapid and straightforward method for the early diagnosis of atelectasis. This is especially true among ICUs patients with respiratory and hemodynamic parameters within a normal range, and where the direct signs of atelectasis mostly appear on chest radiograph crack deviations (3, 4, 26), parenchymal turbidities, linear boundary, and vascular displacements, among which increased density of dysfunctional lung areas is the most-obvious manifestation of atelectasis. In order to help ICUs doctors diagnose atelectasis early using portable chest X-ray, we established a model of bedside chest X-rays for detecting atelectasis by applying transfer learning based on the ResNet50

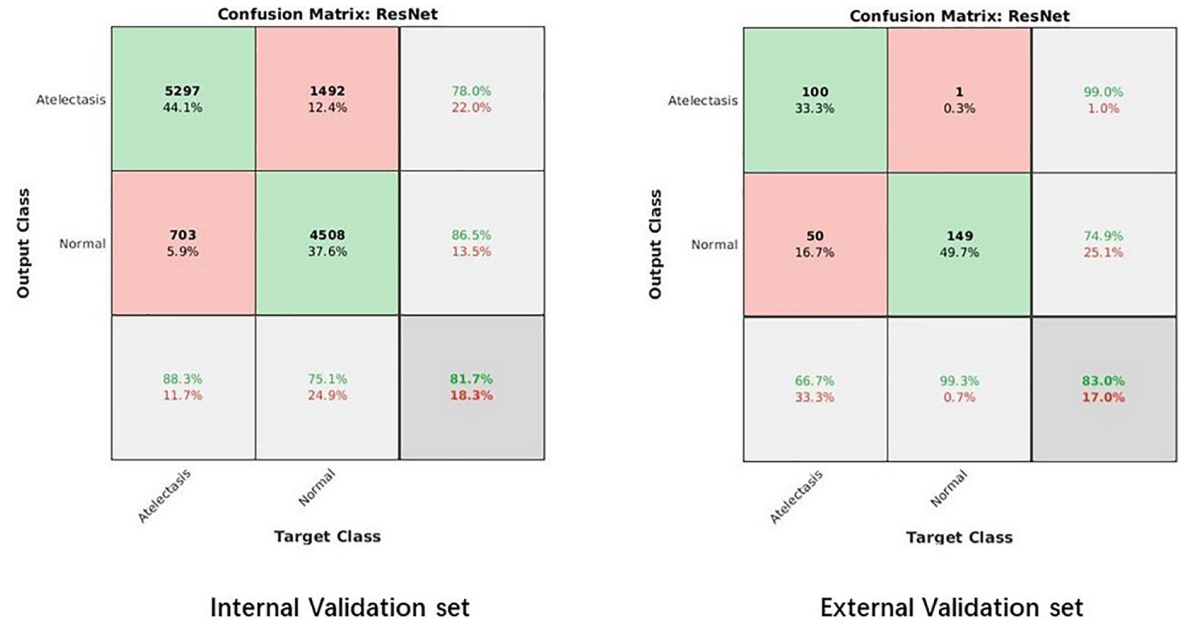



FIGURE 4
The confusion matrix of the internal and external validation sets.

TABLE 2 Accuracy, sensitivity, specificity, AUC scores of internal and external datasets.

| Datasets | Accuracy | Sensitivity | Specificity | AUC scores |
|---|---|---|---|---|
| Internal validation datasets | 81.70% | 75.10% | 88.30% | 89.31% |
| External validation datasets | 85.30% | 70.70% | 100.00% | 98.39% |

convolutional network, and used the explicit atelectasis data extracted from the MIMIC-CXR-JPG database as an internal testing cohort, which yielded an accuracy of 81.70%. This result was further externally verified using ICUs atelectasis image data rescreened and relabeled by doctors at The First Affiliated Hospital of Jinan University, achieving an accuracy rate of 86.90%. Obviously, the process of data relabeling is one of the main reasons for the increased accuracy of external validation in this study compared with previous studies.

Data enhancement and transfer learning were simultaneously adopted in this study to improve the accuracy of image classification and avoid overfitting. The ResNet50 network model parameters were obtained through the migration study less, high precision, deep residual layer network structure is complex, which solves the problem of low efficiency based on large training data sets and makes training more precise. Calculating the specificity, sensitivity, accuracy, and AUC revealed that the training model was highly robust (27), which provided external verification, with values of 100.00, 70.70, 86.90, and 98.39%, respectively. Finally, the features extracted from the training images were visualized by a heat map displaying the lung and the region near the heart. In addition, the novelty of this study was highlighted by the application of the transfer learning method to chest X-ray atelectasis examinations, and its reliable external validation.

This study had some limitations. Firstly, we used all the atelectasis image data sets during 2011–2017 in the MIMIC-CXR-JPG database, which is large but only provided relatively limited patient information, such as gender, age, and diagnostic test, and the clinical backgrounds of patients were unavailable. Atelectasis diagnoses could therefore only be labeled according to the diagnostic test, and whether it was associated with other pulmonary complications remains to be determined. We will further attempt to establish a more practical model combined with the experience of clinical practice, use more diverse neural network learning algorithms and network models, and make horizontal comparison with other more advanced networks, to classify atelectasis in more detail, with a view to providing greater assistance in early clinical intervention, diagnosis and treatment. Secondly, the case-control design used in this study artificially increased the prevalence of atelectasis by using positive data collected from the MIMIC-CXR-JPG database and The First Affiliated Hospital of Jinan University, thus overestimating the positive predictive value compared with the clinical reality (28). In addition, the data sets of internal and external validation

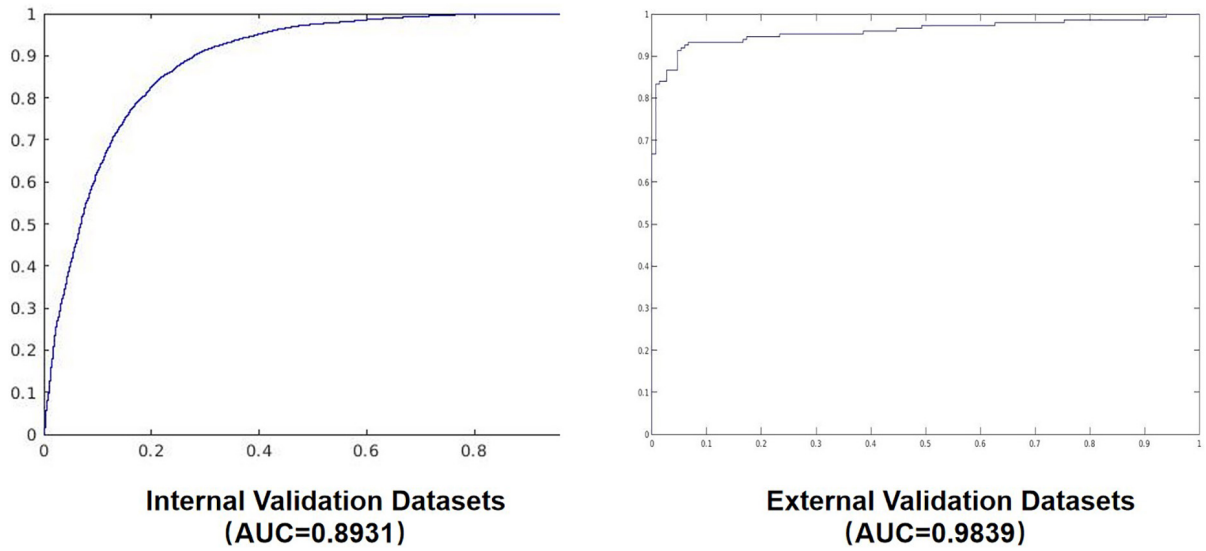



FIGURE 5
The AUC diagram of the internal and external validation sets.

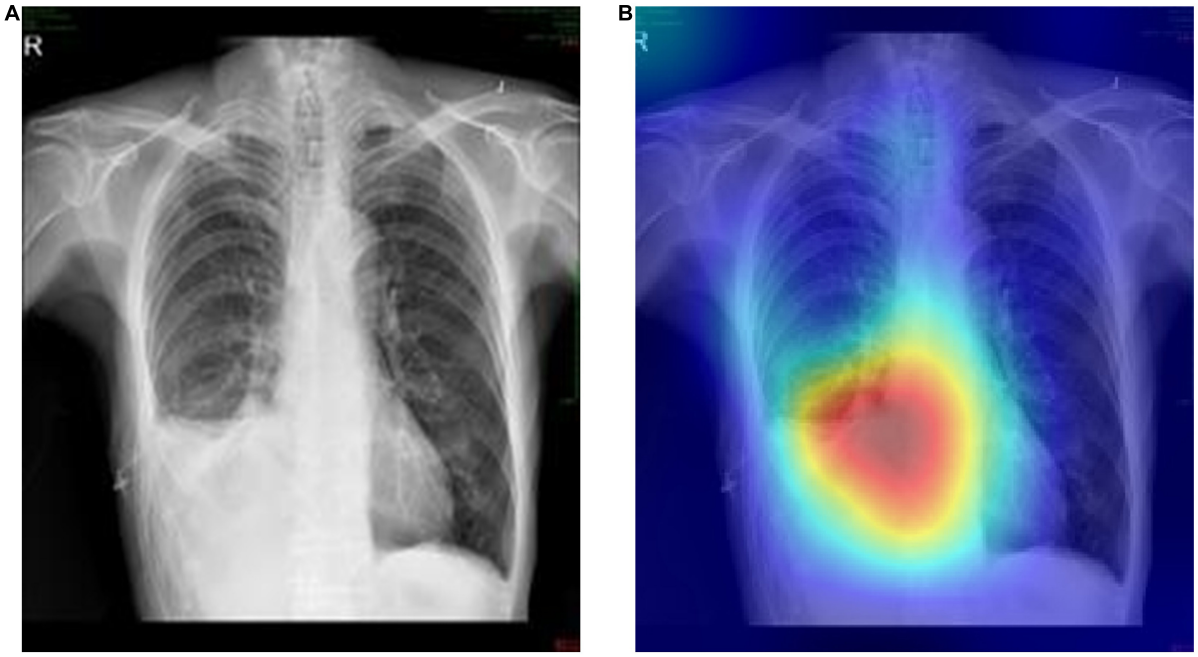



FIGURE 6
Representative cases in the external testing sets. **(A)** Example of a true-positive case, that is the original chest radiograph of atelectasis with pleural effusion in ICUs. **(B)** Grad-CAM heatmap of class activation derived from model prediction classification.

included in this paper are random and uniform data, which satisfies the ideal comparison of data to a certain extent. Queue design will therefore be used in the future to obtain more-reliable actual tags. Thirdly, our internal and external testing cohorts were basically derived from portable chest X-ray images from ICUs and the results might not be applicable to outpatients and general patients. Moreover, the image data in the MIMIC-CXR-JPG database (6, 29) were derived from foreign databases, and there were some differences in diagnostic reporting and standards. There was some heterogeneity in the atelectasis data obtained from The First Affiliated Hospital of Jinan University based on the external

validation, and so the results obtained should be considered exploratory only.

In the future, we will attempt to explore more cutting-edge and optimized AI models for portable chest X-ray diagnoses of acute and severe pulmonary complications such as atelectasis, and further promote precision medicine (30, 31), to allow the application of machine learning in clinical imaging diagnosis to realize human-machine mutual assistance and true generalization.

## Conclusion

In summary, this study found that when detecting atelectasis, a model obtained by training with sufficiently large data sets exhibited better external verification and can better help ICUs doctors to diagnose atelectasis and implement interventions early.

## Data availability statement

The raw data supporting the conclusions of this article will be made available by the authors, without undue reservation.

## Ethics statement

The studies involving human participants were reviewed and approved by the IRB of the First Affiliated Hospital of Jinan University. The patients/participants provided their written informed consent to participate in this study. Written informed consent was obtained from the individual(s), and minor(s)' legal guardian/next of kin, for the publication of any potentially identifiable images or data included in this article.

## Author contributions

XH created the study protocol, completed the main experimental training, and wrote the first manuscript draft. BL provided clinical guidance and critically revised the manuscript. TH assisted with the study design and performed data collection. SY and WW participated in the analysis and interpretation of data. HY assisted with manuscript revision and data confirmation. JL contributed to data interpretation and manuscript revision. All authors contributed to the article and approved the submitted version.

## Funding

This study was supported by the Guangdong Provincial Key Laboratory of Traditional Chinese Medicine Information (2021B1212040007).

## Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

## Publisher's note



## References

1. Restrepo RD, Braverman J. Current challenges in the recognition, prevention and treatment of perioperative pulmonary atelectasis. *Expert Rev Respir Med.* (2015) 9:97–107. doi: 10.1586/17476348.2015.996134

2. MacMahon H, Giger M. Portable chest radiography techniques and teleradiology. *Radiol Clin North Am.* (1996) 34:1–20.

3. Marini JJ. Acute lobar atelectasis. *Chest.* (2019) 155:1049–58. doi: 10.1016/j.chest.2018.11.01

4. Woodring JH, Reed JC. Radiographic manifestations of lobar atelectasis. *J Thorac Imaging.* (1996) 11:109–44. doi: 10.1097/00005382-199621000-00003

5. Moses DA. Deep learning applied to automatic disease detection using chest X-rays. *J Med Imag Radiat Oncol.* (2021) 65:498–517. doi: 10.1111/1754-9485.13273

6. Johnson AEW, Pollard TJ, Berkowitz SJ, Greenbaum NR, Lungren MP, Deng CY, et al. MIMIC-CXR, a de-identified publicly available database of chest radiographs with free-text reports. *Sci Data.* (2019) 6:317. doi: 10.1038/s41597-019-0322-0

7. Wu WT, Li YJ, Feng AZ, Li L, Huang T, Xu AD, et al. Data mining in clinical big data: the frequently used databases, steps, and methodological models. *Mil Med Res.* (2021) 8:44. doi: 10.1186/s40779-021-00338-z

8. Yang J, Li Y, Liu Q, Li L, Feng A, Wang T, et al. Brief introduction of medical database and data mining technology in big data era. *J Evid Based Med.* (2020) 13:57–69. doi: 10.1111/jebm.12373

9. Peng Y, Wang X, Lu L, Bagheri M, Summers R, Lu Z. Negbio:a high-performance tool for negation and uncertainty detection in radiology reports. *AMIA Jt Summits Transl Sci Proc.* (2018) 2017:188–96.

10. Irvin J, Rajpurkar P, Ko M, Yu Y, Ciurea-Ilcus S, Chute C, et al. "Chexpert:A Large Chest Radiograph Dataset with Uncertainty Labels and Expert Comparison," in *Proceedings of the ThirtyThird AAAI Conference on Artificial Intelligence*, Hawaii, HI (2019). p. 590–7.

11. Sirazitdinov I, Kholiavchenko M, Kuleev R, Ibragimov B. "Data Augmentation for Chest Pathologies Classification," in *Proceedings of the*

*2019 IEEE 16th International Symposium on Biomedical Imaging (ISBI 2019)*, (Piscataway, NJ: IEEE) (2019). p. 1216–9. doi: 10.1109/ISBI.2019.8759573

12. He KM, Zhang XY, Ren SQ, Sun J. "Deep residual learning for image recognition," in *Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, (Piscataway, NJ: IEEE) (2016). p. 770–8. doi: 10.1109/CVPR.2016.90

13. Abiyev RH, Ma'aitah MKS. Deep convolutional neural networks for chest diseases detection. *J Healthc Eng.* (2018) 2018:4168538. doi: 10.1155/2018/4168538

14. Simonyan K, Zisserman A. Very deep convolutional networks for large-scale image recognition. *arXiv* [Preprint] (2014). doi: 10.48550/arXiv.1409.1556

15. Jason Yosinski JC, Bengio Y, Lipson H. How transferable are features in deep neural networks? *Adv Neural Inf Process Syst.* (2014) 27:3320–28.

16. Matsumoto T, Kodera S, Shinohara H, Yamaguchi T, Higashikuni Y, Kiyosue A, et al. Diagnosing heart failure from chest x-ray images using deep learning. *Int. Heart J.* (2020) 61:781–6. doi: 10.1536/ihj.19-714

17. Ayan EÜH. "Diagnosis of pneumonia from chest x-ray images using deep learning," in *Proceedings of the2019 scientific meet-ing on Electrical-Electronics & Biomedical Engineering and Computer Science (EBBT)*, (Piscataway, NJ: IEEE) (2019). p. 1–5. doi: 10.1109/EBBT.2019.8741582

18. Majeed T, Rashid R, Ali D, Asaad A. Issues associated with deploying CNN transfer learning to detect COVID-19 from chest X-rays. *Phys Eng Sci Med.* (2020) 43:1289–303. doi: 10.1007/s13246-020-00934-8

19. Selvaraju RR, Cogswell M, Das A, Vedantam R, Parikh D, Batra D. Grad-CAM: visual explanations from deep networks *via* gradient-based localization. *Int J Comput Vis.* (2020) 128:618–26. doi: 10.1109/ICCV.2017.74

20. Selvaraju RR, Vedantam R, Cogswell M, Parikh D, Batra D. Grad-CAM: why did you say that?visual explanations from deep networks *via* gradientbased localization. *arXiv* [Preprint] (2016).

21. Zeiler MD, Fergus R. Visualizing and understanding convolutional networks. In: Fleet D, Pajdla T, Schiele B, Tuytelaars T editors. *Computer Vision–ECCV 2014*. New York, NY: Springer Publishing (2014). doi: 10.1007/978-3-319-10590-1_53

22. Hoshikawa Y, Tochii D. Postoperative atelectasis and pneumonia after general thoracic surgery. *Kyobu Geka.* (2017) 70:649–55.

23. Smetana GW. Postoperative pulmonary complications: an update on risk assessment and reduction. *Cleve Clin J Med.* (2009) 76(Suppl. 4):S60–5. doi: 10.3949/ccjm.76.s4.10

24. Ray K, Bodenham A, Paramasivam E. Pulmonary atelectasis in anaesthesia and critical care. *Continuing Educ Anaesth Critic Care Pain.* (2014) 14:236–45. doi: 10.1093/bjaceaccp/mkt064

25. Tarbiat M, Bakhshaei MH, Khorshidi HR, Manafi B. Portable chest radiography immediately after post-cardiac surgery; an essential tool for the early diagnosis and treatment of atelectasis: a case report. *Tanaffos.* (2020) 19:418–21.

26. Hobbs BB, Hinchcliffe A, Greenspan RH. Effects of acute lobar atelectasis on pulmonary hemodynamics. *Invest Radiol.* (1972) 7:1–10.

27. Arle JE, Mei L, Carlson KW. Robustness in neural circuits. In: Makarov SN, Noetscher GM, Nummenmaa A editors. *Brain and Human Body Modeling 2020: Computational Human Models Presented at EMBC 2019 and the BRAIN Initiative(R) 2019 Meeting*. Cham: Springer Publishing (2021). p. 213–29. doi: 10.1007/978-3-030-45623-8

28. Thian YL, Ng D, Hallinan JTPD, Jagmohan P, Sia SY, Tan CH, et al. Deep learning systems for pneumothorax detection on chest radiographs: a multicenter external validation study. *Radiol Artif Intell.* (2021) 3:e200190. doi: 10.1148/ryai.2021200190

29. Zhao QY, Wang H, Luo JC, Luo MH, Liu LP, Yu SJ, et al. Development and validation of a machine-learning model for prediction of extubation failure in intensive care units. *Front Med.* (2021) 8:676343. doi: 10.3389/fmed.2021.676343

30. Lee JG, Jun S, Cho YW, Lee H, Kim GB, Seo JB, et al. Deep learning in medical imaging: general overview. *Korean J Radiol.* (2017) 18:570–84. doi: 10.3348/kjr.2017.18.4.570

31. Hamet P, Tremblay J. Artificial intelligence in medicine. *Metabolism.* (2017) 69S:S36–40. doi: 10.1016/j.metabol

frontiers | Frontiers in Medicine

TYPE Original Research
PUBLISHED 25 July 2022
DOI 10.3389/fmed.2022.942356



OPEN ACCESS

EDITED BY
Zhongheng Zhang,
Sir Run Run Shaw Hospital, China

REVIEWED BY
Batuhan Bardak,
Tobb University of Economics and Technology, Turkey
Zhongheng Zhang,
Sir Run Run Shaw Hospital, China

*CORRESPONDENCE
Yue Guan
23395691@qq.com
Jingwen Wang
wangjingwen8021@163.com

†These authors have contributed equally to this work and share first authorship

SPECIALTY SECTION
This article was submitted to Intensive Care Medicine and Anesthesiology, a section of the journal Frontiers in Medicine

RECEIVED 12 May 2022
ACCEPTED 30 June 2022
PUBLISHED 25 July 2022

CITATION
Cui C, Mu F, Tang M, Lin R, Wang M, Zhao X, Guan Y and Wang J (2022) A prediction and interpretation machine learning framework of mortality risk among severe infection patients with pseudomonas aeruginosa. *Front. Med.* 9:942356. doi: 10.3389/fmed.2022.942356



# A prediction and interpretation machine learning framework of mortality risk among severe infection patients with pseudomonas aeruginosa

Chen Cui†, Fei Mu†, Meng Tang†, Rui Lin, Mingming Wang, Xian Zhao, Yue Guan* and Jingwen Wang*

Department of Pharmacy, Xijing Hospital, Fourth Military Medical University, Xi'an, China

Pseudomonas aeruginosa is a ubiquitous opportunistic bacterial pathogen, which is a leading cause of nosocomial pneumonia. Early identification of the risk factors is urgently needed for severe infection patients with P. aeruginosa. However, no detailed relevant investigation based on machine learning has been reported, and little research has focused on exploring relationships between key risk clinical variables and clinical outcome of patients. In this study, we collected 571 severe infections with P. aeruginosa patients admitted to the Xijing Hospital of the Fourth Military Medical University from January 2010 to July 2021. Basic clinical information, clinical signs and symptoms, laboratory indicators, bacterial culture, and drug related were recorded. Machine learning algorithm of XGBoost was applied to build a model for predicting mortality risk of P. aeruginosa infection in severe patients. The performance of XGBoost model (AUROC = 0.94 ± 0.01, AUPRC = 0.94 ± 0.03) was greater than the performance of support vector machine (AUROC = 0.90 ± 0.03, AUPRC = 0.91 ± 0.02) and random forest (AUROC = 0.93 ± 0.03, AUPRC = 0.89 ± 0.04). This study also aimed to interpret the model and to explore the impact of clinical variables. The interpretation analysis highlighted the effects of age, high-alert drugs, and the number of drug varieties. Further stratification clarified the necessity of different treatment for severe infection for different populations.

KEYWORDS

**machine learning, interpretation, stratification analysis, Pseudomonas aeruginosa, severe infection, risk factors**

## Introduction

Pseudomonas aeruginosa (P. aeruginosa), a ubiquitous Gram-negative pathogen, can colonize almost any part of the human body (1). More than 50% of severe acute and chronic hospital-acquired infections are caused by P. aeruginosa (2), such as ventilator-associated pneumonia and catheter infections in immunocompromised patients (3–5). It contributes to mortality rates as high as 13.5% in ventilation-associated pneumonia

caused by P. aeruginosa (6). The most common cause of death from cystic fibrosis is P. aeruginosa lung infections (7).

Pseudomonas aeruginosa infection with diverse pathological background exerts a heavy health burden for modern society. Thus, there is an urgent need to identify the mortality risk factors of infection for severe infection patients with P. aeruginosa early. A retrospective study has shown that APACHE II score and septic shock are critical factors for mortality in P. aeruginosa bacteremia, and combination therapy does not significantly reduce overall 14-day mortality (8). Several other studies have analyzed the risk factors for mortality of P. aeruginosa using logistic regression, such as age, sex, ICU admission, glucocorticoid use, inappropriate treatment regimens, mechanical ventilators, the use of a central venous catheter, and a higher APACHE II score (8–10). In a multi-center study, risk factors for mortality of community-acquired P. aeruginosa included previous pseudomonas infection/colonization, tracheostomy, bronchiectasis, invasive respiration and/or vasopressor therapy (IRVS), and very severe chronic obstructive pulmonary disease (COPD) (11). Using of previous antibiotic and ICU admission is important risk factors for drug-resistant P. aeruginosa (12), which increases the number of days in hospital stays and all-cause mortality in hospitalized patients significantly (13). Most of studies above used traditional logistic regression to predict the risk factors of P. aeruginosa infection, and there was no research for identification of the mortality risk prediction of P. aeruginosa infection in severe patients.

Machine learning is a data-driven computing method, which does a lot of work based on big data. While machine learning has been demonstrated in a few different fields, it has only recently been gaining popularity in the field of medicine. Compared to logistic regression, machine learning methods are often more comprehensive, accurate, and rapid in clinical risk prediction (14). Various machine learning methods have been widely used in constructing prediction models of disease risk, such as gastrointestinal bleeding risk assessment, prediction of mortality in intensive care units, and sepsis-associated thrombocytopenia (15–17). Ma et al. used an unsupervised learning algorithm to classify septic shock into five phenotypes, investigate the associated risk factors, and determine the best treatment strategy for these phenotypes (18). However, there has not yet been a machine learning method for the mortality risk of severe infection patients with P. aeruginosa.

In this study, we proposed a mortality risk prediction framework for severe infection patients with P. aeruginosa infection based on machine learning. Our framework focused on decision support and model interpretation. Based on XGBoost algorithm and electronic medical records (EMR) data, we built a machine learning model with good predictive performance using grid searching and cross-validation (19). Furthermore, the SHapley Additive exPlanation (SHAP) values were used to explain the prediction model from a global perspective for overcoming the shortcomings of machine learning models (20). It has the advantage of providing more details about the relationship between predictive variables and outcomes, and describing in detail the relationship between clinical factors and risks. The interpretative analysis revealed key clinical features of the risk of mortality P. aeruginosa infection in severe patients. Finally, we conducted a stratified analysis of patients from three aspects: infection site, advanced age, and the number of intravenous drug varieties. The results have some implications for P. aeruginosa clinical practice. Our study enables accurate predictions of the risk of mortality P. aeruginosa infection in severe patients, as well as interpretation of key variables that can support clinical decision making more accurately and effectively.

# Materials and methods

## Patient selection

The study was conducted at the Xijing Hospital of the Fourth Military Medical University, and a total of 571 patients with severe infections were included in the study between January 2010 and July 2021. There were 338 patients in the death group and 233 patients in the control group. Our study was approved by the domestic ethics committee with the approval number KY20212130-C-1. This study is a retrospective, observational study design that does not require informed consent. The collected research data were de-identified and analyzed anonymously.

## Data collection

Data collected using EMR at the First Affiliated Hospital of Fourth Military Medical University: basic information: age, sex, etc.; drug related: number of drug varieties, number of antibiotics drugs varieties, high-alert medication, etc.; clinical signs and symptoms: headache, cough, temperature, etc.; laboratory indicators: white blood cell count, absolute neutrophil value, etc.; bacterial culture: blood culture, urine culture, etc. All data collected are provided in the Supplementary Section (Supplementary Table 1). Here, high-alert medication refers to drugs that may cause serious injury or death to patients due to improper use of medication errors (17). According to the severity of adverse consequences that may be caused by their clinical use, high-alert medication is divided into 3 grades: A, B, and C. For the specific classification of high-alert medication, please refer to the recommended list of high-alert medication in China recommended by the Chinese Pharmaceutical Association (https://www.cpa.org.cn/index.php?do=info&cid=75676) and the management of high-alert medication in Xijing Hospital of the Fourth Military

Medical University. The details of high-alert medication can be found in the Supplementary Section (Supplementary Table 2).

## Inclusion criteria and exclusion criteria

Inclusion criteria: From January 2010 to July 2021, hospitalized patients with severe infection who associated with P. aeruginosa infection; Diagnosis of severe infection with P. aeruginosa, severe infection was defined as requiring at least 3 days of intravenous antibiotic therapy and at least 3 days of hospitalization after the diagnosis of confirmed infection. The ICD code for the diagnosis of severe infection in this study is shown in Supplementary Table 3. The P. aeruginosa infection was defined by combining the patient's clinical symptoms, signs, laboratory indicators, microbial culture, imageology, etc. Culture specimens of microorganisms come from different sites of infection, such as blood culture, urine culture, and sputum culture. The result of the patient's treatment was death or recovery.

Exclusion criteria: Non-P. aeruginosa infection; Patients with incomplete data and medical record information (the missing value of laboratory indicators exceeds 50%, incomplete medical history, no medication records); Some comorbidities such as autoimmune diseases (systemic lupus erythematosus, ANCN-associated vasculitis, rheumatoid arthritis, etc.,), malignant tumors (stomach cancer, ovarian cancer, lung cancer, etc.,) were excluded; Suspected contaminated specimens (the same sample culture of 3 or more pathogenic bacteria); Non-infected or colonized patient, such as the patient's clinical symptoms, signs, laboratory indicators, imageology were not abnormal; Hospitalization for less than 3 days.

## Preprocessing and imputation of clinical variables

All the clinical variables we collected could be divided into numerical and categorical variables according to clinical significance, and longitudinal and non-longitudinal variables according to whether repeated monitoring occurred during admission. Then, the categorical variables were converted into one-hot vectors. For clinical longitudinal variables, we extracted the maximum increase and maximum decrease during hospitalization for each variable. For laboratory longitudinal variables, we extracted the slope of all laboratory variables over time, the maximum increase and decrease during hospitalization. Finally, we got 91 variables in total (including derived variables). A detailed description and classification of all variables can be found in the Supplementary Section (Supplementary Table 1).

Outliers were detected using the interquartile range (IQR). As a threshold, the 2 times of IQR were used, and points exceeding this threshold (the upper quartile + 2 times of IQR, or the lower quartile – 2 times of IQR) were defined as outliers. Data points out of the valid value threshold were identified as outliers. The excluded outliers were modified as the nearest threshold.

Variables which had more than 50% missing values were deleted, while variables which had less than 20% missing values were replaced by the median values. Multivariate imputation by chained equations (MICE) was used to impute missing values while loss rates of variables were between 20 and 50%.

Finally, the $z$-score normalization was only performed for the all continuous values used by Support Vector Machine (SVM) (21). Since tree-based models such as XGBoost did not require standardization, the $z$-score normalization step was omitted when interpreting XGBoost, LightGBM (22), CatBoost (23), and Random Forests (RF) (24).

## Model algorithm

The XGBoost is a scalable end-to-end tree boosting system, which implements machine learning algorithms in a gradient enhancement framework that is efficient, flexible, and portable. It could be used for handling sparse data, and solving many data science problems quickly and accurately. The XGBoost has been widely used by data scientists to obtain state-of-the-art results in many machine learning challenges. The equations were as follow:

$$\mathcal{L}(\varnothing) = \sum_{i}^{n} l\left(\hat{y}_i, y_i\right) + \sum_{j}^{k} \Omega(f_j) \tag{1}$$

Here, $l$ is a loss function that measures the differences between the prediction $\hat{y}_i$and the target $y_i$. The $\Omega$ penalizes the complexity of the model.

In order to minimize the $\mathcal{L}$, the function could be write as:

$$\tilde{\mathcal{L}}^{(t)} = \sum_{i=1}^{n} \left[ g_i f_t\left(X_i\right) + \frac{1}{2} h_i f_t^2\left(X_i\right) \right] + \Omega(f_t) \tag{2}$$

$$g_i = \partial_{\hat{y}^{(t-1)}} l\left(y_i, \hat{y}^{(t-1)}\right) \tag{3}$$

$$h_i = \partial^2_{\hat{y}^{(t-1)}} l(y_i, \hat{y}^{(t-1)}) \tag{4}$$

Here, all XGBoost models were implemented by using XGBoost (version 1.5.1). All codes were implemented using Python 3.7.9.

## Method comparison

In order to evaluate the performance of our model, we compared the XGBoost with LightGBM, CatBoost, SVM, and

RF methods. All models have been optimized by grid searching to adjust hyperparameters. The detailed hyperparameters of XGBoost were described in Section 3.2. We selected the best model for predicting mortality risk for patients with severe P. aeruginosa infection.

The different parameters of LightGBM, CatBoost, SVM, and RF are summarized in Supplementary Section (Supplementary Tables 4–7). LightGBM and CatBoost were implemented by lightgbm 3.3.2 and catboost 1.0.6 in Python 3.7.9. The SVM and RF models were implemented by using scikit-learn. All code was implemented using Python 3.7.9.

## Evaluation metrics

The performance of the machine learning classifier was assessed using accuracy (ACC), receiver operator characteristics (ROC) curve, precision recall (PR) curve, area under the receiver operator characteristics curve (AUROC), and area under the precision recall curve (AUPRC), as defined by the following metrics:

$$ACC = \frac{TP + TN}{TP + TN + FP + FN} \tag{5}$$

$$Recall = True\ Positive\ Rate = \frac{TP}{TP + FN} \tag{6}$$

$$Precision = \frac{TP}{TP + FP} \tag{7}$$

$$False\ Positive\ Rate = \frac{FP}{TN + FP} \tag{8}$$

where TP, TN, FP, and FN refer to true positives, true negatives, false positives, and false negatives, respectively. Here, a "positive" label means that the outcome indicator of sample is death.

## Interpretation algorithm

In order to interpret the prediction results of XGBoost, Shapley additive explanations values were introduced, which unify Shapley regression values (20), Shapley sampling values, local interpretable model-agnostic explanations (LIME) (25), and other three existing additive feature attribution methods (DeepLIFT) (26), layer-wise relevance propagation (27), and quantitative input influence. Shapley values were defined as a class of additive feature attribution methods, which have an explanation model that is a liner function of binary variables as follow:

$$g\left(z'\right) = \varnothing_0 + \sum_{i=1}^{M} \varnothing_i z_i' \tag{9}$$

Where $z' \in \{0,1\}^M$, $M$ is the number of simplified input feature, and $\varnothing_i \in \mathbb{R}$. $\varnothing_0$ is the constant of the interpretation model, $\varnothing_i$ is the predicted mean value of all training samples, and is the attribution value of each feature.

## Statistical analysis

In this paper, two independent-sample $t$-tests were used for the statistical analysis. A $p$-value of less than 0.05 was considered significant. All statistical analyses were performed using Scipy 1.7.2.

# Results

## General information

A total of 571 hospitalized patients infected with P. aeruginosa were included in this study. The flow chart of this study is shown in Figure 1. In terms of the source of infection, pulmonary infections accounted for the highest percentage of 455 cases (80%), followed by bloodstream infections with 57 cases (10%) and skin and soft tissue infection with 54 cases (9%). A detailed description of the clinical characteristics of the whole cohort is provided in Table 1.

## Model optimization and performance

To optimize the XGBoost model, the dataset was divided into five sets. One of the five sets was selected as test set, the rest four sets were selected as training set. We explored different hyperparameters through a grid search, such as the maximum depth, the number of estimators, and learning rate. We considered the maximum depth with 2, 4, 8, 16, and 32, the number of estimators with 5, 10, 15, 20, 25, and 30, and the learning rate with 0.01, 0.05, 0.1, 0.2, 0.3, and 0.5. The best models with different hyperparameters were selected according to the mean performance based on cross validation.

The ROC curve and PR curve of three machine learning models are shown in Figure 2. The AUC of 5-fold cross validations were between 0.90 and 0.96 and PR of 5-fold cross validations was between 0.91 and 0.97. The ACC, AUROC, and AUPRC of mean performance of 5-fold cross validation were displayed in Table 2. The results shown that XGBoost had better prediction ability than other methods.

## Model interpretation

Although the XGBoost model can achieve good predictions performance, the lacking of interpretation limits the application in clinical practice. To facilitate interpretation of the prediction model, an artificial intelligence SHAP values for global model interpretation were introduced (20). Compared with traditional

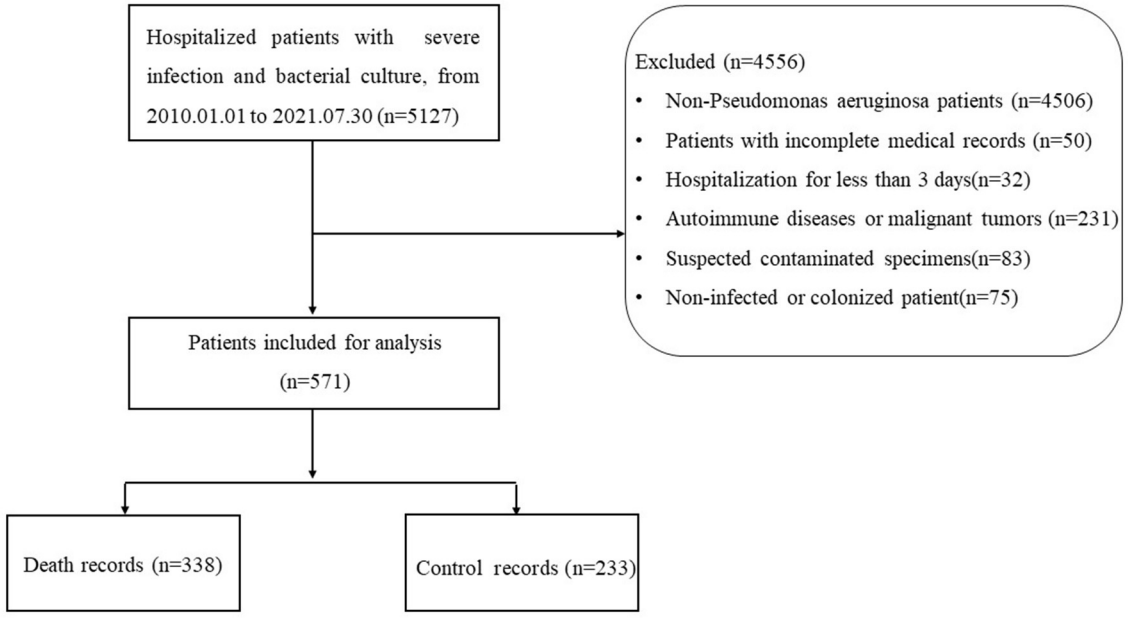



FIGURE 1
Flow chart of the study.

feature importance methods (such as decision tree weights importance), SHAP values have better consistency and can show the positive or negative relationship of each predictive variable with respect to the target.

## Importance of clinical variables

According to the importance and impacts of variables on model prediction, a bee swarm plot was formed for each feature. As shown in Figure 3, a series bee swarm plots were listed in their order of importance.

We found that older (red) patients had a higher risk of mortality than younger (blue) patients (large on the horizontal). Similarly, patients who used more types of high-alert medications and more types of intravenous drugs had higher risk of mortality than those who used fewer types. The lower the maximum decrease in respiratory rate (significantly lower than normal), the higher the risk of mortality. In addition, patients who underwent drainage (red) had a lower risk of mortality than patients who did not undergo drainage (blue). It is important to emphasize that all effects only describe the behavior of the model and are not causality in the real world.

## Detailed dependencies of variables

To further elucidate the detailed relationship between mortality risk and clinical variables, SHAP interaction values were used to reveal the dependencies relationships based on the key feature of importance the bee swarm plots. Here, each point corresponds to a sample of infected patients, and each scatter plot shows the effect of features on SHAP interaction values. The results were shown in SHAP dependence plots (Figure 4). By analyzing the dependencies factors, it was found that the risk of mortality was significantly higher in patients with higher maximum increases in urea and creatinine when the number of intravenous drugs was higher.

# Evaluation on different pathogens infection

In addition to P. aeruginosa, Klebsiella pneumoniae (K. pneumoniae) is major hospital-acquired pathogen, causing pneumonia, urinary tract infection, intra-abdominal infection, and bacteremia in immunocompromised patients (28).

Here, we build a clinical dataset of K. pneumoniae infections as an external validation for testing and discussing the generalization performance of our model. The hyperparameters of model were obtained from the best performance in Section 3.2. Five sub-models trained on the 5-fold cross validation were used in the external validation set. The average performance of each sub-model on these external test sets is shown in Table 3. We can find that the performance of model on K. pneumoniae still had some degree of predictive ability, but it is a little worse than prediction for infection patients with P. aeruginosa. It suggested that our

TABLE 1 Characteristics of patients at baseline and clinical outcomes.

| Categories | Variables | Total ($n = 571$) |
|---|---|---|
| **Basic information** | Age (years) [median (IQR)] | 64 (47–81) |
| | Male [No. (%)] | 428 (74.86%) |
| | Hosp (days) [median (IQR)] | 23 (13–40) |
| | Drug Allergy [No. (%)] | 69 (12%) |
| | Smoking [No. (%)] | 105 (18%) |
| | Alcohol User [No. (%)] | 55 (10%) |
| **Drug related** | Number of Drug Varieties [median (IQR)] | 52 (39-66) |
| | Number of Intravenous Drugs Varieties [median (IQR)] | 7 (4-10) |
| **Clinical signs and symptoms** | Headache [No. (%)] | 91 (16%) |
| | Cough [No. (%)] | 365 (64%) |
| | Expectoration [No. (%)] | 322 (56%) |
| | Sore Throat [No. (%)] | 15 (3%) |
| | Hemoptysis [No. (%)] | 7 (1%) |
| | Dyspnea [No. (%)] | 149 (26%) |
| | Vomiting [No. (%)] | 187 (33%) |
| | Diarrhea [No. (%)] | 76 (13%) |
| | Lymphadenopathy [No. (%)] | 14 (2%) |
| | Drainage [No. (%)] | 222 (39%) |
| | Tracheotomy [No. (%)] | 104 (18%) |
| | Endotracheal Intubation [No. (%)] | 150 (26%) |
| | Central Venous Catheter [No. (%)] | 43 (8%) |
| | Indwelling Catheter [No. (%)] | 302 (53%) |
| | PICC Catheter [No. (%)] | 141 (25%) |
| | Temperature (°C) [median (IQR)] | 36.9 (36.5–37.6) |
| | Respiratory Rate ($min^{-1}$) [median (IQR)] | 21.0 (19.0–25.0) |
| | Heart Rate ($min^{-1}$) [median (IQR)] | 89.0 (78.0–105.0) |
| | DBP (mmHg) [median (IQR)] | 68.0 (60.0–76.0) |
| | SBP (mmHg) [median (IQR)] | 116.0 (102.0–129.0) |
| **Bacterial culture** | Blood [No. (%)] | 57 (10%) |
| | Urine [No. (%)] | 16 (3%) |
| | Phlegm [No. (%)] | 455 (80%) |
| | Secretions [No. (%)] | 54 (9%) |
| | Cerebrospinal Fluid [No. (%)] | 7 (1%) |
| | Feces [No. (%)] | 0 (0%) |
| | Number of Concurrent Infection [No. (%)] | 399 (70%) |
| **Laboratory Indicators** | WBC($\times 10^9$/L) [median (IQR)] | 10.08 (6.9–14.39) |
| | NEUT# ($\times 10^9$/L) [median (IQR)] | 7.96 (5.12–11.82) |
| | NEUT% [median (IQR)] | 0.83 (0.74–0.89) |
| | RBC ($\times 10^{12}$/L) [median (IQR)] | 3.19 (2.77–3.66) |
| | PLA ($\times 10^9$/L) [median (IQR)] | 166.0 (91.0–258.0) |
| | HGB (g/L) [median (IQR)] | 95.0 (84.0–110.0) |
| | ALT (IU/L) [median (IQR)] | 29.0 (17.0–57.0) |
| | AST (IU/L) [median (IQR)] | 31.0 (20.0–55.0) |
| | DBIL (μmol/L) [median (IQR)] | 8.4 (4.6–16.0) |
| | CREA (μmol/L) [median (IQR)] | 78.0 (59.0–115.0) |
| | Urea (mmol/L) [median (IQR)] | 8.87 (5.7–15.0) |
| | ALB (g/L) [median (IQR)] | 31.6 (28.5–34.8) |

*(Continued)*

TABLE 1 Continued

| Categories | Variables | Total ($n = 571$) |
|---|---|---|
| | SAA (mg/L) [median (IQR)] | 202.0 (72.1–421.0) |
| | ESR (mm/h) [median (IQR)] | 56.0 (26.25–84.5) |
| | CRP (mg/L) [median (IQR)] | 60.5 (25.25–116.15) |
| | IL-6 (pg/mL) [median (IQR)] | 56.96 (25.03–139.2) |
| | PCT (ng/mL) [median (IQR)] | 0.95 (0.31–3.42) |

NEUT# represents the neutrophil count.

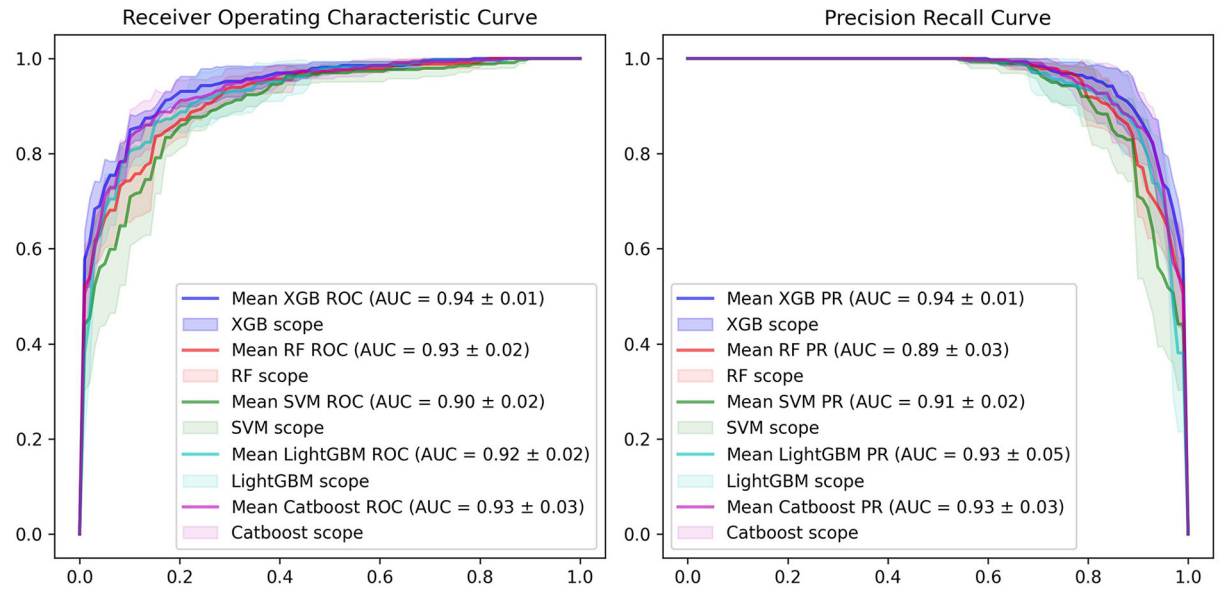



FIGURE 2
Receiver operator characteristics (ROC) curve and precision recall (PR) curve of five machine learning models.

TABLE 2 Methods comparison based on AUROC and AUPRC.

| Method | ACC | AUROC | AUPRC |
|---|---|---|---|
| XGBoost | 0.88 ± 0.02 | 0.94 ± 0.01 | 0.94 ± 0.03 |
| LightGBM | 0.86 ± 0.05 | 0.92 ± 0.02 | 0.93 ± 0.05 |
| CatBoost | 0.86 ± 0.02 | 0.93 ± 0.03 | 0.93 ± 0.03 |
| Random Forest | 0.86 ± 0.03 | 0.93 ± 0.03 | 0.89 ± 0.04 |
| Support Vector Machine | 0.84 ± 0.03 | 0.90 ± 0.03 | 0.91 ± 0.02 |

prediction model had some capacity for prediction on different pathogens infection.

## Stratification analysis

### Stratified analysis of infection sites

Figures 5A,B shows the number and proportion of P. aeruginosa cultured at different infection sites. The percentage of P. aeruginosa cultured in phlegm, others, and blood was higher. Here, others refer to infection sites except blood, urine, phlegm, secretions, cerebrospinal fluid, and feces. Compared with the control group, more P. aeruginosa was cultured in phlegm of patients in the death group. And the proportion of P. aeruginosa cultured in sputum and urine, sputum and blood at the same time was significantly higher than the death group. At the same time, we analyzed the association between the infection site and the number of concurrent infections. Figure 5C suggests that there was a statistically significant difference between the number of co-infections in the death group and the control group ($p < 0.01$), and the death group was often accompanied by 0–3 co-infections, when infection site was blood. The association between the infection site and the number of high warning drugs used was also analyzed. Figure 5D shows that when P. aeruginosa was detected in sputum culture or blood culture, the number of A-Alert drug use was more in the death group than in the control group, and the difference was significant (all $p < 0.01$).

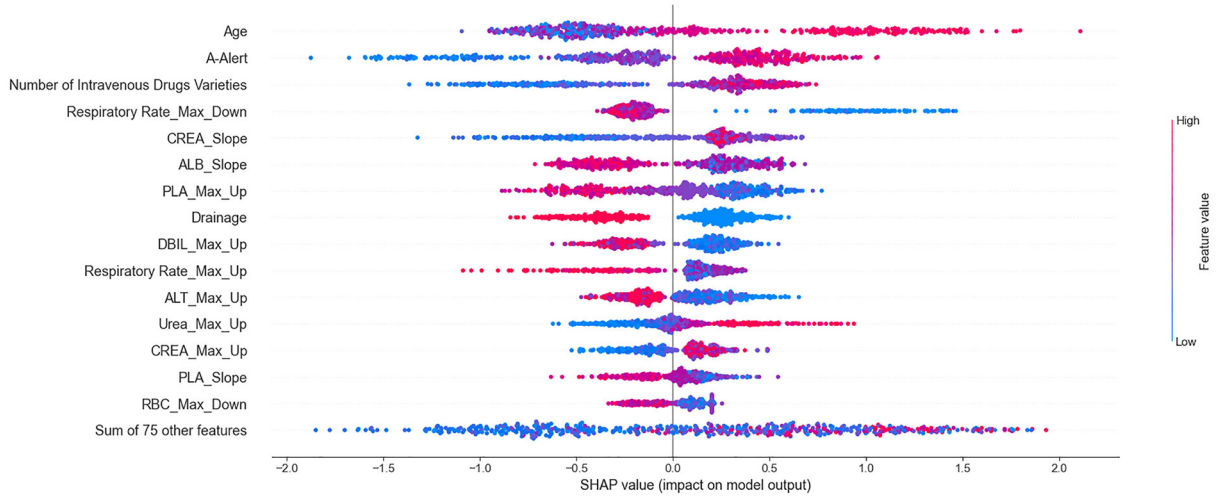



FIGURE 3
Summarize bee swarm plots for top 15 clinical variables of SHAP values. In a bee swarm plot, each point corresponding to a sample of single P. aeruginosa infected patient of data set. The position of each point on the horizontal axis indicated the effect of that feature on the model prediction, and the color of a point reflected the eigenvalue of the case. For binary variables (such as drainage or not), red dots and bule dots correspond to 1 and 0 respectively. For numerical variables (such as age), the color of dots represented high and low values, respectively. Overlapping points that fall in the same horizontal position will be scattered vertically to show the density.

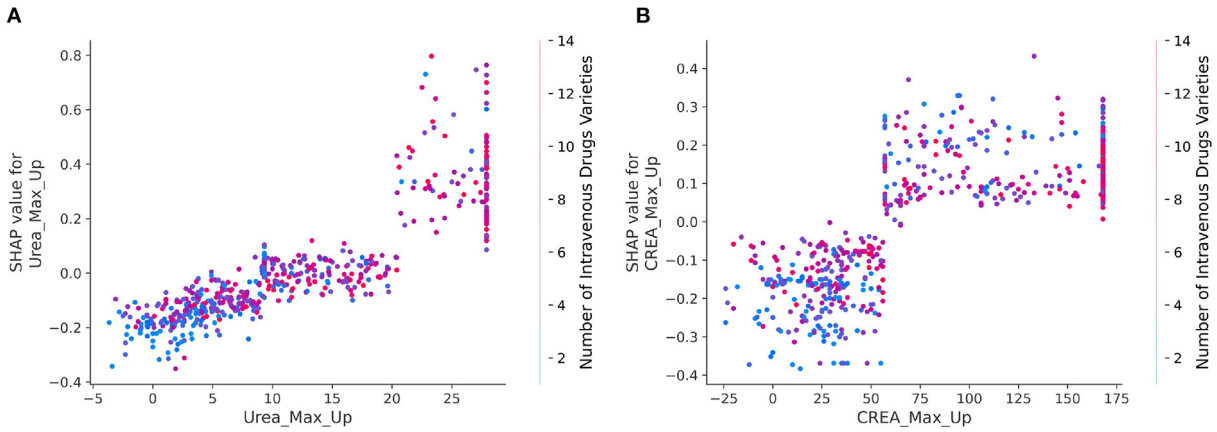



FIGURE 4
**(A,B)** SHAP dependence plots for interaction of crucial clinical variables. The x axis represents the eigenvalue of the axis title, and the y axis indicates the corresponding SHAP value, representing the contribution of this feature to prediction results of model. The color of every dot reflects the eigenvalues of right axis title. The larger the value of the x-coordinate of the sample point, the variable of x-axis is more large. The larger the value of the y coordinate of the sample point, the greater risk of mortality of the sample point, and the redder the color of the sample point, the higher the value of the right index.

## Stratified analysis of age

These results in Figures 6A,B suggested an increasing trend in the number of deaths with increasing age. When the patients were older than 75, the maximum decrease in respiratory rate in the death group was significantly different from that in the control group (Figure 6C, $p < 0.01$). Figure 6D shows that when patients were older than 18, the maximum increase in platelets was significantly lower in the death group than that in the control group (all $p < 0.01$). Figure 6E shows that the older the age, the greater the number of A-alert drugs was used. And when the patients were younger than 75, the number of A-alert drugs used in the death group was significantly different from the control group (all $p < 0.05$ or $p < 0.01$). While the patients were older than 75, the number of B-alert drugs used in the death

TABLE 3 Performance of model on external validation sets.

| Infection | ACC | AUROC | AUPRC |
|---|---|---|---|
| P. aeruginosa | 0.88 ± 0.02 | 0.94 ± 0.01 | 0.94 ± 0.03 |
| K. pneumoniae | 0.85 ± 0.04 | 0.91 ± 0.03 | 0.92 ± 0.05 |

group was significantly different from that of the control group (Figure 6F, $p < 0.01$).

### Stratified analysis of drug varieties

Figure 7 shows the association between the number of drug varieties and the maximum increase in creatinine, the maximum increase in urea, the maximum decrease in respiratory rate, and the maximum decrease in diastolic blood pressure. The results in Figures 7C,D show that the higher the number of intravenous drug varieties, the more significant the maximum increase was in creatinine and urea ($p < 0.01$ or $p < 0.05$). Figure 7E reveals the association between the number of intravenous drug varieties and the maximum decrease of respiratory rate. When the number of intravenous drug varieties was $< 7$, the maximum decrease of respiratory rate in the death group was significantly smaller than that of the control group ($p < 0.01$). Figure 7F suggests that when the number of intravenous drug varieties was $< 10$, the maximum decrease in diastolic blood pressure in the death group was statistically significantly different from that of the control group (all $p < 0.01$), and close attention should be paid to patients whose maximum decreases in diastolic blood pressure were small and the number of intravenous drug varieties was more than 10.

## Discussion

P. aeruginosa infection constitutes a major clinical challenge (29). Therefore, it is of great significance to predict the risk factors of mortality for P. aeruginosa in severe patients. In this study, we assessed the risk factors of 571 patients with severe infection with P. aeruginosa, such as 338 deaths and 233 cures. A prediction model for mortality risk of P. aeruginosa in severe patients was established. Compared to some other machine learning algorithms, the XGBoost model achieved the best performance in ACC, AUROC, and AUPRC. Furthermore, in order to indicate the relationship between the clinical variables and the risk of mortality, the SHAP values were introduced to evaluated the importance of clinical variables in predictor.

The most obvious finding to emerge from the analysis above was that advanced age was one of the mortality risk factors for P. aeruginosa infection in severe patients, which was consistent with the results of a previous study (10). Then, the number of high-alert medication types and intravenous drug types were risk factors for mortality from P. aeruginosa infection, which had not been described in previous severe infection studies. The combination of many drugs is likely to cause some side effects on patients, and studies have shown that paying attention to high-risk drugs will greatly reduce hospitalization, disability, death, and other conditions (30). When the number of high-alert medication types and intravenous drugs types is too much, it reminds clinicians to pay more attention to the medication situation of patients. Timely adjustments of medication regimen are expected to improve the prognosis and reduce mortality of patients. Drainage is also a mortality risk factor for P. aeruginosa infection in severe patients. The results of this study indicate that patients who have been drained have a lower risk of mortality than patients who do not have been drained. This result is in accord with the fact that drainage is conducive to the timely discharge of purulent secretions, effusions, blood, and exudates from the wound. Drainage might possess dual roles in clinical treatment, one in assessing the condition patients and the other in facilitating wound healing. For abscess without effective drainage, the minimum effective concentrations for antimicrobial activity may not be reached.

In addition, we further stratified to explore the relationships between the site of infection, age stratification, and the number of medication species with important variables, laying the foundation for future variable interaction studies. The conclusion also verified that the risk of death from blood culture with P. aeruginosa was higher than other sites, and it was consistent with our common knowledge. Simultaneously, with patients' ages increasing and the higher the number of intravenous drug varieties used, the number of deaths showed an increasing trend. It suggested that risk factors such as advanced age and the number of drug varieties used need to be actively paid attention to for patients infected with P. aeruginosa, especially when the age was greater than 75 and the number of drug species was $>$ 10. These conclusions were preliminary and needed to be further validated.

This retrospective study still had several limitations. Firstly, this was a single-center study and, therefore, has all the limitations inherent in such a study design. The distribution and characteristics of the clinical data used in this study could vary among different regions. In future works, integrating more data and having more precise estimates are possible. The clinical data from the multi-centric will help researchers to build more generalization and prospective prediction models. Future works should be used in elucidate the diversity of AMP resistance mechanisms in more realistic clinical settings. In future works, our model should be used in the multi-centric study or other clinical datasets such as MIMIC III (31) or a critical care database involving patients with infection (32). Secondly, our model was built

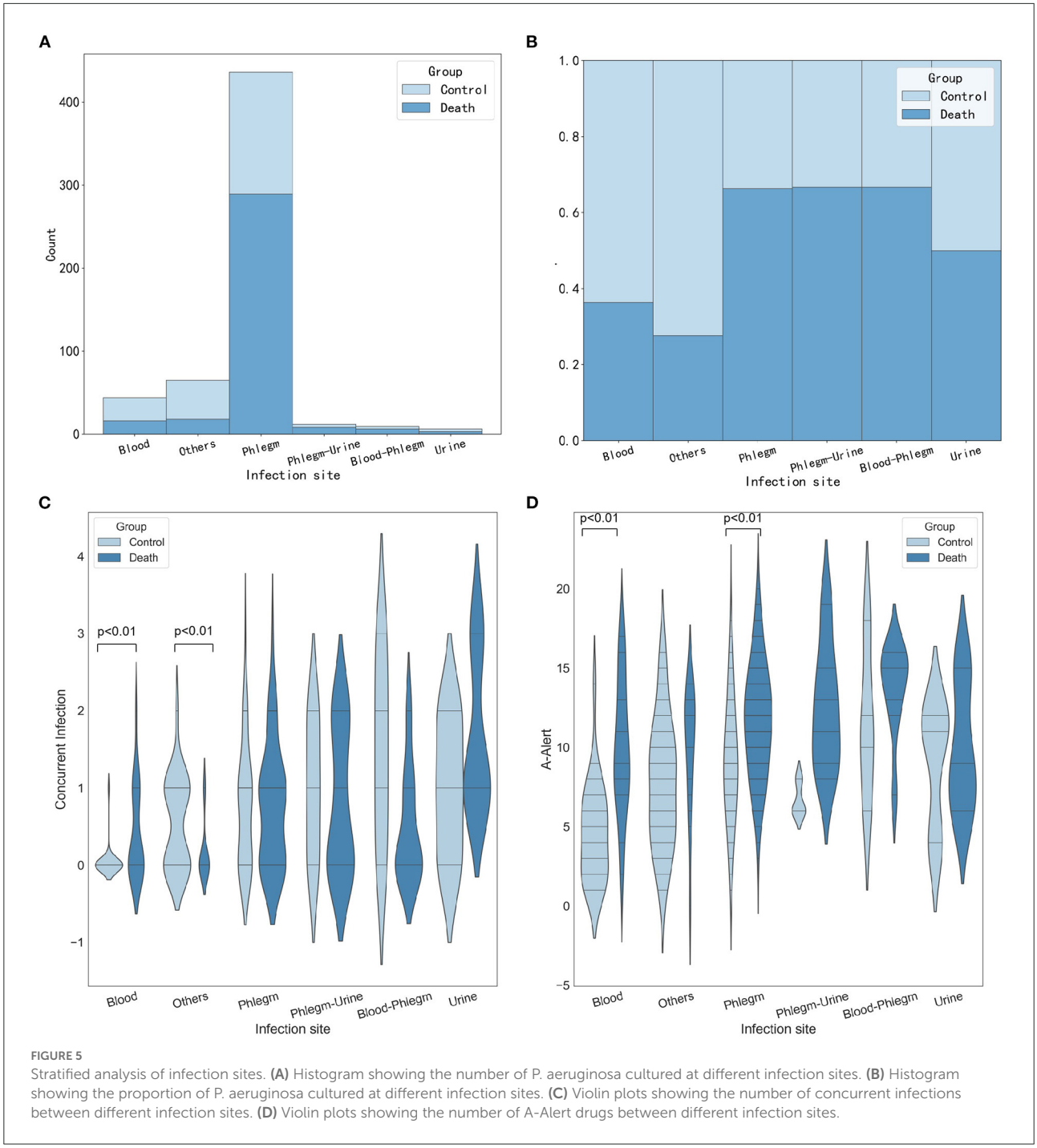



FIGURE 5
Stratified analysis of infection sites. **(A)** Histogram showing the number of P. aeruginosa cultured at different infection sites. **(B)** Histogram showing the proportion of P. aeruginosa cultured at different infection sites. **(C)** Violin plots showing the number of concurrent infections between different infection sites. **(D)** Violin plots showing the number of A-Alert drugs between different infection sites.

based on the clinical data from patients with P. aeruginosa infection. Although it showed some predictive capacity in testing on clinical data from patients with Klebsiella pneumoniae infection, it still has significant shortcomings compared to the performance in P. aeruginosa infection. And it needs to be improved in future studies to expand the application of the model. Thirdly, the clinical data used in this study only included part of structured clinical information. Other informative data, such as nursing notes and radiology reports were not used. More detailed clinical data such as drug dosage or time, mechanical ventilation time and effectiveness evaluation will provide a new perspective for deep analysis and interpretation. Finally, since only the correlation rather than the causal relationship between the predictors and risk

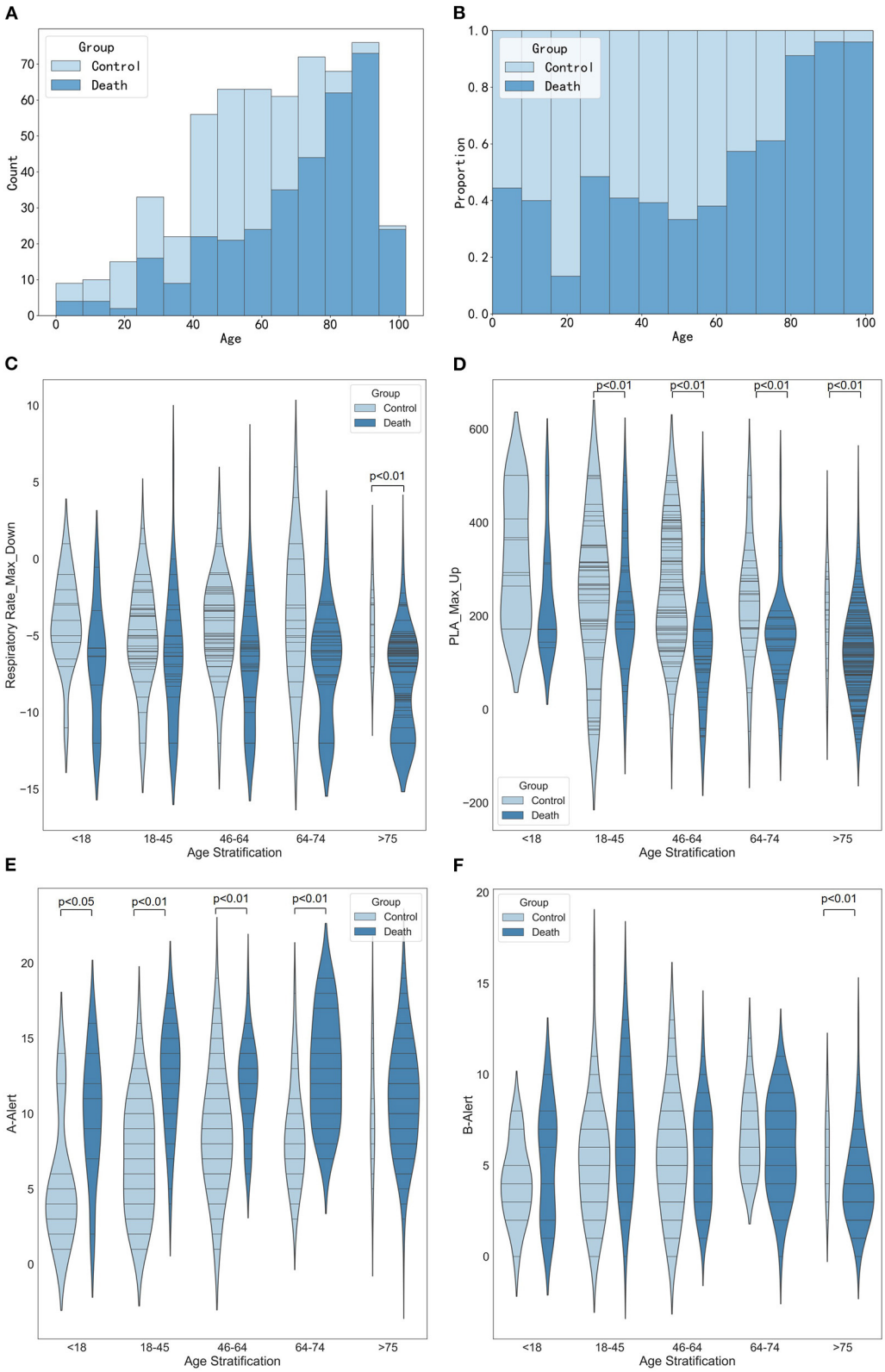



FIGURE 6
Stratified analysis of age. **(A)** Histogram showing the number of age stratification. **(B)** Histogram showing the proportion of age stratification. **(C)** Violin plots showing the maximum decrease in respiratory rate between different age stratification. **(D)** Violin plots showing the maximum increase in platelets between different age stratification. **(E)** Violin plots showing the number of A-alert drugs between different age stratification. **(F)** Violin plots showing the number of B-alert drugs between different age stratification.

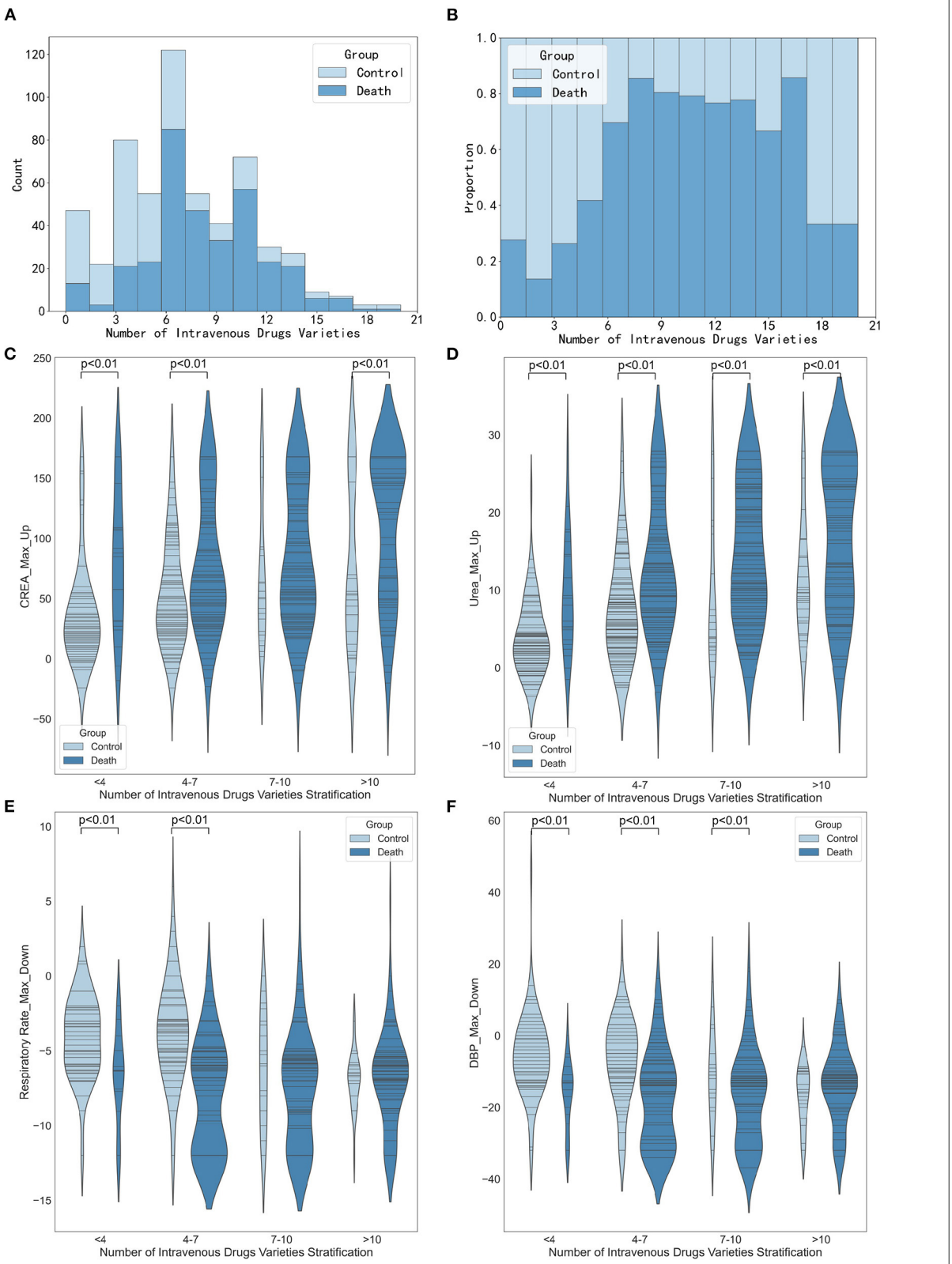



FIGURE 7
Stratified analysis of intravenous drugs varieties. **(A)** Histogram showing the number of intravenous drugs varieties stratification. **(B)** Histogram showing the proportion of intravenous drugs varieties stratification. **(C)** Violin plots showing the maximum increase in creatinine between different intravenous drugs varieties stratification. **(D)** Violin plots showing the number of maximum increase in urea between different intravenous drugs varieties stratification. **(E)** Violin plots showing the maximum decrease in respiratory rate between different medication varieties stratification. **(F)** Violin plots showing the maximum decrease in diastolic diastolic blood pressure between different medication varieties stratification.

outcome was considered in this study, our conclusions still required further prospective trials to evaluate. More in-depth investigation of the causal relationship between the clinical feature and risk is essential for supporting clinical control and decision.

## Data availability statement

The original contributions presented in the study are included in the article/Supplementary material, further inquiries can be directed to the corresponding authors.

## Ethics statement

The studies involving human participants were reviewed and approved by Institutional Medical Ethical Committee of the First affiliated Hospital of Air Force Medical University, China (approval No. KY20212130-C-1) on August 30, 2021. Written informed consent for participation was not required for this study in accordance with the national legislation and the institutional requirements.

## Author contributions

MW and XZ contributed to data collecting and processing. CC contributed to modeling and calculations. FM and MT performed the results analysis. CC, FM, and MT drafted the manuscript. RL contributed to manuscript revision. JW and YG provided overall supervision and undertook the responsibility of submitting the manuscript for publication. All authors discussed and commented on the manuscript.

## Funding

This research was financially supported by the National Natural Science Foundation of China (Nos. 72074218, 81903837, and 81774190).

## Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

## Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations or those of the publisher, the editors, and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

## Supplementary material

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fmed.2022.942356/full#supplementary-material

## References

1. Schütz C, Ho D-K, Hamed MM, Abdelsamie AS, Röhrig T, Herr C, et al. A new pqsr inverse agonist potentiates tobramycin efficacy to eradicate pseudomonas aeruginosa biofilms. *Adv Sci (Weinh).* (2021) 8:e2004369. doi: 10.1002/advs.202004369

2. Tuon FF, Dantas LR, Suss PH, Tasca Ribeiro VS. Pathogenesis of the biofilm: a review. *Pathogens.* (2022) 11:300. doi: 10.3390/pathogens11030300

3. Laborda P, Sanz-Garcia F, Hernando-Amado S, Martinez JL. Pseudomonas aeruginosa: an antibiotic resilient pathogen with environmental origin. *Curr Opin Microbiol.* (2021) 64:125–32. doi: 10.1016/j.mib.2021.09.010

4. Cabot G, Lopez-Causape C, Ocampo-Sosa AA, Sommer LM, Dominguez MA, Zamorano L, et al. Deciphering the resistome of the widespread pseudomonas aeruginosa sequence type 175 international high-risk clone through whole-genome sequencing. *Antimicrob Agents Chemother.* (2016) 60:7415–23. doi: 10.1128/AAC.01720-16

5. Chastre J, Fagon J-Y. Ventilator-associated pneumonia. *Am J Respir Crit Care Med.* (2002) 165:867–903. doi: 10.1164/ajrccm.165.7.2105078

6. Yang F, Gu J, Yang L, Gao C, Jing H, Wang Y, et al. Protective efficacy of the trivalent pseudomonas aeruginosa vaccine candidate pcrv-opri-hcp1 in murine pneumonia and burn models. *Sci Rep.* (2017) 7:3957. doi: 10.1038/s41598-017-04029-5

7. Bricio-Moreno L, Sheridan VH, Goodhead I, Armstrong S, Wong JKL, Waters EM, et al. Evolutionary trade-offs associated with loss of pmrb function in host-adapted pseudomonas aeruginosa. *Nat Commun.* (2018) 9:2635. doi: 10.1038/s41467-018-04996-x

8. Kim YJ, Jun YH, Kim YR, Park KG, Park YJ, Kang JY, et al. Risk factors for mortality in patients with pseudomonas aeruginosa bacteremia; retrospective study of impact of combination antimicrobial therapy. *BMC Infect Dis.* (2014) 14:161. doi: 10.1186/1471-2334-14-161

9. Ababneh MA. Rababa'h AM, Almomani BA, Ayoub AM, Al-Azzam SI. *A ten-year surveillance of p aeruginosa bloodstream infections in a tertiary care hospital: trends and risk factors for mortality. Int J Clin Pract.* (2021) 75:e14409. doi: 10.1111/ijcp.14409

10. Babich T, Naucler P, Valik JK, Giske CG, Benito N, Cardona R, et al. Risk factors for mortality among patients with pseudomonas aeruginosa bacteraemia: a retrospective multicentre study. *Int J Antimicrob Agents.* (2020) 55:105847. doi: 10.1016/j.ijantimicag.2019.11.004

11. Restrepo MI, Babu BL, Reyes LF, Chalmers JD, Soni NJ, Sibila O, et al. Burden and risk factors for pseudomonas aeruginosa community-acquired pneumonia: a multinational point prevalence study of hospitalised patients. *Eur Respir J.* (2018) 52:1701190. doi: 10.1183/13993003.01190-2017

12. Raman G, Avendano EE, Chan J, Merchant S, Puzniak L. Risk factors for hospitalized patients with resistant or multidrug-resistant pseudomonas aeruginosa infections: a systematic review and meta-analysis. *Antimicrob Resist Infect Control.* (2018) 7:79. doi: 10.1186/s13756-018-0370-9

13. Nathwani D, Raman G, Sulham K, Gavaghan M, Menon V. Clinical and economic consequences of hospital-acquired resistant and multidrug-resistant pseudomonas aeruginosa infections: a systematic review and meta-analysis. *Antimicrob Resist Infect Control.* (2014) 3:32. doi: 10.1186/2047-2994-3-32

14. Churpek MM, Yuen TC, Winslow C, Meltzer DO, Kattan MW, Edelson DP. Multicenter comparison of machine learning methods and conventional regression for predicting clinical deterioration on the wards. *Crit Care Med.* (2016) 44:368–74. doi: 10.1097/CCM.0000000000001571

15. Shung DL, Au B, Taylor RA, Tay JK, Laursen SB, Stanley AJ, et al. Validation of a machine learning model that outperforms clinical risk scoring systems for upper gastrointestinal bleeding. *Gastroenterology.* (2020) 158:160–7. doi: 10.1053/j.gastro.2019.09.009

16. 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) 3:42–52. doi: 10.1016/S2213-2600(14)70239-5

17. Jiang X, Wang Y, Pan Y, Zhang W. Prediction models for sepsis-associated thrombocytopenia risk in intensive care units based on a machine learning algorithm. *Front Med (Lausanne).* (2022) 9:837382. doi: 10.3389/fmed.2022.837382

18. Ma P, Liu J, Shen F, Liao X, Xiu M, Zhao H, et al. Individualized resuscitation strategy for septic shock formalized by finite mixture modeling and dynamic treatment regimen. *Crit Care.* (2021) 25:243. doi: 10.1186/s13054-021-03682-7

19. Chen T, Guestrin C. *Xgboost: A Scalable Tree Boosting System. Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. San Francisco, USA: Association for Computing Machinery (2016). p. 785–94.

20. Lundberg SM, Lee S-I. A unified approach to interpreting model predictions. In: *Proceedings of the 31st International Conference on Neural Information Processing Systems*. Long Beach, California: Curran Associates Inc. (2017). p. 4768–77.

21. Cortes C, Vapnik V. Support-vector networks. *Mach Learn.* (1995) 20:273–97. doi: 10.1007/BF00994018

22. Ke G, Meng Q, Finley T, Wang T, Chen W, Ma W, et al. Lightgbm: A highly efficient gradient boosting decision tree. In: *Proceedings of the 31st International Conference on Neural Information Processing Systems*. Long Beach, California: Curran Associates Inc. (2017). p. 3149–9657.

23. Prokhorenkova L, Gusev G, Vorobev A, Dorogush AV, Gulin A. Catboost: Unbiased Boosting with Categorical Features. In: *Proceedings of the 32nd International Conference on Neural Information Processing Systems; Montréal, Canada: Curran Associates Inc*. (2018). p. 6639–49.

24. Breiman L. Random forests. *Mach Learn.* (2001) 45:5–32. doi: 10.1023/A:1010933404324

25. Ribeiro MT, Singh S, Guestrin C. "Why should I trust you?": explaining the predictions of any classifier. In: *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining; San Francisco, California, USA: Association for Computing Machinery*. (2016). p. 1135–44.

26. Shrikumar A, Greenside P, Shcherbina A, Kundaje A. Learning important features through propagating activation differences. In: *Proceedings of the 34th International Conference on Machine Learning*. Sydney: JMLR.org (2017). p. 3145–53.

27. Datta A, Sen S, Zick Y. Algorithmic transparency via quantitative input influence: theory and experiments with learning systems. In: *IEEE Symposium on Security and Privacy (SP)*. San Jose (2016). p. 598–617.

28. Bengoechea JA, Sa Pessoa J. Klebsiella pneumoniae infection biology: living to counteract host defences. *FEMS Microbiol Rev.* (2019) 43:123–44. doi: 10.1093/femsre/fuy043

29. Hernandez-Jimenez P, Lopez-Medrano F, Fernandez-Ruiz M, Silva JT, Corbella L, San-Juan R, et al. Derivation of a score to predict infection due to multidrug-resistant pseudomonas aeruginosa: a tool for guiding empirical antibiotic treatment. *J Glob Antimicrob Resist.* (2022) 29:215–21. doi: 10.1016/j.jgar.2022.03.014

30. Saedder EA, Brock B, Nielsen LP, Bonnerup DK, Lisby M. Identifying high-risk medication: a systematic literature review. *Eur J Clin Pharmacol.* (2014) 70:637–45. doi: 10.1007/s00228-014-1668-z

31. Johnson AEW, Pollard TJ, Shen L, Lehman L-WH, Feng M, Ghassemi M, et al. Mimic-Iii, a freely accessible critical care database. *Sci Data.* (2016) 3:160035. doi: 10.1038/sdata.2016.35

32. Xu P, Chen L, Zhu Y, Yu S, Chen R, Huang W, et al. Critical care database comprising patients with infection. *Front Public Health.* (2022) 10:852410. doi: 10.3389/fpubh.2022.852410

frontiers | Frontiers in Medicine

TYPE Original Research
PUBLISHED 18 August 2022
DOI 10.3389/fmed.2022.853989



OPEN ACCESS

EDITED BY
Nan Liu,
National University of Singapore, Singapore

REVIEWED BY
Ashraf Roshdy,
Alexandria University, Egypt
Siqi Li,
Duke-NUS Medical School, Singapore

*CORRESPONDENCE
Zhenjie Hu
syicu@vip.sina.com

SPECIALTY SECTION
This article was submitted to Intensive Care Medicine and Anesthesiology, a section of the journal Frontiers in Medicine

RECEIVED 24 January 2022
ACCEPTED 04 August 2022
PUBLISHED 18 August 2022

CITATION
Li B, Huo Y, Zhang K, Chang L, Zhang H, Wang X, Li L and Hu Z (2022) Development and validation of outcome prediction models for acute kidney injury patients undergoing continuous renal replacement therapy. *Front. Med.* 9:853989. doi: 10.3389/fmed.2022.853989



# Development and validation of outcome prediction models for acute kidney injury patients undergoing continuous renal replacement therapy

Bo Li, Yan Huo, Kun Zhang, Limin Chang, Haohua Zhang, Xinrui Wang, Leying Li and Zhenjie Hu*

Intensive Care Unit, Hebei Medical University Fourth Affiliated Hospital and Hebei Provincial Tumor Hospital, Shijiazhuang, China

**Object:** This study aimed to develop and validate a set of practical predictive tools that reliably estimate the 28-day prognosis of acute kidney injury patients undergoing continuous renal replacement therapy.

**Methods:** The clinical data of acute kidney injury patients undergoing continuous renal replacement therapy were extracted from the Medical Information Mart for Intensive Care IV database with structured query language and used as the development cohort. An all-subset regression was used for the model screening. Predictive models were constructed *via* a logistic regression, and external validation of the models was performed using independent external data.

**Results:** Clinical prediction models were developed with clinical data from 1,148 patients and validated with data from 121 patients. The predictive model based on seven predictors (age, vasopressor use, red cell volume distribution width, lactate, white blood cell count, platelet count, and phosphate) exhibited good predictive performance, as indicated by a C-index of 0.812 in the development cohort, 0.811 in the internal validation cohort and 0.768 in the external validation cohort.

**Conclusions:** The model reliably predicted the 28-day prognosis of acute kidney injury patients undergoing continuous renal replacement therapy. The predictive items are readily available, and the web-based prognostic calculator (https://libo220284.shinyapps.io/DynNomapp/) can be used as an adjunctive tool to support the management of patients.

KEYWORDS

**acute kidney injury, continuous renal replacement therapy, prediction model, nomogram, validation**

# 1. Introduction

Acute kidney injury (AKI) is a critical comorbidity and a global health problem with high morbidity and high mortality (1–3). In the intensive care unit (ICU), the morbidity can be as high as 50% (4). Since there are no specific drugs for AKI, renal replacement therapy (RRT) plays a major role in treatment (2). Although there is currently no evidence that continuous RRT (CRRT) is superior to intermittent RRT (IRRT) (5–7), CRRT is often preferred for hemodynamically unstable patients (2). However, among these patients, even with appropriate CRRT, there is still very high mortality (8), and the cost of treatment is often high. Thus, it is important to develop reliable tools that can inform expectations regarding outcomes and decisions regarding treatment.

Clinical predictive models can estimate the probability of a patient's outcome through the statistical implementation of a series of clinical characteristics of the patient, and may be helpful for patient management as a decision support tool (9). Currently, the most widely used outcome prediction models in the ICU are the Acute Physiology and Chronic Health Evaluation II (APACHE II) classification system (10) and the Sepsis-related Organ Failure Assessment (SOFA) score (11). However, these models do not focus on outcome prediction in AKI patients undergoing CRRT. Several prediction models have been published (12, 13), but there are some limitations in clinical practice, such as improper variable selection strategies, difficulty of use in clinical settings and a lack of generalizability to different settings. Therefore, there is an urgent need develop an easy-to-use predictive tool that supports clinical decision-making.

We developed and validated outcome prediction models of AKI patients treated with CRRT.

# 2. Methods

## 2.1. Data source

The development cohort included 1148 patients who were recruited from Medical Information Mart for Intensive Care IV (MIMIC IV version 1.0) (14, 15). MIMIC IV is a relational database containing the real information of patients admitted to the ICUs of Beth Israel Deaconess Medical Center in Boston, MA, USA, from 2008 to 2019. The principal investigator completed the Human Research Course (Record ID: 37097306) and obtained access to this database, and the project was approved by the institutional review boards of the Computational Physiology Laboratory of the Massachusetts Institute of Technology and Beth Israel Deaconess Medical Center and was granted a waiver of informed consent. All data were extracted with structured query language (SQL) from BigQuery.

The validation cohort included 121 patients treated in the Department of Intensive Care Unit, Fourth Hospital of Hebei Medical University, Shijiazhuang, China. This study was approved by the Ethics Committee of the Fourth Hospital of Hebei Medical University (approval number: 2021KS034).

## 2.2. Patient involvement

The inclusion criteria in this study were as follows: (1) AKI patients meeting the KDIGO-AKI criteria; and (2) patients who received CRRT after diagnosis. Patients younger than 18 years were excluded, and when the same patient were admitted multiple times, only data for the first admission was included. In addition, in the validation cohort, patients whose family members voluntarily stopped treatment within 24 h after receiving CRRT were also excluded.

## 2.3. Diagnosis and outcomes

AKI was defined as any of the following Kidney Disease Improving Global Outcomes (KDIGO) criteria (16): increase in SCr $\geq 0.3 mg/dl$ ($\geq 26.5 mol/l$) within 48 h; increase in SCr $\geq 1.5$ times baseline, which is known or presumed to have occurred within the prior 7 days; or a urine volume $< 0.5 ml/kg/h$ for 6 h.

The primary outcome was defined as death within 28 days after receiving CRRT. Patients in the validation cohort whose family members voluntarily stopped treatment for more than 24 h were considered dead.

## 2.4. Variable extraction

The following variables were extracted from the relevant literature and clinical records:

**Demographic characteristics:** Age (17–21), sex (20, 21), height, and weight (21).

**Comorbidities:** Congestive heart failure (CHF) (18), atrial fibrillation (AF), chronic liver disease (CLD), chronic obstructive pulmonary disease (COPD), chronic coronary syndrome (CCS) (18), hypertension, diabetes, and malignant cancer (18, 19).

**Last vital signs within 2 h prior to receiving CRRT:** Heart rate (HR) (18), mean arterial pressure (MAP) (18, 21), and temperature (T).

**Results of the last laboratory test within 24 h prior to receiving CRRT:** White blood cell count (WBC), hemoglobin (HB) (17, 20), red cell volume distribution width (RDW) (22), platelet count (PLT) (18, 20, 21), sodium (20), potassium (20), calcium, phosphate (18, 23, 24), total bilirubin (TBIL) (18, 20, 21), albumin (18, 20, 21), creatinine (18, 21), baseline creatinine (20, 21), pH (17, 20), oxygenation index (21), base

excess (20), and lactate (20, 21). Oxygenation index is calculated by equation $PaO_2/FiO_2$.

**Interventions 24 h prior to receiving CRRT:** Mechanical ventilation (18, 20, 21), vasopressor use (20, 21), sedative use, and analgesic use.

Central venous pressure (CVP) (missing rate: 74.7%), mean platelet volume (25) (missing rate: 100%), troponin (missing rate: 73.9%), N-terminal pro B type natriuretic peptide (NT-proBNP) (missing rate: 97.4%), and creatine kinase (missing rate: 70.2%) were not extracted due to excessive amounts of missing data (missing rate > 50%), and there appears to be no evidence of their relationship with prognosis in this group of patients.

## 2.5. Handling of missing data

In the development cohort, there were missing data for most variables. Variables with excessive amounts of missing data were excluded. We assumed that the data were missing at random and filled in missing data using multiple imputation with chained equations. We performed fifty multiple imputations and merged the dataset into the development dataset. All analyses were performed with R software (version 4.1.1; R Foundation for Statistical Computing).

## 2.6. Model development

We used a Q-Q plot to assess the normality of the continuous variables, and cubic spline functions were used to assess the linearity of the relationship. Continuous variables that did not conform to normal or linear distributions were converted to categorical covariates based on their clinical significance. The continuous variables are expressed as the mean (standard deviation), and the categorical covariates are reported as numbers and percentages.

All variables were included in the logistic regression model, and we added an interaction term between mechanical ventilation and oxygenation index. The variables were screened using an all-subset regression, with the best model judged by adjusting the r-squared and Bayesian information criterion (BIC). The screened models were tested for multicollinearity by calculating the variance inflation factor (VIF).

Finally, we used the best model to construct a nomogram that could provide clinicians with an intuitive and quantitative tool for predicting the outcomes of AKI patients undergoing CRRT.

## 2.7. Model validation

The model discrimination was evaluated with the C-index and area under the receiver operator characteristic curve (AUC). The model calibration was evaluated with Brier scores and calibration plots. Decision curve analysis (DCA) curves were used to assess the clinical applicability of the model (26, 27).

Internal validation was performed with the enhanced bootstrap technique, in which regression models were fitted in 1,000 bootstrap replicates, drawn with replacement from the development cohort. The model was refitted in each bootstrap replicate and tested using the original sample to estimate optimism in the model performance. External validation was performed with the validation cohort.

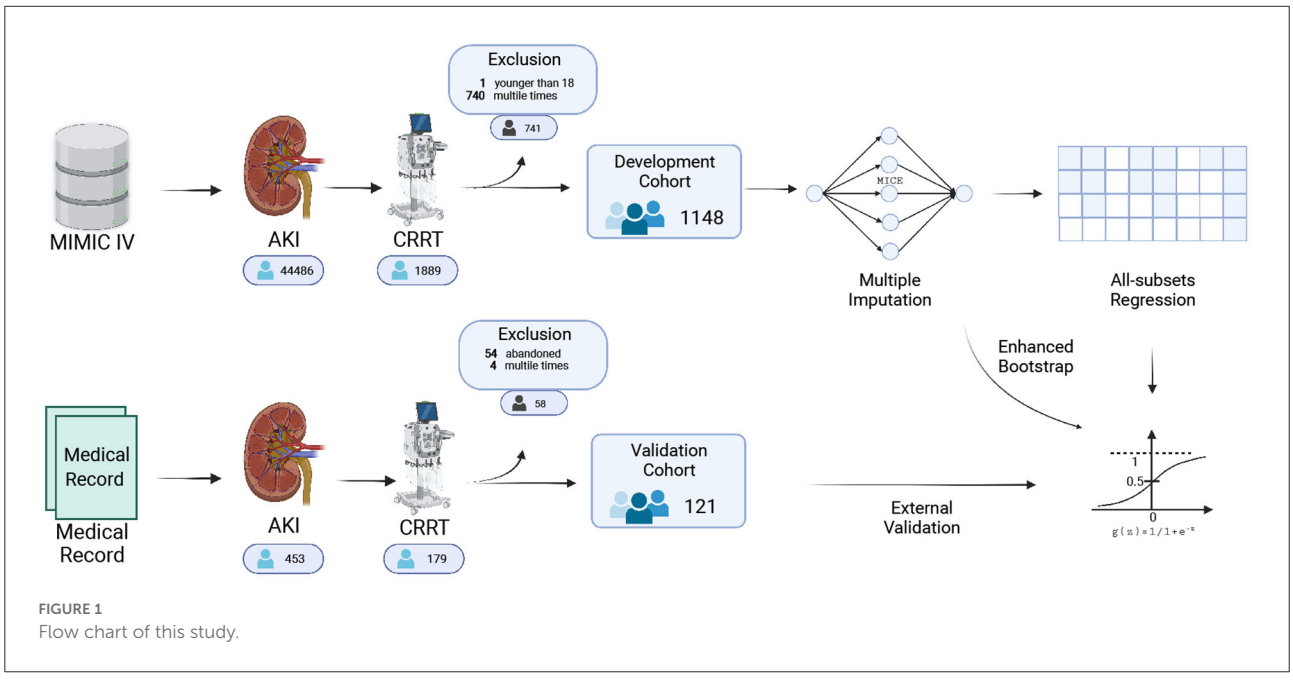



FIGURE 1
Flow chart of this study.

TABLE 1 Clinical characteristics of the development cohort.

| | Overall | Survival | Death |
|---|---|---|---|
| *N* | 1,148*50 | 662*50 | 486*50 |
| Sex (male) [*n*(%)] | 34,800 (60.6) | 19,950 (60.3) | 14,850 (61.1) |
| Age [mean (SD)] | 63.17 (14.61) | 62.60 (14.68) | 63.95 (14.48) |
| BMI [mean (SD)] | 31.37 (8.28) | 31.02 (8.06) | 31.84 (8.56) |
| CHF [*n* (%)] | 12,150 (21.2) | 7,150 (21.6) | 5,000 (20.6) |
| AF [*n* (%)] | 16,200 (28.2) | 9,850 (29.8) | 6,350 (26.1) |
| CLD [*n* (%)] | 9,350 (16.3) | 4,550 (13.7) | 4,800 (19.8) |
| COPD [*n* (%)] | 7,600 (13.2) | 4,300 (13.0) | 3,300 (13.6) |
| CCS [*n* (%)] | 16,200 (28.2) | 9,900 (29.9) | 6,300 (25.9) |
| Hypertension [*n* (%)] | 6350 (11.1) | 3,100 (9.4) | 3,250 (13.4) |
| Diabetes [*n* (%)] | 23,500 (40.9) | 14,950 (45.2) | 8,550 (35.2) |
| Malignant cancer [*n* (%)] | 7,650 (13.3) | 3600 (10.9) | 4,050 (16.7) |
| MAP (*mmHg*) [mean (SD)] | 73.71 (13.79) | 75.83 (14.13) | 70.83 (12.76) |
| HR (*bpm*) [mean (SD)] | 88.66 (19.44) | 85.79 (19.17) | 92.56 (19.11) |
| Temperature (°C) [*n* (%)] | | | |
| <36.0 | 7,210 (12.6) | 3515 (10.6) | 3695 (15.2) |
| [36.0, 37.5] | 41,331 (72.0) | 24,776 (74.9) | 16,555 (68.1) |
| [37.5, 38.0] | 4,330 (7.5) | 2,368 (7.2) | 1,962 (8.1) |
| ≥38.0 | 4,529 (7.9) | 2441 (7.4) | 2,088 (8.6) |
| WBC ($*10^9/L$) [n(%)] | | | |
| <4.0 | 2,535 (4.4) | 1,120 (3.4) | 1,415 (5.8) |
| [4.0, 10.0] | 17,198 (30.0) | 12,220 (36.9) | 4,978 (20.5) |
| [10.0, 40.0] | 36,139 (63.0) | 19,188 (58.0) | 16,951 (69.8) |
| ≥40.0 | 1528 (2.7) | 572 (1.7) | 956 (3.9) |
| Hemoglobin ($g/dL$) [mean (SD)] | 9.29 (1.75) | 9.31 (1.72) | 9.26 (1.78) |
| RDW (%) [mean (SD)] | 17.05 (2.67) | 16.67 (2.34) | 17.57 (2.99) |
| PLT ($*10^9/L$) [n(%)] | | | |
| >150 | 24,543 (42.8) | 16,170 (48.9) | 8,373 (34.5) |
| ≤150 | 11,659 (20.3) | 7,170 (21.7) | 4,489 (18.5) |
| ≤100 | 14,989 (26.1) | 7,372 (22.3) | 7,617 (31.3) |
| ≤50 | 6209 (10.8) | 2,388 (7.2) | 3821 (15.7) |
| Sodium ($mmol/L$) [n(%)] | | | |
| <135.0 | 19,410 (33.8) | 11,644 (35.2) | 7,766 (32.0) |
| [135.0,145.0] | 33,517 (58.4) | 20,001 (60.4) | 13,516 (55.6) |
| >145.0 | 4473 (7.8) | 1455 (4.4) | 3,018 (12.4) |
| Potassium ($mmol/L$) [mean (SD)] | 4.77 (0.97) | 4.70 (0.94) | 4.87 (1.00) |
| Calcium ($mmol/L$) [n(%)] | | | |
| <2.25 | 45,339 (79.0) | 26,116 (78.9) | 19,223 (79.1) |
| [2.25, 2.75] | 11,499 (20.0) | 6,674 (20.2) | 4,825 (19.9) |
| >2.75 | 562 (1.0) | 310 (0.9) | 252 (1.0) |
| Phosphate ($mmol/L$) [mean (SD)] | 2.08 (0.79) | 1.95 (0.75) | 2.26 (0.82) |
| Total bilirubin ≥ 17.1 $\mu mol/L$ [n(%)] | 36,292 (63.2) | 19,085 (57.7) | 17,207 (70.8) |
| Albumin ($g/dL$) [mean(SD)] | 2.91 (0.73) | 2.98 (0.71) | 2.82 (0.74) |
| Creatinine/Baseline creatinine [n(%)] | | | |
| <1.5 | 4,578 (8.0) | 3,363 (10.2) | 1,215 (5.0) |
| ≥1.5 | 5,527 (9.6) | 3,653 (11.0) | 1,874 (7.7) |
| ≥2.0 | 12,982 (22.6) | 7,150 (21.6) | 5,832 (24.0) |
| ≥3.0 | 34,313 (59.8) | 18,934 (57.2) | 15,379 (63.3) |
| pH [mean (SD)] | 7.31 (0.11) | 7.34 (0.10) | 7.28 (0.12) |
| Oxygenation index [*n* (%)] | | | |
| ≤100 | 6,409 (11.2) | 3,086 (9.3) | 3323 (13.7) |
| [100, 200] | 22,440 (39.1) | 12,171 (36.8) | 10,269 (42.3) |
| [200, 300] | 18,407 (32.1) | 11,345 (34.3) | 7,062 (29.1) |
| >300 | 10,144 (17.7) | 6498 (19.6) | 3646 (15.0) |
| Base excess ($mmol/L$) [mean (SD)] | −5.48 (6.38) | −3.89 (5.67) | −7.64 (6.65) |
| Lactate ($mmol/L$) [mean (SD)] | 3.93 (4.27) | 2.54 (2.61) | 5.83 (5.25) |
| Mechanical ventilation use [*n* (%)] | 18,250 (31.8) | 9,500 (28.7) | 8,750 (36.0) |
| Vasopressor use [*n* (%)] | 34,900 (60.8) | 15,200 (45.9) | 19,700 (81.1) |
| Sedative use [*n* (%)] | 38,800 (67.6) | 20,600 (62.2) | 18,200 (74.9) |
| Analgesic use [*n* (%)] | 42,900 (74.7) | 22,700 (68.6) | 20,200 (83.1) |

*BMI, body mass index; CHF, congestive heart failure; AF, atrial fibrillation; CLD, chronic liver disease; COPD, chronic obstructive pulmonary disease; CCS, chronic coronary syndromes; MAP, mean arterial pressure; HR, heart rate; WBC, white blood cells count; RDW, red cell volume distribution width; PLT, platelet count.

# 3. Results

## 3.1. Model development

In total 1,148 patients from the MIMIC IV database were eventually included in our study (Figure 1). The 50 datasets obtained by multiple imputation techniques were merged into the final development cohort (Table 1). The best models were screened by adjusting the r-squared value and BIC (Figure 2).

The VIFs of the screened variables were all <5. Seven variables (age, vasopressor use, RDW, lactate, WBC, PLT, and phosphate) were finally included in our model, which was used to plot the nomogram (Figure 3) and make the web-based prognostic calculator (Figure 4, https://libo220284.shinyapps.io/DynNomapp/).

The predictive performance of our model as measured by the C-index was 0.812 (Table 2 and Figure 5A) in the development cohort, indicating that the model had relatively good discriminative capacity. Our model showed high agreement between the actual and predicted probabilities in the development cohort, with a Brier score of 0.173 (Table 2 and Figure 5B). In addition, the DCA curve demonstrated that our model was clinically useful in the development cohort (Figures 5C,D).

## 3.2. Internal validation

Our model also achieved good internal validation performance after 1,000 bootstrap replicates, with a C-index of 0.811 and a Brier score of 0.173 (Table 2).

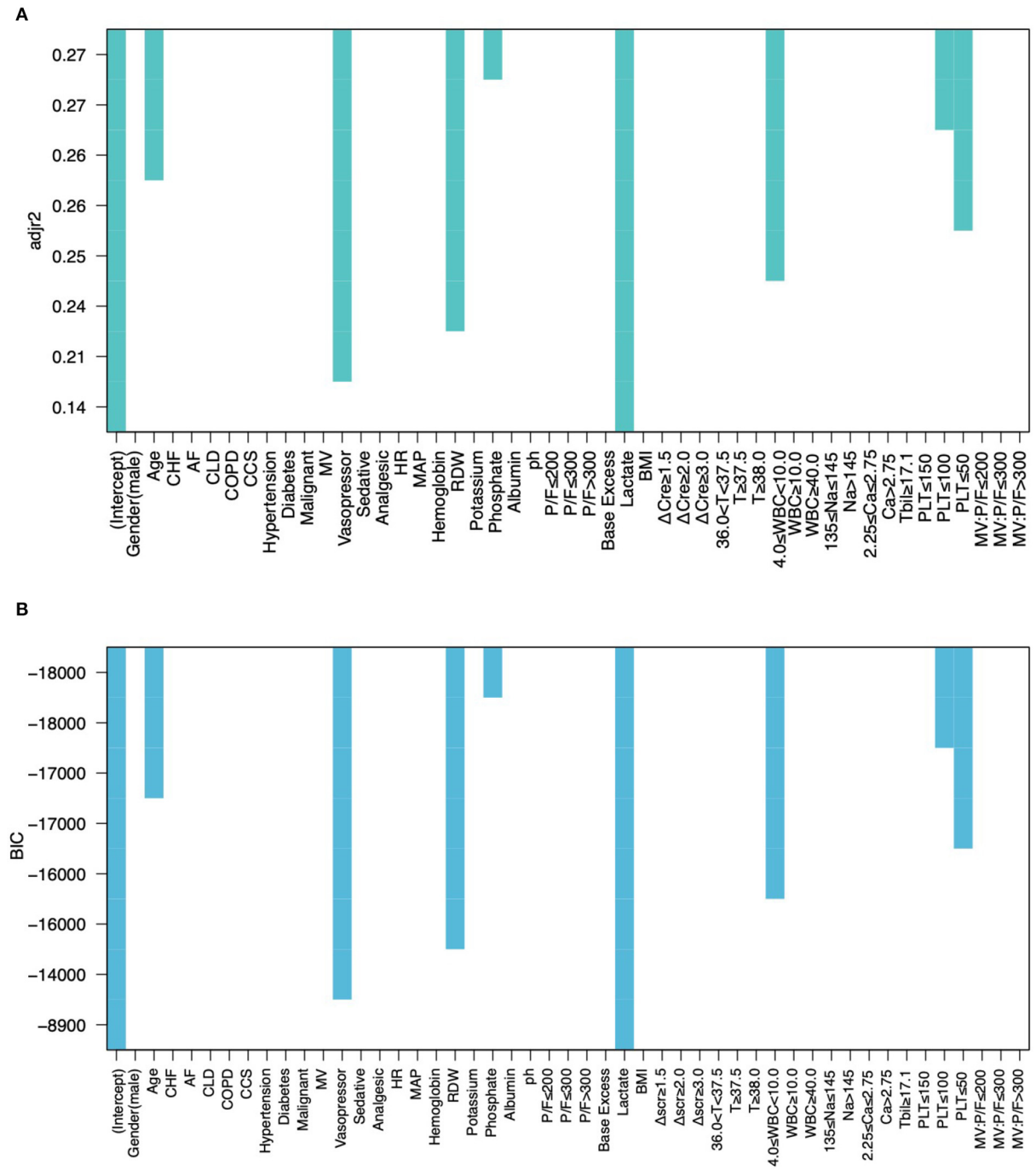



FIGURE 2
All-subsets regression by adjusting the r-squared value **(A)** and BIC **(B)**.

## 3.3. External validation

In total 121 patients were eventually included in the external validation cohort (Table 3 and Figure 1). The predictive performance of the nomogram as measured by the C-index was 0.768 (Table 2 and Figure 6A) in the external validation cohort, indicating that the model had relatively good discriminative capacity and generalizability in different settings. The nomogram also showed acceptable agreement between the actual and predicted probabilities in the external validation cohort, with a Brier score of 0.202 (Table 2 and Figure 6B). In addition, the DCA curve demonstrated

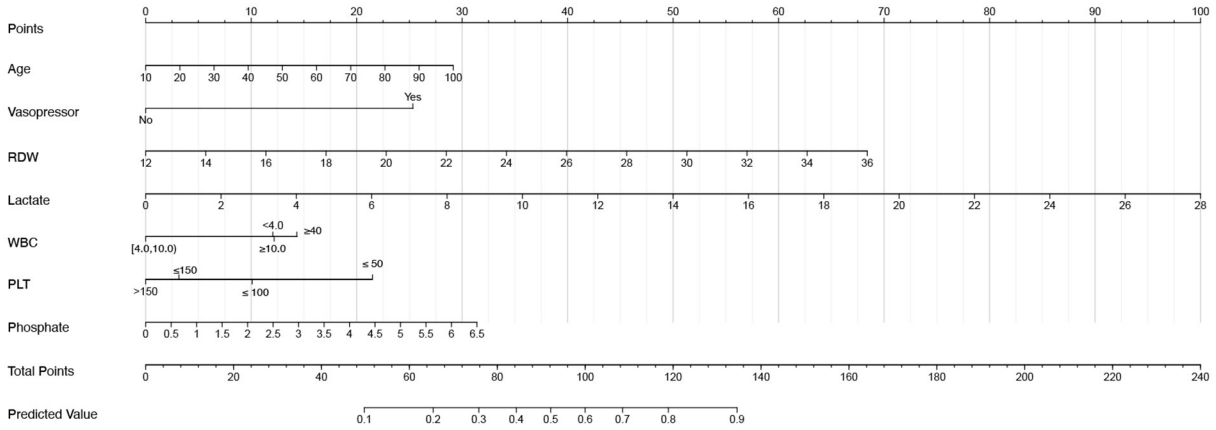



FIGURE 3
The nomogram for acute kidney injury patients undergoing continuous renal replacement therapy.

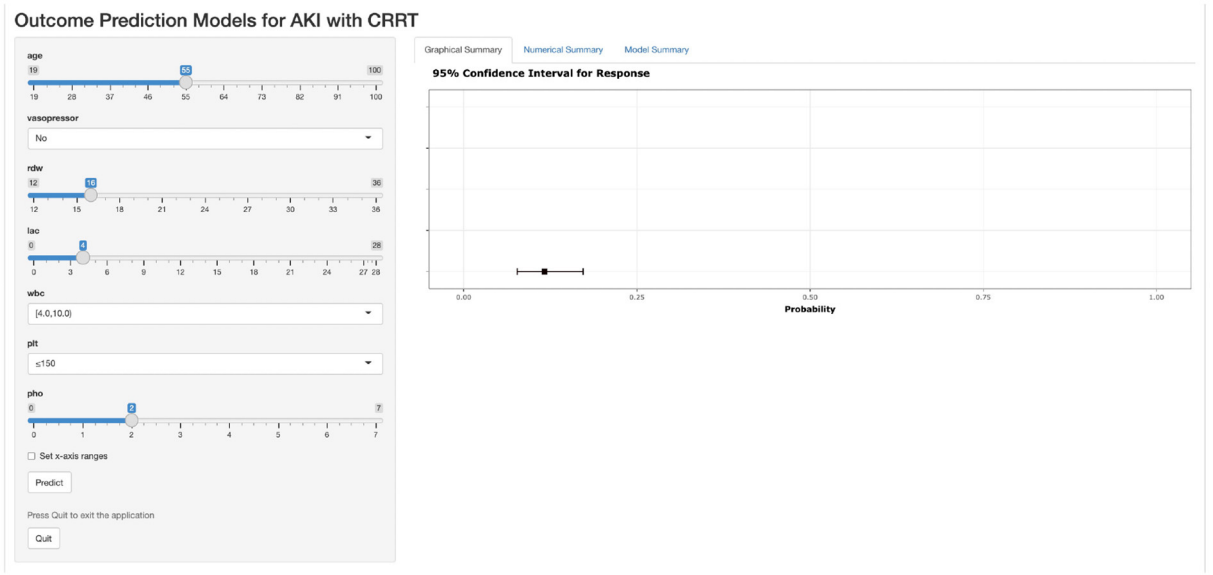



FIGURE 4
The web-based prognostic calculator.

that our model was clinically useful in different settings (Figures 6C,D).

## 4. Discussion

AKI is common in the ICU, and although a subset of small studies has shown that preventive measures, and the rapid identification of AKI can lead to improved outcomes (28–30), patients entering the ICU often already have AKI, thus in clinical practice in the ICU, ICU physicians tend to focus more on the treatment and prognosis of AKI than on the prevention and diagnosis of AKI.

TABLE 2 The performance in model development, internal validation, and external validation.

| | C-index | Brier score |
|---|---|---|
| Development | 0.812 | 0.173 |
| Internal validation | 0.811 | 0.173 |
| External validation | 0.768 | 0.202 |

CRRT plays an important role in the management of AKI in the ICU. Since not all patients with AKI ultimately benefit from CRRT, patients, their relatives and clinicians need reliable

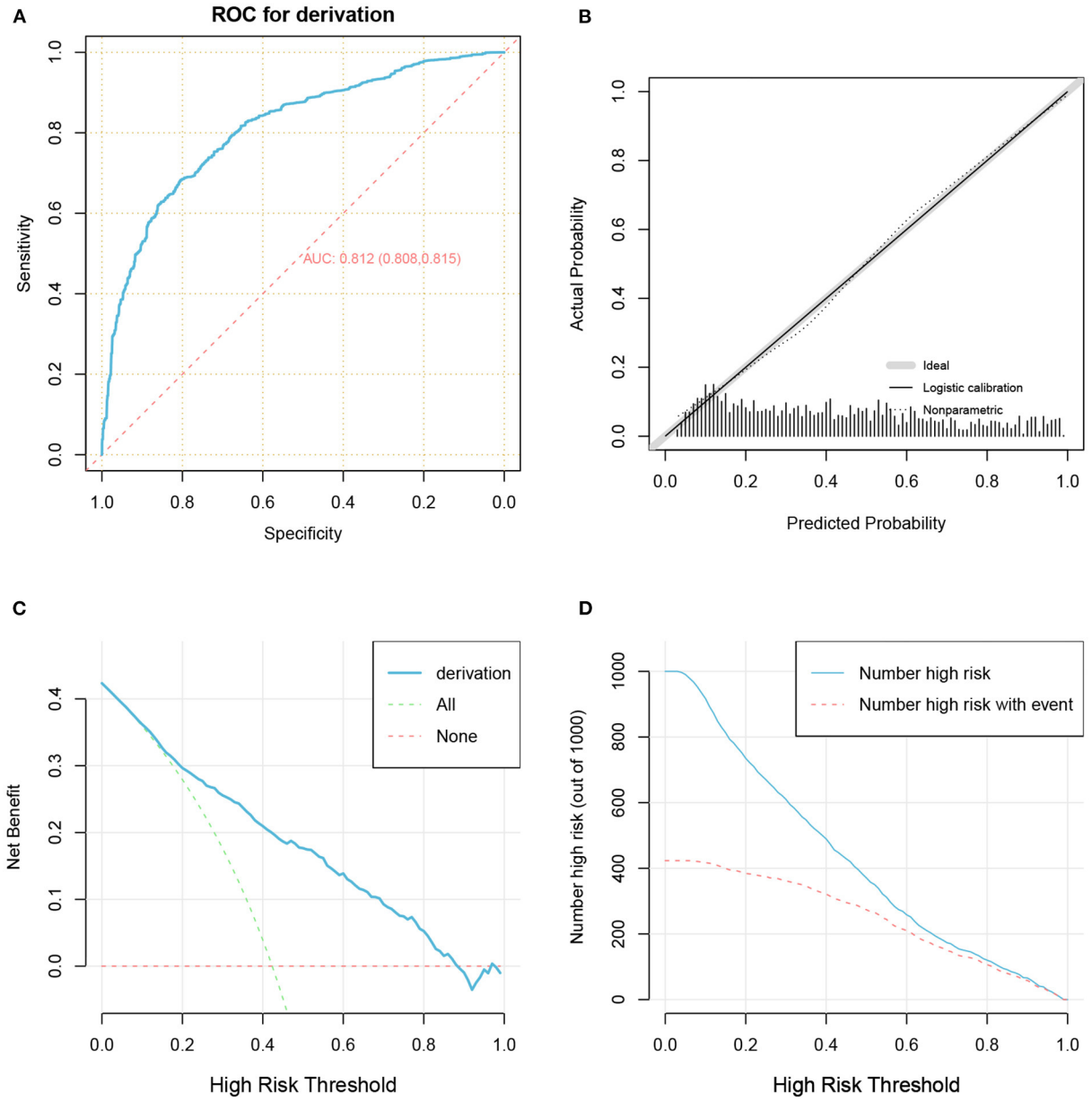



FIGURE 5
The receiver operator characteristic curve **(A)**, calibration plots **(B)**, decision curve analysis curves **(C)**, and clinical impact curve **(D)** for model in the development cohort.

information regarding prognosis such that they can effectively participate in shared decision-making. This is important because they are unlikely to rely solely on clinician experience and intuition when making treatment decisions.

With the widespread use of electronic medical record systems in clinical settings, "big data" and clinical medicine are becoming inseparable. From the perspectives of volume, speed, and diversity, the ICU is a wonderful combination of "big data" and clinical medicine (31). In such an era of big data, the organic combination of medical informatics and big data analytics provides a fertile new ground for analyzing the management of AKI (32, 33). Prediction tools provide an opportunity to improve AKI management in the era of big data.

Numerous predictive models of acute kidney injury are available (34), but few models are available for patients with AKI who are receiving CRRT (18, 35–37). Therefore, we aimed to obtain a reliable tool to predict the 28-day mortality in this group of patients. It is essential to clarify that although the use

TABLE 3 The predictor items of external validation cohort.

| | Overall | Survival | Death |
|---|---|---|---|
| *N* | 121 | 63 | 58 |
| Age [mean (SD)] | 62.56 (14.79) | 57.32 (14.18) | 68.26 (13.36) |
| Vasopressor use [*n* (%)] | 88 (72.7) | 36 (57.1) | 52 (89.7) |
| RDW (%) [median [IQR]] | 14.90 [13.90, 16.20] | 14.50 [13.70, 15.35] | 15.60 [14.60, 17.12] |
| Lactate (*mmol/L*) [mean (SD)] | 3.80 (3.83) | 2.89 (2.43) | 4.79 (4.74) |
| WBC ($*10^9/L$) [*n* (%)] | | | |
| <4.0 | 4 (3.3) | 2 (3.2) | 2 (3.4) |
| [4.0, 10.0] | 28 (23.1) | 17 (27.0) | 11 (19.0) |
| [10.0, 40.0] | 87 (71.9) | 42 (66.7) | 45 (77.6) |
| ≥40.0 | 2 (1.7) | 2 (3.2) | 0 (0.0) |
| PLT ($*10^9/L$) [*n* (%)] | | | |
| >150 | 53 (43.8) | 30 (47.6) | 23 (39.7) |
| ≤150 | 24 (19.8) | 11 (17.5) | 13 (22.4) |
| ≤100 | 23 (19.0) | 13 (20.6) | 10 (17.2) |
| ≤50 | 21 (17.4) | 9 (14.3) | 12 (20.7) |
| Phosphate (*mmol/L*) [mean (SD)] | 1.69 (0.70) | 1.59 (0.68) | 1.81 (0.71) |

*RDW, red cell volume distribution width; WBC, white blood cells count; PLT, platelet count.

of Major Adverse Kidney Events (MAKE) has been suggested as a composite endpoint for such studies (38). Such a composite endpoint was also used in the SEA-MAKE score developed by Sukmark et al. (39). Twenty-eight day mortality was chosen as the single endpoint in this study. The primary considerations are as follows: First, the significant advantage of the composite endpoints is that it increases the number of events, but in patients with AKI undergoing CRRT, mortality would have been high enough and a better solution might have been to use a multivariate outcome with different outcomes, but due to the limitations of the study, this issue needs to be considered in future studies. Second, we did not know which predictors contributed to each component of the composite outcomes. Finally, even with the current definition of MAKE, death is still the most serious and important outcome of a concern. Therefore, mortality was ultimately chosen as the outcome variable in this study.

Ultimately, the prediction models performed robustly in a validation cohort from different geographical regions, time periods, and settings of care. The predictors in our model are readily available, and the nomogram and web-based prognostic calculator could facilitate clinical adoption.

## 4.1. Comparison with previous studies

Several prediction models of the outcome of AKI patients with CRRT have been developed, although their clinical use is rare.

Kim et al. (12) developed the MOSAIC model for patients with AKI undergoing CRRT. Unfortunately, this model only incorporated APACHE II outcomes and SOFA scores, and although these data were extremely accessible, they did not consider several other indicators that have predictive value and are readily available. A study by Oh et al. (22) showed that RDW was an independent predictor of the 28-day mortality in patients with AKI receiving CRRT. Phosphate reflected disease severity and predicted mortality in AKI patients undergoing CRRT in the studies by Jung et al. (23, 24). Both RDW and phosphate were included in our study. In addition, we considered additional comorbidities and laboratory indicators.

Machine learning algorithms have also been applied to predict outcomes in AKI patients undergoing CRRT (13). Machine learning algorithms appear to provide better predictive performance than traditional models, but their hard-to-interpret nature may also lead to overestimation of model accuracy and exaggeration of actual performance (40). We chose the more robust logistic regression model in our study. Our model did not perform worse than machine learning algorithms.

The HELENICC score is an excellent model for predicting mortality in patients with sepsis-related AKI undergoing CRRT (41), but not all patients with AKI undergoing CRRT have sepsis, and we hope that our model will be useful for clinical decision making in a larger number of patients with AKI.

The greatest advantage of this study over previous studies is that the external validation was based on completely independent data, and good model performance was achieved. This finding demonstrates the good generalizability of our model.

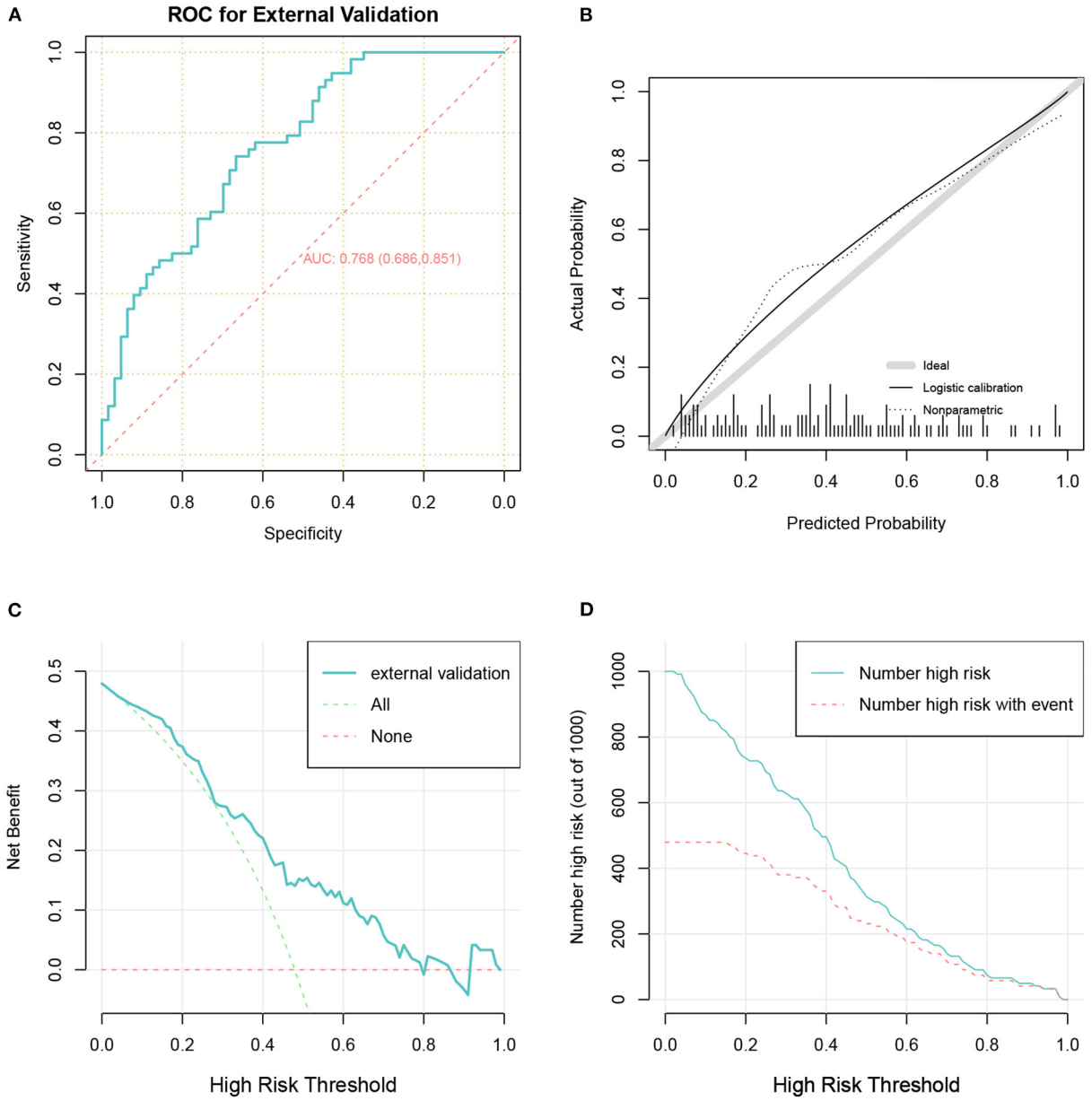



FIGURE 6
The receiver operator characteristic curve **(A)**, calibration plots **(B)**, decision curve analysis curves **(C)**, and clinical impact curve **(D)** for model in the validation cohort.

## 4.2. Implications for clinical practice

As statistician Professor Efron stated, in the absence of genius-level insight, statistical estimation theory is intended as an instrument for peering through noisy data and discerning a smooth underlying truth (42). Our models are not solely designed to predict patient outcomes but to offer new possibilities for clinicians and patient families to participate in shared decision-making regarding patient care.

We were able to quickly assess the risk of patient death with the nomogram and web-based prognostic calculator in this study, but some challenges exist.

On the one hand, although ICU physicians readily accept data-driven advice in their interactions with smart devices and the Internet, they remain cautious regarding the advice such technology provides in clinical decision-making (43). Even when models conclude that some AKI patients will not be able to reverse their deterioration even with CRRT, ICU physicians

still prefer to treat them to the fullest. Physicians are always concerned that they are doing too little, and sometimes they are willing to do more than resuscitation interventions knowing that a treatment does not fundamentally change the patient's outcome (44). Using such technologies in clinical work must provide actionable information for the right patient at the right time. For example, outcomes can be predictive information to help clinicians make clinical decisions with some basis of reference. In addition, many factors that influence clinical decisions, including clinical, social, and personal factors, are not necessarily reflected in the digital record, thus any predictive results need to be evaluated, interpreted, and fleshed out by the clinician before any action is taken. Therefore, it is still the clinician who makes the final decision. Of course, this also requires critical care physicians to have some ability to interpret and use these results (43).

On the other hand, no medical practice is immune to ethical considerations, and the application of these technologies to the management of critically ill patients is fundamentally a medical practice for patients. This also requires compliance with medical ethical requirements.

It is important to emphasize that the inappropriate use of these technologies can cause harm to patients (45). Therefore, we must be cautious and ensure that it can be reasonably and safely tested and used in critically ill patients (46).

### 4.3. Weaknesses of the study

There are potential limitations in our study.

First, missing data are unavoidable in retrospective studies. Rather than excluding all patients with missing data from the analysis, we used multiple imputation to reduce the impact of missing data. With theoretical and empirical evidence of the technique's superiority to traditional complete case analysis, multiple interpolation has become widely accepted and is increasingly used (47, 48).

Second, because our development cohort was derived from the MIMIC-IV database, variables with significant predictive value that are easily accessible, such as the mean platelet volume and some widely reported biomarkers, were not included in our study. Han et al. (25) showed that the mean platelet volume may be an inexpensive and useful predictor of the 28-day all-cause mortality in AKI patients requiring CRRT. The predictive value of biomarkers such as tissue inhibitor metalloproteinase-2 (TIMP-2), insulin-like growth factor-binding protein 7 (IGFBP7) and neutrophil gelatinase-associated lipocalin (NGAL) has also been widely reported (49, 50). Unfortunately, these variables were not available in the MIMIC-IV database. These variables may need to be considered in future model updates.

Finally, our model seems to underestimate the mortality rate of patients. However, the performance during model development, internal validation, and external validation was in the acceptable range. Importantly, our validation cohort was completely independent of the development cohort in both time and space.

## 5. Conclusion

The prediction model we developed based on data from 1,148 patients from the MIMIC IV database reliably estimated outcomes in a fully independent validation cohort containing data from 121 patients. The predictor items are readily available, and the nomogram and the web-based prognostic calculator offer new possibilities for shared clinical decision-making between clinicians and patient families.

## Data availability statement

The original contributions presented in the study are included in the article/Supplementary material, further inquiries can be directed to the corresponding author/s.

## Ethics statement

The studies involving human participants were reviewed and approved by the Ethics Committee of the Fourth Hospital of Hebei Medical University. Written informed consent for participation was not required for this study in accordance with the national legislation and the institutional requirements.

## Author contributions

Study design: BL, YH, KZ, and ZH. Data collection: BL, LC, HZ, XW, and LL. Data analysis and drafting of the manuscript: BL. Data interpretation: BL, YH, and KZ. Revising the manuscript content: YH and KZ. Approving the final version of the manuscript: ZH. All authors contributed to the article and approved the submitted version.

## Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

## Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those

of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

## Supplementary material

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fmed.2022.853989/full#supplementary-material

## References


1. Hoste EAJ, Kellum JA, Selby NM, Zarbock A, Palevsky PM, Bagshaw SM, et al. Global epidemiology and outcomes of acute kidney injury. *Nat Rev Nephrol*. (2018) 14:607–25. doi: 10.1038/s41581-018-0052-0

2. Negi S, Koreeda D, Kobayashi S, Yano T, Tatsuta K, Mima T, et al. Acute kidney injury: epidemiology, outcomes, complications, and therapeutic strategies. *Semin Dial*. (2018) 31:519–27. doi: 10.1111/sdi.12705

3. Ronco C, Bellomo R, Kellum JA. Acute kidney injury. *Lancet*. (2019) 394:1949–64. doi: 10.1016/S0140-6736(19)32563-2

4. Hoste EAJ, Bagshaw SM, Bellomo R, Cely CM, Colman R, Cruz DN, et al. Epidemiology of acute kidney injury in critically ill patients: the multinational AKI-EPI study. *Intens Care Med*. (2015) 41:1411–23. doi: 10.1007/s00134-015-3934-7

5. Rabindranath K, Adams J, Macleod AM, Muirhead N. Intermittent versus continuous renal replacement therapy for acute renal failure in adults. *Cochrane Database Syst Rev*. (2007) CD003773. doi: 10.1002/14651858.CD003773.pub3

6. Schneider AG, Bellomo R, Bagshaw SM, Glassford NJ, Lo S, Jun M, et al. Choice of renal replacement therapy modality and dialysis dependence after acute kidney injury: a systematic review and meta-analysis. *Intens Care Med*. (2013) 39:987–97. doi: 10.1007/s00134-013-2864-5

7. Nash DM, Przech S, Wald R, O'Reilly D. Systematic review and meta-analysis of renal replacement therapy modalities for acute kidney injury in the intensive care unit. *J Crit Care*. (2017) 41:138–44. doi: 10.1016/j.jcrc.2017.05.002

8. Uchino S, Bellomo R, Morimatsu H, Morgera S, Schetz M, Tan I, et al. Continuous renal replacement therapy: a worldwide practice survey. The beginning and ending supportive therapy for the kidney (B.E.S.T. Kidney) investigators. *Intensive Care Med*. (2007) 33:1563–70. doi: 10.1007/s00134-007-0754-4

9. Steyerberg EW, Moons KGM, van der Windt DA, Hayden JA, Perel P, Schroter S, et al. Prognosis research strategy (PROGRESS) 3: prognostic model research. *PLoS Med*. (2013) 10:e1001381. doi: 10.1371/journal.pmed.1001381

10. Knaus WA, Draper EA, Wagner DP, Zimmerman JE. APACHE II: a severity of disease classification system. *Crit Care Med*. (1985) 13:818–29. doi: 10.1097/00003246-198510000-00009

11. Vincent JL, Moreno R, Takala J, Willatts S, De Mendonça A, Bruining H, et al. The SOFA (sepsis-related organ failure assessment) score to describe organ dysfunction/failure. On behalf of the working group on sepsis-related problems of the european society of intensive care medicine. *Intens Care Med*. (1996) 22:707–10. doi: 10.1007/BF01709751

12. Kim Y, Park N, Kim J, Kim DK, Chin HJ, Na KY, et al. Development of a new mortality scoring system for acute kidney injury with continuous renal replacement therapy. *Nephrology*. (2019) 24:1233–40. doi: 10.1111/nep.13661

13. Kang MW, Kim J, Kim DK, Oh KH, Joo KW, Kim YS, et al. Machine learning algorithm to predict mortality in patients undergoing continuous renal replacement therapy. *Crit Care*. (2020) 24:42. doi: 10.1186/s13054-020-2752-7

14. Johnson A, Bulgarelli L, Pollard T, Horng S, Celi LA, Mark R. MIMIC-IV (Version 1.0). *PhysioNet*. (2021). doi: 10.13026/s6n6-xd98

15. Goldberger AL, Amaral LA, Glass L, Hausdorff JM, Ivanov PC, Mark RG, et al. PhysioBank, PhysioToolkit, and PhysioNet: components of a new research resource for complex physiologic signals. *Circulation*. (2000) 101:E215–20. doi: 10.1161/01.CIR.101.23.e215

16. Levin A, Stevens PE, Bilous RW, Coresh J, Francisco ALMD, Jong PED, et al. Kidney disease: improving global outcomes (KDIGO) CKD work group. KDIGO 2012 clinical practice guideline for the evaluation and management of chronic kidney disease. *Kidney Int Suppl*. (2013) 3:1–150. doi: 10.1038/kisup.2012.73

17. Lines SW, Cherukuri A, Murdoch SD, Bellamy MC, Lewington AJP. The outcomes of critically ill patients with acute kidney injury receiving renal replacement therapy. *Int J Artif Organs*. (2011) 34:2–9. doi: 10.5301/IJAO.2011.6312

18. Demirjian S, Chertow GM, Zhang JH, O'Connor TZ, Vitale J, Paganini EP, et al. Model to predict mortality in critically ill adults with acute kidney injury. clinical journal of the american society of nephrology: *CJASN*. (2011) 6:2114–20. doi: 10.2215/CJN.02900311

19. Stads S, Fortrie G, van Bommel J, Zietse R, Betjes MGH. Impaired kidney function at hospital discharge and long-term renal and overall survival in patients who received CRRT. *Clin J Am Soc Nephrol*. (2013) 8:1284–91. doi: 10.2215/CJN.06650712

20. De Corte W, Dhondt A, Vanholder R, De Waele J, Decruyenaere J, Sergoyne V, et al. Long-term outcome in ICU patients with acute kidney injury treated with renal replacement therapy: a prospective cohort study. *Crit Care*. (2016) 20:256. doi: 10.1186/s13054-016-1409-z

21. Katayama S, Uchino S, Uji M, Ohnuma T, Namba Y, Kawarazaki H, et al. Factors predicting successful discontinuation of continuous renal replacement therapy. *Anaesth Intens Care*. (2016) 44:453–7. doi: 10.1177/0310057X1604400401

22. Oh HJ, Park JT, Kim JK, Yoo DE, Kim SJ, Han SH, et al. Red blood cell distribution width is an independent predictor of mortality in acute kidney injury patients treated with continuous renal replacement therapy. *Nephrol Dial Transpl*. (2012) 27:589–94. doi: 10.1093/ndt/gfr307

23. Jung SY, Kim H, Park S, Jhee JH, Yun HR, Kim H, et al. Electrolyte and mineral disturbances in septic acute kidney injury patients undergoing continuous renal replacement therapy. *Medicine*. (2016) 95:e4542. doi: 10.1097/MD.0000000000004542

24. Jung SY, Kwon J, Park S, Jhee JH, Yun HR, Kim H, et al. phosphate is a potential biomarker of disease severity and predicts adverse outcomes in acute kidney injury patients undergoing continuous renal replacement therapy. *PLoS ONE*. (2018) 13:e0191290. doi: 10.1371/journal.pone.0191290

25. Han JS, Park KS, Lee MJ, Kim CH, Koo HM, Doh FM, et al. Mean platelet volume is a prognostic factor in patients with acute kidney injury requiring continuous renal replacement therapy. *J Crit Care*. (2014) 29:1016–21. doi: 10.1016/j.jcrc.2014.07.022

26. Vickers AJ, Elkin EB. Decision curve analysis: a novel method for evaluating prediction models. *Med Decis Making*. (2006) 26:565–74. doi: 10.1177/0272989X06295361

27. Kerr KF, Brown MD, Zhu K, Janes H. Assessing the clinical impact of risk prediction models with decision curves: guidance for correct interpretation and appropriate use. *J Clin Oncol*. (2016) 34:2534–40. doi: 10.1200/JCO.2015.65.5654

28. Meersch M, Schmidt C, Hoffmeier A, Van Aken H, Wempe C, Gerss J, et al. Prevention of cardiac surgery-associated AKI by implementing the KDIGO guidelines in high risk patients identified by biomarkers: the PrevAKI randomized controlled trial. *Intens Care Med*. (2017) 43:1551–61. doi: 10.1007/s00134-016-4670-3

29. Göcze I, Jauch D, Götz M, Kennedy P, Jung B, Zeman F, et al. Biomarker-guided intervention to prevent acute kidney injury after major surgery: the prospective randomized BigpAK study. *Ann Surg*. (2018) 267:1013–20. doi: 10.1097/SLA.0000000000002485

30. Selby NM, Casula A, Lamming L, Stoves J, Samarasinghe Y, Lewington AJ, et al. An organizational-level program of intervention for AKI: a pragmatic stepped wedge cluster randomized trial. *J Am Soc Nephrol*. (2019) 30:505–15. doi: 10.1681/ASN.2018090886

31. Sanchez-Pinto LN, Luo Y, Churpek MM. Big data and data science in critical care. *Chest*. (2018) 154:1239–48. doi: 10.1016/j.chest.2018.04.037

32. Sutherland SM, Goldstein SL, Bagshaw SM. Acute kidney injury and big data. *Contrib Nephrol*. (2018) 193:55–67. doi: 10.1159/000484963

33. Sutherland SM, Chawla LS, Kane-Gill SL, Hsu RK, Kramer AA, Goldstein SL, et al. Utilizing electronic health records to predict acute kidney injury risk and outcomes: workgroup statements from the 15(Th) ADQI consensus conference. *Can J Kidney Health Dis*. (2016) 3:11. doi: 10.1186/s40697-016-0099-4

34. Hodgson LE, Sarnowski A, Roderick PJ, Dimitrov BD, Venn RM, Forni LG. Systematic review of prognostic prediction models for acute kidney injury (AKI) in general hospital populations. *BMJ Open*. (2017) 7:e016591. doi: 10.1136/bmjopen-2017-016591

35. Koyner JL, Adhikari R, Edelson DP, Churpek MM. Development of a multicenter ward-based AKI prediction model. *Clin J Am Soc Nephrol*. (2016) 11:1935–43. doi: 10.2215/CJN.00280116

36. Malhotra R, Kashani KB, Macedo E, Kim J, Bouchard J, Wynn S, et al. A risk prediction score for acute kidney injury in the intensive care unit. *Nephrol Dial Transpl*. (2017) 32:814–22. doi: 10.1093/ndt/gfx026

37. Bhatraju PK, Zelnick LR, Katz R, Mikacenic C, Kosamo S, Hahn WO, et al. A prediction model for severe aki in critically ill adults that incorporates clinical and biomarker data. *Clin J Am Soc Nephrol*. (2019) 14:506–14. doi: 10.2215/CJN.04100318

38. Leaf DE, Waikar SS. End points for clinical trials in acute kidney injury. *Am J Kidney Dis*. (2017) 69:108–16. doi: 10.1053/j.ajkd.2016.05.033

39. Sukmark T, Lumlertgul N, Praditpornsilpa K, Tungsanga K, Eiam-Ong S, Srisawat N. SEA-MAKE score as a tool for predicting major adverse kidney events in critically ill patients with acute kidney injury: results from the SEA-AKI study. *Ann Intens Care*. (2020) 10:42. doi: 10.1186/s13613-020-00657-9

40. Obermeyer Z, Emanuel EJ. Predicting the future - big data, machine learning, and clinical medicine. *N Engl J Med*. (2016) 375:1216–9. doi: 10.1056/NEJMp1606181

41. da Hora Passos R, Ramos JGR, Mendonça EJB, Miranda EA, Dutra FRD, Coelho MFR, et al. A clinical score to predict mortality in septic acute kidney injury patients requiring continuous renal replacement therapy: the HELENICC score. *BMC Anesthesiol*. (2017) 17:21. doi: 10.1186/s12871-017-0312-8

42. Efron B. Prediction, estimation, and attribution. *J Am Stat Assoc*. (2020) 115:636–55. doi: 10.1080/01621459.2020.1762613

43. Verghese A, Shah NH, Harrington RA. What this computer needs is a physician: humanism and artificial intelligence. *JAMA*. (2018) 319:19–20. doi: 10.1001/jama.2017.19198

44. Gawande A. *Being Mortal: Illness, Medicine, and What Matters in the End*. London: Profile Books (2014). p. 282.

45. Han YY, Carcillo JA, Venkataraman ST, Clark RSB, Watson RS, Nguyen TC, et al. Unexpected increased mortality after implementation of a commercially sold computerized physician order entry system. *Pediatrics*. (2005) 116:1506–12. doi: 10.1542/peds.2005-1287

46. Ghassemi M, Celi LA, Stone DJ. State of the art review: the data revolution in critical care. *Crit Care*. (2015) 19:118. doi: 10.1186/s13054-015-0801-4

47. Bounthavong M, Watanabe JH, Sullivan KM. Approach to addressing missing data for electronic medical records and pharmacy claims data research. *Pharmacotherapy*. (2015) 35:380–7. doi: 10.1002/phar.1569

48. Austin PC, White IR, Lee DS, van Buuren S. Missing data in clinical research: a tutorial on multiple imputation. *Can J Cardiol*. (2020) 37:1322–31. doi: 10.1016/j.cjca.2020.11.010

49. Xie Y, Ankawi G, Yang B, Garzotto F, Passannante A, Breglia A, et al. Tissue inhibitor metalloproteinase-2 (TIMP-2) IGF-binding protein-7 (IGFBP7) levels are associated with adverse outcomes in patients in the intensive care unit with acute kidney injury. *Kidney Int*. (2019) 95:1486–93. doi: 10.1016/j.kint.2019.01.020

50. Kümpers P, Hafer C, Lukasz A, Lichtinghagen R, Brand K, Fliser D, et al. Serum neutrophil gelatinase-associated lipocalin at inception of renal replacement therapy predicts survival in critically ill patients with acute kidney injury. *Crit Care*. (2010) 14:R9. doi: 10.1186/cc8861

frontiers | Frontiers in Medicine

TYPE Original Research
PUBLISHED 22 August 2022
DOI 10.3389/fmed.2022.907126



OPEN ACCESS

EDITED BY
Zhongheng Zhang,
Sir Run Run Shaw Hospital, China

REVIEWED BY
Sangseok Lee,
Inje University Sanggye Paik Hospital, South Korea
Morteza Sadeghi,
University of Isfahan, Iran
Keyi Yu,
Peking Union Medical College Hospital (CAMS), China

*CORRESPONDENCE
Wen-Kuei Chang
wkchang@vghtpe.gov.tw

†These authors have contributed equally to this work

SPECIALTY SECTION
This article was submitted to Intensive Care Medicine and Anesthesiology, a section of the journal Frontiers in Medicine

RECEIVED 20 May 2022
ACCEPTED 29 July 2022
PUBLISHED 22 August 2022

CITATION
Li Y-S, Chang K-Y, Lin S-P, Chang M-C and Chang W-K (2022) Group-based trajectory analysis of acute pain after spine surgery and risk factors for rebound pain.
*Front. Med.* 9:907126.
doi: 10.3389/fmed.2022.907126



# Group-based trajectory analysis of acute pain after spine surgery and risk factors for rebound pain

Yi-Shiuan Li[1,2†], Kuang-Yi Chang[1,2†], Shih-Pin Lin[1,2], Ming-Chau Chang[2,3] and Wen-Kuei Chang[1,2*]

[1]Department of Anesthesiology, Taipei Veterans General Hospital, Taipei, Taiwan, [2]School of Medicine, National Yang Ming Chiao Tung University, Taipei, Taiwan, [3]Department of Orthopedics, Taipei Veterans General Hospital, Taipei, Taiwan

**Background:** This retrospective study was designed to explore the types of postoperative pain trajectories and their associated factors after spine surgery.

**Materials and methods:** This study was conducted in a single medical center, and patients undergoing spine surgery with intravenous patient-controlled analgesia (IVPCA) for postoperative pain control between 2016 and 2018 were included in the analysis. Maximal pain scores were recorded daily in the first postoperative week, and group-based trajectory analysis was used to classify the variations in pain intensity over time and investigate predictors of rebound pain after the end of IVPCA. The relationships between the postoperative pain trajectories and the amount of morphine consumption or length of hospital stay (LOS) after surgery were also evaluated.

**Results:** A total of 3761 pain scores among 547 patients were included in the analyses and two major patterns of postoperative pain trajectories were identified: Group 1 with mild pain trajectory (87.39%) and Group 2 with rebound pain trajectory (12.61%). The identified risk factors of the rebound pain trajectory were age less than 65 years (odds ratio [OR]: 1.89; 95% *CI*: 1.12–3.20), female sex (*OR*: 2.28; 95% *CI*: 1.24–4.19), and moderate to severe pain noted immediately after surgery (*OR*: 3.44; 95% *CI*: 1.65–7.15). Group 2 also tended to have more morphine consumption ($p < 0.001$) and a longer length of hospital stay ($p < 0.001$) than Group 1.

**Conclusion:** The group-based trajectory analysis of postoperative pain provides insight into the patterns of pain resolution and helps to identify unusual courses. More aggressive pain management should be considered in patients with a higher risk for rebound pain after the end of IVPCA for spine surgery.

KEYWORDS

**group-based trajectory analysis, spine surgery, patient-controlled analgesia (PCA), rebound pain, multimodal analgesia (MMA)**

# Introduction

The indications for spine surgeries vary from herniated disks, spondylolisthesis, fractures, and tumors to scoliosis correction surgeries, and most of these patients need decompression and spine fusion surgeries. As the knowledge of spinal biomechanics, imaging diagnostics, and medical technology is improving over time, the complexity and diversity of spine surgery are increasing as well (1). Although these complex surgeries may benefit those suffering from spinal disease (2, 3), intense pain following the procedures, especially in the immediate and early postoperative period (4–6), often results in clinical problems such as delayed recovery induced by a reduction in patient mobility (7–9). As a result, effective postoperative pain control is of paramount importance and has been connected with better surgical outcomes (10, 11), reduced length of hospital stays (LOSs) (10, 11), lower incidence of chronic postsurgical pain (12), and decreased opioid dependence (7, 13). However, how to well control acute pain after spine surgery remains a major challenge for clinicians (1, 6, 7, 14, 15).

Intravenous patient-controlled analgesia (IVPCA) is a common and effective method to relieve acute pain after spine surgery (16–18) and it optimizes the delivery of analgesics and minimizes the interindividual variability in pharmacokinetics and pharmacodynamics (19). While some studies emphasized the importance of multimodal analgesia in spine surgery (1, 6, 14), IVPCA remains the gold standard for postoperative pain control for spine surgery worldwide (15–18). In addition, IVPCA provides better analgesia after spine surgery than conventional as-needed analgesic regimens do and improves patient satisfaction in the early postoperative days as well (20). However, moderate to severe rebound pain after the discontinuation of IVPCA was noted in other types of surgeries (21), and it is not clear whether this phenomenon also exists in patients receiving IVPCA for pain control after spine surgery. Accordingly, we hypothesized that some patients undergoing spine surgery were at risk of having rebound pain after the end of IVPCA and that there were risk factors associated with the development of rebound pain and designed this retrospective study to investigate these issues. The group-based trajectory analysis was used to classify the variations in postoperative pain scores over time and identify patients with rebound pain after discontinuing IVPCA. The risk factors of rebound pain were also explored, and the influence of rebound pain trajectory on the total amount of IVPCA consumption and LOS after surgery were evaluated as well.

# Materials and methods

## The inclusion and exclusion criteria

This study was approved by the Institutional Review Board of Taipei Veterans General Hospital, Taipei, Taiwan (IRB-TPEVGH no. 2020-01-003AC). Written informed consent was waived and all the included patients were de-identified before analysis. We carefully reviewed the electronic medical records of patients receiving spine surgery and postoperative IVPCA for postoperative pain control in our hospital from January 2016 to December 2018 and collected all records. Those with severe postoperative complications, less than three pain assessments in the first postoperative week, IVPCA use of fewer than 48 h, age $< 20$ years old, staged surgery, re-operation, or missing key data, such as operation records, were excluded from the analysis.

## Anesthesia method and pain management

In this study, all patients were administered general anesthesia with fentanyl (2–3 μg/kg) followed by propofol (1–2 mg/kg) and cisatracurium (0.2 mg/kg) or rocuronium (0.8 mg/kg) for induction. After endotracheal intubation, general anesthesia was maintained using desflurane or sevoflurane with the aforementioned neuromuscular blocking agents. Toward the end of the surgery, the inhalation agent was tapered off and the residual neuromuscular block was reversed with neostigmine and atropine. All patients were transferred to our post-anesthesia care unit where an infusion pump for IVPCA was connected to the patients with a loading dose of morphine of 2–4 mg and a bolus dose of 1 mg. No adjunct analgesics, such as acetaminophen and non-steroidal anti-inflammatory drugs, were administered with IVPCA. After the discontinuation of IVPCA on the fourth postoperative day (POD 4), pain management was shifted to oral medications, including Ultracet (acetaminophen 325 mg + tramadol 37.5 mg) every 6 h and 25 mg diclofenac every 8 h as needed.

## Data collection and endpoints

After surgery, patient-reported pain scores on a numeric rating scale (NRS) from 0 to 10 for no pain to the worst pain were recorded by the nurses in charge at least one time per day. Postoperative maximal daily pain scores were collected in series and used in the trajectory analyses. Patient attributes, such as age, sex, body mass index (BMI), and comorbidities, surgical features, such as surgical time and blood loss, PCA pump settings, and LOS after surgery were collected. The primary

endpoint was the patterns of postoperative pain trajectories, and the secondary endpoints were the total amount of PCA consumption and LOS after surgery.

## Statistical analysis

Group-based trajectory analysis was employed to categorize the variations in postoperative pain over time and the technical details refer to Jones et al. (22). The numbers and features of postoperative trajectories were decided by comparing the Bayesian information criteria of different models and examining the generated trajectories and estimated parameters (23, 24). Two main patterns of pain trajectories were identified, and we compared patient characteristics between the two groups with the Student's *t*-test, the Mann–Whitney *U*–tests, or the chi-squared tests as appropriate. The relationships between the types of postoperative pain trajectories and collected variables were evaluated and presented as odds ratios (OR) with 95% confidence intervals (CI) as well. Backward model selection with an exit criterion of significance level greater than 0.05 was performed to determine the final model for the prediction of postoperative pain trajectories. In addition, a simplified risk scoring system was developed to predict a rebound pain trajectory after the discontinuation of IVPCA for spine surgery. The area under the receiver operating characteristic (ROC) curve (AUC) was used to assess the predictive power of the final model and the simplified risk scoring system. Besides, linear backward regression analysis with an exit significance level of 0.05 was used to select independent predictors of total morphine consumption and log-transformed LOS after surgery. The adjusted association between the types of pain trajectories and total morphine consumption or LOS was also evaluated after the final predictive models were determined. A *p*-value less than 0.05 was considered statistically significant in this study. All the analyses were conducted using SAS software, version 9.4 (SAS Institute Inc., Cary, NC, United States).

# Results

## Analysis of postoperative pain trajectories

There were 547 patients with a pain score of 3,761 included in the analysis, and the average maximal pain scores on the first five PODs ranged between 2.98 and 3.33 (Figure 1, blue line). The mean morphine consumption was 52.6 mg and the median LOS was 7 days. The two postoperative pain trajectory groups were identified after the analysis: Group 1 with a mild pain trajectory (87.4%) and Group 2 (12.6%) with a rebound pain trajectory after the end of IVPCA (Figure 1, black line and red line, respectively). Table 1 shows the comparisons of patient attributes and no significant differences in the surgical features were found between the two groups. However, significant

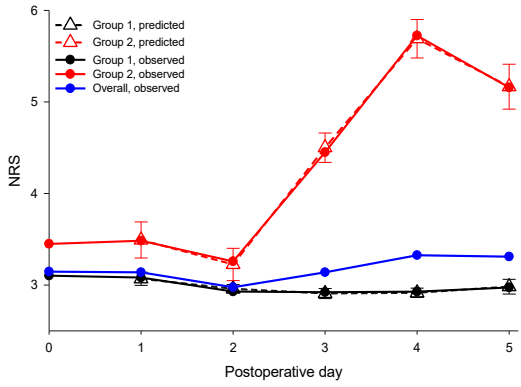



FIGURE 1
Observed and predicted maximal daily pain scores during the first postoperative week were stratified by distinct pain trajectories after spine surgery. NRS, a numeric rating scale for pain intensity. *Solid blue line*: observed overall pain scores during the first postoperative week; *solid black line*: observed pain scores of the mild pain trajectory group; *solid red line*: observed pain scores of the rebound pain trajectory group; dashed black line: predicted pain scores with their 95% confidence interval (*CI*) for the mild pain trajectory group; and *dashed red line*: predicted pain scores with their 95% confidence interval for the rebound pain trajectory group.

**TABLE 1** Comparisons of patient characteristics between the two postoperative pain trajectory groups after spine surgery.

| | Group 1 ($n = 478$) | Group 2 ($n = 69$) | *p* |
|---|---|---|---|
| Age ≤ 65 years | 189 (39.5%) | 38 (55.1%) | 0.011 |
| Sex (women) | 291 (60.9%) | 52 (75.4%) | 0.013 |
| Height (cm) | 157.8 ± 9.1 | 155.9 ± 8.4 | 0.096 |
| Weight (kg) | 67.3 ± 14.6 | 63.7 ± 11.8 | 0.045 |
| Body mass index ≥ 25 kg/m$^2$ | 303 (63.4%) | 42 (60.9%) | 0.390 |
| ASA physical status ≥ 3 | 155 (32.4%) | 21 (30.4%) | 0.428 |
| Creatinine (mg/dl) | 0.89 (0.77–1.07) | 0.84 (0.73–1.13) | 0.388 |
| Maximal NRS before surgery | 2.69 ± 1.02 | 2.86 ± 1.25 | 0.230 |
| Surgical time > 3.5 h | 235 (49.2%) | 39 (56.5%) | 0.155 |
| Surgical blood loss ≥ 500 ml | 219 (45.8%) | 29 (42.0%) | 0.323 |
| Spine segment involved | 3 (2–4) | 3 (2–4) | 0.503 |
| Instrumentation | 411 (86.0%) | 60 (87.0%) | 0.501 |
| Spine involved | | | |
| Thoracic | 49 (10.3%) | 11 (15.9%) | 0.116 |
| Lumbar | 458 (95.8%) | 63 (91.3%) | 0.095 |
| Sacral | 121 (25.3%) | 16 (23.2%) | 0.415 |
| Total IVPCA consumption (ml) | 50.13 ± 26.52 | 69.50 ± 42.55 | < 0.001 |
| Length of hospital stay days | 7 (6–9) | 8 (8–12) | < 0.001 |

Values are mean + SD, count (%) or median (IQR).
IVPCA, intravenous patient-controlled analgesia; ASA, American Society of Anesthesiologists; NRS, a numeric rating scale for pain intensity.

differences in the distributions of age, sex, and body weight were noted between those with rebound pain and their counterparts without it. Moreover, patients in Group 2 also tended to have more morphine consumption and longer LOS after surgery (both $p < 0.001$).

## Factors associated with rebound pain trajectory after the end of intravenous patient-controlled analgesia

After the group-based trajectory analysis, we identified three factors associated with the rebound pain trajectory, such as age $\leq$ 65 years (adjusted *OR*: 1.89, 95%, *CI*: 1.12–3.20), female sex (*OR*: 2.28, 95% *CI*: 1.24–4.19), and moderate to severe pain (NRS $\geq$ 4) on POD 0 (*OR*: 3.44, 95% *CI*: 1.65–7.15; Table 2). Surgical features and other patient characteristics were not related to the rebound pain trajectory. Moreover, a simplified risk scoring system for predicting rebound pain trajectory after the discontinuation of IVPCA could be developed as the following formula:

$$Risk\ score = 1 * (age \leq 65\ years = 1, > 65 = 0) + 1 * (female = 1, male = 0) + 2 * (Moderate\ to\ severe\ pain\ on\ POD\ 0 = 1, no\ to\ mild\ pain = 0)$$

Figure 2 illustrates the estimated probabilities of rebound pain trajectory at distinct risk scores. The risk of developing rebound pain after the end of IVPCA ranged from 5.6 to 42.4% for patients with no to all three risk factors. Figure 3 depicts the ROC curves of the original model and a simplified scoring system. The predictive power of the two models assessed by areas under ROC curves was similar (0.64).

## Predictors of total morphine consumption after surgery

After the backward model selection processes, five independent predictors of increased morphine consumption were identified, such as age $\leq$ 65, male sex (both $p < 0.001$), greater preoperative pain ($p = 0.001$), more spine segment involved ($p = 0.009$), and rebound pain trajectory ($p < 0.001$; Table 3). On average, those with the rebound pain trajectory consumed 17.9 mg more morphine during their IVPCA course

TABLE 2 Risk factors of rebound pain trajectory after the discontinuation of IVPCA following spine surgery.

| | β | SE (β ) | OR | 95% CI | *p* | Simplified risk score |
|---|---|---|---|---|---|---|
| Age ≤ 65 vs. > 65 | 0.64 | 0.27 | 1.89 | 1.12∼3.20 | 0.018 | 1 |
| Sex (women vs. men) | 0.82 | 0.31 | 2.28 | 1.24∼4.19 | 0.008 | 1 |
| NRS ≥ 4 on POD 0 | 1.23 | 0.37 | 3.44 | 1.65∼7.15 | 0.001 | 2 |

OR, odds ratio; CI, confidence interval; NRS, a numeric rating scale for pain intensity after surgery; POD, postoperative day.

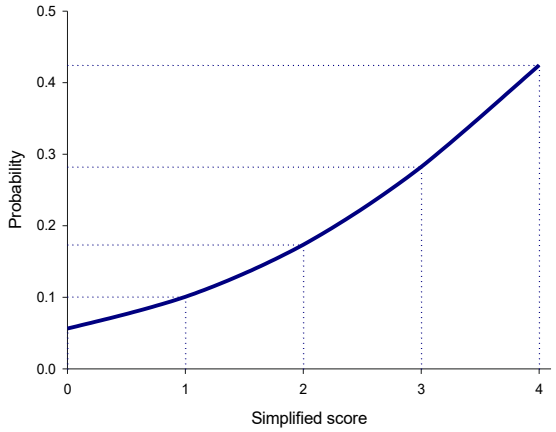



FIGURE 2
Predicted probability of the rebound pain trajectory for the simplified risk scoring systems after the discontinuation of intravenous patient-controlled analgesia (IVPCA) for spine surgery. The probability of developing rebound pain after the end of IVPCA for spine surgery increased gradually from 5.6% for the simplified score of 0–42.4% for the score of 4.

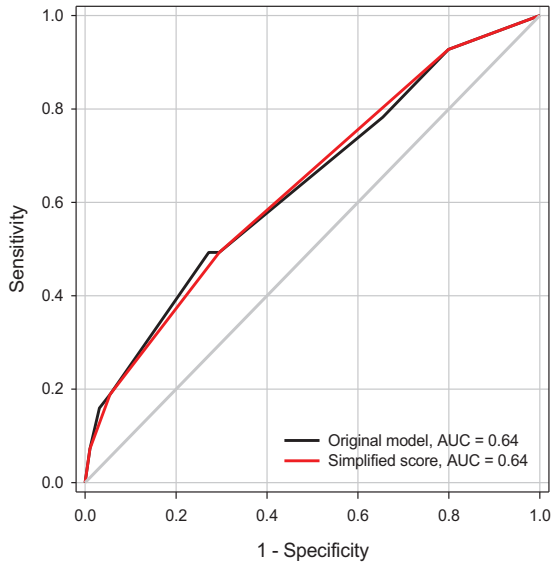



FIGURE 3
A receiver operating characteristic (ROC) curve analysis of predictive power for the selected model and the simplified risk scoring system for the rebound pain trajectory. AUC, area under ROC curve.

TABLE 3 Predictors of total IVPCA consumption after spine surgery.

| | β | SE | Std β | *p* |
|---|---|---|---|---|
| Pain trajectory (Group 2 vs. Group 1) | 17.93 | 3.59 | 0.20 | < 0.001 |
| Age (≤65 vs. > 65 years) | 13.36 | 2.45 | −0.22 | < 0.001 |
| Sex (women vs. men) | −9.61 | 2.46 | −0.16 | < 0.001 |
| NRS on POD 0 | 3.64 | 1.12 | 0.13 | 0.001 |
| Spine segment involved | 2.42 | 0.92 | 0.11 | 0.009 |
| *Constant* | 33.45 | 4.81 | – | < 0.001 |

β, regression coefficients; SE, standard error; std β, standardized regression coefficients; Group 1, mild pain trajectory; Group 2, rebound pain trajectory; IVPCA, intravenous patient-controlled analgesia; NRS, a numeric rating scale for pain intensity after surgery; POD, postoperative day.

after controlling for the effects of other predictors in the final model.

TABLE 4 Factors associated with length of hospital stay (LOS)* after spine surgery.

| | β | SE | Std β | *p* | *exp*(β ) |
|---|---|---|---|---|---|
| Pain trajectory (Group 2 vs. Group 1) | 0.16 | 0.05 | 0.13 | 0.001 | 1.172 |
| Surgical time > 3.5 h | 0.19 | 0.03 | 0.22 | < 0.001 | 1.204 |
| Lumbar spine involved | −0.42 | 0.08 | −0.21 | < 0.001 | 0.660 |
| Maximal NRS before surgery | 0.06 | 0.02 | 0.14 | < 0.001 | 1.058 |
| ASA physical status ≥3 | 0.10 | 0.04 | 0.11 | 0.005 | 1.106 |
| Spine segment involved | 0.03 | 0.01 | 0.10 | 0.014 | 1.034 |
| *Constant* | 2.10 | 0.10 | | < 0.001 | 8.144 |

*Length of hospital stay is log-transformed in the analysis.
β, regression coefficients; SE, standard error of regression coefficients; std β, standardized regression coefficients; *exp*(β), exponentiated regression coefficients; Group 1, mild pain trajectory; Group 2, rebound pain trajectory; ASA, American Society of Anesthesiologists; IVPCA, intravenous patient-controlled analgesia; NRS, a numeric rating scale for pain intensity.

## Factors related to length of hospital stays after surgery

There were six factors associated with LOS after surgery, such as surgical time, lumbar spine involved, preoperative pain (all $p < 0.001$), American Society of Anesthesiologists (ASA) physical status ($p = 0.005$), spine segment involved ($p = 0.014$), and rebound pain trajectory (Table 4). On average, patients with the postoperative rebound pain trajectory tended to stay 17.2% longer in hospital ($p = 0.001$) than those with normal pain resolution after the adjustment for the other selected predictors in the model.

# Discussion

This is the first study to describe the phenomenon of rebound pain after the discontinuation of IVPCA for spine surgery. Although Nicholson et al. (25) used "rebound pain" to describe the increase in pain score between 8 and 24 h after surgery in a patient still "receiving PCA," this is totally different from our findings. Approximately one-eighth of the target population experienced this unpleasant journey after the end of IVPCA. With the aid of group-based trajectory analysis, patients with abnormal pain resolution after spine surgery could be recognized and the associated factors could be identified. Regional anesthesia (RA), such as short-lasting spinal anesthesia and peripheral nerve blocks, is widely used in various surgery due to effective pain relief in the early postoperative phase. However, severe pain was noted in up to 40% of patients when the RA wears off, and this phenomenon is known as "rebound pain" (26). Recently, rebound pain was also observed in patients receiving epidural analgesia for video-assisted thoracoscopic surgery (21). All these aforementioned rebound pain phenomena were developed after the transition from an effective analgesic intervention to other routine pain management. These findings highlight the importance of analgesic transition and the necessity of early identification and intervention. Our study provides important clues for clinicians to early detect high-risk patients, and thus, preventive strategies could be initiated in advance to refine the quality of postoperative care and pain management following spine surgery (27).

Several risk factors of rebound pain were identified in this study and among them, younger age (28–31) and female sex (28–30, 32) were associated with analgesic consumption. These two non-modifiable factors were identified as risk factors in rebound pain development in other studies as well (21, 27). In addition, some previous studies revealed that younger age (28–31), female sex (28–30, 32), preoperative NRS (30, 33), and the number of spine segments involvement (4, 34) were associated with higher postoperative pain scores and more analgesic consumption. Although preoperative pain has been proposed as a risk factor for inferior postoperative pain control and more morphine consumption in a previous study (33), our study demonstrated that the postoperative pain on POD 0, rather than the preoperative pain, was an independent predictor of rebound pain trajectory after the end of IVPCA for spine surgery. The discrepancy might result from the difference in outcomes of interest and study population since we focused on the rebound pain trajectory after the end of IVPCA for spine surgery instead of general pain scores observed after surgery. Since the IVPCA remains the gold standard for postoperative pain control in complex spine surgery, the prediction of rebound pain in advance is of paramount importance. In spite of the efforts which have been made to evaluate the effects of surgical time and blood loss and the complexity of the surgery, such as procedure

types and the number of spine segments involved, none of these factors were significantly associated with rebound pain trajectory. A more comprehensive classification of spine surgery might be considered in future studies.

In this study, we used group-based trajectory analysis to model the variations in pain intensity over time and identify distinct patterns of postoperative pain trajectories and their associated factors. Similar to clinical decision-making, this approach directly focuses on postoperative pain observations. Patient characteristics were not involved in the group classification processes but evaluated *post hoc* to avoid untoward interference in trajectory recognition. In addition, the group-based trajectory analysis has a great advantage in handling missing data, which is commonly observed in retrospective studies (35). Furthermore, a simplified risk scoring system was developed based on the estimated results of group-based trajectory analysis. The risk of developing rebound pain after the end of IVPCA could be easily assessed with the help of this system. Among the three risk factors, moderate to severe pain noted immediately after spine surgery despite IVPCA in use should be regarded as an early sign of possible rebound pain after the transition from IVPCA to other analgesic modalities. Once moderate to severe pain is noted after surgery, more aggressive multimodal pain management should be considered to reduce the risk of rebound pain after the end of IVPCA. This scoring system has great potential to be applied in clinical practice to prevent rebound pain after the discontinuation of IVPCA (36, 37) and improve pain control quality following spine surgery. For example, a 70-year-old male patient who is satisfied with IVPCA had no to mild pain on POD 0 after spine surgery, and the simplified risk score of rebound pain is 0; while a 60-year-old female patient who has moderate to severe pain on POD 0 under IVPCA management had the simplified risk score of rebound pain of 4. The probability of developing rebound pain (group 2) after the end of IVPCA in these two patients would be 5.6 and 42.4%, respectively. The clinicians should introduce more vigorous pain management, such as prolonged PCA duration or multiple model pain management control to prevent or manage the rebound pain afterward. However, its validity and clinical utility of this risk scoring system await further investigation.

There were some limitations to our study. First, the impacts of unobserved variables on the patterns of variations in postoperative pain scores over time could not be further evaluated and more covariates should be included in future studies for better prediction of the rebound pain trajectory. Second, the preoperative analgesic prescriptions were not further investigated due to data unavailability. Third, we only evaluated the effects of surgical time, blood loss, instrumentation, and spine segments involved on the risk of having a rebound pain trajectory but did not further assess the associations between different kinds of spine surgical procedures and the incidence of rebound pain since there is still no consensus on the classification of complex spine surgery.

In conclusion, two major patterns of postoperative pain trajectories were recognized in patients receiving IVPCA for spine surgery using group-based trajectory analysis, and about one-eighth of them had a rebound pain trajectory. Three predictors of rebound pain trajectory were identified, namely, younger age, female sex, and moderate to severe pain on POD 0. A simplified risk scoring system was developed based on the analytical results but its clinical utility needs further investigation. Preventive strategies, such as early introduction of more aggressive multimodal analgesia, should be considered in high-risk patients to reduce the incidence of rebound pain since patients with rebound pain trajectory were inclined to have longer hospital stay after surgery and more opioid consumption. Group-based trajectory analysis provides valuable information to categorize variations in postoperative pain over time and detect unusual patterns of pain resolution for further optimization of perioperative pain management. More patient attributes and surgical features should be collected in future studies to further elucidate the underlying mechanism of rebound pain after the end of IVPCA for spine surgery.

## Data availability statement

The raw data supporting the conclusions of this article will be made available by the authors, without undue reservation.

## Ethics statement

The studies involving human participants were reviewed and approved by Institutional Review Board of Taipei Veterans General Hospital. Written informed consent for participation was not required for this study in accordance with the national legislation and the institutional requirements.

## Author contributions

Y-SL contributed to data acquisition and manuscript drafting. W-KC contributed to data validation and draft preparation. S-PL contributed to study coordination and data acquisition. M-CC helped to review and revised the manuscript. K-YC contributed to conceptualization, statistical analysis, and manuscript revision. All authors have read and agreed to the published version of the manuscript.

## Funding

This study was supported by grants from the Taipei Veterans General Hospital, Taipei, Taiwan (V111C-074), the Ministry of Science and Technology, Taipei, Taiwan (MOST 109-2511-H-075-003-MY2), and the Anesthesiology Research and Development Foundation, Taipei, Taiwan (ARDF11101).

## Acknowledgments

We would like to thank Shih-Tien Wang (Department of Orthopedics, Taipei Veterans General Hospital) and Chun-Sung Sung (Department of Anesthesiology, Taipei Veterans General Hospital) for their kind assistance in data collection.

## Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

## Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

## References

1. Waelkens P, Alsabbagh E, Sauter A, Joshi GP, Beloeil H, Prospect Working group** of the European Society of Regional Anaesthesia and Pain therapy (Esra). Pain management after complex spine surgery: a systematic review and procedure-specific postoperative pain management recommendations. *Eur J Anaesthesiol.* (2021) 38:985–94. doi: 10.1097/EJA.0000000000001448

2. Lamperti M, Tufegdzic B, Avitsian R. Management of complex spine surgery. *Curr Opin Anaesthesiol.* (2017) 30:551–6. doi: 10.1097/ACO.0000000000000494

3. Marquez-Lara A, Nandyala SV, Fineberg SJ, Singh K. Current trends in demographics, practice, and in-hospital outcomes in cervical spine surgery: a national database analysis between 2002 and 2011. *Spine (Phila Pa 1976).* (2014) 39:476–81. doi: 10.1097/BRS.0000000000000165

4. Gerbershagen HJ, Aduckathil S, van Wijck AJ, Peelen LM, Kalkman CJ, Meissner W. Pain intensity on the first day after surgery: a prospective cohort study comparing 179 surgical procedures. *Anesthesiology.* (2013) 118:934–44. doi: 10.1097/ALN.0b013e31828866b3

5. Archer K, Bird ML, Haug C, Coronado R, Wegener S, Devin CJ, et al. Patients' experience and expectations of lumbar spine surgery for degenerative conditions: a qualitative study. *Spine J.* (2015) 15:S99–100. doi: 10.1016/j.spinee.2015.07.046

6. Devin CJ, McGirt MJ. Best evidence in multimodal pain management in spine surgery and means of assessing postoperative pain and functional outcomes. *J Clin Neurosci.* (2015) 22:930–8. doi: 10.1016/j.jocn.2015.01.003

7. Debono B, Wainwright TW, Wang MY, Sigmundsson FG, Yang MMH, Smid-Nanninga H, et al. Consensus statement for perioperative care in lumbar spinal fusion: enhanced recovery after surgery (ERAS(R)) society recommendations. *Spine J.* (2021) 21:729–52. doi: 10.1016/j.spinee.2021.01.001

8. Street JT, Lenehan BJ, DiPaola CP, Boyd MD, Kwon BK, Paquette SJ, et al. Morbidity and mortality of major adult spinal surgery. A prospective cohort analysis of 942 consecutive patients. *Spine J.* (2012) 12:22–34. doi: 10.1016/j.spinee.2011.12.003

9. Goldstein CL, Macwan K, Sundararajan K, Rampersaud YR. Perioperative outcomes and adverse events of minimally invasive versus open posterior lumbar fusion: meta-analysis and systematic review. *J Neurosurg Spine.* (2016) 24:416–27. doi: 10.3171/2015.2.SPINE14973

10. Carli F, Kehlet H, Baldini G, Steel A, McRae K, Slinger P, et al. Evidence basis for regional anesthesia in multidisciplinary fast-track surgical care pathways. *Reg Anesth Pain Med.* (2011) 36:63–72. doi: 10.1097/AAP.0b013e31820307f7

11. Lenart MJ, Wong K, Gupta RK, Mercaldo ND, Schildcrout JS, Michaels D, et al. The impact of peripheral nerve techniques on hospital stay following major orthopedic surgery. *Pain Med.* (2012) 13:828–34. doi: 10.1111/j.1526-4637.2012.01363.x

12. Borghi B, D'Addabbo M, White PF, Gallerani P, Toccaceli L, Raffaeli W, et al. The use of prolonged peripheral neural blockade after lower extremity amputation: the effect on symptoms associated with phantom limb syndrome. *Anesth Analg.* (2010) 111:1308–15. doi: 10.1213/ANE.0b013e3181f4e848

13. Schoenfeld AJ, Nwosu K, Jiang W, Yau AL, Chaudhary MA, Scully RE, et al. Risk factors for prolonged opioid use following spine surgery, and the association with surgical intensity, among opioid-naive patients. *J Bone Joint Surg Am.* (2017) 99:1247–52. doi: 10.2106/JBJS.16.01075

14. Dunn LK, Durieux ME, Nemergut EC. Non-opioid analgesics: novel approaches to perioperative analgesia for major spine surgery. *Best Pract Res Clin Anaesthesiol.* (2016) 30:79–89. doi: 10.1016/j.bpa.2015.11.002

15. Dietz N, Sharma M, Adams S, Alhourani A, Ugiliweneza B, Wang D, et al. Enhanced recovery after surgery (ERAS) for spine surgery: a systematic review. *World Neurosurg.* (2019) 130:415–26. doi: 10.1016/j.wneu.2019.06.181

16. Reynolds RA, Legakis JE, Tweedie J, Chung Y, Ren EJ, Bevier PA, et al. Postoperative pain management after spinal fusion surgery: an analysis of the efficacy of continuous infusion of local anesthetics. *Global Spine J.* (2013) 3:7–14. doi: 10.1055/s-0033-1337119

17. Javed T, Ahad B, Singh P, Ahmad R. A prospective randomized study to compare tramadol and morphine for postoperative analgesia in spine surgeries using intravenous patient controlled analgesia. *Int J Res Med Sci.* (2017) 5:3350–3354. doi: 10.18203/2320-6012.ijrms20173140

18. Venkatraman R, Pushparani A, Balaji R, Nandhini P. Comparison of low dose intravenous fentanyl and morphine infusion for postoperative analgesia in spine fusion surgeries - a randomized control trial. *Braz J Anesthesiol.* (2021) 71:339–44. doi: 10.1016/j.bjane.2020.12.013

19. Macintyre PE. Safety and efficacy of patient-controlled analgesia. *Br J Anaesth.* (2001) 87:36–46. doi: 10.1093/bja/87.1.36

20. McNicol ED, Ferguson MC, Hudcova J. Patient controlled opioid analgesia versus non-patient controlled opioid analgesia for postoperative pain. *Cochrane Database Syst Rev.* (2015) 6:CD003348.

21. Chang WK, Li YS, Wu HL, Tai YH, Lin SP, Chang KY. Group-based trajectory analysis of postoperative pain in epidural analgesia for video-assisted thoracoscopic surgery and risk factors of rebound pain. *J Chin Med Assoc.* (2022) 85:216–221. doi: 10.1097/JCMA.0000000000000647

22. Nagin DS, Jones BL, Passos VL, Tremblay RE. Group-based multi-trajectory modeling. *Stat Methods Med Res.* (2018) 27:2015–23. doi: 10.1177/0962280216673085

23. Chang WK, Tai YH, Lin SP, Wu HL, Tsou MY, Chang KY. An investigation of the relationships between postoperative pain trajectories and outcomes after surgery for colorectal cancer. *J Chine Med Assoc.* (2019) 82:865–71. doi: 10.1097/JCMA.0000000000000166

24. Teng WN, Wu HL, Tai YH, Lei HJ, Tsou MY, Chang KY. Group-based trajectory analysis of postoperative pain and outcomes after liver cancer surgery. *J Chine Med Assoc.* (2021) 84:95–100. doi: 10.1097/JCMA.0000000000000446

25. Nicholson T, Maltenfort M, Getz C, Lazarus M, Williams G, Namdari S. Multimodal pain management protocol versus patient controlled narcotic analgesia for postoperative pain control after shoulder arthroplasty. *Arch Bone Jt Surg.* (2018) 6:196–202.

26. Lavand'homme P. Rebound pain after regional anesthesia in the ambulatory patient. *Curr Opin Anaesthesiol.* (2018) 31:679–84. doi: 10.1097/ACO.0000000000000651

27. Barry GS, Bailey JG, Sardinha J, Brousseau P, Uppal V. Factors associated with rebound pain after peripheral nerve block for ambulatory surgery. *Br J Anaesth.* (2021) 126:862–71. doi: 10.1016/j.bja.2020.10.035

28. Mei W, Seeling M, Franck M, Radtke F, Brantner B, Wernecke KD, et al. Independent risk factors for postoperative pain in need of intervention early after awakening from general anaesthesia. *Eur J Pain.* (2010) 14:e1–7. doi: 10.1016/j.ejpain.2009.03.009

29. Murray AA, Retief FW. Acute postoperative pain in 1 231 patients at a developing country referral hospital: incidence and risk factors. *South Afr J Anaesth Analg.* (2015) 22:19–24. doi: 10.1080/22201181.2015.1115608

30. Gerbershagen HJ, Pogatzki-Zahn E, Aduckathil S, Peelen LM, Kappen TH, van Wijck AJ, et al. Procedure-specific risk factor analysis for the development of severe postoperative pain. *Anesthesiology.* (2014) 120:1237–45. doi: 10.1097/ALN.0000000000000108

31. Ip HY, Abrishami A, Peng PW, Wong J, Chung F. Predictors of postoperative pain and analgesic consumption: a qualitative systematic review. *Anesthesiology.* (2009) 111:657–77. doi: 10.1097/ALN.0b013e3181aae87a

32. Aubrun F, Salvi N, Coriat P, Riou B. Sex- and age-related differences in morphine requirements for postoperative pain relief. *Anesthesiology.* (2005) 103:156–60. doi: 10.1097/00000542-200507000-00023

33. Yang MMH, Hartley RL, Leung AA, Ronksley PE, Jetté N, Casha S, et al. Preoperative predictors of poor acute postoperative pain control: a systematic review and meta-analysis. *BMJ Open.* (2019) 9:e025091. doi: 10.1136/bmjopen-2018-025091

34. Chaterjee S, Ghosh S, Ray S, Rudra A. Pain management after spinal surgery. *Indian J Pain.* (2015) 29:14. doi: 10.4103/0970-5333.145916

35. Nagin DS, Odgers CL. Group-based trajectory modeling in clinical research. *Annu Rev Clin Psychol.* (2010) 6:109–38. doi: 10.1146/annurev.clinpsy.121208.131413

36. Marshall K, McLaughlin K. Pain management in thoracic surgery. *Thorac Surg Clin.* (2020) 30:339–46. doi: 10.1016/j.thorsurg.2020.03.001

37. Muñoz-Leyva F, Cubillos J, Chin KJ. Managing rebound pain after regional anesthesia. *Korean J Anesthesiol.* (2020) 73:372. doi: 10.4097/kja.20436

frontiers | Frontiers in Medicine

TYPE Original Research
PUBLISHED 23 August 2022
DOI 10.3389/fmed.2022.960296



OPEN ACCESS

EDITED BY
Qinghe Meng,
Upstate Medical University,
United States

REVIEWED BY
Khin Wee Lai,
University of Malaya, Malaysia
Jinghua Wang,
Tianjin Medical University General Hospital, China

*CORRESPONDENCE
Stefan Hegselmann
stefan.hegselmann@uni-muenster.de

SPECIALTY SECTION
This article was submitted to Intensive Care Medicine and Anesthesiology, a section of the journal Frontiers in Medicine

RECEIVED 02 June 2022
ACCEPTED 03 August 2022
PUBLISHED 23 August 2022

CITATION
Hegselmann S, Ertmer C, Volkert T, Gottschalk A, Dugas M and Varghese J (2022) Development and validation of an interpretable 3 day intensive care unit readmission prediction model using explainable boosting machines. *Front. Med.* 9:960296. doi: 10.3389/fmed.2022.960296



# Development and validation of an interpretable 3 day intensive care unit readmission prediction model using explainable boosting machines

Stefan Hegselmann[1]*, Christian Ertmer[2], Thomas Volkert[2], Antje Gottschalk[2], Martin Dugas[3] and Julian Varghese[1]

[1]Institute of Medical Informatics, University of Münster, Münster, Germany, [2]Department of Anesthesiology, Intensive Care and Pain Medicine, University Hospital Münster, Münster, Germany, [3]Institute of Medical Informatics, Heidelberg University Hospital, Heidelberg, Germany

**Background:** Intensive care unit (ICU) readmissions are associated with mortality and poor outcomes. To improve discharge decisions, machine learning (ML) could help to identify patients at risk of ICU readmission. However, as many models are black boxes, dangerous properties may remain unnoticed. Widely used *post hoc* explanation methods also have inherent limitations. Few studies are evaluating inherently interpretable ML models for health care and involve clinicians in inspecting the trained model.

**Methods:** An inherently interpretable model for the prediction of 3 day ICU readmission was developed. We used explainable boosting machines that learn modular risk functions and which have already been shown to be suitable for the health care domain. We created a retrospective cohort of 15,589 ICU stays and 169 variables collected between 2006 and 2019 from the University Hospital Münster. A team of physicians inspected the model, checked the plausibility of each risk function, and removed problematic ones. We collected qualitative feedback during this process and analyzed the reasons for removing risk functions. The performance of the final explainable boosting machine was compared with a validated clinical score and three commonly used ML models. External validation was performed on the widely used Medical Information Mart for Intensive Care version IV database.

**Results:** The developed explainable boosting machine used 67 features and showed an area under the precision-recall curve of $0.119 \pm 0.020$ and an area under the receiver operating characteristic curve of $0.680 \pm 0.025$. It performed on par with state-of-the-art gradient boosting machines ($0.123 \pm 0.016$, $0.665 \pm 0.036$) and outperformed the Simplified Acute Physiology Score II ($0.084 \pm 0.025$, $0.607 \pm 0.019$), logistic regression ($0.092 \pm 0.026$, $0.587 \pm 0.016$), and recurrent neural networks ($0.095 \pm 0.008$, $0.594 \pm 0.027$). External validation confirmed that explainable boosting machines ($0.221 \pm 0.023$, $0.760 \pm 0.010$) performed similarly to gradient boosting machines ($0.232 \pm 0.029$, $0.772 \pm 0.018$). Evaluation of the model inspection showed that explainable boosting machines can be useful to detect and remove problematic risk functions.

**Conclusions:** We developed an inherently interpretable ML model for 3 day ICU readmission prediction that reached the state-of-the-art performance of black box models. Our results suggest that for low- to medium-dimensional datasets that are common in health care, it is feasible to develop ML models that allow a high level of human control without sacrificing performance.

KEYWORDS

**intensive care unit, readmission, artificial intelligence, machine learning, explainable AI, interpretable machine learning, doctor-in-the-loop, human evaluation**

# Introduction

Discharge decisions in an intensive care unit (ICU) are complex and require consideration of several aspects (1). Discharging a patient too early can lead to the deterioration of the patient's health status that requires subsequent ICU readmission. This is associated with mortality and poor outcomes such as an increased length of ICU stay (2–4). A study conducted in 105 ICUs in the United States in 2013 found a median ICU readmission rate of 5.9% (5). Identified risk factors include admission origin, comorbidities, physiological abnormalities, and age (4, 6, 7). However, incorporating all available information appropriately for interpretation of an individual patient case can be challenging for clinicians (8).

Machine learning (ML) can automatically detect patterns in large quantities of data and has already shown the potential to transform health care (9). However, many ML models are considered black boxes, since they can be too complex for humans to understand (10). Studies have found that ML models contained an unnoticed racial bias (11) or relied on dangerous correlations (12), which can cause distrust among stakeholders, preventing their adoption (13). Interpretable ML could alleviate these issues by providing human-understandable explanations, enabling users to ensure properties such as fairness or robustness (14). Many studies have used so-called *post hoc* explanation methods such as local interpretable model-agnostic explanations (15) or Shapley additive explanations (16), which provide an explanation for a single prediction (17–19). However, *post hoc* methods have several shortcomings with respect to robustness and adversarial attacks (20–22) limiting their usefulness in health care settings (23). Hence, in this work, we used inherently interpretable or transparent models (10, 24) that allow humans to inspect and understand the entire model before using it for predictions.

A research gap exists owing to the lack of studies about transparent ML models for health care that include human evaluations. A recent review on explainable artificial intelligence using electronic health records showed that only nine out of 42 studies used inherently interpretable models (25). Applications included mortality prediction, disease classification, risk stratification, and biomedical knowledge discovery. However, only three studies reported human expert confirmation of their results, which is considered essential for a meaningful evaluation of interpretable ML (14). For ICU readmission prediction, we identified two papers (26, 27) that explicitly developed interpretable models based on rule sets and logistic regression (LR). However, no human validation of the results was performed.

In this study, we aimed to develop an inherently interpretable explainable boosting machine (EBM) model for the prediction of 3 day ICU readmission. We involved clinicians in the development process to inspect and verify the entire model. The validation process was evaluated to determine its effect and reveal possible issues. Second, the resulting EBM model was compared with different baseline and state-of-the-art black box ML models to assess the effect of transparency on performance.

# Materials and methods

## Study setting and preregistration

This study was approved by the ethics review board of the medical chamber Westfalen-Lippe (reference number: 2020-526-f-S). We provided the TRIPOD (Transparent Reporting of a Multivariable Prediction Model for Individual Prognosis or Diagnosis) checklist (28) in Supplementary material 1. This work was preregistered online (29); however, it had two deviations: a readmission interval of 3 days instead of 7 days was considered to exclude fewer patients with insufficient follow-ups. Also, we only performed external validation for the final performance results, which we considered most relevant. An overview of all steps conducted for this study can be found in Figure 1. All code for preprocessing the data, training the models, and inspecting the final EBM model is publicly available (30, 31).

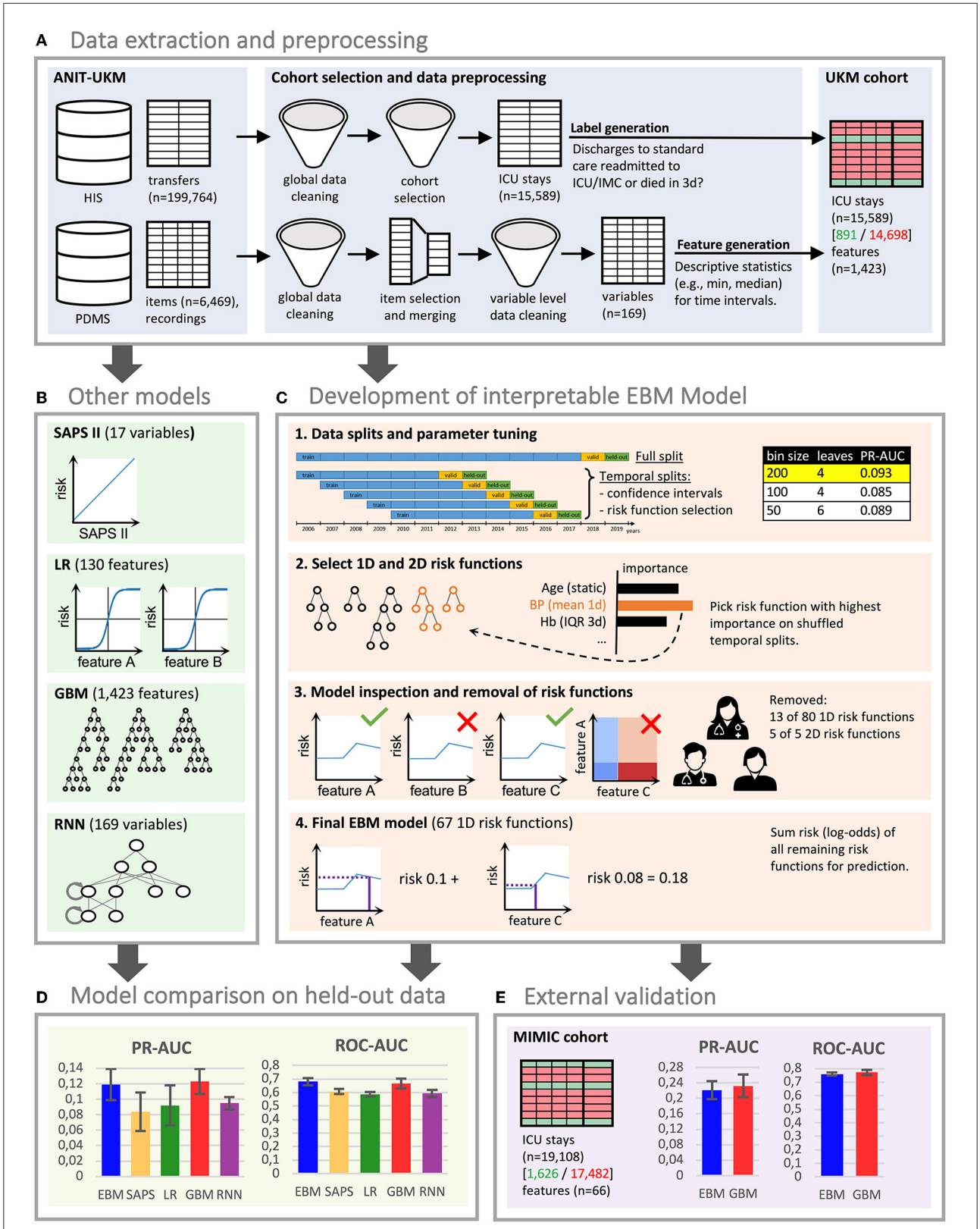



FIGURE 1
Flowchart of the study. **(A)** We created a local cohort for the development of machine learning (ML) models. Information on intensive care unit (ICU) transfers was extracted from the hospital information system (HIS), and ICU data was extracted from the patient data management system (PDMS). Extensive preprocessing was applied to clean the data. We generated labels for 3 day ICU readmission and descriptive statistics as

*(Continued)*

FIGURE 1 (Continued)
features. **(B)** Four ML models were developed for comparison. For LR, we also performed feature selection. The RNN directly uses the time series data. **(C)** The development of the EBM model involved four steps [see 1–4 in **(C)**]. We conducted parameter tuning for EBM (and our other models) and performed greedy risk function selection based on the importance determined on the temporal splits. In step 3, we inspected the model with a team of clinicians to identify and remove problematic risk functions. The remaining risk functions were used for the predictions. **(D)** We evaluated all models for their area under the precision-recall curve (PR-AUC) and area under the receiver operating characteristic curve (ROC-AUC) on the hold-out split. **(E)** External validation for the EBM and GBM models was performed on the Medical Information Mart for Intensive Care (MIMIC) version IV. **(D,E)** Error bars were determined with the standard deviation on five temporal splits. EBM, explainable boosting machine; SAPS II, Simplified Acute Physiology Score II; LR, logistic regression; GBM, gradient boosting machine; RNN, recurrent neural network.

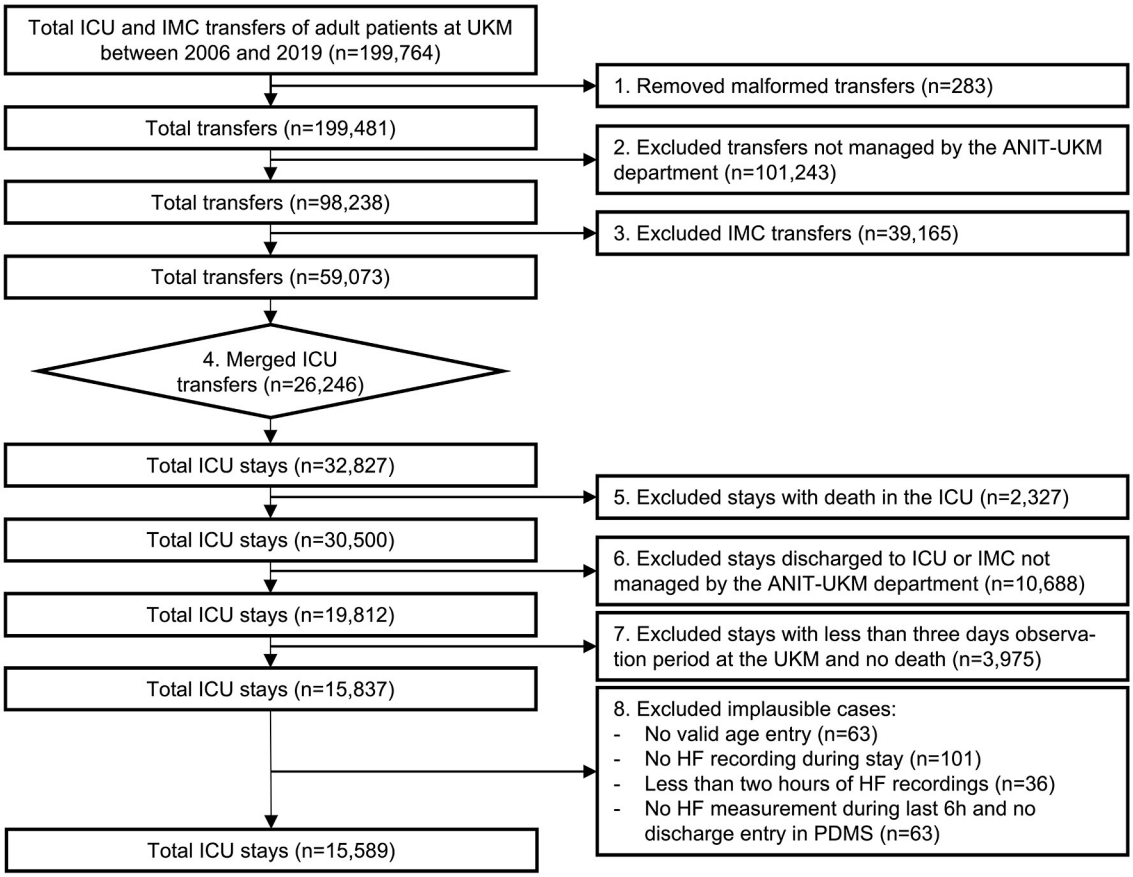



FIGURE 2
Flowchart of the cohort selection for the University Hospital Münster (UKM) cohort. Transfers to ICU and IMC wards of the UKM between 2006 and 2019 served as initial data. We included four ICUs managed by the ANIT-UKM department. Transfers had to be merged using a manual procedure to obtain consecutive ICU stays. Patients who died in the ICU and those who were discharged to an external ICU or IMC were excluded. We required an observation period of at least 3 days to ensure readmission to an ICU in the UKM. Lastly, implausible cases were removed.

## Cohort

We included all ICU patients managed by the Department of Anesthesiology, Intensive Care and Pain Medicine at the University Hospital Münster (ANIT-UKM) who were discharged to standard care and had a follow-up period of at least 3 days (see Figure 2). Initially, all ICU and intermediate care (IMC) transfers of adult patients between 2006 and 2019 were retrieved from the hospital information system (HIS; ORBIS, Dedalus Healthcare Group; $n = 199{,}764$). First, 283 entries were removed because of ambiguous discharge dates, overlapping hospital stays, or overlapping transfers that could not be delineated. Next, transfers not managed by the ANIT-UKM ($n = 101{,}243$) and IMC transfers ($n = 39{,}165$) were excluded. In step 4, we merged consecutive transfers ($n = 26{,}246$) into a single ICU stay. Some entries ($n = 147$) contained artifacts with short intervals between two transfers, and we designed a stepwise procedure to decide whether a discharge occurred. Next, we excluded ICU stays that ended with the death of the patient ($n = 2{,}327$) or a discharge to an external

ICU or IMC unit ($n = 10{,}688$). We used the same procedure as in step 4 to identify artifacts ($n = 67$) and to distinguish consecutive transfers and readmissions to an external ICU. We then excluded all ICU stays without a 3 day follow-up period at the UKM to ensure that all patients with worsening conditions who were included were transferred to an observed ICU ($n = 3{,}975$). This also excluded patients who were transferred to an external facility or home, which introduced a selection bias. However, we reckoned that ensuring a complete observation interval outweighed this effect. Lastly, we removed implausible cases with no age entry ($n = 63$) or that had only very few heart frequency recordings ($n = 200$); thus, 15,589 ICU stays were included.

ICU patients who were readmitted to any ICU ($n = 822$) or IMC unit ($n = 31$) or died within 3 days ($n = 38$) were labeled as true (Supplementary material 2). Patient deaths were also labeled to obtain a consistent outcome. Patients who were discharged to standard care and underwent a planned procedure with a subsequent re-admission to an ICU or IMC unit incorrectly received a positive label. However, we considered this effect to be small. To verify our cohort selection and labeling procedure, we sampled 20 positive stays stratified across wards and verified them using additional clinical information.

Table 1 summarizes the key characteristics of the resulting UKM cohort. The ICU patients of the included stays had a mean age of 63.33 ± 14.73 years, and more than two-thirds of them were male ($n = 10{,}670$). ICU patients with 3 day readmission or who died after discharge showed several differences: the patients were 3 years older on average, the proportion of male patients further increased from 68.3 to 70.8%, and the mean length of the previous ICU stay was approximately 13.5 hours longer. Supplementary material 2 contains an overview of the included ICUs.

TABLE 1 Overview of the UKM cohort.

| Characteristic | All ICU stays | No 3 day readmission or death after ICU discharge | 3 day readmissions or death after ICU discharge |
|---|---|---|---|
| Number of ICU stays, *n* (%) | 15,589 (100.0) | 14,698 (94.3) | 891 (5.7) |
| Number of patients, *n* (%) | 14,188 (100.0) | 13,349 (94.1) | 839 (5.9) |
| Age, mean ± SD, years | 63.33 ± 14.73 | 63.16 ± 14.77 | 66.08 ± 13.85 |
| Female sex, *n* (%) | 4,919 (100.0) | 4,659 (94.7) | 260 (5.3) |
| Male sex, *n* (%) | 10,670 (100.0) | 10,039 (94.1) | 631 (5.9) |
| Length of ICU stay, mean ± SD, days | 3.70 ± 8.08 | 3.67 ± 8.11 | 4.23 ±7.53 |
| ICU at discharge | ICU 1 ($n = 4{,}063$) | ICU 1 ($n = 3{,}820$) | ICU 1 ($n = 243$) |
| | ICU 2 ($n = 6{,}402$) | ICU 2 ($n = 6{,}035$) | ICU 2 ($n = 367$) |
| | ICU 3 ($n = 1{,}034$) | ICU 3 ($n = 960$) | ICU 3 ($n = 74$) |
| | ICU 4 ($n = 4{,}090$) | ICU 4 ($n = 3{,}883$) | ICU 4 ($n = 207$) |

The key characteristics of all included ICU stays and the ICU stays divided by their labels. This information is based on ICU stays, so a single patient can be considered more than once.

## Variables and features

We included data that was routinely collected in the ICU for our analysis. For this purpose, 6,496 item definitions with 651,258,647 time-stamped recordings were extracted from the patient data management system (PDMS; Quantitative Sentinel, GE Healthcare) of the ANIT-UKM (see the flow chart in Supplementary material 2). We excluded all variables that were not collected during the study period ($n = 1{,}322$), derived variables computed using formulas in the PDMS ($n = 1{,}029$), and clinical notes because of the highly heterogeneous data quality ($n = 777$). We also excluded clinically irrelevant variables ($n = 1{,}979$) such as device-specific or billing information. The remaining 1,362 variables were processed in consultation with a senior physician who had extensive experience with the PDMS. For 802 non-medication variables, we determined the coverage across the study period and generated descriptive statistics to exclude irrelevant variables ($n = 522$). Of the resulting 280 variables, 70 were included directly, and 210 were further processed and merged into 50 variables. For medications, we assigned World Health Organization Anatomical Therapeutic Chemical (ATC) codes to all entries. We defined 44 clinically relevant medication categories within the ATC hierarchy and merged the respective variables. All medication variables that were not assigned to any category were excluded ($n = 187$). In addition, we manually determined five medication categories as additional variables for therapeutic and prophylactic antithrombotic agents and equivalence dosages of cardiac stimulants, norepinephrine and dopamine, and glucocorticoids, which we considered clinically relevant. Hence, we included 120 non-medication and 49 medication variables (Supplementary material 2). Further data cleaning methods are described in Supplementary material 2.

We assigned variables to nine different classes according to their data and generated respective features for each class (see Supplementary material 2). This was particularly important for time series data since EBM models cannot handle it. We featurized time series data *via* median, interquartile range (IQR), minimum, maximum, and linear trend for different time windows. We defined three time horizons (high, medium, and low) based on the median sampling interval of a variable that used different time windows before ICU discharge (high: 4, 12, and 24 hours; medium: 12, 24 hours, and 3 days; low: 1, 3, and 7 days). Hence, we generated 15 features for each time series variable. Patient flows, medications, and interventions were always considered as low time horizon. For patient flows,

we extrapolated the daily flow. For medications, we used a binary indicator and the number of administered drugs. For interventions, we also used a binary indicator and the interval since it was last performed. For static data, we used the last value from the most appropriate time interval (patient history, hospital stay, and ICU stay). Four additional features were created manually, which results in a total of 1,423 features. A list of all variables, feature classes, and their respective features is given in Supplementary material 2.

## Explainable boosting machines and baseline models

EBMs belong to the class of generalized additive models (32). A generalized additive model (33) models a label $\hat{y}$ by a bias term $\beta_0$ and a sum of features transformed by shape functions $f_i(x_i)$. The label $\hat{y}$ can optionally be transformed by a link function $g$ (see equation 1). EBMs add additional shape functions for the interactions of two variables $f_{i,j}(x_i, x_j)$ (34) and use the logit link function for dichotomous classifications analogous to LR (see equation 2, note that the logit function was moved to the right side).

$$g(\hat{y}) = \beta_0 + \sum_i f_i(x_i) \tag{1}$$

$$\hat{y} = logit^{-1}\left(\beta_0 + \sum_i f_i(x_i) + \sum_{i \neq j} f_{i,j}(x_i, x_j)\right) \tag{2}$$

In this study, the shape functions $f_i(x_i)$ and $f_{i,j}(x_i, x_j)$ of EBMs are also called one- (1D) and two-dimensional (2D) risk functions, because each of them models the log-odds of being readmitted to the ICU within 3 days. Different methods can be used to estimate the risk functions (33). EBMs use boosted decision trees that allow versatile function shapes that have shown optimal performance across several tasks (35). By visualizing the learned risk functions, EBMs can be inspected and owing to their modularity, inappropriate functions can be removed. Also, for a given input, contributions of each risk function can be used as an explanation of a prediction. A study that applied them in two health care tasks highlighted their potential to identify and remove spurious correlations (12). Moreover, an evaluation revealed that physicians can grasp the concept of EBMs and feel confident working with them (36). In this work, we compared to the validated Simplified Acute Physiology Score (SAPS) II, LR with feature selection, gradient boosting machines (GBMs), and recurrent neural networks (RNNs) with long short-term memory units for comparison (Supplementary material 2). We selected 130 features for the LR model, and we conjectured that inspecting this model requires a similar effort as inspecting our EBM model with at most 100 risk functions. Hence, the LR model serves as an interpretable baseline of the same complexity. GBMs and RNNs are both considered black box models owing to their complexity.

## Development of the EBM model with a limited number of risk functions

For our experiments, we used the area under the precision-recall curve (PR-AUC) as the primary performance indicator due to the label imbalance. We also reported the area under the receiver operating characteristic curve (ROC-AUC) since it is commonly reported in the medical literature. We selected the two most recent years for validation and hold-out data to simulate a real-world deployment (17). Five temporal splits were used for risk function selection and estimation of the standard deviation as pseudo-confidence intervals (Supplementary material 2).

To limit the model size and allow inspection in a reasonable amount of time, we performed automatic risk function selection of at most 80 1D and 20 2D functions based on their importance. To obtain good parameters, we first performed tuning based on the PR-AUC on the train and validation data of the full split (Supplementary material 2). We did this in three steps: we performed parameter tuning on all features, we estimated the 80 most important 1D risk functions approximately, and performed another parameter tuning for these 80 risk functions. Next, we used these parameters for risk function selection in a greedy stepwise forward procedure based on their mean importance on the five temporal splits (Supplementary material 2). We used the temporal splits to get more robust estimates and to prevent overfitting on the full split. A random 85% training and 15% validation split were used for each temporal split because a subset of variables was only collected for some years, which led to a biased weight estimate when using training and validation data based on years. Importance was calculated as the mean absolute log-odds score of a risk function. Finally, we chose the risk function selection with the highest PR-AUC performance on the full validation split. We repeated the same procedure for 2D risk functions on the features of the included 1D risk functions. This is coherent with the EBMs training algorithm, which first trains 1D functions and then adds 2D functions for the residuals.

## Inspection of the EBM model by a multidisciplinary team

The goal of the EBM model inspection was to identify the risk functions that should not remain in the final prediction model. The model was inspected by a team of three individuals: a senior physician working at the included ICUs, a senior

physician responsible for the data infrastructure at the ANIT-UKM, and the developer of the EBM model with a machine learning and health care background. They discussed and determined potential problems of the risk functions a priori to agree on a common set of exclusion criteria. For each risk function, they discussed its main properties and agreed on its content, then they determined if any of the identified problems applied, and then they decided if the problems justified the exclusion of a risk function. We recorded the identified problems for all risk functions (Supplementary material 3) and collected qualitative feedback during the EBM model inspection (Supplementary material 2).

### External validation on the medical information mart for intensive care version IV database

We used the Medical Information Mart for Intensive Care (MIMIC) version IV database for external validation (37, 38). It contains 76,540 ICU stays of 53,150 patients admitted to the Beth Israel Deaconess Medical Center between 2008 and 2019. After applying a similar cohort selection and labeling procedures, we included 19,108 ICU stays, of which 1,626 (8.5%) were labeled positively (Supplementary material 2). For performance comparison, we resampled negative instances to obtain the same positive rate as in the UKM cohort. We extracted 41 variables responsible for the 67 features used in the final EBM model from MIMIC-IV. Only a single variable could not be created. We also performed external validation with the GBM model, as it performed best in the model comparison. However, we only used the variables of the EBM model because extracting all variables from the MIMIC-IV database was not feasible. Both models were trained again on the MIMIC-IV data.

## Results

### Development of the EBM model with a limited number of risk functions

We first performed parameter tuning for an EBM with all features (Supplementary material 2). The best EBM with 1,423 1D risk functions achieved a PR-AUC of 0.151 ± 0.028 and a ROC-AUC of 0.652 ± 0.034 on the hold-out split. Next, we performed risk function selection based on the five temporal splits. Supplementary material 2 contains the performance for different numbers of risk functions and bin sizes. The best EBM model had a bin size of 200 and contained 80 1D risk functions. It achieved a PR-AUC of 0.130 ± 0.021 and a ROC-AUC of 0.681 ± 0.026. We repeated the same procedure for the 2D risk functions. We added five 2D functions with a bin size of four. The resulting model showed a decreased performance, with a PR-AUC of 0.113 ± 0.018 and ROC-AUC of 0.646 ± 0.01. The 85 most important risk functions of the resulting EBM model and their respective variables, features, and relative importance (variance) are listed in Table 2. The five 2D risk functions yielded the highest importance, followed by the 1D functions for endotracheal tube, age, antithrombotic agents in a prophylactic dosage, partial thromboplastin time, and $O_2$ saturation. The graphical representations of all risk functions are given in Supplementary material 3.

### Inspection of the EBM model by a multidisciplinary team

The resulting EBM model was inspected by a multidisciplinary team including two clinicians to identify and remove problematic risk functions. A priori to the model inspection, they identified four potential problems that they assigned to risk functions during the inspection:

- It encodes health care disparities that should not be reproduced ($n = 0$)
- It contains undesirable artifacts from the data generation process ($n = 8$)
- It contradicts medical knowledge ($n = 13$)
- It is not interpretable so that its effect cannot be clearly determined ($n = 17$).

The model inspection took 4 hours, that is, approximately 3 minutes per function. Not all risk functions with a problem were excluded, so we assigned the risk functions into three classes: included without problems ($n = 52$), included with problems ($n = 15$), and excluded with problems ($n = 18$). Most functions were excluded owing to the lack of interpretability ($n = 10$), followed by undesirable artifacts ($n = 6$) and contradictions of medical knowledge ($n = 6$). More than one problem could be assigned to each risk function. Five functions for partial thromboplastin time (PTT) were excluded because of artifacts. Using the feature histograms, the team recognized a change in the PTT measurement procedure since 2019, invalidating the risk functions learned on the training data. Also, all 2D risk functions were labeled as not interpretable and were excluded from the model. Figure 3 shows two included 1D risk functions and three 1D and one 2D functions that were excluded because of different problems. After model inspection, the EBM contained 67 1D risk functions. It achieved a PR-AUC of 0.119 ± 0.020 and a ROC-AUC of 0.680 ± 0.025 on the hold-out data. Hence, inspection decreased the PR-AUC and increased the ROC-AUC compared with a model trained on all 1D risk functions.

We collected qualitative feedback from the team during model inspection (Supplementary material 2). A major problem

TABLE 2 Overview of the variables and features of the risk functions included in the final EBM model ordered by importance.

| No. | Variable(s) | Feature(s) | Relative importance % | Excluded during model inspection |
|---|---|---|---|---|
| 1 | Age [years], Base Excess (BE) [mmol/L] | Static per patient, IQR 3 days | 4.20 | X |
| 2 | Drugs for constipation, Leucocytes [thousand/μL] | Unique 1 day, median 1 day | 3.52 | X |
| 3 | Blood volume out [mL], Procalcitonin [ng/mL] | Extrapolate 7 days, maximum 7 days | 2.57 | X |
| 4 | Hematocrit [%], Blood volume out [mL] | Maximum 3 days, extrapolate 3 days | 2.19 | X |
| 5 | Leucocytes [thousand/μL], Blood volume out [mL] | Median 1 day, extrapolate 3 days | 1.87 | X |
| 6 | Endotracheal tube (tubus) exists | Days since last application | 1.71 | |
| 7 | Age [years] | Static per patient | 1.70 | |
| 8 | Antithrombotic agents prophylactic dosage | Days since last application | 1.65 | |
| 9 | Partial thromboplastin time (PTT) [s] | Maximum 1 day | 1.63 | X |
| 10 | $O_2$ saturation [%] | Minimum 12 hours | 1.58 | |
| 11 | Blood volume out [mL] | Extrapolate 7 days | 1.52 | |
| 12 | Gamma-GT [U/L] | Median 7 days | 1.46 | |
| 13 | Chloride [mmol/L] | Trend per day 3 days | 1.40 | |
| 14 | Heart rate [bpm] | Minimum 4 hours | 1.39 | |
| 15 | Partial thromboplastin time (PTT) [s] | Maximum 3 days | 1.37 | X |
| 16 | Chloride [mmol/L] | Minimum 1 day | 1.37 | |
| 17 | Hemoglobin [mmol/L] | Maximum 3 days | 1.30 | |
| 18 | Length of stay before ICU [days] | Manually added | 1.28 | |
| 19 | Hematocrit [%] | Maximum 3 days | 1.26 | |
| 20 | Calcium [mmol/L] | Trend per day 3 days | 1.26 | X |
| 21 | Estimated glomerular filtration rate (eGFR) ml/min/1.73 $m^2$ | Trend per day 7 days | 1.24 | |
| 22 | Richmond agitation sedation (RAS) scale | Maximum 3 days | 1.24 | |
| 23 | Urine volume out [mL] | Extrapolate 1 day | 1.24 | |
| 24 | Thrombocytes [thousand/μL] | Trend per day 7 days | 1.24 | |
| 25 | Blood volume out [mL] | Extrapolate 3 days | 1.23 | |
| 26 | $paO_2/FiO_2$ [mmHg/$FiO_2$] | Median 1 day | 1.21 | |
| 27 | pH | Trend per day 3 days | 1.21 | |
| 28 | Phosphate [mg/dL] | Minimum 7 days | 1.20 | |
| 29 | pH | Median 1 day | 1.20 | |
| 30 | Body core temperature [°C] | Minimum 1 day | 1.18 | X |
| 31 | Creatine kinase (CK) [U/L] | Minimum 7 days | 1.15 | |
| 32 | Richmond agitation sedation (RAS) scale | Trend per day 12 hours | 1.13 | X |
| 33 | Potassium [mmol/L] | Median 1 day | 1.13 | |
| 34 | Glasgow coma scale (GCS) score | Minimum 3 days | 1.11 | |
| 35 | Body core temperature [°C] | Median 1 day | 1.10 | |
| 36 | Base excess (BE) [mmol/L] | IQR 3 days | 1.10 | X |
| 37 | Blood urea nitrogen [mg/dL] | Minimum 3 days | 1.10 | |
| 38 | $paO_2/FiO_2$ [mmHg/$FiO_2$] | Trend per day 3 days | 1.09 | |
| 39 | Drugs for constipation | Unique 1 day | 1.09 | |
| 40 | Urine volume out [mL] | Extrapolate 7 days | 1.09 | |
| 41 | Partial thromboplastin time (PTT) [s] | Minimum 7 days | 1.07 | X |
| 42 | Diastolic blood pressure [mmHg] | Median 1 day | 1.06 | |
| 43 | Partial pressure of oxygen ($pO_2$) [mmHg] | Minimum 12 hours | 1.06 | |
| 44 | Creatine kinase-MB (CK-MB) [U/L] | Maximum 3 days | 1.05 | |
| 45 | Richmond agitation sedation (RAS) scale | Maximum 1 day | 1.05 | |
| 46 | Partial thromboplastin time (PTT) [s] | Minimum 3 days | 1.05 | X |
| 47 | Systolic blood pressure [mmHg] | IQR 12 hours | 1.05 | |
| 48 | $paO_2/FiO_2$ [mmHg/$FiO_2$] | Median 3 days | 1.04 | |
| 49 | Creatine kinase (CK) [U/L] | Median 7 days | 1.04 | X |

*(Continued)*

TABLE 2 Continued

| No. | Variable(s) | Feature(s) | Relative importance % | Excluded during model inspection |
|---|---|---|---|---|
| 50 | Lactate [mmol/L] | Maximum 3 days | 1.04 | |
| 51 | Creatine kinase-MB (CK-MB) [U/L] | Median 3 days | 1.04 | |
| 52 | Lactate [mmol/L] | Minimum hours | 1.00 | |
| 53 | Phosphate [mg/dL] | Maximum 1 day | 1.00 | |
| 54 | Partial thromboplastin time (PTT) [s] | Maximum 7 days | 0.98 | X |
| 55 | Partial pressure of carbon dioxide ($PCO_2$) [mmHg] | Median 1 day | 0.98 | |
| 56 | Base excess (BE) [mmol/L] | Trend per day 3 days | 0.97 | |
| 57 | Glucose [mg/dL] | Median 3 days | 0.97 | |
| 58 | Base excess (BE) [mmol/L] | Minimum hours | 0.96 | |
| 59 | Methemoglobinemia (MetHb) [%] | Minimum hours | 0.96 | |
| 60 | Is on automatic ventilation | Days since last application | 0.95 | |
| 61 | Body core temperature [°C] | Minimum 4 hours | 0.95 | X |
| 62 | Partial pressure of carbon dioxide ($PCO_2$) [mmHg] | IQR 1 day | 0.95 | |
| 63 | Sodium [mmol/L] | Median 3 days | 0.93 | |
| 64 | Leucocytes [thousand/μL] | Median 1 day | 0.92 | |
| 65 | Sodium [mmol/L] | Trend per day 3 days | 0.92 | |
| 66 | Procalcitonin [ng/mL] | Maximum 7 days | 0.91 | |
| 67 | Base excess (BE) [mmol/L] | Median hours | 0.91 | |
| 68 | Mean blood pressure [mmHg] | Median 4 hours | 0.87 | |
| 69 | Leucocytes [thousand/μL] | Trend per day 3 days | 0.84 | X |
| 70 | pH | Median 3 days | 0.84 | |
| 71 | Bilirubin total [mg/dL] | Maximum 7 days | 0.84 | |
| 72 | Partial pressure of oxygen ($pO_2$) [mmHg] | IQR hours | 0.84 | |
| 73 | Base excess (BE) [mmol/L] | IQR 1 day | 0.83 | |
| 74 | Body core temperature [°C] | Trend per day 1 day | 0.83 | |
| 75 | C-reactive protein [mg/dL] | Maximum 3 days | 0.83 | |
| 76 | Heart rate [bpm] | Minimum 1 day | 0.82 | |
| 77 | Hematocrit [%] | Median hours | 0.80 | |
| 78 | Partial pressure of carbon dioxide ($PCO_2$) [mmHg] | Minimum 3 days | 0.76 | |
| 79 | Mean blood pressure [mmHg] | Median hours | 0.72 | |
| 80 | Calcium [mmol/L] | Maximum 1 day | 0.69 | |
| 81 | Estimated respiratory rate | Median 1 day | 0.68 | |
| 82 | pH | IQR 1 day | 0.67 | |
| 83 | Leucocytes [thousand/μL] | IQR 3 days | 0.63 | |
| 84 | Heart rate [bpm] | IQR 4 hours | 0.60 | |
| 85 | Reduced hemoglobin (RHb) | Median hours | 0.60 | X |

These risk functions were selected from a total of 1,423 based on their importance on the five-temporal splits. Risk functions 1–5 are two-dimensional, and the remaining functions are one-dimensional. The relative importance was determined on the final training split. The last column indicates whether a risk function was excluded during the model inspection by a team of physicians. Visualizations of all risk functions and the detailed reasons for exclusion are given in the supplement.

was drawing the line for risk function exclusion. Most functions partially fulfilled at least one problem. The team agreed to exclude a risk function when a problem was clearly present and would have a considerable impact on patients; that is, value ranges with many patients affected. Still, many functions could be assigned to either category (comments 1–3). The team stated that it was difficult to consider the cohort reduced to a single independent risk function (comments 4–7). This is against clinical practice, where several patient measurements are integrated. Also, only examining patient features at the time of discharge was hard, since usually the whole patient history is factored in (comment 8). In addition, the team members tended to construct explanations for risk functions without clear evidence (comment 9). Moreover, values outside the usual value ranges and IQR and trend features were more difficult to understand (comments 10 and 11). In particular, the 2D functions posed a problem because the combinations of features were uncommon in clinical practice. Even though

it was possible to grasp the content of the risk function, it was difficult to infer its clinical implications that led to exclusion (comment 12). There was a tendency to rely more on the model to derive useful relationships when a risk function was less interpretable (comment 13). In addition to that, we collected general properties that hindered or supported interpretability, which confirmed previous findings (36).

## Performance of EBM compared to baseline models

After the risk function selection and model inspection, the EBM model contained 67 1D risk functions. It achieved a PR-AUC of 0.119 ± 0.020 and a ROC-AUC of 0.680 ± 0.025 (Figure 4). For recall values of 0.4, 0.5, 0.6, and 0.8 the precision values were 0.130 ± 0.032, 0.111 ± 0.019, 0.105 ± 0.013, and 0.082 ± 0.005. Utilizing SAPS II in the last 24 hours showed an inferior performance of 0.084 ± 0.025 (PR-AUC) and 0.607 ± 0.019 (ROC-AUC). Also, LR with 130 selected features and the RNN achieved a lower performance, with a PR-AUC of 0.092 ± 0.026 and 0.095 ± 0.008 and a ROC-AUC of 0.587 ± 0.016 and 0.594 ± 0.027. Both were placed between the EBM and SAPS II for PR-AUC and below SAPS II for ROC-AUC. The latter could be due to the optimization of PR-AUC during parameter tuning and variable selection. The GBM trained on all 1,423 features achieved a PR-AUC of 0.123 ± 0.016 and a ROC-AUC of 0.665 ± 0.036. Hence, it performed similarly to the developed EBM model with 67 1D risk functions.

## External validation on the medical information mart for intensive care version IV database

The final EBM model for the UKM cohort used 67 features generated by 42 variables. We extracted 41 of those variables from MIMIC-IV. Variables were collected differently for the MIMIC cohort (Supplementary material 2). The EBM for external validation contained 66 1D risk functions. For the GBM model, we generated all the features of the 41 variables, resulting in 515 features. The EBM and GBM performed similarly on MIMIC-IV, with a PR-AUC of 0.221 ± 0.023 and 0.232 ± 0.029 and a ROC-AUC of 0.760 ± 0.010 and 0.772 ± 0.018 (Figure 4). This performance was much higher than that for the UKM cohort, which we mainly attributed to the better data quality of MIMIC-IV.

# Discussion

This study showed that for the prediction of 3 day ICU readmission, a transparent EBM model containing only 67 risk functions performed on par with state-of-the-art GBMs trained on 1,423 features and outperformed RNNs trained on time series data. Both the GBMs and RNNs can be considered black box models owing to their complexity. Hence, we found additional evidence that in a health care setting with structured data, a simple and inherently interpretable model can be sufficient for competitive prediction performance (10). The final model achieved a PR-AUC of 0.119 ± 0.020 and a ROC-AUC of 0.680 ± 0.025. External validation on the MIMIC-IV database showed improved EBM results of a PR-AUC of 0.221 ± 0.023 and a ROC-AUC of 0.760 ± 0.010 and confirmed that they performed similarly to the GBMs. Our results are consistent with those of previous studies, showing that EBMs outperformed LR and were on par with random forests and boosting methods (12, 34). However, in contrast to the existing work, adding 2D risk functions lead to lower performance on the hold-out data. Several risk functions of the final EBM model are consistent with the main risk factors reported in the literature (4, 6, 7), such as age, length of hospital stay before ICU admission, disease severity (e.g., based on the GCS score), physiological state (e.g., heart rate), and need for organ support (e.g., presence of an endotracheal tube). In our study, many concepts had much finer granularity; for example, several variables captured the physiological state of the patient. We also note that some known risk factors were available features but did not end up in the final model. Among those are sex, admission origin, and use of vasopressors. However, some information might be mediated through other variables. For example, blood loss is usually a clear indicator of a past surgery and might contain additional information, making it more relevant than a simple indicator for surgery. The overall predictive performance for 3 day ICU readmissions was relatively low. This is probably due to the limitations regarding data quality, which are supported by the higher performance on MIMIC-IV. MIMIC-IV was created in several iterations and integrated the feedback of many researchers, which led to higher data quality. Moreover, the prediction of ICU readmission prediction is a difficult task, and only a few readmissions are preventable (39). Still, we think that an EBM model for the prediction of 3 day ICU trained on a local cohort can offer useful insights for decision-making in the ICU.

Several studies on ICU readmission prediction have been conducted (26, 40–52), and we identified two systematic reviews (53, 54). Most of them also used MIMIC (38), not the most recent version IV, for model development or validation. The readmission intervals ranged from 48 hours (46, 47, 52) to 72 hours (26, 50, 51), 7 days (48), 30 days (40, 44, 49), and anytime until hospital discharge (41–43). A single study considers multiple intervals of 24 hours, 72 hours, 7 days, 30 days, and anytime (45). We chose an ICU readmission interval of 3 days because clinicians at the ANIT-UKM expressed that it would include relevant medical conditions that they could act upon before discharging a patient and, hence, would

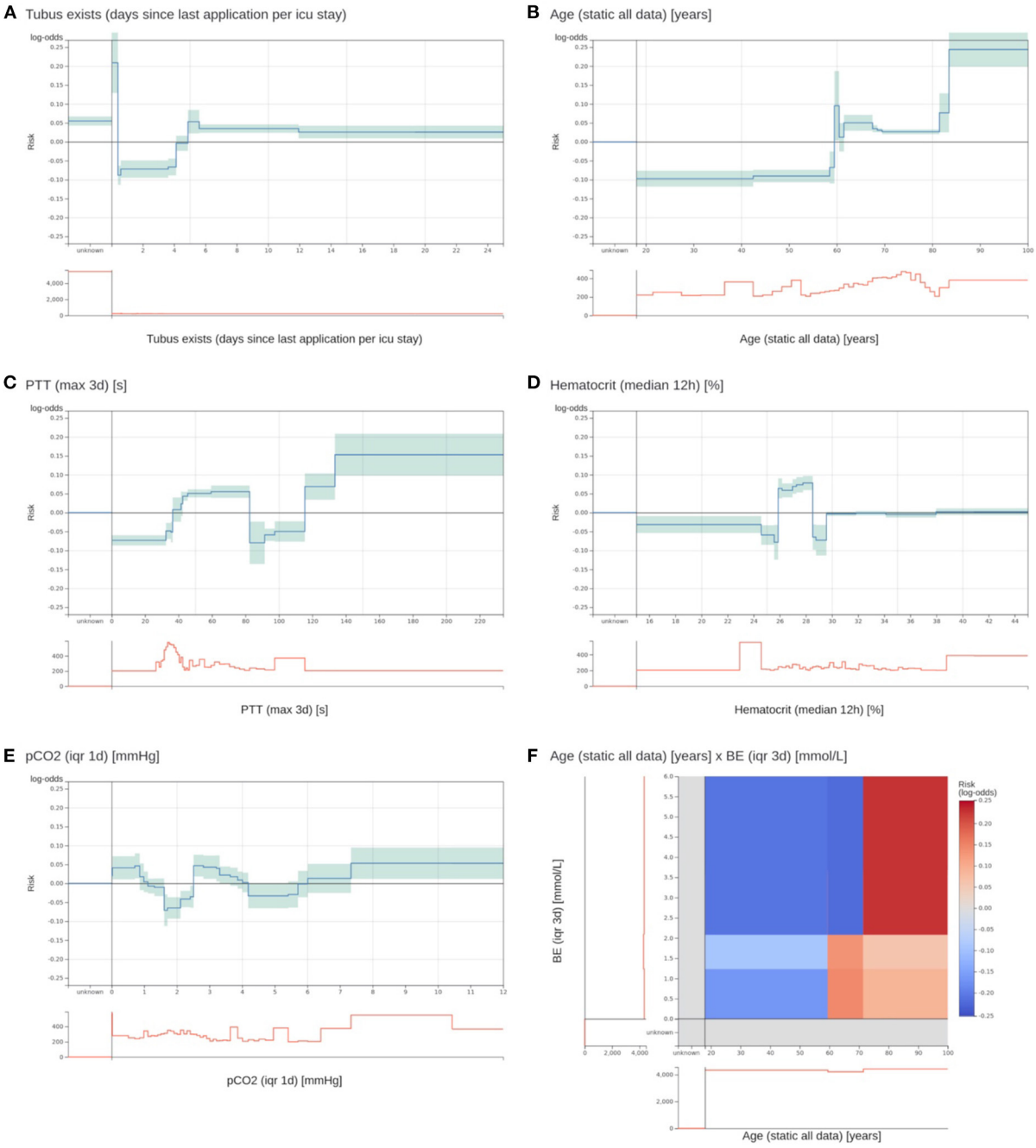



FIGURE 3
Two most important risk functions and four excluded risk functions of the EBM model. **(A,B)** Two most important risk functions that are included in the EBM model. **(A)** Contains the number of days since the last existence of an endotracheal tube. Patients that have an endotracheal tube immediately before discharge have a highly increased risk. Lower risk is assigned to values between 0.4 and 4.1 days. Also, patients with no endotracheal tube (unknown) receive an increased risk. **(B)** The risk function for age shows an increased risk for higher age values. There is a peak at 60 years with no obvious explanation. **(C)** A maximum PTT value over the last 3 days before discharge between 82.5 and 115.5 s gets a lower risk for 3 day ICU readmission. It was identified that this is an artifact of the previous procedure to determine the PTT for cardiac surgery patients. This will not generalize for future data. **(D)** For a median hematocrit between 24.875 and 28.525%, the model determined an elevated risk. For slightly lower and higher values, the risk is negative. This is against common medical knowledge, where a decreasing hematocrit value should be associated with increased risk. **(E)** The interquartile range (IQR) of the partial pressure of carbon dioxide (pCO2) over the last day before discharge receives an increased risk for values between 0 and 0.863 and 2.513 and 3.313 mmHg. However, the interpretation of this behavior and determining its clinical implications was impossible. **(F)** The 2D risk function for age and the IQR of the base excess (BE) over 3 days. Patients over 71.5 years have a high risk for a high IQR of the BE. Patients between 59.5 and 71.5 have only a slightly increased risk for low IQR values, and younger patients have a decreased risk across all BE values. The team excluded it due to a lack of interpretability.

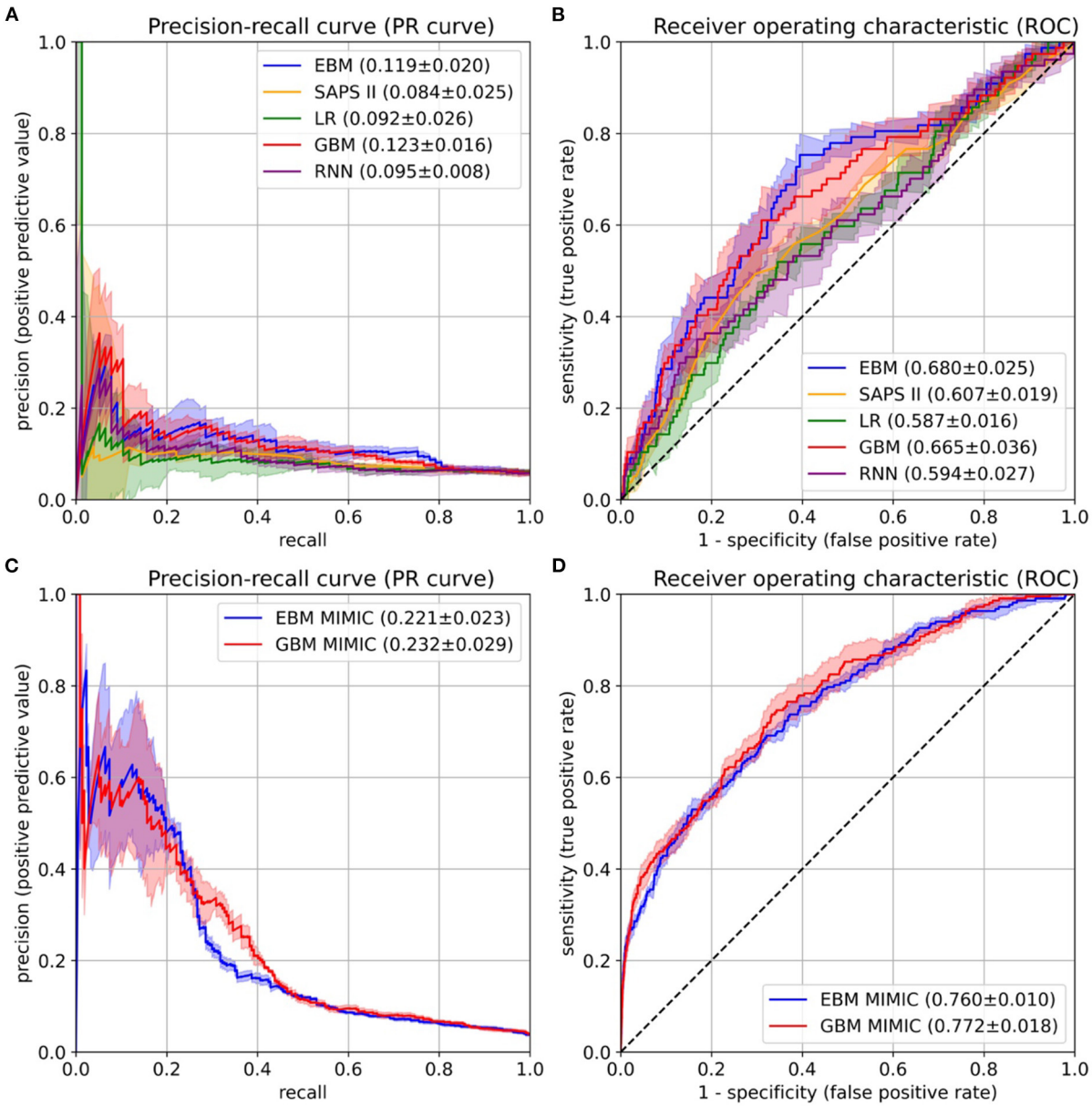



FIGURE 4
Performance evaluation on the University Hospital Münster (UKM) cohort **(A,B)** and external validation on the Medical Information Mart for Intensive Care version IV (MIMIC-IV) database **(C,D)**. **(A)** The area under the precision-recall curve (PR-AUC) was considered the most relevant performance indicator owing to the imbalanced label distribution. We optimized the PR-AUC during the parameter tuning and selection procedures for all models. The differences between models are relatively small. The explainable boosting machines (EBMs) and gradient boosting machines (GBMs) show the highest PR-AUC. **(B)** The area under the receiver operating characteristic curve (ROC-AUC) was determined as an additional performance measure. Again, the EBM and GBM models performed best. **(C,D)** The same performance indicators were determined on the MIMIC-IV database. Both models again showed similar results. The confidence intervals for all curves were determined with the standard deviation on the five temporal splits.

be most useful in practice. Also, we considered it a good trade-off between having sufficient follow-up and preventing exclusion of patients due to loss of follow-up (see step 7 in Figure 2). Previous studies have tested many models, and two (26, 27) mentioned the goal of developing interpretable models, but no validation by humans was performed. All studies reported ROC-AUC, which ranged from 0.64 (52) to 0.91 (42). Unfortunately, comparing the performance with the existing work is impossible for two reasons. First, we considered PR-AUC due to the label imbalance of ICU readmissions and

optimized it in our experiments. However, none of the existing studies have reported this performance measure. One study contained a precision-recall curve (47), but no area under the curve. Second, we created a custom UKM cohort, and we used MIMIC-IV for external validation. None of the identified studies used these data. If the ROC-AUC is considered as a performance measure, our results are in the lower spectrum of the reported models. However, we did not optimize for it in our experiments.

A main goal of this study was to involve clinicians in the model development process to inspect the learned EBM and remove problematic risk functions. This approach showed mixed results. On the one hand, our collaboration confirmed that clinicians can easily grasp the concept of EBMs (36), making them a useful transparent model candidate for health care applications (55). Like LR, which is well-known in the medical domain, feature contributions are summed to a total log-odds score. This modularity also allowed to focus on a single risk function at a time. Confidence intervals and histograms over patient densities further helped to assess the relevance of function segments. For instance, it was possible to ignore fluctuations of risk functions in regions with few patients. In addition, our model development process enabled discussions with clinicians and encouraged a critical review of the model. Several aspects were raised for the first time, such as the problem with PTT measurements. Hence, with EBMs, stakeholders can be involved in the development process to establish trust, which could ultimately lead to higher adoption rates (13). Moreover, we identified and removed 18 risk functions due to the lack of interpretability, undesirable data artifacts, and contradiction of medical knowledge. This demonstrates the capability of EBMs to enable the identification and removal of undesirable components. This would have been impossible with a black box ML model (10, 12). Lastly, model inspection led to a performance increase on the hold-out data, which suggests better generalization.

However, we also observed several shortcomings during the model inspection. Of the 85 risk functions, 33 were labeled as problematic, of which 17 were not interpretable. Reducing a patient cohort to one or two features and considering a fixed time interval before discharge are counter to typical clinical practice, where many variables are usually integrated over a long time horizon. Thus, it was often difficult to create an intuition about the effect of certain risk functions. Also, for meaningful interpretation of EBMs, it is necessary to understand the model inputs (24, 55). In particular, interpretability was hindered by variables and descriptive statistics that are less common in clinical practice. One workaround would be to let clinicians choose interpretable features a priori. In addition, the shapes of risk functions sometimes showed a fluctuating behavior (36). We already increased the bin size to prevent these artifacts, but some still occurred in the final model. Another major issue was drawing the line between the inclusion and exclusion of risk functions. Most functions showed problematic behaviors. Thus, we decided to exclude only functions with a problem that affected a considerable part of the cohort. However, this decision rule is vague, and we expect low interrater reliability. We think it could be helpful to have a clear application scenario to determine more specific rules for exclusion. Moreover, we observed that it was more difficult to justify the exclusion of less interpretable functions and that the team relied on the EBM algorithm to find relevant associations in the data (56, 57).

This work has limitations. Even though the prediction of ICU readmission is a relevant medical problem, it can be difficult to turn predictions into actions when institutional factors such as insufficient ICU beds must be considered. No multicenter cohort was used for the development and validation of our prediction model, so the external validity of our results is low. Also, the data quality of the local cohort was limited, and our experiments only focused on a single interpretable model. External validation on the MIMIC-IV database was only performed for two models, and no in-depth analysis was performed for the improved performance. Moreover, interpretability should be evaluated in the context of its end task (14). Ideally, this could be increased trust leading to higher adoption of the system or even improved patient outcomes. We limited our analysis to prediction performance, the identification of problematic risk functions, and qualitative feedback. Moreover, no rigorous set of rules has been established for model inspection, so the process would likely exhibit low interrater reliability. The confidence intervals of the performance were only estimated on five temporal splits, and our EBM did not outperform the existing ML models by a large margin. Lastly, automatic risk function selection for EBMs might have removed important confounders, making it impossible to detect them during the model inspection.

## Conclusion

We demonstrated a procedure to develop a transparent EBM model for the prediction of 3 day ICU readmission that involved clinicians to inspect and verify the learned model. The EBM performed on par with or outperformed state-of-the-art black box ML models such as GBMs and RNNs. This suggests that a simple inherently interpretable model might suffice for clinical use in cases with low- to medium-dimensional data, while allowing a high level of human control. Evaluation of the model inspection revealed that an EBM model can facilitate a critical review with clinicians and enables identification of problematic components.

## Data availability statement

The patient datasets in this article are not readily available to protect patient privacy. The MIMIC-IV dataset used for external

validation is available from https://doi.org/10.13026/s6n6-xd98. All the code used for the experiments is available from https://doi.org/10.5281/zenodo.5627167.

## Ethics statement

The studies involving human participants were reviewed and approved by the ethical review board of the medical chamber Westfalen-Lippe approved this study (reference number: 2020-526-f-S). Written informed consent for participation was not required for this study in accordance with the national legislation and the institutional requirements.

## Author contributions

SH designed the study, developed the code for all experiments, and wrote the manuscript. SH, TV, and CE performed the data pre-processing, cohort selection, and experiments. All authors provided critical feedback and helped shape the research and manuscript. All authors contributed to the article and approved the submitted version.

## Funding

This work was supported by the German Research Foundation (Deutsche Forschungsgemeinschaft, DFG Grants DU 352/11-1 and DU 352/11-2).

## Acknowledgments

The authors would like to thank Oliver Wenning for helping with the software development and Monica Agrawal for proofreading the article. This manuscript has been released as a preprint at medRxiv (58).

## Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

## Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

## Supplementary material

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fmed.2022.960296/full#supplementary-material

## References

1. Kramer AA, Higgins TL, Zimmerman JE. Can this patient be safely discharged from the ICU? *Intensive Care Med.* (2016) 42:580–2. doi: 10.1007/s00134-015-4148-8

2. Rosenberg AL, Watts C. Patients readmitted to ICUs: a systematic review of risk factors and outcomes. *Chest.* (2000) 118:492–502. doi: 10.1378/chest.118.2.492

3. Renton J, Pilcher DV, Santamaria JD, Stow P, Bailey M, Hart G, et al. Factors associated with increased risk of readmission to intensive care in Australia. *Intensive Care Med.* (2011) 37:1800. doi: 10.1007/s00134-011-2318-x

4. Kramer AA, Higgins TL, Zimmerman JE. Intensive care unit readmissions in U.S. hospitals: patient characteristics, risk factors, and outcomes. *Crit Care Med.* (2012) 40:3–10. doi: 10.1097/CCM.0b013e31822d751e

5. Kramer AA, Higgins TL, Zimmerman JE. The association between ICU readmission rate and patient outcomes. *Crit Care Med.* (2013) 41:24–33. doi: 10.1097/CCM.0b013e3182657b8a

6. Ponzoni CR, Corrêa TD, Filho RR, Serpa Neto A, Assunção MSC, Pardini A, et al. Readmission to the intensive care unit: incidence, risk factors, resource use, and outcomes. A retrospective cohort study. *Ann Am Thorac Soc.* (2017) 14:1312–9. doi: 10.1513/AnnalsATS.201611-851OC

7. Santamaria JD, Duke GJ, Pilcher DV, Cooper DJ, Moran J, Bellomo R. Readmissions to intensive care: a prospective multicenter study in Australia and New Zealand. *Crit Care Med.* (2017) 45:290–7. doi: 10.1097/CCM.0000000000002066

8. Wright MC, Dunbar S, Macpherson BC, Moretti EW, Fiol GD, Bolte J, et al. Toward designing information display to support critical care. *Appl Clin Inform.* (2016) 07:912–29. doi: 10.4338/ACI-2016-03-RA-0033

9. Rajkomar A, Dean J, Kohane I. Machine learning in medicine. *N Engl J Med.* (2019) 380:1347–58. doi: 10.1056/NEJMra1814259

10. Rudin C. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nat Mach Intell.* (2019) 1:206–15. doi: 10.1038/s42256-019-0048-x

11. Obermeyer Z, Powers B, Vogeli C, Mullainathan S. Dissecting racial bias in an algorithm used to manage the health of populations. *Science.* (2019) 366:447–53. doi: 10.1126/science.aax2342

12. Caruana R, Lou Y, Gehrke J, Koch P, Sturm M, Elhadad N. Intelligible models for healthcare: predicting pneumonia risk and hospital 30-day readmission. In: *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining - KDD '15.* Sydney, NSW: ACM Press (2015). p. 1721–30. doi: 10.1145/2783258.2788613

13. Wiens J, Saria S, Sendak M, Ghassemi M, Liu VX, Doshi-Velez F, et al. Do no harm: a roadmap for responsible machine learning for health care. *Nat Med.* (2019) 25:1337–40. doi: 10.1038/s41591-019-0548-6

14. Doshi-Velez F, Kim B. Towards a rigorous science of interpretable machine learning. *arXiv.* (2017) arXiv170208608. Available online at: http://arxiv.org/abs/1702.08608

15. Ribeiro MT, Singh S, Guestrin C. "Why should i trust you?": explaining the predictions of any classifier. *arXiv*. (2016) arXiv:160204938. Available online at: http://arxiv.org/abs/1602.04938

16. Lundberg SM, Lee SI. A unified approach to interpreting model predictions. In: Guyon I, Luxburg UV, Bengio S, Wallach H, Fergus R, Vishwanathan S, et al, editors. *Advances in Neural Information Processing Systems 30*. Curran Associates Inc. (2017). p. 4765–74. Available online at: http://papers.nips.cc/paper/7062-a-unified-approach-to-interpreting-model-predictions.pdf

17. Hyland SL, Faltys M, Hüser M, Lyu X, Gumbsch T, Esteban C, et al. Early prediction of circulatory failure in the intensive care unit using machine learning. *Nat Med.* (2020) 26:364–73. doi: 10.1038/s41591-020-0789-4

18. Thorsen-Meyer HC, Nielsen AB, Nielsen AP, Kaas-Hansen BS, Toft P, Schierbeck J, et al. Dynamic and explainable machine learning prediction of mortality in patients in the intensive care unit: a retrospective study of high-frequency data in electronic patient records. *Lancet Dig Health.* (2020) 2:e179–91. doi: 10.1016/S2589-7500(20)30018-2

19. Lauritsen SM, Kristensen M, Olsen MV, Larsen MS, Lauritsen KM, Jørgensen MJ, et al. Explainable artificial intelligence model to predict acute critical illness from electronic health records. *Nat Commun.* (2020) 11:3852. doi: 10.1038/s41467-020-17431-x

20. Alvarez-Melis D, Jaakkola TS. On the robustness of interpretability methods. *arXiv*. (2018) arXiv:180608049. Available online at: http://arxiv.org/abs/1806.08049

21. Slack D, Hilgard S, Jia E, Singh S, Lakkaraju H. Fooling LIME and SHAP: adversarial attacks on post hoc explanation methods. In: *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*. New York, NY: ACM (2020). p. 180–6. doi: 10.1145/3375627.3375830

22. Laugel T, Lesot MJ, Marsala C, Renard X, Detyniecki M. The dangers of post-hoc interpretability: unjustified counterfactual explanations. *arXiv*. (2019) arXiv:190709294. doi: 10.24963/ijcai.2019/388

23. Ghassemi M, Oakden-Rayner L, Beam AL. The false hope of current approaches to explainable artificial intelligence in health care. *Lancet Dig Health.* (2021) 3:e745–50. doi: 10.1016/S2589-7500(21)00208-9

24. Lipton ZC. The mythos of model interpretability. *arXiv*. (2017) arXiv:160603490. Available online at: http://arxiv.org/abs/1606.03490

25. Payrovnaziri SN, Chen Z, Rengifo-Moreno P, Miller T, Bian J, Chen JH, et al. Explainable artificial intelligence models using real-world electronic health record data: a systematic scoping review. *J Am Med Inform Assoc.* (2020) 27:1173–85. doi: 10.1093/jamia/ocaa053

26. Fialho AS, Cismondi F, Vieira SM, Reti SR, Sousa JMC, Finkelstein SN. Data mining using clinical physiology at discharge to predict ICU readmissions. *Expert Syst Appl.* (2012) 39:13158–65. doi: 10.1016/j.eswa.2012.05.086

27. Badawi O, Breslow MJ. Readmissions and death after ICU discharge: development and validation of two predictive models. *PLoS ONE.* (2012) 7:e0048758. doi: 10.1371/journal.pone.0048758

28. Collins GS, Reitsma JB, Altman DG, Moons KGM. Transparent reporting of a multivariable prediction model for individual prognosis Or diagnosis (TRIPOD): the TRIPOD statement. *Ann Intern Med.* (2015) 162:55. doi: 10.7326/M14-0697

29. Hegselmann S. *Interpretable 7-Day Intensive Care Unit Readmission Prediction Using Machine Learning: a Retrospective Study*. OSF Registries (2020). Available online at: https://osf.io/9r7gt (accessed March 9, 2021).

30. Hegselmann S. *stefanhgm/Interpretable-3-Day-ICU-Readmission-Prediction: Initial Version Used for all Experiments in the Paper*. Zenodo (2021). Available online at: https://zenodo.org/record/5627167 (accessed October 30, 2021).

31. Hegselmann S. *stefanhgm/EBM-Java-UI: Initial Version of EBM-Java-UI*. Zenodo (2021). Available online at: https://zenodo.org/record/5541444 (accessed September 30, 2021).

32. Nori H, Jenkins S, Koch P, Caruana R. InterpretML: a unified framework for machine learning interpretability. *arXiv*. (2019) arXiv:190909223. Available online at: http://arxiv.org/abs/1909.09223

33. Hastie T, Tibshirani R. Generalized additive models. *Stat Sci.* (1986) 1:297–310. doi: 10.1214/ss/1177013604

34. Lou Y, Caruana R, Gehrke J, Hooker G. Accurate intelligible models with pairwise interactions. In: *Proceedings of the 19th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining - KDD '13*. Chicago, IL: ACM Press (2013). p. 623. doi: 10.1145/2487575.2487579

35. Lou Y, Caruana R, Gehrke J. Intelligible models for classification and regression. In: *Proceedings of the 18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining—KDD '12*. Beijing: ACM Press (2012). p. 150. doi: 10.1145/2339530.2339556

36. Hegselmann S, Volkert T, Ohlenburg H, Gottschalk A, Dugas M, Ertmer C. An evaluation of the doctor-interpretability of generalized additive models with interactions. In: *Machine Learning for Healthcare Conference*. PMLR (2020). p. 46–79. Available online at: http://proceedings.mlr.press/v126/hegselmann20a.html (accessed March 9, 2021).

37. Johnson A, Bulgarelli L, Pollard T, Horng S, Celi LA, Mark R. *MIMIC-IV*. PhysioNet. Available online at: https://physionet.org/content/mimiciv/1.0/ (accessed August 23, 2021).

38. Goldberger AL, Amaral LAN, Glass L, Hausdorff JM, Ivanov PCh, Mark RG, et al. PhysioBank, PhysioToolkit, and PhysioNet: components of a new research resource for complex physiologic signals. *Circulation.* (2000) 101:E215–20. doi: 10.1161/01.CIR.101.23.e215

39. Al-Jaghbeer MJ, Tekwani SS, Gunn SR, Kahn JM. Incidence and etiology of potentially preventable ICU readmissions. *Crit Care Med.* (2016) 44:1704–9. doi: 10.1097/CCM.0000000000001746

40. Barbieri S, Kemp J, Perez-Concha O, Kotwal S, Gallagher M, Ritchie A, et al. Benchmarking deep learning architectures for predicting readmission to the ICU and describing patients-at-risk. *Sci Rep.* (2020) 10:1111. doi: 10.1038/s41598-020-58053-z

41. Lin Y, Wu JY, Lin K, Hu YH, Kong GL. Prediction of intensive care unit readmission for critically ill patients based on ensemble learning. *Beijing Da Xue Xue Bao Yi Xue Ban.* (2021) 53:566–72. doi: 10.19723/j.issn.1671-167X.2021.03.021

42. Loreto M, Lisboa T, Moreira VP. Early prediction of ICU readmissions using classification algorithms. *Comput Biol Med.* (2020) 118:103636. doi: 10.1016/j.compbiomed.2020.103636

43. Rojas JC, Carey KA, Edelson DP, Venable LR, Howell MD, Churpek MM. Predicting intensive care unit readmission with machine learning using electronic health record data. *Ann Am Thorac Soc.* (2018) 15:846–53. doi: 10.1513/AnnalsATS.201710-787OC

44. Lin YW, Zhou Y, Faghri F, Shaw MJ, Campbell RH. Analysis and prediction of unplanned intensive care unit readmission using recurrent neural networks with long short-term memory. *PLoS ONE.* (2019) 14:e0218942. doi: 10.1371/journal.pone.0218942

45. Pakbin A, Rafi P, Hurley N, Schulz W, Harlan Krumholz M, Bobak Mortazavi J. Prediction of ICU readmissions using data at patient discharge. In: *2018 40th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*. Honolulu, HI: IEEE (2018). p. 4932–5. doi: 10.1109/EMBC.2018.8513181

46. 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:e017199. doi: 10.1136/bmjopen-2017-017199

47. McWilliams CJ, Lawson DJ, Santos-Rodriguez R, Gilchrist ID, Champneys A, Gould TH, et al. Towards a decision support tool for intensive care discharge: machine learning algorithm development using electronic healthcare data from MIMIC-III and Bristol, UK. *BMJ Open.* (2019) 9:e025925. doi: 10.1136/bmjopen-2018-025925

48. Ouanes I, Schwebel C, Français A, Bruel C, Philippart F, Vesin A, et al. A model to predict short-term death or readmission after intensive care unit discharge. *J Crit Care.* (2012) 27:422.e1–9. doi: 10.1016/j.jcrc.2011.08.003

49. Xue Y, Klabjan D, Luo Y. Predicting ICU readmission using grouped physiological and medication trends. *Artif Intell Med.* (2019) 95:27–37. doi: 10.1016/j.artmed.2018.08.004

50. Curto S, Carvalho JP, Salgado C, Vieira SM, Sousa JMC. Predicting ICU readmissions based on bedside medical text notes. In: *2016 IEEE International Conference on Fuzzy Systems (FUZZ-IEEE)*. Vancouver, BC: IEEE (2016). p. 2144-a—51-h. doi: 10.1109/FUZZ-IEEE.2016.7737956

51. Abu-Awwad R, Buran G. Predictors of early readmission to the intensive care unit. *Chest.* (2012) 142:280A. doi: 10.1378/chest.1390058

52. Brown SES, Ratcliffe SJ, Kahn JM, Halpern SD. The epidemiology of intensive care unit readmissions in the United States. *Am J Respir Crit Care Med.* (2012) 185:955–64. doi: 10.1164/rccm.201109-1720OC

53. Hosein FS, Bobrovitz N, Berthelot S, Zygun D, Ghali WA, Stelfox HT. A systematic review of tools for predicting severe adverse events following patient discharge from intensive care units. *Crit Care.* (2013) 17:R102. doi: 10.1186/cc12747

54. Markazi-Moghaddam N, Fathi M, Ramezankhani A. Risk prediction models for intensive care unit readmission: A systematic review of methodology and applicability. *Austral Crit Care.* (2019) 33:367–74. doi: 10.1016/j.aucc.2019.05.005

55. Tonekaboni S, Joshi S, McCradden MD, Goldenberg A. What clinicians want: contextualizing explainable machine learning for clinical end use. In: *Proceedings*

*of the 4th Machine Learning for Healthcare Conference*. PMLR (2019). p. 359–80. Available online at: https://proceedings.mlr.press/v106/tonekaboni19a.html

56. Kaur H, Nori H, Jenkins S, Caruana R, Wallach H, Wortman Vaughan J. Interpreting interpretability: understanding data scientists' use of interpretability tools for machine learning. In: *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*. Honolulu, HI: ACM (2020). p. 1–14. doi: 10.1145/3313831.3376219

57. Poursabzi-Sangdeh F, Goldstein DG, Hofman JM, Vaughan JW, Wallach H. Manipulating and measuring model interpretability. *arXiv*. (2021) arXiv:180207810. doi: 10.1145/3411764.3445315

58. Hegselmann S, Ertmer C, Volkert T, Gottschalk A, Dugas M, Varghese J. Development and validation of an interpretable 3-day intensive care unit readmission prediction model using explainable boosting machines. *medRxiv*. (2021). doi: 10.1101/2021.11.01.21265700

frontiers | Frontiers in Medicine

TYPE Original Research
PUBLISHED 28 September 2022
DOI 10.3389/fmed.2022.933037



OPEN ACCESS

EDITED BY
Longxiang Su,
Peking Union Medical College Hospital (CAMS), China

REVIEWED BY
Fady Alnajjar,
United Arab Emirates University, United Arab Emirates
Prashant Nasa,
NMC Specialty Hospital Al Nahda, United Arab Emirates

*CORRESPONDENCE
Baohua Liu
baohualiu@bjmu.edu.cn
Yong Jiang
jy78@vip.sina.com

SPECIALTY SECTION
This article was submitted to Intensive Care Medicine and Anesthesiology, a section of the journal Frontiers in Medicine

RECEIVED 30 April 2022
ACCEPTED 01 September 2022
PUBLISHED 28 September 2022

CITATION
Deng Y, Liu S, Wang Z, Wang Y, Jiang Y and Liu B (2022) Explainable time-series deep learning models for the prediction of mortality, prolonged length of stay and 30-day readmission in intensive care patients. *Front. Med.* 9:933037. doi: 10.3389/fmed.2022.933037



# Explainable time-series deep learning models for the prediction of mortality, prolonged length of stay and 30-day readmission in intensive care patients

Yuhan Deng[1], Shuang Liu[1], Ziyao Wang[1], Yuxin Wang[1], Yong Jiang[2,3]* and Baohua Liu[1]*

[1]School of Public Health, Peking University, Beijing, China, [2]Department of Neurology, Beijing Tiantan Hospital, Capital Medical University, Beijing, China, [3]China National Clinical Research Center for Neurological Diseases, Beijing, China

**Background:** In-hospital mortality, prolonged length of stay (LOS), and 30-day readmission are common outcomes in the intensive care unit (ICU). Traditional scoring systems and machine learning models for predicting these outcomes usually ignore the characteristics of ICU data, which are time-series forms. We aimed to use time-series deep learning models with the selective combination of three widely used scoring systems to predict these outcomes.

**Materials and methods:** A retrospective cohort study was conducted on 40,083 patients in ICU from the Medical Information Mart for Intensive Care-IV (MIMIC-IV) database. Three deep learning models, namely, recurrent neural network (RNN), gated recurrent unit (GRU), and long short-term memory (LSTM) with attention mechanisms, were trained for the prediction of in-hospital mortality, prolonged LOS, and 30-day readmission with variables collected during the initial 24 h after ICU admission or the last 24 h before discharge. The inclusion of variables was based on three widely used scoring systems, namely, APACHE II, SOFA, and SAPS II, and the predictors consisted of time-series vital signs, laboratory tests, medication, and procedures. The patients were randomly divided into a training set (80%) and a test set (20%), which were used for model development and model evaluation, respectively. The area under the receiver operating characteristic curve (AUC), sensitivity, specificity, and Brier scores were used to evaluate model performance. Variable significance was identified through attention mechanisms.

**Results:** A total of 33 variables for 40,083 patients were enrolled for mortality and prolonged LOS prediction and 36,180 for readmission prediction. The rates of occurrence of the three outcomes were 9.74%, 27.54%, and 11.79%, respectively. In each of the three outcomes, the performance of RNN, GRU, and LSTM did not differ greatly. Mortality prediction models, prolonged LOS prediction models, and readmission prediction models achieved AUCs of 0.870 ± 0.001, 0.765 ± 0.003, and 0.635 ± 0.018, respectively. The top

significant variables co-selected by the three deep learning models were Glasgow Coma Scale (GCS), age, blood urea nitrogen, and norepinephrine for mortality; GCS, invasive ventilation, and blood urea nitrogen for prolonged LOS; and blood urea nitrogen, GCS, and ethnicity for readmission.

**Conclusion:** The prognostic prediction models established in our study achieved good performance in predicting common outcomes of patients in ICU, especially in mortality prediction. In addition, GCS and blood urea nitrogen were identified as the most important factors strongly associated with adverse ICU events.

KEYWORDS

**intensive care unit (ICU), mortality, length of stay, readmission, prognostic prediction, deep learning**

# Introduction

Patients in the intensive care unit (ICU) are usually critically ill, presenting a high mortality risk compared with other departments in the hospital (1). In addition, readmission and prolonged length of stay (LOS) are both common clinical outcomes indicating patients' health conditions (2, 3), critical care quality (4, 5), and medical efficiency (6). Thus, early identification of seriously ill patients and those with prolonged LOS and readmission risk and subsequent management is exceedingly important in improving patient outcomes and providing optimal allocation of medical resources.

However, traditional scoring systems, even some machine learning methods in predicting these outcomes, especially in stratifying the risk of readmission, have shown only modest results (7–10). Although part of the existing work based on machine learning models seems promising (11–13), few of them are able to take advantage of the characteristics of features collected in the ICU, which are time-series forms. Presently, these time-series problems can be approached with deep learning-based models, such as recurrent neural network (RNN) and its derived models, namely, gated recurrent unit (GRU) (14) and long short-term memory (LSTM) (15), which can learn valuable information from a large number of rapidly changing variables, making it possible to make full use of ICU data collected at a high frequency (16). Based on these advanced models, several studies have conducted prognostic prediction of patients in ICU, but most were disease-specific or ICU-specific (17–20), the clinical use of which was restricted to a specific group. To the best of our knowledge, no studies have ever predicted common outcomes while maximizing the value of these models of patients in general ICU. Furthermore, because of the complexity of these deep learning models, they are not easy to interpret, which restricts their practical application to clinical decisions (21, 22). Therefore, transparency and explainability must be considered when constructing prediction models. Recently, several methods have been introduced to improve model interpretability; among them, attention mechanisms seem to be one of the most prospective approaches (23), which have been proven to provide the foundation for clinical interpretation (24). Through explainable prediction models, significant factors can be identified at an early stage to help clinicians offer better medical interventions.

In this study, we aimed to apply three time-series deep learning models for predicting three common ICU outcomes, namely, mortality, prolonged LOS, and readmission, of patients in ICU from the Medical Information Mart for Intensive Care-IV (MIMIC-IV) database and identified predictors of high importance based on attention mechanisms to facilitate model interpretability.

# Materials and methods

## Data source and study participants

Patient information was extracted from the MIMIC-IV database (25) to conduct a retrospective cohort study. The MIMIC-IV database contains real medical records with comprehensive information for each patient, ranging from demographic information, vital signs, and laboratory tests to medication administration. All patient information was collected from those who were admitted to the emergency departments and ICU of a tertiary academic medical center in Boston, MA, United States, from 2008 to 2019. The database involves a total of 53,150 patients admitted to the ICU, and all patients' information was de-identified.

A total of 40,083 patients were included in our study. Patients were excluded for the following reasons: (1) age $\leq$ 18 years or $\geq$ 90 years and (2) stay in the ICU for

less than 24 h. In addition, we only included the first admission record if a patient was admitted to the ICU more than once, so the admission records and subject IDs corresponded.

## Predictors and outcomes

We extracted the following data from the MIMIC-IV database upon the initial 24 h of ICU admission and the last 24 h before discharge, and all of the variables were selected according to three conventional scoring systems [APACHE II (26), SOFA (27) and SAPS II (28)]: (1) basic information: age, sex, admission type, ethnicity; (2) diagnosis: AIDS, hematologic malignancy, metastatic cancer; (3) laboratory measurements: serum sodium, serum potassium, serum creatinine, hematocrit, white blood cell count, blood urea nitrogen (BUN), serum bicarbonate, bilirubin, platelets; (4) vital signs: temperature, mean arterial pressure, systolic blood pressure, heart rate, respiratory rate, PaO2, Glasgow coma score (GCS); (5) medication administration: dopamine, dobutamine, epinephrine, norepinephrine; (6) output: urinary output; (7) surgical procedures: invasive mechanical ventilation, non-invasive mechanical ventilation.

Three primary outcomes were needed for prediction in our study. One is the occurrence of death in the hospital, which was defined as whether the patient died during hospitalization, and this information can be extracted from *hospital_expire_flag* in the *admissions table* in the MIMIC-IV database. Another is the occurrence of prolonged LOS, a binary variable with a cutoff point of 75th percentile LOS of the study participants, which was 4 days in our study. Thus, patients with LOS for more than 4 days were labeled as 1, and those with LOS for less than 4 days were labeled as 0. Prolonged LOS information was calculated from the *icustays table*. The other outcome is readmission, which was defined as whether the patient was recorded as having full-cause readmission within 30 days after hospital discharge.

Data extracted from the initial 24 h after ICU admission were used to predict mortality and prolonged LOS, while data derived from the last 24 h before discharge were used to predict the risk of 30-day readmission.

## Data preprocessing and statistical analysis

Continuous variables are presented as the means ± SDs or medians and interquartile ranges and are compared using Student's *t*-test or Wilcoxon rank-sum test according to their normality test results. Categorical variables are presented as counts and percentages and compared through the Chi-square test or Fisher's exact test with significant $p$-values $< 0.05$.

According to recording frequencies, predictors can be classified into dynamic predictors and static predictors. Dynamic variables were those recorded more than once during ICU hospitalization, mostly consisting of vital signs and laboratory tests. Static variables, which included demographic information such as age, sex, and admission type, were all constant and did not change over time. The initial 24 h of ICU admission and the last 24 h before discharge were divided into a time-series of 24 steps, and all variables were obtained for each 1 h window to generate a complete dataset. For static variables, the same value of each patient was recorded 24 times. For dynamic variables, if a variable was recorded more than once in an hour, its mean value was used for aggregation, and then the last observation carried forward (LOCF) was conducted to impute missing values of time-series data. After the first missingness imputation, variables with missing rates of more than 30% were excluded. All categorical variables were one-hot encoded, so the final number of predictors was 33.

All participants were randomly split into a training set (80%) and a test set (20%). The mean value of each continuous variable in the training set was used to impute the remaining missing values in both the training set and the test set. Three deep learning models, RNN, GRU, and LSTM, were used for model development in the training set, and model performance was evaluated in terms of AUC, sensitivity, specificity, and Brier score in the test set. Variable importance according to the attention mechanism was also produced from the test set.

All data analysis procedures were conducted with SAS 9.4 and Python 3.7.

## Recurrent neural network

The mechanism of RNN to tackle time-series problems is that it includes a hidden layer, which incorporates information from all former steps, and with the extension of each time step, the hidden layer iteratively updates, and stores new memory. As shown in **Figure 1A**, $X_t$ represents input variables of the present time step, while $H_{t-1}$ is the hidden layer of the previous time step, two of which co-determine the hidden layer $H_t$ of the present time step, so $H_t$ contains all information of both the previous time steps and the present time step.

## Gated recurrent unit

Gated recurrent unit enriches the structure of RNN with gating systems (an update gate and a reset gate) to solve the problem of too much information kept in the hidden layer when time sequences are too long, in which the update gate ($Z_t$) decides how much information to forget and how much information to keep and the reset gate ($R_t$) determines how much information on former steps to forget, as shown in **Figure 1B**.

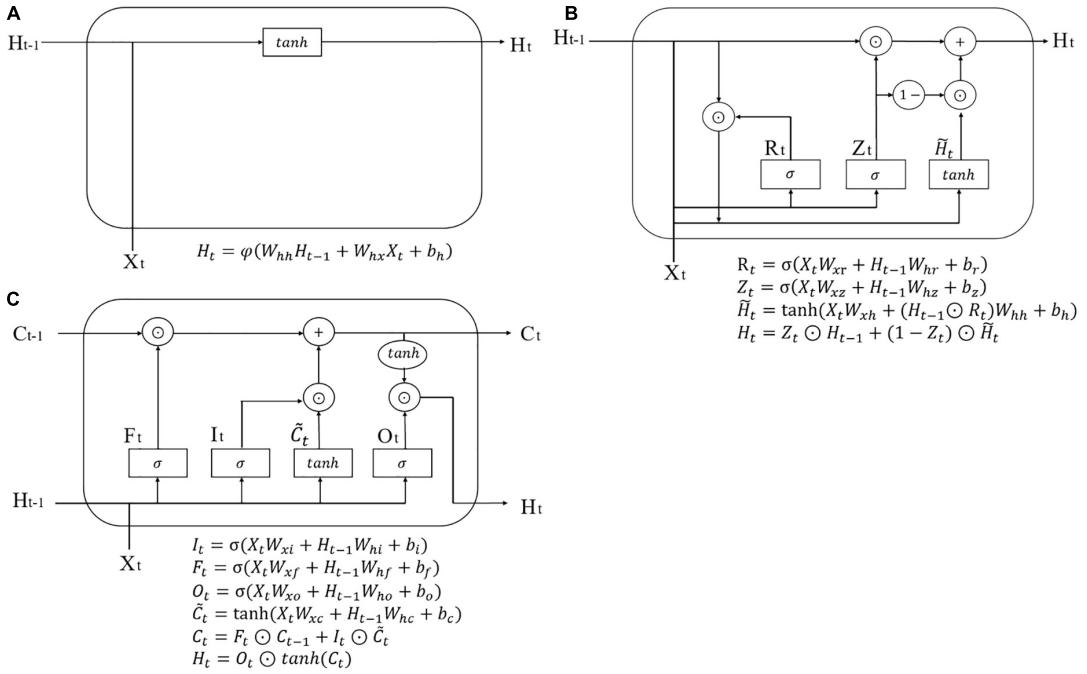



FIGURE 1
Model diagram of a single cell. **(A)** RNN; **(B)** GRU; **(C)** LSTM.

## Long short-term memory

Long short-term memory is more complicated than GRU. It has three gates, an input gate ($I_t$), a forget gate ($F_t$), and an output gate ($O_t$) addition with a memory cell $C_t$. The three gates are all generated by $X_t$ and $X_{t-1}$, and they separately decide how much present input information to keep, how much previous information to forget, and how much total information to output. The schematic diagram of an LSTM cell is shown in Figure 1C.

## Attention mechanism

Considering the complexity of the three deep learning models, especially LSTM, which has relatively more parameters, it would be very difficult to explain the contribution of each variable from these prediction models. Hence, an additional layer was added to each of the three models at the level of input variables; specifically, each variable of each time step ($33 \times 24$ time-specific variables in all) was given an attention weight, which can be represented as $a_t = softmax(x_tW_t)$, and the sum of the weight of each time step was equal to 1 ($|a_t| = 1$), so the new input variable was represented as $X_{new} = A \odot X$. As a result, we ignored the possibly different contributions of each time step but focused on the contribution of each variable. Through the aggregation of all time steps, the global contribution of each variable can be generated.

# Results

## Patient characteristics

A total of 40,083 patients were included in our study for the prediction of mortality and prolonged LOS after excluding those who did not meet the selection criteria, and 36,180 of them were included to predict readmission, as shown in Figure 2. Among these patients, 3,903 (9.74%) deaths occurred during hospitalization, and 11,038 (27.54%) underwent prolonged LOS. After excluding 3,903 patients who died in the hospital, 4,268 (11.79%) were readmitted to the hospital within 30 days after discharge. The comparison of basic information of these patients stratified by outcomes is shown in Table 1. Patients with in-hospital death, compared with those without, were older ($P < 0.001$), comprised more women ($P < 0.001$) and more other or unknown ethnicity ($P < 0.001$), and were more likely to be admitted to the emergency room and transferred from the hospital ($P < 0.001$), had a longer LOS in the ICU ($P < 0.001$), and were more likely to be diagnosed with metastatic cancer ($P < 0.001$) and hematologic malignancy ($P < 0.001$). Patients with prolonged LOS were also comprised of more women ($P < 0.015$) and other or unknown ethnicity ($P < 0.001$),

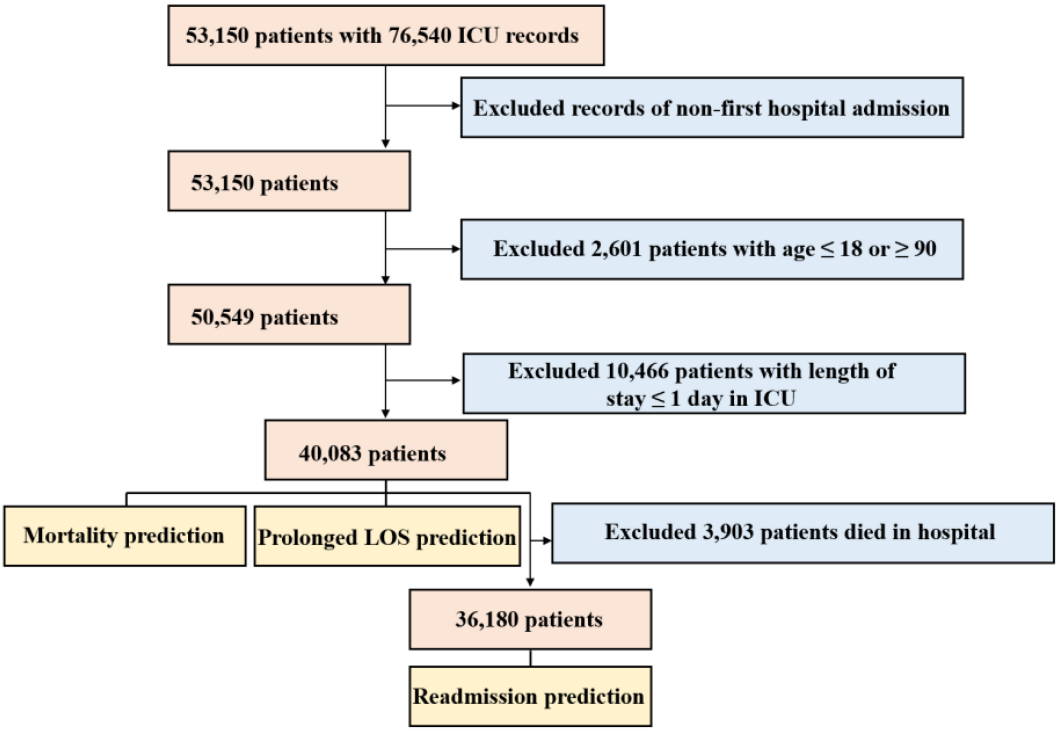



FIGURE 2
Flow chart depicting the inclusion of study participants.

more transferred from the hospital ($P < 0.001$), and more were diagnosed with hematologic malignancy ($P < 0.048$), while fewer were diagnosed with metastatic cancer ($P = 0.025$). Patients who were readmitted were also older ($P < 0.001$), comprised of more white people and fewer other or unknown ethnicity ($P < 0.001$), more were transferred from the hospital ($P < 0.001$) and diagnosed with metastatic cancer ($P < 0.001$) and hematologic malignancy ($P < 0.001$). The diagnosis of AIDS showed similar results between both patients with and without in-hospital death ($P = 0.777$), prolonged LOS ($P = 0.985$), and readmission ($P = 0.146$).

## Model performance

The receiver operating characteristic (ROC) curves of the three prediction models in predicting in-hospital mortality, prolonged LOS, and 30-day readmission are shown in Figures 3A–C. The AUCs of RNN, GRU, and LSTM in predicting mortality were 0.862 ± 0.001, 0.870 ± 0.001, and 0.869 ± 0.002, respectively, and those in prolonged LOS prediction were 0.761 ± 0.002, 0.757 ± 0.011, and 0.765 ± 0.003, respectively. The AUCs of readmission prediction reached only 0.625 ± 0.008, 0.631 ± 0.011, and 0.635 ± 0.018 for the three deep learning models. Other performance metrics, namely, sensitivity, specificity, and Brier score, are shown in Table 2.

## Variable significance

The significance of the variables is shown in Figures 4–6. All three prediction models (RNN, GRU, and LSTM) indicated the important roles of GCS, age, blood urea nitrogen, and administration of norepinephrine in predicting mortality. GCS, invasive ventilation, and blood urea nitrogen were all among the top five significant predictors for prolonged LOS prediction. Blood urea nitrogen, GCS score, and ethnicity were strong predictors for 30-day readmission prediction.

# Discussion

In this study, three time-series deep learning models were applied to predict in-hospital mortality, prolonged LOS, and 30-day readmission with conventional and easily available variables in ICU settings, and influential factors associated with the three outcomes were identified through attention mechanisms to enhance model interpretability.

Our study focused on the outcome prediction of general patients without distinguishing their diseases, and the results showed in-hospital mortality of 9.74%, a prolonged LOS of 27.54%, and 30-day readmission of 11.79%, which were roughly consistent with previous studies (29, 30). For better practical use in clinical settings, we only included variables

TABLE 1 Characteristics of study participants grouped by outcomes.

| Characteristic | Total ($N$ = 40,083) | Outcome 1 | | | Outcome 2 | | | Total ($N$ = 36,180) | Outcome 3 | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | | Death ($N$ = 3,903) | Survival ($N$ = 36,180) | $P$-value | PLOS ($N$ = 11,038) | Non-PLOS ($N$ = 29,045) | $P$-value | | Readmission ($N$ = 4,268) | Non-readmission ($N$ = 31,912) | $P$-value |
| **Age/year, Mean ± SD** | 63.6 ± 16.1 | 68.5 ± 14.7 | 63.1 ± 16.2 | < **0.001** | 63.7 ± 16.0 | 63.6 ± 16.2 | 0.444 | 63.1 ± 16.2 | 64.7 ± 15.4 | 62.9 ± 16.3 | < **0.001** |
| **Sex, n (%)** | | | | < **0.001** | | | **0.015** | | | | 0.412 |
| Male | 23,096 (57.6) | 2,131 (54.6) | 20,965 (57.9) | . | 6,253 (56.6) | 16,843 (58.0) | | 20,965 (57.9) | 2,498 (58.5) | 18,467 (57.9) | |
| Female | 16,987 (42.4) | 1,772 (45.4) | 15,215 (42.1) | . | 4,785 (43.4) | 12,202 (42.0) | | 15,215 (42.1) | 1,770 (41.5) | 13,445 (42.1) | |
| **Ethnicity, n (%)** | | | | < **0.001** | | | < **0.001** | | | | < **0.001** |
| White | 26,768 (66.8) | 2,307 (59.1) | 24,461 (67.6) | . | 7,044 (63.8) | 19,724 (67.9) | | 24,461 (67.6) | 2,998 (70.2) | 21,463 (67.3) | |
| Black American | 3,540 (8.8) | 289 (7.4) | 3,251 (9.0) | . | 934 (8.5) | 2,606 (9.0) | | 3,251 (9.0) | 394 (9.2) | 2,857 (9.0) | |
| Asian | 1,178 (2.9) | 116 (3.0) | 1,062 (2.9) | | 291 (2.6) | 887 (3.1) | | 1,062 (2.9) | 125 (2.9) | 937 (2.9) | |
| Hispanic | 1,423 (3.6) | 103 (2.6) | 1,320 (3.6) | | 373 (3.4) | 1,050 (3.6) | | 1,320 (3.6) | 138 (3.2) | 1,182 (3.7) | |
| Others/Unknown | 7,174 (17.9) | 1,088 (27.9) | 6,086 (16.8) | | 2,396 (21.7) | 4,778 (16.5) | | 6,086 (16.8) | 613 (14.4) | 5,473 (17.2) | |
| **Admission location, n (%)** | | | | < **0.001** | | | < **0.001** | | | | < **0.001** |
| Emergency room | 17,587 (43.9) | 2,024 (51.9) | 15,563 (43.0) | . | 4,862 (44.0) | 12,725 (43.8) | | 15,563 (43.0) | 1,915 (44.9) | 13,648 (42.8) | |
| Physician referral | 10,154 (25.3) | 412 (10.6) | 9,742 (26.9) | | 2,073 (18.8) | 8,081 (27.8) | | 9,742 (26.9) | 870 (20.4) | 8,872 (27.8) | |
| Transfer from hospital | 9,946 (24.8) | 1,236 (31.7) | 8,710 (24.1) | . | 3,511 (31.8) | 6,435 (22.2) | | 8,710 (24.1) | 1,213 (28.4) | 7,497 (23.5) | |
| Others | 2,396 (6.0) | 231 (5.9) | 2,165 (6.0) | | 592 (5.4) | 1,804 (6.2) | | 2,165 (6.0) | 270 (6.3) | 1,895 (5.9) | |
| **LOS/day, Mean ± SD** | 4.1 ± 5.3 | 6.2 ± 6.8 | 3.9 ± 5.0 | < **0.001** | 9.6 ± 7.6 | 2.0 ± 0.8 | < **0.001** | 3.9 ± 5.0 | 5.3 ± 6.9 | 3.7 ± 4.7 | < **0.001** |
| **Metastatic cancer, n (%)** | | | | < **0.001** | | | **0.025** | | | | < **0.001** |
| Yes | 4,715 (11.8) | 776 (19.9) | 3,939 (10.9) | | 1,234 (11.2) | 3,481 (12.0) | | 3,939 (10.9) | 552 (12.9) | 3,387 (10.6) | |
| No | 35,368 (88.2) | 3,127 (80.1) | 32,241 (89.1) | . | 9,804 (88.8) | 25,564 (88.0) | | 32,241 (89.1) | 3,716 (87.1) | 28,525 (89.4) | |
| **Hematologic malignancy, n (%)** | | | | < **0.001** | | | **0.048** | | | | < **0.001** |
| Yes | 1,278 (3.2) | 257 (6.6) | 1,021 (2.8) | | 383 (3.5) | 895 (3.1) | | 1,021 (2.8) | 165 (3.9) | 856 (2.7) | |
| No | 38,805 (96.8) | 3,646 (93.4) | 35,159 (97.2) | . | 10,655 (96.5) | 28,150 (96.9) | | 35,159 (97.2) | 4,103 (96.1) | 31,056 (97.3) | |
| **AIDS, n (%)** | | | | 0.777 | | | 0.985 | | | | 0.146 |
| Yes | 47 (0.1) | 4 (0.1) | 43 (0.1) | | 13 (0.1) | 34 (0.1) | | 43 (0.1) | 2 (0.0) | 41 (0.1) | |
| No | 40,036 (99.9) | 3,899 (99.9) | 36,137 (99.9) | . | 11,025 (99.9) | 29,011 (99.9) | | 36,137 (99.9) | 4,266 (100) | 31,871 (99.9) | |

PLOS, prolonged length of stay; non-PLOS, non-prolonged length of stay; AIDS, acquired immune deficiency syndrome. The bold font designates the statistically significant variables with $p$ value less than 0.05.





10.3389/fmed.2022.933037

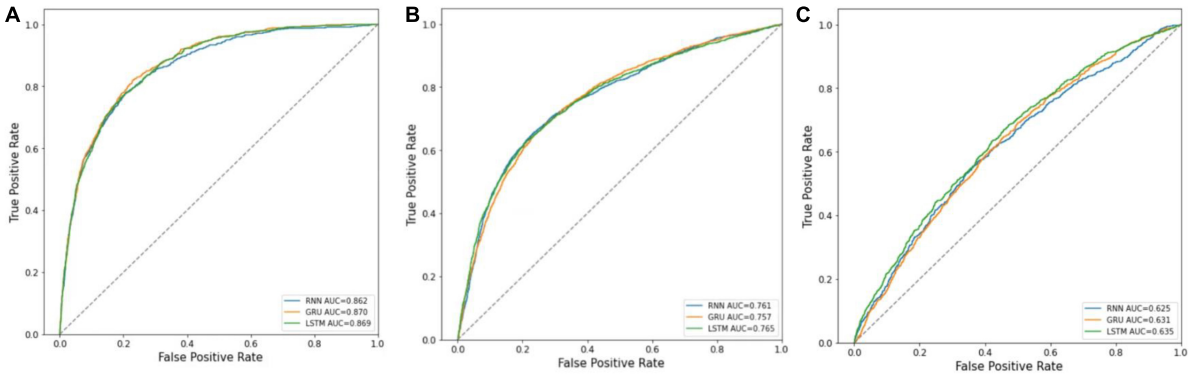



FIGURE 3
ROC curves of RNN, GRU, and LSTM. **(A)** Mortality prediction; **(B)** prolonged LOS prediction; **(C)** 30-day readmission prediction.

TABLE 2 Model performance in predicting hospital mortality, PLOS, and 30-day readmission of patients in ICU.

| Performance | Mortality prediction | | | PLOS prediction | | | 30-day readmission prediction | | |
|---|---|---|---|---|---|---|---|---|---|
| | RNN | GRU | LSTM | RNN | GRU | LSTM | RNN | GRU | LSTM |
| AUC | 0.862 ± 0.001 | **0.870 ± 0.001** | 0.869 ± 0.002 | 0.761 ± 0.002 | 0.757 ± 0.011 | **0.765 ± 0.003** | 0.625 ± 0.008 | 0.631 ± 0.011 | **0.635 ± 0.018** |
| Sensitivity | 0.787 ± 0.012 | **0.796 ± 0.015** | 0.790 ± 0.020 | 0.651 ± 0.009 | **0.666 ± 0.018** | 0.655 ± 0.027 | 0.658 ± 0.036 | 0.652 ± 0.083 | **0.691 ± 0.064** |
| Specificity | **0.786 ± 0.011** | 0.782 ± 0.012 | 0.783 ± 0.017 | **0.771 ± 0.009** | 0.741 ± 0.012 | 0.760 ± 0.024 | **0.567 ± 0.039** | 0.541 ± 0.072 | 0.524 ± 0.061 |
| Brier Score | **0.073 ± 0.003** | 0.087 ± 0.006 | 0.082 ± 0.010 | **0.169 ± 0.006** | 0.204 ± 0.019 | 0.185 ± 0.014 | 0.105 ± 0.001 | 0.105 ± 0.002 | **0.104 ± 0.009** |

AUC, area under the curve; PLOS, prolonged length of stay; RNN, recurrent neural network; GRU, gated recurrent unit; LSTM, long short-term memory. The bold font represents the best score of the three models.

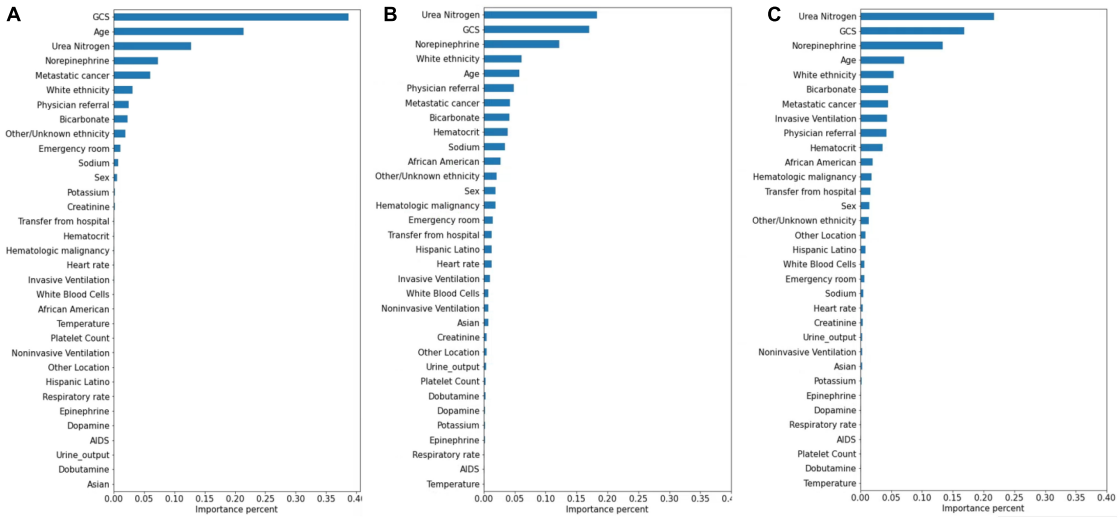



FIGURE 4
Variable importance generated by mortality prediction models. **(A)** RNN; **(B)** GRU; **(C)** LSTM.

that are commonly used and easily available according to three traditional scoring systems [APACHE II (26), SOFA (27), and SAPS II (28)] and collected within 24 h, so compared with other similar studies, the number of variables in this study was relatively small, which partly explained the not very outstanding performance of our prediction models. For example, in Golas's study, 3,512 variables were included (31) and in Sherman's study, 165 variables were included (32), while in our study, only 33 variables were included, which were all among the common clinical measurement indicators.

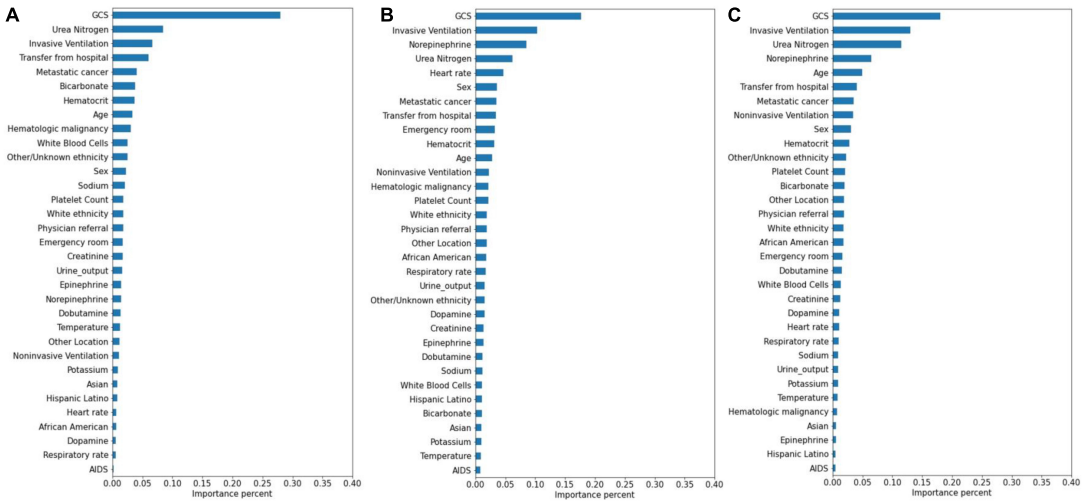



FIGURE 5
Variable importance generated by prolonged LOS prediction models. **(A)** RNN; **(B)** GRU; **(C)** LSTM.

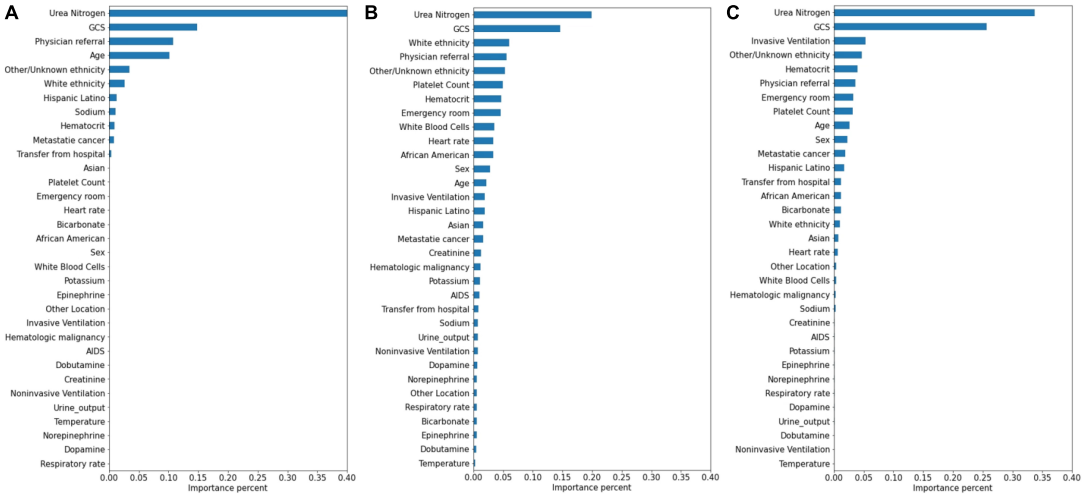



FIGURE 6
Variable importance generated by 30-day readmission prediction models. **(A)** RNN; **(B)** GRU; **(C)** LSTM.

Specifically, the values of AUC indicated good discrimination capability in mortality prediction, moderate in prolonged LOS, and not as good in readmission. All prediction models were trained with a 24-h time window, which was a comprehensive consideration of various conditions, such as the significance of each period, the complexity of variable availability, and the missingness rate. Nevertheless, the length of the time window can also have a certain impact on model performance. In Na's study, the best-performing model (GRU) was trained with 8/16/24/48-h time windows, and the overall tendency indicated that the extended time window corresponded to better predictive performance (33). In addition, the performance of the readmission prediction model may be strongly affected by the period of readmission, ranging from 24 h to 30 days in existing studies (34–36); usually, the shorter the time interval is the better the prediction capability. Thus, using a relatively narrow time window, which is 24 h, to predict long-term outcomes theoretically resulted in a weak predictive capability. However, the result is still competitive in all three outcome predictions because of the application of deep learning models with a small quantity of time-series variables (8, 9, 31, 37).

The results of the performances of the three deep learning modes (RNN, GRU, and LSTM) did not differ greatly in predicting outcomes, and this was inconsistent with what was obtained by Na's study (33). For a similar task (mortality

prediction using RNN, GRU, and LSTM with variables collected within a 48-h observation window), GRU and LSTM performed better than RNN. In their study, the observation window was double-length, which may be related to the difference in the results. The superiority of LSTM and GRU is that their additional gate systems can better select important information stored in hidden layers on each time step, so when the time window is too short, the information contained is more likely to be undiscardable so that the advantages of LSTM and GRU cannot be reflected (23).

Attention mechanisms allowed us to identify important features used by three different models in prediction, and the influential variables of each outcome selected by different deep learning models also did not differ greatly. The GCS was identified as the top important factor for mortality, prolonged LOS, and readmission prediction, and the same results can also be extracted from other similar studies. For example, some studies have concluded that GCS is an independent mortality-related factor and has the most significant feature importance in some specific diseases (38, 39). This variable was also demonstrated to be one of the most important determinants of prolonged LOS in patients with traumatic brain injury (40). Moreover, in Oh's study, 2.28-fold higher unplanned 2-day readmissions were associated with GCS scores less than 13 (41). A lower GCS score indicates more severely impaired consciousness, which may lead to a poor outcome if timely medical intervention is not conducted (42). Age was also demonstrated to have a strong relationship with in-hospital mortality in the ICU by previous studies (43, 44), with a higher mortality rate occurring among elderly patients. These patients generally have reduced immunity, underlying chronic diseases, and worse recovery ability, which may complicate their health status and result in adverse outcomes (45, 46). In Martin's study, BUN was discovered to have a significant association with 28-day mortality (47), and Jamshid's study identified BUN as one of the factors with the highest predictive values to predict the risk of mortality from patients with severe COVID-19 (48), which also provides support for our results. BUN was also identified as a significant variable for prolonged LOS and readmission prediction, and the same results can also be found in homogeneous studies (49, 50). The increased level of BUN is associated with kidney damage, which is supported by multiple mechanisms (51). We also included some medication administration information following SOFA scoring systems (27), and the results showed that norepinephrine, which was recommended as first-line therapy for cardiogenic shock (52), had decisive implications on mortality prediction. This result was also generated by Lu's study, which concluded that patients in cardiogenic shock treated with norepinephrine had significantly increased short-term mortality rates (53). These patients, especially those in refractory shock, usually had an extremely poor prognosis, which lead to higher mortality (54). We also found that invasive ventilation was a decisive predictor for prolonged LOS, a risk factor also suggested by a meta-analysis containing 28 articles (3). In the prediction of readmission, the results showed that ethnicity was a decisive predictor, with the white people owning an increased probability for readmission and other/unknown ethnicity decreasing. In Mukhopadhyay's study, the results also showed that ethnicity was independently associated with hospital readmissions (55).

There are several limitations to our study. First, we excluded some variables that may have predictive values because of high missingness rates, such as the mean arterial blood pressure and bilirubin, and the insurance variable, which may influence LOS, was also not included considering that more than half of the insurance type was labeled "Others." Second, as a single-center study, the generalizability and representation of our conclusion still need to be demonstrated by other data sources. Third, the alternative variables may still be not comprehensive. For example, the diagnosis at ICU admission was not considered a predictor in our study, which may affect the application and generalization of this model in different patient groups. More variables that are easily available need to be explored to further improve model performance.

## Conclusion

Three time-series deep learning models were applied for the prediction of three common ICU outcomes, namely, mortality, prolonged LOS, and readmission. The prediction models reached good performance, especially in mortality prediction, which is of great value in clinical settings considering the conventional and easily available variables incorporated. Our results also indicate that GCS and blood urea nitrogen were highly associated with adverse outcomes of patients in ICU, and focusing on these variables can better assist clinical decisions.

## Data availability statement

All data analyzed in this study were obtained from the MIMIC-IV database, which can be found at https://physionet.org/about/database/.

## Ethics statement

Ethical review and informed consent were not required for this study as the study database, the MIMIC-IV database is publicly available, and all patient data are de-identified.

## Author contributions

YD designed the study and wrote the manuscript draft. BL and YJ critically revised the manuscript. SL assisted with the study protocol and data analysis. ZW contributed to manuscript editing and model explanation. YW helped with manuscript revision. All authors read and approved the final manuscript.

## Funding

This study received financial support from the National Key Research and Development Program of China (No. 2018YFC1311700).

## Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

## Publisher's note



## References

1. Awad A, Bader-El-Den M, McNicholas J, Briggs J. Early hospital mortality prediction of intensive care unit patients using an ensemble learning approach. *Int J Med Inform.* (2017) 108:185–95. doi: 10.1016/j.ijmedinf.2017.10.002

2. Rosenberg AL, Watts C. Patients readmitted to ICUs: a systematic review of risk factors and outcomes. *Chest.* (2000) 118:492–502. doi: 10.1378/chest.118.2.492

3. Peres IT, Hamacher S, Oliveira FLC, Thomé AMT, Bozza FA. What factors predict length of stay in the intensive care unit? Systematic review and meta-analysis. *J Crit Care.* (2020) 60:183–94. doi: 10.1016/j.jcrc.2020.08.003

4. Ramos JGR, Forte DN. Accountability for reasonableness and criteria for admission, triage and discharge in intensive care units: an analysis of current ethical recommendations. *Rev Bras Ter Intensiva.* (2021) 33:38–47. doi: 10.5935/0103-507X.20210004

5. Kılıç M, Yüzkat N, Soyalp C, Gülhaş N. Cost analysis on intensive care unit costs based on the length of stay. *Turk J Anaesthesiol Reanim.* (2019) 47:142. doi: 10.5152/TJAR.2019.80445

6. Verburg IWM, Atashi A, Eslami S, Holman R, Abu-Hanna A, de Jonge E, et al. Which models can i use to predict adult icu length of stay? A systematic review. *Crit Care Med.* (2017) 45:e222–31. doi: 10.1097/CCM.0000000000002054

7. Lee H, Lim CW, Hong HP, Ju JW, Jeon YT, Hwang JW, et al. Efficacy of the APACHE II score at ICU discharge in predicting post-ICU mortality and ICU readmission in critically ill surgical patients. *Anaesth Intensive Care.* (2015) 43:175–86. doi: 10.1177/0310057X1504300206

8. Wu J, Lin Y, Li P, Hu Y, Zhang L, Kong G. Predicting prolonged length of icu stay through machine learning. *Diagnostics.* (2021) 11:2242. doi: 10.3390/diagnostics11122242

9. Lineback CM, Garg R, Oh E, Naidech AM, Holl JL, Prabhakaran S. Prediction of 30-day readmission after stroke using machine learning and natural language processing. *Front Neurol.* (2021) 12:649521. doi: 10.3389/fneur.2021.649521

10. Su L, Xu Z, Chang F, Ma Y, Liu S, Jiang H, et al. Early prediction of mortality, severity, and length of stay in the intensive care unit of sepsis patients based on sepsis 3.0 by machine learning models. *Front Med.* (2021) 28:664966. doi: 10.3389/fmed.2021.664966

11. Loreto M, Lisboa T, Moreira VP. Early prediction of ICU readmissions using classification algorithms. *Comput Biol Med.* (2020) 118:103636. doi: 10.1016/j.compbiomed.2020.103636

12. Hou N, Li M, He L, Xie B, Wang L, Zhang R, et al. Predicting 30-days mortality for MIMIC-III patients with sepsis-3: a machine learning approach using XGboost. *J Transl Med.* (2020) 18:462. doi: 10.1186/s12967-020-02620-5

13. Alsinglawi B, Alshari O, Alorjani M, Mubin O, Alnajjar F, Novoa M, et al. An explainable machine learning framework for lung cancer hospital length of stay prediction. *Sci Rep.* (2022) 12:607. doi: 10.1038/s41598-021-04608-7

14. Chung J, Gulcehre C, Cho K, Bengio Y. Empirical evaluation of gated recurrent neural networks on sequence modeling. *Neural Evol Comput.*. (2014):doi: 10.48550/arXiv.1412.3555

15. Hochreiter S, Schmidhuber J. Long short-term memory. *Neural Comput.* (1997) 9:1735–80. doi: 10.1162/neco.1997.9.8.1735

16. Nemati S, Holder A, Razmi F, Stanley MD, Clifford GD, Buchman TG. An interpretable machine learning model for accurate prediction of sepsis in the ICU. *Crit Care Med.* (2018) 46:547–53. doi: 10.1097/CCM.0000000000002936

17. Zhao Y, Zhang R, Zhong Y, Wang J, Weng Z, Luo H, et al. Statistical analysis and machine learning prediction of disease outcomes for COVID-19 and pneumonia patients. *Front Cell Infect Microbiol.* (2022) 12:838749. doi: 10.3389/fcimb.2022.838749

18. Sun Y, Kaur R, Gupta S, Paul R, Das R, Cho SJ, et al. Development and validation of high definition phenotype-based mortality prediction in critical care units. *JAMIA Open.* (2021) 4:ooab004. doi: 10.1093/jamiaopen/ooab004

19. Wernly B, Mamandipoor B, Baldia P, Jung C, Osmani V. Machine learning predicts mortality in septic patients using only routinely available ABG variables: a multi-centre evaluation. *Int J Med Inform.* (2021) 145:104312. doi: 10.1016/j.ijmedinf.2020.104312

20. Maheshwari S, Agarwal A, Shukla A, Tiwari R. A comprehensive evaluation for the prediction of mortality in intensive care units with LSTM networks: patients with cardiovascular disease. *Biomed Tech.* (2020) 65:435–46. doi: 10.1515/bmt-2018-0206

21. London AJ. Artificial intelligence and black-box medical decisions: accuracy versus explainability. *Hastings Cent Rep.* (2019) 49:15–21. doi: 10.1002/hast.973

22. Lim WX, Chen Z, Ahmed A. The adoption of deep learning interpretability techniques on diabetic retinopathy analysis: a review. *Med Biol Eng Comput.* (2022) 60:633–42. doi: 10.1007/s11517-021-02487-8

23. Gandin I, Scagnetto A, Romani S, Barbati G. Interpretability of time-series deep learning models: a study in cardiovascular patients admitted to Intensive care unit. *J Biomed Inform.* (2021) 121:103876. doi: 10.1016/j.jbi.2021.103876

24. Song H, Rajan D, Thiagarajan JJ, Spanias A. Attend and diagnose: clinical time series analysis using attention models. *arXiv* [Preprint]. (2017):doi: 10.48550/arXiv.1711.03905

25. Johnson, A, Bulgarelli L, Pollard T, Horng S, Celi LA, Mark R. MIMIC-IV (version 0.4). *PhysioNet.* (2020). doi: 10.13026/a3wn-hq05

26. Knaus WA, Draper EA, Wagner DP, Zimmerman JE. APACHE II: a severity of disease classification system. *Crit Care Med.* (1985) 13:818–29.

27. Vincent JL, Moreno R, Takala J, Willatts S, De Mendonça A, Bruining H, et al. The SOFA (Sepsis-related Organ Failure Assessment) score to describe organ dysfunction/failure. on behalf of the working group on sepsis-related problems of the european society of intensive care medicine. *Intensive Care Med.* (1996) 22:707–10. doi: 10.1007/BF01709751

28. Le Gall JR, Lemeshow S, Saulnier F. A new simplified acute physiology score (SAPS II) based on a European/North American multicenter study. *JAMA.* (1993) 270:2957–63. doi: 10.1001/jama.270.24.2957

29. Grieshop S. Post-intensive care syndrome. *Am J Crit Care.* (2022) 31:145. doi: 10.4037/ajcc2022899

30. Hunter A, Johnson L, Coustasse A. Reduction of intensive care unit length of stay: the case of early mobilization. *Health Care Manag.* (2014) 33:128–35. doi: 10.1097/HCM.0000000000000006

31. Golas SB, Shibahara T, Agboola S, Otaki H, Sato J, Nakae T, et al. A machine learning model to predict the risk of 30-day readmissions in patients with heart failure: a retrospective analysis of electronic medical records data. *BMC Med Inform Decis Mak.* (2018) 18:44. doi: 10.1186/s12911-018-0620-z

32. Sherman E, Alejo D, Wood-Doughty Z, Sussman M, Schena S, Ong CS, et al. Leveraging machine learning to predict 30-day hospital readmission after cardiac surgery. *Ann Thorac Surg.* (2021). [Epub ahead of print]. doi: 10.1016/j.athoracsur.2021.11.011

33. Na Pattalung T, Ingviya T, Chaichulee S. Feature explanations in recurrent neural networks for predicting risk of mortality in intensive care patients. *J Pers Med.* (2021) 11:934. doi: 10.3390/jpm11090934

34. Ofoma UR, Chandra S, Kashyap R, Herasevich V, Ahmed A, Gajic O, et al. . Findings from the implementation of a validated readmission predictive tool in the discharge workflow of a medical intensive care unit. *Ann Am Thorac Soc.* (2014) 11:737–43. doi: 1513/AnnalsATS.201312-436OC

35. Martin LA, Kilpatrick JA, Al-Dulaimi R, Mone MC, Tonna JE, Barton RG, et al. Predicting ICU readmission among surgical ICU patients: development and validation of a clinical nomogram. *Surgery.* (2019) 165:373–80. doi: 10.1016/j.surg.2018.06.053

36. Salet N, Stangenberger VA, Eijkenaar F, Schut FT, Schut MC, Bremmer RH, et al. Identifying prognostic factors for clinical outcomes and costs in four high-volume surgical treatments using routinely collected hospital data. *Sci Rep.* (2022) 12:5902. doi: 10.1038/s41598-022-09972-6

37. Li F, Xin H, Zhang J, Fu M, Zhou J, Lian Z. Prediction model of in-hospital mortality in intensive care unit patients with heart failure: machine learning-based, retrospective analysis of the MIMIC-III database. *BMJ Open.* (2021) 11:e044779. doi: 10.1136/bmjopen-2020-044779

38. Hu H, Lai X, Tan C, Yao N, Yan L. Factors associated with in-patient mortality in the rapid assessment of adult earthquake trauma patients. *Prehosp Disaster Med.* (2022) 37:299–305. doi: 10.1017/S1049023X22000693

39. Li X, Wang L, Zhang J, He M, Xu J. XGBoost machine learning algorism performed better than regression models in predicting mortality of moderate to severe traumatic brain injury. *World Neurosurg.* (2022) 163:e617–22. doi: 10.1016/j.wneu.2022.04.044

40. Tardif PA, Moore L, Boutin A, Dufresne P, Omar M, Bourgeois G, et al. Hospital length of stay following admission for traumatic brain injury in a Canadian integrated trauma system: a retrospective multicenter cohort study. *Injury.* (2017) 48:94–100. doi: 10.1016/j.injury.2016.10.042

41. Oh TK, Song IA, Jeon YT. Impact of Glasgow Coma Scale scores on unplanned intensive care unit readmissions among surgical patients. *Ann Transl Med.* (2019) 7:520. doi: 10.21037/atm.2019.10.06

42. Jones C. Glasgow coma scale. *Am J Nurs.* (1979) 79:1551–3.

43. Kukoè A, Mihelèiæ A, Miko I, Romiæ A, Pražetina M, Tipura D, et al. Clinical and laboratory predictors at ICU admission affecting course of illness and mortality rates in a tertiary COVID-19 center. *Heart Lung.* (2022) 53:1–10. doi: 10.1016/j.hrtlng.2022.01.013

44. Arvaniti K, Dimopoulos G, Antonelli M, Blot K, Creagh-Brown B. Epidemiology and age-related mortality in critically Ill patients with intra-abdominal infection or sepsis: an international cohort study. *Int J Antimicrob Agents.* (2022) 20:106591. doi: 10.1016/j.ijantimicag.2022.106591

45. Sofu H, Üçpunar H, Çamurcu Y, Duman S, Konya MN, Gürsu S, et al. Predictive factors for early hospital readmission and 1-year mortality in elder patients following surgical treatment of a hip fracture. *Ulus Travma Acil Cerrahi Derg.* (2017) 23:245–50. doi: 10.5505/tjtes.2016.84404

46. Nasa P, Juneja D, Singh O. Severe sepsis and septic shock in the elderly: an overview. *World J Crit Care Med.* (2012) 1:23–30. doi: 10.5492/wjccm.v1.i1.23

47. Harazim M, Tan K, Nalos M, Matejovic M. Blood urea nitrogen - independent marker of mortality in sepsis. *Biomed Pap Med Fac Univ Palacky Olomouc Czech Repub.* (2022). [Epub ahead of print]. doi: 10.5507/bp.2022.015

48. Jamshidi E, Asgary A, Tavakoli N, Zali A, Setareh S, Esmaily H, et al. Using Machine Learning to Predict Mortality for COVID-19 Patients on Day 0 in the ICU. *Front Digit Health.* (2022) 3:681608. doi: 10.3389/fdgth.2021.681608

49. Omar HR, Guglin M. Longer-than-average length of stay in acute heart failure : determinants and outcomes. *Herz.* (2018) 43:131–9. doi: 10.1007/s00059-016-4532-3

50. Gao S, Yin G, Xia Q, Wu G, Zhu J, Lu N, et al. Development and validation of a nomogram to predict the 180-day readmission risk for chronic heart failure: a multicenter prospective study. *Front Cardiovasc Med.* (2021) 8:731730. doi: 10.3389/fcvm.2021.731730

51. Li X, Zheng R, Zhang T, Zeng Z, Li H, Liu J. Association between blood urea nitrogen and 30-day mortality in patients with sepsis: a retrospective analysis. *Ann Palliat Med.* (2021) 10:11653–63. doi: 10.21037/apm-21-2937

52. Thiele H, Ohman EM, de Waha-Thiele S, Zeymer U, Desch S. Management of cardiogenic shock complicating myocardial infarction: an update 2019. *Eur Heart J.* (2019) 40:2671–83. doi: 10.1093/eurheartj/ehz363

53. Lu X, Wang X, Gao Y, Walline JH, Yu S, Ge Z, et al. Norepinephrine use in cardiogenic shock patients is associated with increased 30 day mortality. *ESC Heart Fail.* (2022) 9:1875–83. doi: 10.1002/ehf2.13893

54. Singer KE, Sussman JE, Kodali RA, Winer LK, Heh V, Hanseman D, et al. Hitting the vasopressor ceiling: finding norepinephrine associated mortality in the critically Ill. *J Surg Res.* (2021) 265:139–46. doi: 10.1016/j.jss.2021.03.042

55. Mukhopadhyay A, Mohankumar B, Chong LS, Hildon ZJL, Tai BC, Quek SC. Factors and experiences associated with unscheduled 30-day hospital readmission: a mixed method study. *Ann Acad Med Singap.* (2021) 50:751–64. doi: 10.47102/annals-acadmedsg.2020522

frontiers | Frontiers in Medicine

TYPE Original Research
PUBLISHED 18 October 2022
DOI 10.3389/fmed.2022.974025



OPEN ACCESS

EDITED BY
Qinghe Meng,
Upstate Medical University,
United States

REVIEWED BY
Chen Hui,
The First Affiliated Hospital of Soochow University, China
Xiao Long,
Peking Union Medical College Hospital (CAMS), China

*CORRESPONDENCE
Penglin Ma
mapenglin1@163.com

†These authors have contributed equally to this work and share first authorship

SPECIALTY SECTION
This article was submitted to Intensive Care Medicine and Anesthesiology, a section of the journal Frontiers in Medicine

RECEIVED 20 June 2022
ACCEPTED 26 September 2022
PUBLISHED 18 October 2022

CITATION
Qiang W, Xiao C, Li Z, Yang L, Shen F, Zeng L and Ma P (2022) Impactful publications of critical care medicine research in China: A bibliometric analysis. *Front. Med.* 9:974025. doi: 10.3389/fmed.2022.974025



# Impactful publications of critical care medicine research in China: A bibliometric analysis

Wei Qiang[1†], Chuan Xiao[2†], Zhe Li[3†], Li Yang[1], Feng Shen[2], Lin Zeng[4] and Penglin Ma[5*]

[1]Department of Library, Guizhou Medical University, Guiyang, China, [2]Department of Intensive Care Unit, The Affiliated Hospital of Guizhou Medical University, Guiyang, China, [3]Department of Critical Care Medicine, Renji Hospital, School of Medicine, Shanghai Jiao Tong University, Shanghai, China, [4]Research Center of Clinical Epidemiology, Peking University Third Hospital, Beijing, China, [5]Department of Critical Care Medicine, Guiqian International General Hospital, Guiyang, China

**Background:** Although publications have been increasing rapidly, the research quality has yet to improve in the field of critical care medicine (CCM) in China. This study aimed at investigating the current status of and the influential factors for impactful publications in CCM research by Chinese authors.

**Methods:** Publications by authors with the affiliation of critical care medicine department or intensive care unit (CCM/ICU) in Chinese as well as American hospitals from 2001 to 2020 were retrieved from the Web of Science Core Collection (WoSCC) database for this bibliometric analysis. Moreover, statistical analyses to test factors affecting impactful publications by Chinese authors were performed.

**Results:** Of 13,487 articles retrieved by this search strategy, 6,622 were published by Chinese authors as first or corresponding authors. The annual publications by Chinese authors have been rapidly increasing from 2001 to 2020, and so did the citations to these articles. However, the proportion in the world of publications by Chinese authors was much less than that by American authors each year [*M* (IQR): 1.85 (9.592) vs. 27.77 (7.3), $p < 0.001$]. In addition, impactful articles were significantly less published by Chinese than by American authors, including articles either in journals with a high impact factor ($p < 0.001$) or in the top 10 journals in the field of CCM (5.4 vs 13.4%, $p < 0.001$), and articles with high citation frequency as well ($p < 0.001$). Moreover, the percentage of impactful publications by Chinese authors was likely associated with academic background and regions of the author's affiliations, funds support, public health events of COVID-19, and collaboration between authors.

**Conclusion:** Our results demonstrated that CCM research in China grew rapidly in the recent 20 years. However, the impactful publications remained limited, largely owing to the shortage of comprehensive research training, inactive collaboration, and underfunded CCM research.

KEYWORDS

**China, research, impactful publications, critical care medicine, factors, bibliometric analysis**

# Introduction

Critical care medicine (CCM) in China was seen to make great progress in the past two decades (1). It plays an important role not only in the management of critical illness in hospitals but also in increasing actions of public health emergencies and natural disasters. However, the achievement of research was not consistent with clinical practice in the field of CCM in China. Based on bibliometric analysis, Li et al. (2) reported that the number of publications on CCM was much less in China than that in the United States and other developed countries from 2000 to 2010. Moreover, the research with high quality were mostly concentrated in Taiwan and Hong Kong. In fact, the majority of articles in this field from mainland China were less impactful during this period (3).

To promote the scientific research in CCM, experts from the Chinese Society of Critical Care Medicine established the China Critical Care Clinical Trials Group (CCCCTG), comprising intensivists from 24 ICUs from 21 provinces in China, which joined the Global Sepsis Alliance (GSA) in 2010. In addition, there were more and more activities specific to scientific research training, for example, the Conference of Critical Care Research Forum (CCCRF), Salon for Young Critical Care Investigators, and Critical Care Research Campaign, etc (4). Accordingly, the number of publications on CCM from China has been increasing rapidly over the last decade (5–8). Meanwhile, the research quality has yet to improve. An updated bibliometric analysis showed that China contributed only 1% of the top 2,000 highly cited articles on critical care, as of 13 February 2018 (9). In addition, there never was an article on CCM from China with annual citations over 100 before 2018 (10). These data suggest that problems remain in promoting the quality of research on CCM in China. Notably, the barriers were under-investigated. Therefore, this study aimed at investigating the current status of and the influential factors for impactful publications in CCM research from 2001 to 2020 by Chinese authors, who reported the affiliation of Critical Care Medicine department or intensive care unit (CCM/ICU) in Chinese hospitals, through a bibliometric and visualized analysis.

# Methods

## Data sources and search strategies

Web of Science Core Collection (WoSCC) database is one of the most comprehensive, systematic, and authoritative databases, which has been successfully used for bibliometric analysis (11, 12). Publications by authors reporting the affiliation of CCM/ICU in Chinese hospitals from 2001 to 2020 were retrieved for this bibliometric analysis. The search strategy was "Address: (Chinese OR China OR CN) AND (Intense Care Unit OR Crit Care OR ICU OR intensive care OR critical care) NOT Address: (Respiratory OR Pulmonary OR PCCM)." The data set retrieved from the WoSCC database was transformed into an Excel version. The collected articles were further screened by the first or corresponding authors who reported the affiliation of CCM/ICU in Chinese hospitals. Being a comparator, data regarding publications from CCM/ICU in American hospitals were collected by the same search strategy, but "American OR America OR US" replaced "Chinese OR China OR CN." Time windows were unified as "1 January 2001 to 31 December 2020"; and there were no language or article type restrictions. All data were collected online on 1 May 2022 and no ethical proof was required.

## Data collection

Data regarding publications retrieved from WoSCC included title, keywords, authors, affiliations and regions, journal, date of publication, funding, citations, etc. Data were extracted by two authors (QW and ZL) independently and the agreement of the results was 98%, showing significant consistency. All data were saved in a text or excel format for further analysis.

## Bibliometric analysis

All downloaded documents were imported to the Web of Science-Incites Research Performance Analysis Platform (WoS-Incites, https://incites.clarivate.com/), VOSviewer (version 1.6.15), and Microsoft Excel 2019. WoS-Incites were used to analyze the number of publications, impact factors of the journals, citation frequency, characteristics of the authors, and their affiliations. VOS viewer 1.6.15 (Leiden University, Leiden, The Netherlands) was used to analyze and visualize co-authorship of authors, institutes, countries, and co-occurrence analysis of keywords (13). Microsoft Excel 2019 was used to diagrammatize results from WoS-Incites (14).

## Statistical analysis

Continuous variables were expressed as mean ± standard deviation (mean ± SD) or median [interquartile range; *M* (IQR)] depending on whether they followed a normal distribution. Differences between groups were compared by Student's *t*-test or the Wilcoxon rank sum test based on data distribution. Categorical variables were described using cases and percentages or proportions. And differences between groups were compared by the chi-square test or Fisher's exact probability method. Two-sided $p$-values $< 0.05$ were considered statistically significant.

# Results

## Publications and citations

### Publications

There were 13,487 articles published in international peer-reviewed journals listed on the Science Citation Index (SCI) from 2001 to 2020 reporting one author at least with the affiliation of CCM/ICU in Chinese hospitals. The number of annual publications was over 100 in 2008 and has been rapidly increasing since 2008 (Figure 1 inner).

Out of the 13,487 publications, 6,622 were further retrieved by the first or corresponding author who reported the affiliation of CCM /ICU in Chinese hospitals (Figure 1). Stratified with the impact factors (IFs), 531 (8.02%), 4,685 (70.75%), 1,192 (18.00%), and 214 (3.23%) out of 6,622 articles were published in journals without IF, IF $\leq$ 5, IF between 5–10 (5 $<$ IF $<$ 10) and IF $\geq$ 10, respectively (Supplementary Figure S1). Notably, the number of annual publications in journals with IF $>$ 5 would not exceed 100 until 2017, while publications were over 50 in journals with IF $\geq$ 10 till 2020 (Figure 1).

### Publications by Chinese vs. American authors

The proportion of publications in the world on CCM research from 2001 to 2020 by Chinese authors was much less than that by American authors [*M* (IQR): 1.9 (0.4, 10.0) vs 27.8 (25.6, 32.9), $p < 0.001$, Table 1]. However, the proportions contributed by Chinese authors increased yearly, while a decreased trend was found in that by American authors in this study period (Supplementary Table S1). As shown in Supplementary Figure S2, the proportion of publications by American authors always ranked first in the world each year from 2001 to 2020. The ranking of publications by Chinese authors has entered the top 10 since 2012 (rank ninth) and kept the second place since 2014. Significantly, the number and percentage (the number/the total) of publications in the top 10 high impactful journals in the field of CCM (including NEJM; JAMA, BMJ, Am J Resp Crit Care Med, Intensive Care Med, Critical Care Med, Ann Intensive Care, the detailed data are shown in Supplementary Table S2) were also much less by Chinese authors than by American authors in these two decades [358 (5.4%) vs 3,060 (13.4%), $p < 0.001$, Table 1].

## The keywords in publications

A total number of 10,980 and 21,689 keywords were identified from 6,622 and 22,819 publications by first and corresponding authors with affiliations of CCM/ICU in Chinese and American hospitals, respectively. The top 724 keywords with co-occurrence frequency equal to or over five were selected for co-occurrence network and overlay analysis. The occurrence frequency of keyword was displayed in circle size, as shown in Figures 2A,B. It was shown that "Sepsis" was only the same one out of the top five keywords (ranked by the occurrence frequency) in publications by either Chinese authors ["Sepsis" (642), "Acute lung injury" (412), "Inflammation" (289), "Apoptosis" (268), "Mortality" (232)] or America authors ["Pediatric" (1,402), "Trauma" (624), "Sepsis" (605), "Critical care" (584), "Intensive care unit" (546); Figures 2A,B, Supplementary Table S3], respectively. In addition, eight keywords including "Sepsis," "Septic shock," "Acute kidney injury," Acute lung injury," "Mechanical ventilation," "Inflammation," "Mortality," and "Intensive care unit" were shared in the top 20 keywords by both Chinese and American publications (Supplementary Table S3). Moreover, it was demonstrated that seven and 13 keywords of Chinese publications, in comparison with 0 and 20 keywords of American publications, were categorized as basic researches and clinical researches, respectively ($p = 0.008$, Table 1).

In addition, the overlay analysis of the keywords represented the trends of topics in Chinese and American publications between 2001 and 2020. The circles of keywords were marked on colors from blue to yellow to display the overlay visual map of the keywords over time, which was quantitively calculated by the average publication year of the articles in which the keyword appeared (Avg.pub.year) (15). It was demonstrated that the Avg.pub.year of "nuclear factor-kappa B (2014.29, Ranked eighth)" and "ischemia/reperfusion injury (2013.75, Ranked sixth)" in publications by Chinese authors were about 7–9 years delayed from that of the similar keywords "nuclear factor-kappa B (2005.44, Ranked first)" and "reperfusion (2006.10, Ranked fourth)" by authors from America among the top 10 earliest research topics (Supplementary Figures S3A,B, Supplementary Table S4). As shown in Supplementary Table S4, the latest research topics were similar in publications by authors from China and America, that mainly focused on COVID-19, and the Avg.pub.years of these hot topics were from 2019 to 2020.

### Citations

The citations to articles published by the first or corresponding authors with the affiliation of CCM /ICU in Chinese hospitals each year from 2001 to 2020 are also shown in Figure 1. Similar to the trend of publications, the total citations to these articles kept a rapid growth year by year from 2012, despite a rollback in 2018 (Figure 1). Out of the top 10 highly cited articles by Chinese authors, only one was not related to COVID-19 as shown in Supplementary Table S5 (16). The total citations of the top 10 highly cited articles in the world ranged from 3,116 to 10,788 from 2001 to 2020 (Supplementary Table S5), of which there was only one

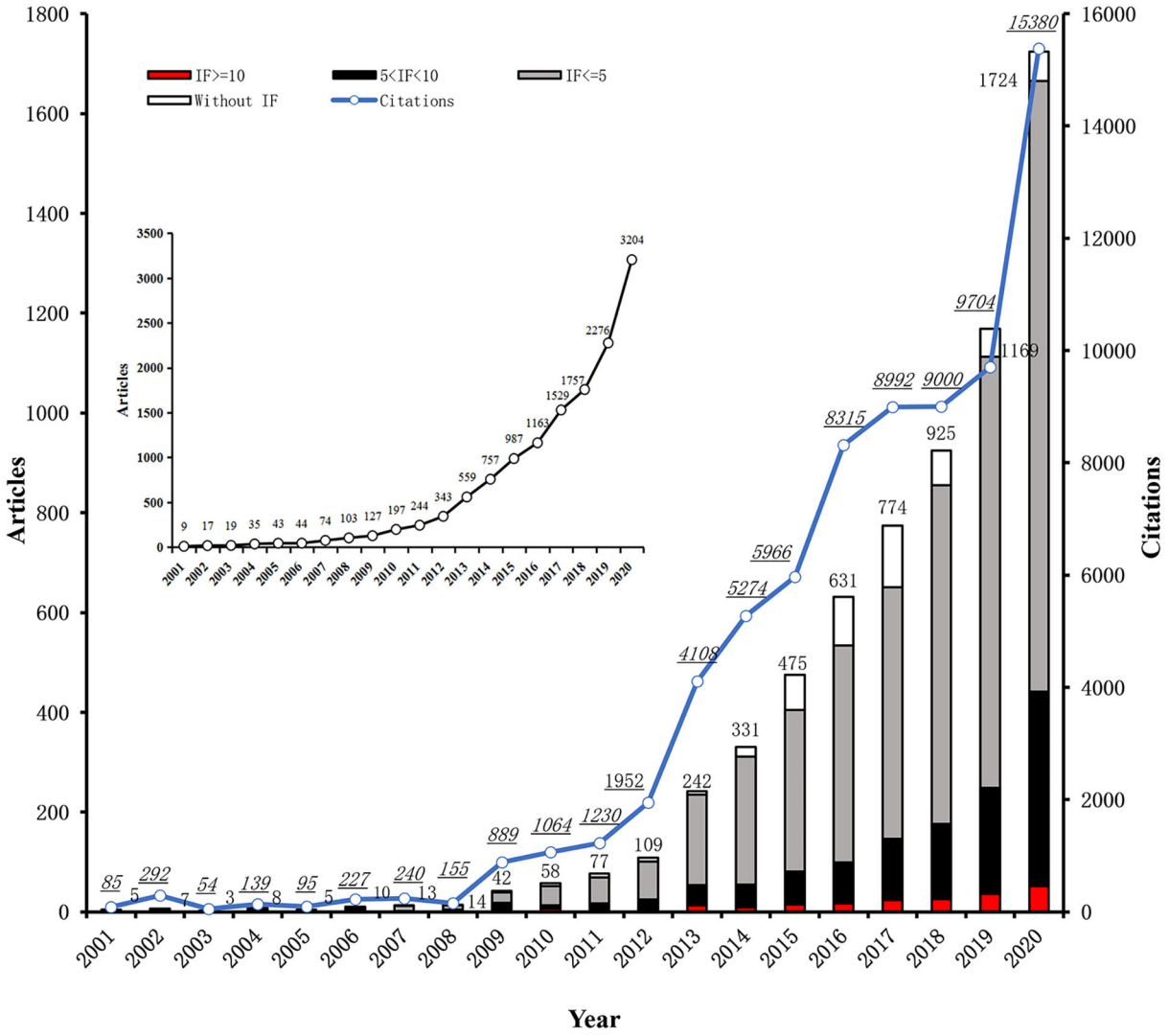



FIGURE 1
Annual publications and the cumulative citations from 2001 to 2020.The X-axis represented each year from 2001 to 2020, the Y-axis represented the number of annual publications (bar) and the cumulative citations (blue line) to all articles till the year published by first and corresponding authors from affiliations of Critical Care Medicine (CCM) department or intensive care unit (ICU) in Chinese hospitals. These annual publications were stratified by the impact factor (IF) of the journals publishing these articles, marked on red (IF $\geq$ 10), black (5 < IF < 10), gray (IF $\leq$ 5) and white (without IF) as well. The inner figure showed the trend of annual publications with one author at least on author list from affiliations of CCM department or ICU in Chinese hospitals.

article published by the Chinese author (17). Significantly, the percentage of articles published by Chinese authors was much lower than that of American authors in the top 10, 100, and 1,000 highly cited articles in the world, as shown in Table 1 ($p < 0.001$). Additionally, either the average citation frequency (citations/articles) in the two decades [*M* (IQR): 17.0 (11.9, 18.4) vs. 27.8 (15.9, 36.9), $p = 0.012$] or the individual citation frequency in WOSCC [citations of individual article; *M* (IQR): 5.0 (2.0, 14.0) vs. 8.0 (1.0, 24.0), $p < 0.001$] was significantly lower in Chinese authors publications (Table 1).

## Factors barred to or facilitated the impactful publications in Chinese CCM research

### Academic background of the authors' affiliations

As shown in Supplementary Figure S4, 69.55% (491/706) of the first and corresponding authors reported affiliations of CCM/ ICU in Chinese hospitals with academic background, including 65.72 and 3.82% of them affiliated with university (or college) and research institutes, respectively. Meanwhile,

TABLE 1 Publications and the citations by Chinese vs. American authors from 2001 to 2020.

| | By Chinese authors | By American authors | *p*-Value |
|---|---|---|---|
| Publications in total | 6,622 | 22,819 | — |
| **Publications in journals with, *n* (%)** | | | |
| IF < 5 | 4,669 (70.5) | 13,116 (57.5) | |
| IF ≥ 5 | 1,953 (29.5) | 9,703 (42.5) | <0.001 |
| IF ≥ 10 | 608 (9.2) | 3,066 (13.4) | |
| Publications in top 10 impactful journals linked to CCM[†], *n* (%) | 358 (5.4) | 3,060 (13.4) | <0.001 |
| Yearly proportions of publication in the world (%)[#], *M* (IQR) | 1.9 (0.4, 10.0) | 27.8 (25.6, 32.9) | <0.001 |
| **The top 20 keywords in publications, *n* (%)** | | | |
| Categorized to basic research | 7 (35.0) | 0 (0.0) | 0.008 |
| Categorized to clinical research | 13 (65.0) | 20 (100.0) | |
| Yearly citation frequency (yearly citations/articles, %), *M* (IQR) | 17.0 (11.9, 18.4) | 27.8 (15.9, 36.9) | 0.012 |
| Individual citation frequency in WOSCC[††], *M* (IQR) | 5.0 (2.0, 14.0) | 8.0 (1.0, 24.0) | <0.001 |
| **Highly citated articles[*], *n* (%)** | | | |
| In top 10 | 1 (10.0) | 3 (30.0) | <0.001 |
| In top 100 | 3 (3.0) | 55 (55.0) | |
| In top 1,000 | 21 (2.1) | 660 (66.0) | |

[†]The top 10 impactful journals linked to CCM include NEJM (NEW ENGLAND JOURNAL OF MEDICINE); JAMA (JOURNAL OF THE AMERICAN MEDICAL ASSOCIATION); BMJ (British Medical Journal); Am J Resp Crit Care Med (AMERICAN JOURNAL OF RESPIRATORY AND CRITICAL CARE MEDICINE); Intensive Care Med(INTENSIVE CARE MEDICINE); Critical Care Med (CRITICAL CARE MEDICINE; Ann Intensive Care (ANNALS OF INTENSIVE CARE), the detailed data regarding publications are shown in Supplementary Table S1.

[#]Yearly proportions of publication: the proportions of publications in the world each year from 2001 to 2020 by Chinese vs. American authors on CCM researches, the detailed data are shown in Supplementary Table S1.

[††]Individual citation frequency in WOSCC: Citations of each publication in the database of WOSCC (Web of Science Core Collection).

[*]The highly cited articles: the top 10, 100 and 1,000 highly cited articles in the world.

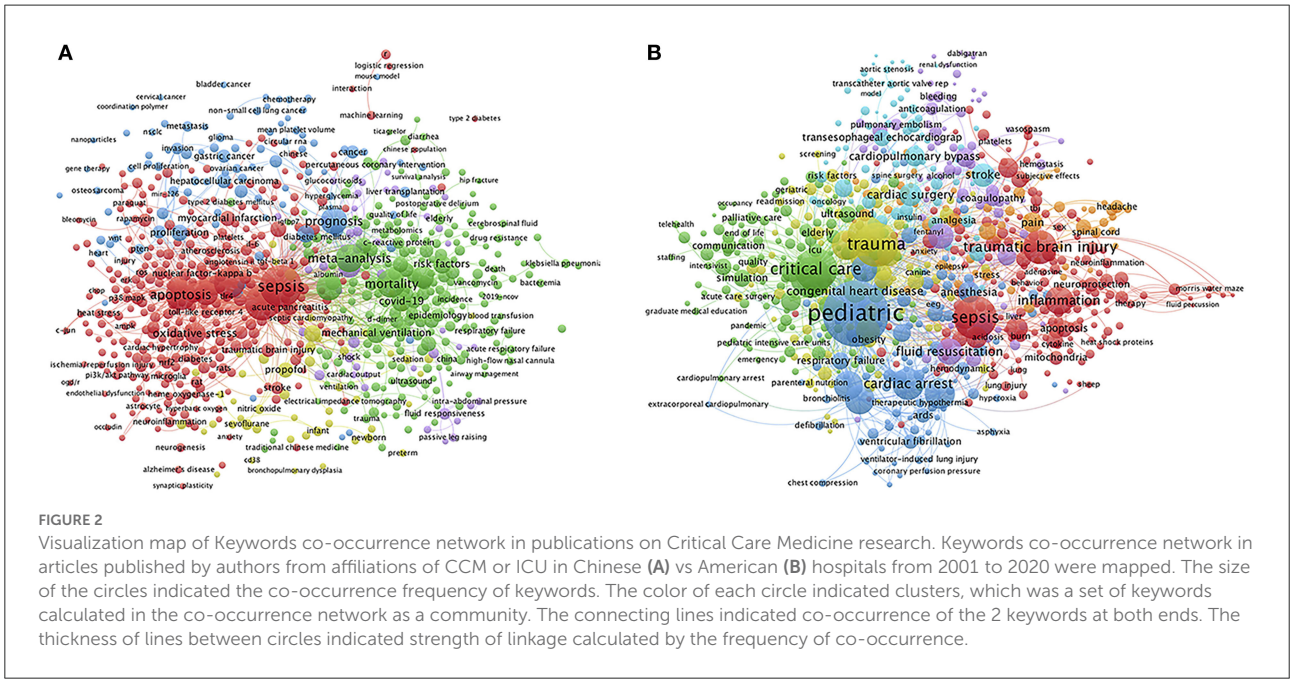



FIGURE 2
Visualization map of Keywords co-occurrence network in publications on Critical Care Medicine research. Keywords co-occurrence network in articles published by authors from affiliations of CCM or ICU in Chinese **(A)** vs American **(B)** hospitals from 2001 to 2020 were mapped. The size of the circles indicated the co-occurrence frequency of keywords. The color of each circle indicated clusters, which was a set of keywords calculated in the co-occurrence network as a community. The connecting lines indicated co-occurrence of the 2 keywords at both ends. The thickness of lines between circles indicated strength of linkage calculated by the frequency of co-occurrence.

only 18.84% of the authors served hospitals without academic background (i.e., hospitals not affiliated with any university, college, or research institutes). Significantly, the percentages of articles published in journals with IF ≥ 5 and IF ≥ 10 by authors from academic hospitals were significantly higher than that from non-academic hospitals ($p < 0.001$, Table 2).

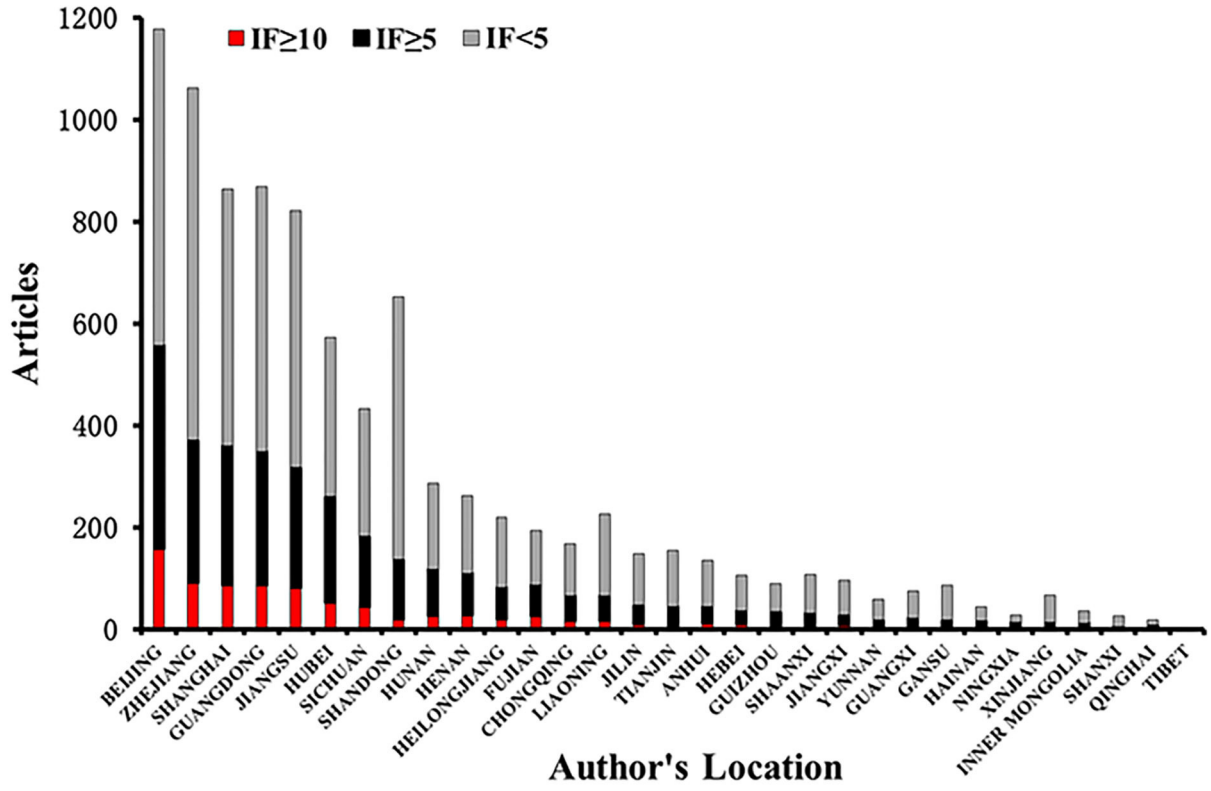



FIGURE 3
Geographical distribution of the publications by Chinese authors. The X-axis were the provinces/municipalities, where the hospitals of the first or corresponding authors of the publications were located. The Y-axis represented the number of their publications from 2001 to 2020, which were stratified by the impact factor (IF) of the journals publishing these articles, marked on red (IF $\geq$ 10), black (IF $<$ 5), and gray (IF $<$ 5) as well. A disequilibrium analysis of the distribution of impactful publications among regions from 2001 to 2020 were performed by using Fisher's exact probability analysis, and the result suggested that P $<$ 0.001, which indicated the number of publications in different impact factor groups varied by regions.

## Geographic distribution of publications by Chinese authors

The affiliations of 6,622 publications were distributed in 31 provinces/municipalities of China (Figure 3). Beijing was the only one out of the 31 provinces/municipalities with publications over 1,000. Meanwhile, there were 13 provinces/municipalities with publications <100 as shown in Figure 3. Significantly, <50 articles were published by authors from Hainan, Tibet, Qinghai, Inner Mongolia, Ningxia, and Shanxi province/municipality in these two decades.

By a Fisher's exact probability analysis, a significant disequilibrium was found in the distribution of the proportions of publications stratified with IF $<$ 5, IF $\geq$ 5, and IF $\geq$ 10 in 31 provinces or municipalities of mainland China (the detailed data are shown in Supplementary Table S6).

## COVID-19-related publications

There were 206 (11.94%) out of 1,724 publications focused on COVID-19 research in 2020. Compared with non–COVID-19 publications in 2020 and in 2019 as well, the percentages of COVID-19-related publications in impactful journals (i.e., IF $\geq$ 5 or IF $\geq$ 10) were significantly increased ($p < 0.001$, Table 2).

## Funds supporting

Out of 6,622 publications by Chinese authors, 2,307 (34.84%) articles were reported with funds support. In comparison with the percentage of publications without fund, the percentage of those with funds supporting was significantly increased in journals with either IF $\geq$ 5 (36.41 vs. 25.77%) or IF $\geq$ 10 (8.58 vs. 9.50%; $p < 0.001$, Table 2).

## Collaboration network analysis of the authors, institutes, and countries

The collaboration network visual map between the authors, the institutes, and the countries in the 13,487 articles was generated by VOS viewer (Figures 4, 5A,B). The total link strength was calculated on the number of publications co-authored by the authors, the institutes, and the countries. Of all 45,266 authors on the author list of the 13,487 articles, 342 who

TABLE 2 Factors affecting the impactful publications in CCM researches by Chinese authors from 2001 to 2020.

| | IF < 5 | IF ≥ 5 | IF ≥ 10 | *p*-Value |
|---|---|---|---|---|
| **Authors from academic hospital*, *n* (%)** | | | | |
| Yes (total = 5,406) | 3,690 (68.3) | 1,716 (31.7) | 518 (9.6) | <0.001 |
| No (total = 1,216) | 980 (80.6) | 236 (19.4) | 90 (7.4) | |
| **Fund for publications#, *n* (%)** | | | | |
| Yes (*n* = 2,307) | 1,467 (63.6) | 840 (36.4) | 198 (8.6) | <0.001 |
| No (*n* = 4,315) | 3,203 (74.2) | 1,112 (25.8) | 410 (9.5) | |
| **Publications related to COVID-19†, *n* (%)** | | | | |
| Yes (in 2020, *n* = 206) | 81 (39.3) | 125 (60.7) | 71 (34.5) | <0.001 |
| No (in 2020, *n* = 1,518) | 1,058 (69.7) | 460 (30.3) | 118 (7.8) | |
| No (in 2019, *n* = 1,169) | 920 (78.7) | 249 (21.3) | 36 (3.1) | |

IF: impact factor of the journals, which published the articles.
*Academic hospital: the word "institute" or "college" or "university" was reported in the affiliation of the first or corresponding author.
#Fund: it was based on the declaration of the article.
†COVID-19: any keywords with regard to COVID-19 (including coronavirus disease, SARS-CoV-2, novel coronavirus pneumonia, etc.) was found in title/abstract.

published 20 or more articles were analyzed. Ranked with the total link strength, the top three authors were "yang, yi" (419), "qiu, haibo" (413), and "liu, ling" (273), who come from the same affiliation, the Department of Critical Care Medicine, Nanjing Zhong da Hospital, School of Medicine, Southeast University, Nanjing. In addition, there were three other authors ("liu, dawei," "long, yun," and "wang, hao") in the top 10 authors with the most collaborations came from the same affiliation too, The Department of Critical Care Medicine, Peking Union Medical College Hospital, as shown in Figure 4, Supplementary Table S7.

Out of the total 6,372 institutes of these authors, 398 published 10 or more articles. The Capital Medical University was the affiliation with the highest collaboration link strength (capital med univ, 1,162), followed by the China Medical University (China med univ, 1,072) and Shanghai Jiao Tong University (shanghai jiao tong univ, 1,037, Figure 5A, Supplementary Table S7). Moreover, authors from 61 countries collaborated with Chinese authors in five or more publications among the 13,487 articles. Authors from America ("usa") collaborated with Chinese authors ("peoples r China") most, followed by authors from "Italy," "England," and "Canada" (Figure 5B, Supplementary Table S7).

## Discussion

It was demonstrated that the publications by the first or corresponding authors with the affiliation of CCM/ICU in Chinese hospitals have been rapidly increasing from 2001 to 2020, and so did the citations to these articles (Figure 1). However, the proportion in the world of publications on CCM research by Chinese authors was much less than that by American authors each year (Table 1). In addition, the number and the percentage of impactful articles were significantly less published by Chinese than by American authors, including articles published in journals with a high impact factor (i.e., IF ≥ 5, IF ≥ 10), articles in the top 10 journals in the field of critical care medicine, and the high frequently cited articles as well (Table 1). Moreover, it was found that several factors likely affected the output of impactful publications in CCM researches by authors with the affiliation of CCM/ICU in Chinese hospitals, such as the academic background of authors affiliations, funds support, public health event of COVID-19, regions of author's affiliation and collaboration between authors (Table 2, Figure 4).

Previous studies suggested that several factors facilitate Chinese CCM research, including rapid economic growth, expansion of ICUs and intensive care practitioners (18), and responses to disasters such as SARS 2003, Wenchuan earthquake in 2008, the outbreak of COVID-19, etc. (1, 19, 20). Furthermore, it was demonstrated that the public health event of COVID-19 was associated with the production of higher impactful publications by Chinese authors in this study (Table 2). Meanwhile, the rapid increase of publication in CCM research in China could be also driven by the academic evaluation system in the past two decades largely. Although there have never been any officially issued rules, in fact, articles, awards, titles, degrees, and honors were highly weighted in the evaluation of an individual or team's competitiveness. Based on the regulations of most medical colleges or institutes, for instance, the candidates were not qualified to apply Doctor of Philosophy (PhD) or Medical Doctor (MD) degree until publishing one article at least in the SCI journal. Notably, reports of this evaluation would be closely tied with professional promotion and appointment. In this study, interestingly, several findings supported this approach, which was not evidenced in the previous studies (9, 10). First, hospitals where the most authors served (69.55%) were affiliated with academic institutes (Supplementary Figure S4). Few of them could be unaffected by this hidden regulation. Moreover, very few authors (18.8%) served nonacademic hospitals. According to the data from Beijing clinical quality control and improvement center for Critical Care Medicine, there are only 25 (29.8%) ICUs (including general, surgical, or medical ICUs) in hospitals with academic backgrounds among a total number of 84 grade II and III hospitals even in Beijing (unreported data). Second, over half of the articles (59.6%, 4,927/8,268) were published by authors from six out of 31 provinces/municipalities including Beijing, Zhejiang, Guangdong, Shanghai, Jiangsu, and Shandong (Figure 3), where medical colleges and research institutes were highly concentrated in China. Finally, the keywords of these articles were linked to lab research rather than clinical topics more frequently (Supplementary Table S3). These findings suggested that the majority of Chinese intensivists working in nonacademic hospitals have not successfully published articles

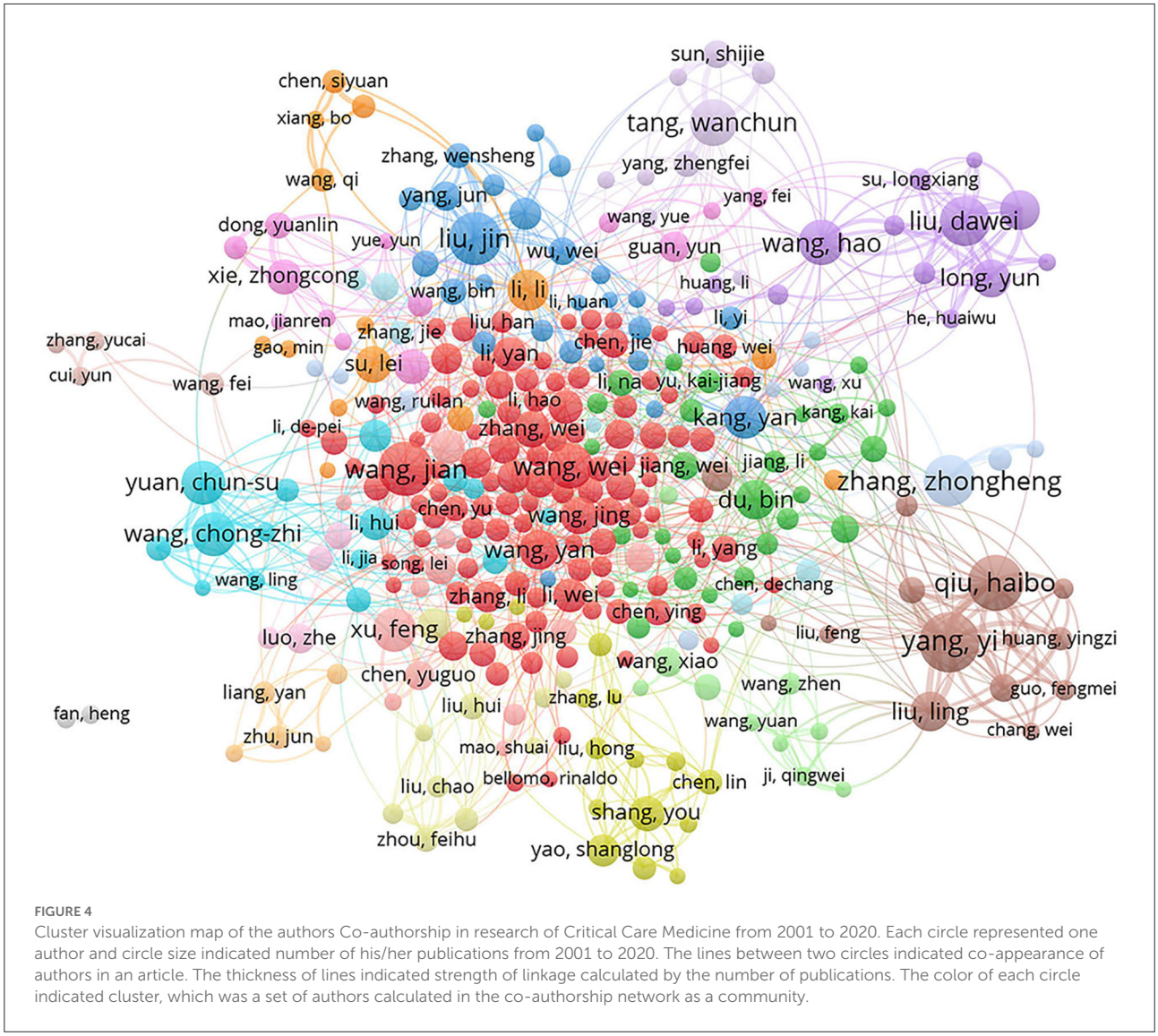



FIGURE 4
Cluster visualization map of the authors Co-authorship in research of Critical Care Medicine from 2001 to 2020. Each circle represented one author and circle size indicated number of his/her publications from 2001 to 2020. The lines between two circles indicated co-appearance of authors in an article. The thickness of lines indicated strength of linkage calculated by the number of publications. The color of each circle indicated cluster, which was a set of authors calculated in the co-authorship network as a community.

in SCI journals during this period. There could be no argument about intensivists in academic hospitals getting better training and having a higher passion for scientific research. But, our results suggested that the research of Chinese intensivists is, partly at least, driven by the academic evaluation system rather than by their interests in questions arising through the day-to-day care of critically ill patients. Fortunately, a special notification was issued by the government for correcting the disadvantage of this evaluation system (https://news.sciencenet.cn/). Hopefully, the researches of Chinese intensivists will be conducted with the impetus to study questions arising through intensive caring. In this way, the production of Chinese intensivists' research will be not only rich but more impactful in future.

A comprehensive training in scientific research is the base for highly impactful publications. Our findings suggested that authors who got better research training probably, for instance, who served in hospitals with academic backgrounds and in cities with more medical colleges/universities/research institutes, be more likely to publish impactful articles (Table 2). To our knowledge, however, there was an acute shortage of training courses specific to critical care research in China. This accounted for the significant difference in publishing impactful articles between Chinese and American authors largely (Table 1).

Collaboration can enhance the power, efficiency, generalizability, and rapid completion of clinical research (21), and hence may improve the research quality large probably. Over the past 5 years, for instance, all 17 randomized controlled trials searched for "sepsis" in the New England Journal of Medicine were interagency collaborations. In addition, only one out of the top 10 most frequently cited articles in the field of

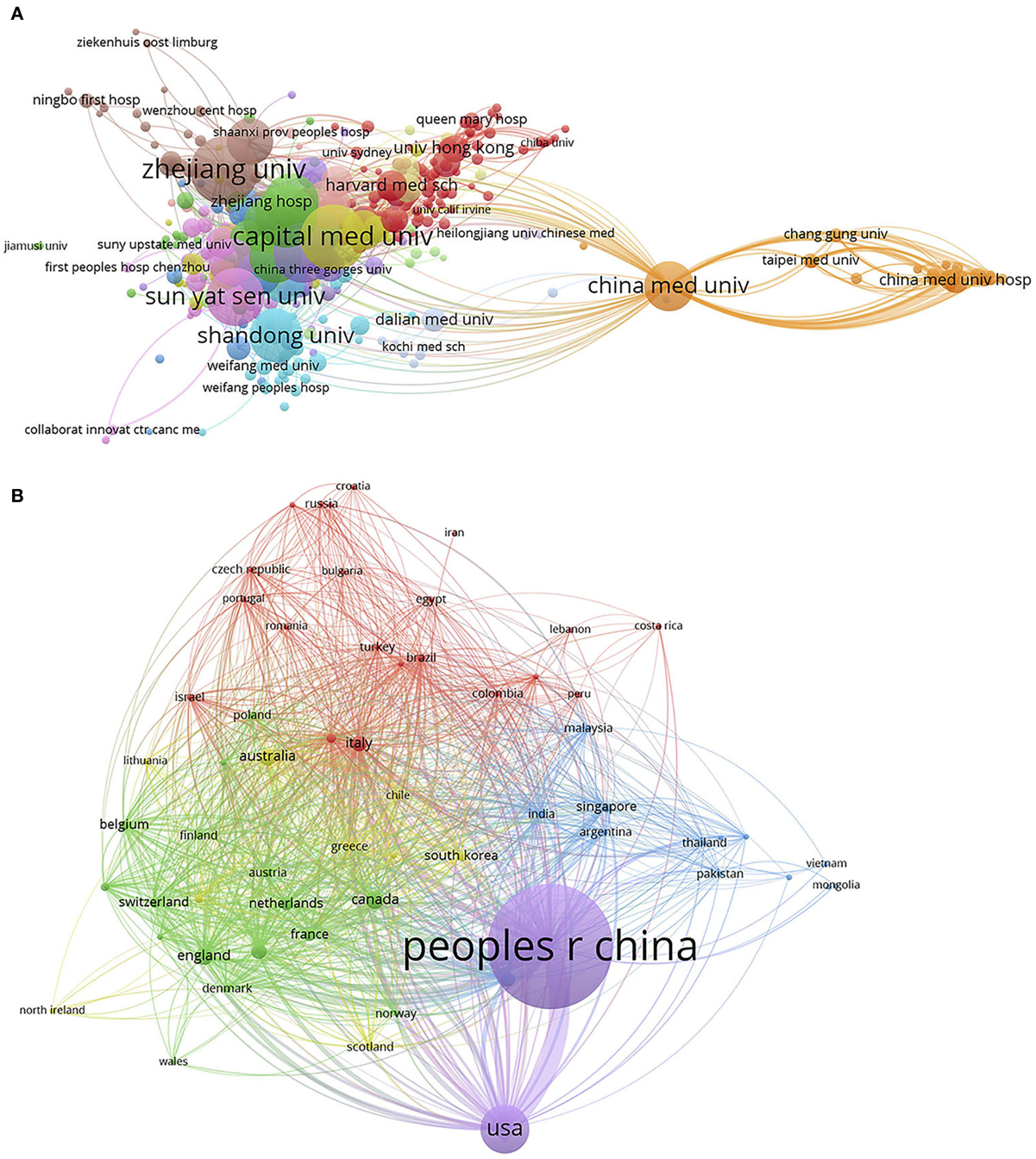



FIGURE 5
National and international Co-authorship network. Cluster regarding collaboration in Critical Care Medicine researches was mapped between authors from different affiliations of China (i.e national collaboration, **A**) or between authors from China and those from other countries (i.e international collaboration, **B**) from 2001 to 2020. Each circle indicated an affiliation of China, or one country. The circle size represented the number of publications and the lines between two circles indicated co-appearance of two affiliations or countries in one article. Color of each circle indicated the cluster. The thickness of lines indicated strength of linkage calculated by the number of publications co-authored by the different affiliations or countries.

CCM was written by Nusbaum independently (9). Significantly, the success of clinical trial groups such as the Canadian Critical Care Trials Group (CCCTG) (22) and Australia and New Zealand Intensive Care Society Clinical Trials Group (ANZICS-CTG) (23) has fueled efforts to build similar collaboration models around the world. In China, an investigator-led group,

China Critical Care Clinical Trials Group (CCCCTG) was launched 20 years ago (24), and was active in Critical Care researches over the ensuing years. By this bibliometric analysis, however, it was revealed that collaborations between either domestic or international authors were limited in CCM researches. Moreover, the results showed that the most frequent collaborations took place among the authors who served in the same ICU (Figure 4, Supplementary Table S7). Therefore, collaboration could be a modifiable factor to promote the research quality of Chinese intensivists in future.

Funding is important to facilitate either basic or clinical medical research. However, CCM research not only in China, but around the world, was under-funded in comparison with other specialties, although critical illnesses became a burden of healthcare increasingly. According to Coopersmith's report, 332 (1.7%) out of 19,257 grants funded by the National Institutes of Health were definitely related to critical care and a maximum of 1,212 (6.3%) grants were possibly related to critical care (25). It was demonstrated that 5,624 (41.6%) out of 13,487 publications reported funding in this study. Additionally, we performed a search on the Website Science net (https://fund.sciencenet.cn/) for grants from catalog of H15 ("acute and intensive care medicine/trauma/burns/plastic surgery") of NSFC (National Natural Science Foundation of China) and successfully applied by the Chinese intensivists from 1 January 2016 to 31 December 2020. Of a total of 1,073 (517.85 million RMB Yuan) funded projects, as shown in Supplementary Table S8, only 141 (13.14%; 6.344 million out of 517.85 million RMB Yuan) led by Chinese intensivists have been approved. Interestingly, rapid growth in clinical trials was found in both websites Clinical Trials (https://clinicaltrials.gov/) and ChiCTR (Chinese Clinical Trail Registry, http://www.chictr.org.cn) registered by Chinese intensivists from 1 January 2016 to 31 December 2020 (Supplementary Figure S5). These findings suggested that multiple resources of funding would be a possible strategy to promote Chinese CCM research in future.

There were several limitations in this study. First, this research was only based on the electronic database of the Web of Science, while other electronic databases were not searched and analyzed, such as PubMed, Embase, and Cochrane Library, especially published in Chinese Literature databases such as CNKI, CQVIP, Wanfang, etc. Second, there were some flaws in our data source. For example, an author signed different names of hospital / institute / university in his / her different published articles, making the system unable to identify the articles published by the same person. Third, the software defaults so that the acronym cannot be changed. and if you want to change it, you may need to do the later stage of photoshop (but this may cause manual revision and non-repeatability of the results). Fourth, there may be differences in data recognition by different software, resulting in possible errors in results. Finally, when calculating clinical registration research items, we cannot completely exclude a very small number of projects led by respiratory and critical illness experts, anesthesiologists, or other emergency department experts from being included in this study.

## Conclusion

This bibliometric analysis demonstrated that CCM research in China grew rapidly in recent 20 years. However, the impactful publications remained limited. The results of this study suggested that the lack of universality, as well as a comprehensive training in scientific researches, inactive collaboration, and underfunded, be the important barriers to the promotion of the quality and quantity of Chinese CCM research.

## Data availability statement

The original contributions presented in the study are included in the article/Supplementary material, further inquiries can be directed to the corresponding author.

## Author contributions

WQ, CX, and ZL contributed to the acquisition, analysis, interpretation of data, statistical analysis, data arrangement, and draft of the manuscript. LY contributed to interpret the results. FS contributed to supervision manuscript. LZ completed all statistical analyses of this study. PM contributed to study concept, supervision, organize the final manuscript, identified as the guarantor of the article, taking responsibility for the integrity of the work, and from inception to published article. All authors contributed to the article and approved the submitted version.

## Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

## Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

## Supplementary material

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fmed.2022.974025/full#supplementary-material

## References


1. Yin H, Wang S, Zhu Y, Zhang R, Ye X, Wei J, et al. The development of critical care medicine in China: from SARS to COVID-19 pandemic. *Crit Care Res Pract.* (2020) 2020:3956732. doi: 10.1155/2020/3956732

2. Li Z, Liao Z, Wu FX, Yang LQ, Sun YM Yu WF. Scientific publications in critical care medicine journals from Chinese authors: a 10-year survey of the literature. *J Trauma.* (2010) 69:E20–3. doi: 10.1097/TA.0b013e3181c45257

3. Ma P, Du B. Critical care research in mainland China: more needed on the international stage. *Intensive Care Med.* (2013) 39:768–70. doi: 10.1007/s00134-013-2853-8

4. Professional Committee of critical Care Medicine of Chinese Pathophysiology Society. (2017). Available online at: http://csccm.org.cn/?p=13723 (accessed May 9, 2022).

5. Cao J, Hu X, Cheng W, Yu L, Tu WJ, Liu Q. Correction to: clinical features and short-term outcomes of 18 patients with corona virus disease 2019 in intensive care unit. *Intensive Care Med.* (2020) 46:1298. doi: 10.1007/s00134-020-06037-y

6. Du M, Garcia JGN, Christie JD, Xin J, Cai G, Meyer NJ, et al. Integrative omics provide biological and clinical insights into acute respiratory distress syndrome. *Intensive Care Med.* (2021) 47:761–71. doi: 10.1007/s00134-021-06410-5

7. Li Y, Li H, Li M, Zhang L, Xie M. The prevalence, risk factors and outcome of cardiac dysfunction in hospitalized patients with COVID-19. *Intensive Care Med.* (2020) 46:2096–8. doi: 10.1007/s00134-020-06205-0

8. Shi L, Xu J, Duan G, Yang H, Wang Y. The pooled prevalence of pulmonary embolism in patients with COVID-19. *Intensive Care Med.* (2020) 46:2089–91. doi: 10.1007/s00134-020-06235-8

9. Zhang Z, Van Poucke S, Goyal H, Rowley DD, Zhong M, Liu N. The top 2,000 cited articles in critical care medicine: a bibliometric analysis. *J Thorac Dis.* (2018) 10:2437–47. doi: 10.21037/jtd.2018.03.178

10. Marshall JC, Kwong W, Kommaraju K, Burns KEA. Determinants of citation impact in large clinical trials in critical care: the role of investigator-led clinical trials groups*. *Crit Care Med.* (2016) 44:663–70. doi: 10.1097/CCM.0000000000001466

11. Wang S, Zhou H, Zheng L, Zhu W, Zhu L, Feng D, et al. Global trends in research of macrophages associated with acute lung injury over past 10 years: a bibliometric analysis. *Front Immunol.* (2021) 12:669539. doi: 10.3389/fimmu.2021.669539

12. Shen L, Wang S, Dai W, Zhang Z. Detecting the interdisciplinary nature and topic hotspots of robotics in surgery: social network analysis and bibliometric study. *J Med Internet Res.* (2019) 21:e12625. doi: 10.2196/12625

13. VOSviewer version 1.6.18. Available online at: https://www.vosviewer.com/ (accessed January 24, 2022).

14. *Clarivate.InCites Benchmarking and Analytics.* Available online at: https://clarivate.com/webofsciencegroup/solutions/incites/ (accessed June 19, 2021).

15. van Eck NJ, Waltman L. Software survey: VOSviewer, a computer program for bibliometric mapping. *Scientometrics.* (2010) 84:523–38. doi: 10.1007/s11192-009-0146-3

16. Su X, Meng ZT, Wu XH, Cui F, Li HL, Wang DX, et al. Dexmedetomidine for prevention of delirium in elderly patients after non-cardiac surgery: a randomised, double-blind, placebo-controlled trial. *Lancet.* (2016) 388:1893–902. doi: 10.1016/S0140-6736(16)30580-3

17. Yang X, Yu Y, Xu J, Shu H, Xia J, Liu H, et al. Clinical course and outcomes of critically ill patients with SARS-CoV-2 pneumonia in Wuhan, China: a single-centered, retrospective, observational study. *Lancet Respir Med.* (2020) 8:475–81. doi: 10.1016/S2213-2600(20)30079-5

18. Liu J, Zhang L, Ma P. A new era of critical care research in China. *J Crit Care.* (2019) 54:20–1. doi: 10.1016/j.jcrc.2019.07.005

19. Metzger JC, Eastman AL, Pepe PE. Year in review 2009: Critical Care–cardiac arrest, trauma and disasters. *Crit Care.* (2010) 14:242. doi: 10.1186/cc9302

20. Du B, Wang C, Singer M. Learning for the next pandemic: the Wuhan experience of managing critically ill people. *BMJ.* (2021) 375:e066090. doi: 10.1136/bmj-2021-066090

21. Marshall JC. Global collaboration in acute care clinical research: opportunities, challenges, and needs. *Crit Care Med.* (2017) 45:311–20. doi: 10.1097/CCM.0000000000002211

22. Marshall JC, Cook DJ. Investigator-led clinical research consortia: the Canadian Critical Care Trials Group. *Crit Care Med.* (2009) 37:S165–72. doi: 10.1097/CCM.0b013e3181921079

23. Jones DA, Cooper DJ, Finfer SR, Bellomo R, Myburgh JA, Higgins A, et al. Advancing intensive care research in Australia and New Zealand: development of the binational ANZIC Research Centre. *Crit Care Resusc.* (2007) 9:198–204. doi: 10.3316/informit.515804578173647

24. Du B, Xi X, Chen D, Peng J, China Critical Care Clinical Trial G. Clinical review: critical care medicine in mainland China. *Critical care.* (2010) 14:206. doi: 10.1186/cc8222

25. Coopersmith CM, Wunsch H, Fink MP, Linde-Zwirble WT, Olsen KM, Sommers MS, et al. A comparison of critical care research funding and the financial burden of critical illness in the United States. *Crit Care Med.* (2012) 40:1072–9. doi: 10.1097/CCM.0b013e31823c8d03

frontiers | Frontiers in Medicine

TYPE Original Research
PUBLISHED 18 November 2022
DOI 10.3389/fmed.2022.935366



OPEN ACCESS

EDITED BY
Qinghe Meng,
Upstate Medical University,
United States

REVIEWED BY
I-Shiang Tzeng,
National Taipei University, Taiwan
Rishikesan Kamaleswaran,
Emory University, United States

*CORRESPONDENCE
Chong-Chi Chiu
chiuchongchi@gmail.com
Chin-Ming Chen
chencm3383@gmail.com

†These authors have contributed equally to this work

SPECIALTY SECTION
This article was submitted to Intensive Care Medicine and Anesthesiology, a section of the journal Frontiers in Medicine

RECEIVED 03 May 2022
ACCEPTED 11 October 2022
PUBLISHED 18 November 2022

CITATION
Liu C-F, Hung C-M, Ko S-C, Cheng K-C, Chao C-M, Sung M-I, Hsing S-C, Wang J-J, Chen C-J, Lai C-C, Chen C-M and Chiu C-C (2022) An artificial intelligence system to predict the optimal timing for mechanical ventilation weaning for intensive care unit patients: A two-stage prediction approach. *Front. Med.* 9:935366. doi: 10.3389/fmed.2022.935366



# An artificial intelligence system to predict the optimal timing for mechanical ventilation weaning for intensive care unit patients: A two-stage prediction approach

Chung-Feng Liu[1†], Chao-Ming Hung[2,3†], Shian-Chin Ko[4], Kuo-Chen Cheng[5], Chien-Ming Chao[6,7], Mei-I Sung[4], Shu-Chen Hsing[4], Jhi-Joung Wang[8,9], Chia-Jung Chen[10], Chih-Cheng Lai[11], Chin-Ming Chen[12]* and Chong-Chi Chiu[2,13,14,15]*

[1]Department of Medical Research, Chi Mei Medical Center, Tainan, Taiwan, [2]Department of General Surgery, E-Da Cancer Hospital, Kaohsiung, Taiwan, [3]College of Medicine, I-Shou University, Kaohsiung, Taiwan, [4]Department of Respiratory Therapy, Chi Mei Medical Center, Tainan, Taiwan, [5]Department of Internal Medicine, Chi Mei Medical Center, Tainan, Taiwan, [6]Department of Intensive Care Medicine, Chi Mei Medical Center, Liouying, Taiwan, [7]Department of Dental Laboratory Technology, Min-Hwei College of Health Care Management, Liouying, Taiwan, [8]Department of Anesthesiology, Chi Mei Medical Center, Tainan, Taiwan, [9]Department of Anesthesiology, National Defense Medical Center, Taipei, Taiwan, [10]Department of Information Systems, Chi Mei Medical Center, Tainan, Taiwan, [11]Division of Hospital Medicine, Department of Internal Medicine, Chi Mei Medical Center, Tainan, Taiwan, [12]Department of Intensive Care Medicine, Chi Mei Medical Center, Tainan, Taiwan, [13]School of Medicine, College of Medicine, I-Shou University, Kaohsiung, Taiwan, [14]Department of Medical Education and Research, E-Da Cancer Hospital, Kaohsiung, Taiwan, [15]Department of General Surgery, Chi Mei Medical Center, Tainan, Taiwan

**Background:** For the intensivists, accurate assessment of the ideal timing for successful weaning from the mechanical ventilation (MV) in the intensive care unit (ICU) is very challenging.

**Purpose:** Using artificial intelligence (AI) approach to build two-stage predictive models, namely, the try-weaning stage and weaning MV stage to determine the optimal timing of weaning from MV for ICU intubated patients, and implement into practice for assisting clinical decision making.

**Methods:** AI and machine learning (ML) technologies were used to establish the predictive models in the stages. Each stage comprised 11 prediction time points with 11 prediction models. Twenty-five features were used for the first-stage models while 20 features were used for the second-stage models. The optimal models for each time point were selected for further practical implementation in a digital dashboard style. Seven machine learning algorithms including Logistic Regression (LR), Random Forest (RF), Support Vector Machines (SVM), K Nearest Neighbor (KNN), lightGBM, XGBoost, and Multilayer Perception (MLP) were used. The electronic medical records of the intubated ICU patients of Chi Mei Medical Center (CMMC) from 2016 to 2019 were included for modeling. Models with the highest area under the receiver

operating characteristic curve (AUC) were regarded as optimal models and used to develop the prediction system accordingly.

**Results:** A total of 5,873 cases were included in machine learning modeling for Stage 1 with the AUCs of optimal models ranging from 0.843 to 0.953. Further, 4,172 cases were included for Stage 2 with the AUCs of optimal models ranging from 0.889 to 0.944. A prediction system (dashboard) with the optimal models of the two stages was developed and deployed in the ICU setting. Respiratory care members expressed high recognition of the AI dashboard assisting ventilator weaning decisions. Also, the impact analysis of with- and without-AI assistance revealed that our AI models could shorten the patients' intubation time by 21 hours, besides gaining the benefit of substantial consistency between these two decision-making strategies.

**Conclusion:** We noticed that the two-stage AI prediction models could effectively and precisely predict the optimal timing to wean intubated patients in the ICU from ventilator use. This could reduce patient discomfort, improve medical quality, and lower medical costs. This AI-assisted prediction system is beneficial for clinicians to cope with a high demand for ventilators during the COVID-19 pandemic.

KEYWORDS

**artificial intelligence, machine learning, intensive care unit, weaning mechanical ventilation, optimal weaning timing**

## Introduction

Mechanical ventilation (MV) is frequently applied in the intensive care unit (ICU). Approximately eight hundred thousand patients receive MV annually in the United States (1). Extubation decision is critical during an ICU stay. An early trial of the weaning process and successful extubation may lower the medical costs and ventilator-related complication rates. Besides, it could improve the patient's prognosis (2–4). Therefore, after the recovery of the critical illness, clinicians should immediately prepare to liberate the patients from MV. Evaluation of an ICU patient's fitness for weaning and subsequent extubation is objectively referred to the airway, respiratory, neurological parameters, *etc.* (5). Most of the times, liberation from MV requires three steps – readiness testing, weaning, and extubating and the process of MV liberation is dynamic and complicated. In daily practice, extubation is usually left to the discretion of the clinician (6); therefore, various protocols for ventilator weaning have been established and assessed to increase the extubation rate (7–18).

Despite following the recommended extubation process established in the American Thoracic Surgery weaning protocol, the failure rate still ranges from 10 to 15% of ICU patients in the United States (19). Truthfully, there has been no significant decrease in extubation failure in the past decades. Therefore, an advanced strategy is mandatory to increase the prediction accuracy (20). Several multivariate outcome prediction models have evolved in many aspects of health care research in these years. They include artificial neural networks (ANN), logistic regression (LR) models, random forest (RF) models, and support vector machines (SVM) (21–26). Machine learning (ML) is a subject of computer science that incorporates numerous components to empower the systems to learn from currently acquired data, predict the outcome, and make changes in action when faced with a new problem. Clinically, ML could increase the prediction rate of successful weaning from ventilatory support. The parameters considered in the prediction of successful weaning and extubation were based on literature (27–34) and clinical experience.

Many studies (35–38) have reported the usefulness of AI in the ICU, such as the early warning systems that predict the risk of physiological deterioration in acutely ill patients, the development of acute respiratory distress syndrome, the early development of sepsis and the pathogen that causes it, and clinical outcome and mortality. However, studies on the utility of AI in predicting the weaning and extubation process among critically ill patients requiring MV are limited (39–43), while those that explore AI's capacity to predict the weaning timing for intubated patients are rare.

This study aims to develop an AI digital dashboard to remind the ICU clinicians of the optimal timing for

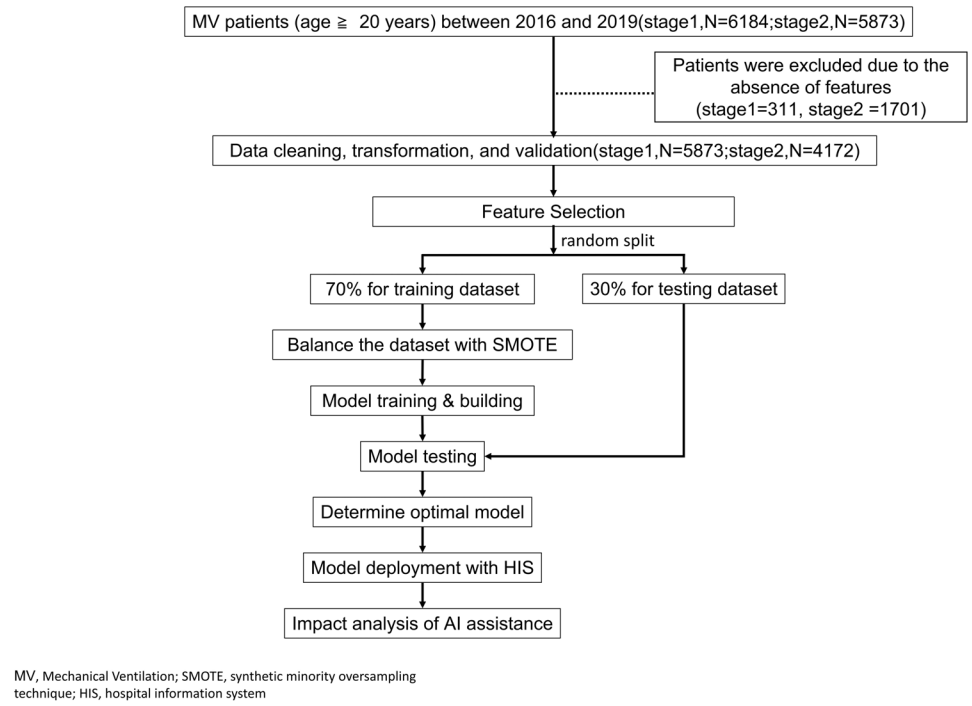



FIGURE 1
Study flow.

weaning initiation, propose an individualized treatment recommendation, and assist in making extubation decisions. Data that can be conveniently collected were chosen as variables for building the prediction model, including patients' characteristics and respiratory pattern parameters during spontaneous breathing trials (SBTs). A preliminary impact analysis was performed after AI assistance to predict successful extubation in ICU patients.

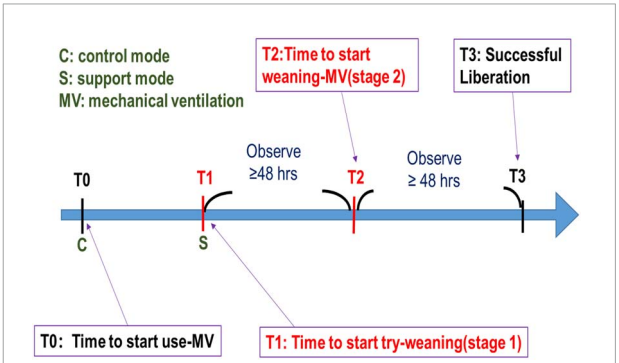



FIGURE 2
Two-stage weaning assessment.

# Materials and methods

## Ethical consideration

This study was reviewed and accepted by the CMMC (IRB Serial No.: 10912-016). The process was performed according to the approved guidelines and regulations, and informed consent was waived from the patients because of the nature of our retrospective study.

## Study design

The study flow chart is demonstrated in Figure 1. In the beginning, we established a professional team, including clinicians, respiratory therapists, data scientists, and information technology engineers, and held regular meetings and discussions. We retrospectively collected data from adult ventilated patients (≥20 years old) who stayed at the ICU of CMMC from January 2016 to December 2019. Patients who signed the DNR (Do not resuscitate) were excluded. According to clinical experience, if the try-weaning timing is appropriate, the success probability of the final complete weaning ventilator will also increase. Therefore, this study divided the complete assessment of ventilator use into two stages: (1) the try-weaning stage and (2) the complete weaning MV stage (Figure 2).

Try-weaning stage means switching the ventilator from control mode to support mode for an ICU patient, while the complete weaning MV stage means transitioning from support mode ventilation to oxygen therapy or extubation for an ICU patient.

## Setting and data source

Chi Mei Medical Center is a large hospital in Tainan, Taiwan with 1288 beds, including 109 ICU beds. It has a comprehensive hospital information system to store each kind of clinical data such as demographics, diagnoses, vital signs, laboratory data, and prescribed medications in the database. Since 2016, CMMC adopted IoT technology to capture parameters from the MV in ICUs automatically per minute. So far, big data from MV was cumulated and ready for further AI and machine learning study.

## Features and outcome variables

The first stage model used 25 features, including primary patient data of age, Acute Physiology and Chronic Health Evaluation II (APACHE II) score, Therapeutic Intervention Scoring System (TISS) score, and the first and last Internet of Thing (IoT) data of the respirator consisting of inspired oxygen fraction ($FiO_2$), positive end-expiratory pressure (PEEP), respiratory rate (RR), minute ventilation (Mv), peak inspiratory pressure (Ppeak), mean airway pressure (mPaw), peripheral oxygen saturation ($SpO_2$), expiratory tidal volume (Vte), heart rate (HR), systolic blood pressure (SBP), and diastolic blood pressure (DBP). Based on clinical experience, the outcome variable was binary coded with 1 (i.e., successful try-weaning), which means that MV was shifted from the control mode to the support mode for at least 48 h, otherwise it was coded with 0.

The second stage model used 20 features, including primary data consisting of age, APACHE II score, and TISS score; and the last respirator IoT data before extubation consisting of $FiO_2$, PEEP, RR, Mv, Ppeak, mPaw, $SpO_2$, pressure support level (PSL), tidal volume with pressure support (PSLvolume), body temperature (BT), HR, SBP, DBP, Glasgow Coma Scale eye-opening (GCS_E), GCS motor response (GCS_M), SBT count during support mode, and sputum suction count within 24 hours before extubation (Suction). The outcome variable was binary coded with 1 (i.e., successful weaning MV), which means weaning from MV for at least 48 h, otherwise coded with 0. This is also accepted as a basis for the provision of government-related health subsidies in Taiwan.

All potential features were selected based on the literature (6, 7, 44, 45), clinic availability and the experience of clinicians. We performed correlation analysis between features and outcomes to assist in feature selection decisions. Features with the raw data were obtained from the hospital information system (HIS) and real-time IoT transferring from ventilators.

## Model building and measurement

Raw data was collected from the electronic medical records of ICU to build the models for stage 1 and stage2. We randomly divided the cleaned data into 70% training and 30% testing

TABLE 1 Stage 1 demography.

| Feature | | Overall | |
|---|---|---|---|
| | | *N*=5873 | |
| Age, mean (SD) | | 64.0 (15.3) | |
| APACHE II score, mean (SD) | | 19.6 (8.5) | |
| TISS score, mean (SD) | | 29.7 (7.9) | |
| IoT data | First* | | Last** |
| $FiO_2$, mean (SD) | 45.4 (20.8) | | 32.3 (15.0) |
| PEEP, mean (SD) | 5.6 (1.4) | | 5.8 (1.5) |
| RR, mean (SD) | 15.8 (4.2) | | 14.0 (4.0) |
| Mv, mean (SD) | 8.8 (2.8) | | 7.9 (2.4) |
| Ppeak, mean (SD) | 24.3 (4.3) | | 23.1 (4.1) |
| mPaw, mean (SD) | 10.6 (2.8) | | 10.1 (2.7) |
| $SpO_2$, mean (SD) | 98.7 (2.3) | | 98.1 (4.7) |
| Vte, mean (SD) | 576.0 (115.5) | | 576.8 (117.8) |
| HR, mean (SD) | 93.8 (21.3) | | 84.4 (21.4) |
| SBP, mean (SD) | 139.4 (37.8) | | 128.8 (33.1) |
| DBP, mean (SD) | 79.2 (21.6) | | 68.9 (18.7) |
| **Outcome** | | | |
| Successful try-weaning within 8 h, *n* (%) | | 1,113 (19.0) | |
| Successful try-weaning within 12 h, *n* (%) | | 1,588 (27.0) | |
| Successful try-weaning within 24 h, *n* (%) | | 2,840 (48.4) | |
| Successful try-weaning within 36 h, *n* (%) | | 3,112 (53.0) | |
| Successful try-weaning within 48 h, *n* (%) | | 3,523 (60.0) | |
| Successful try-weaning within 60 h, *n* (%) | | 3,710 (63.2) | |
| Successful try-weaning within 72 h, *n* (%) | | 3,968 (67.6) | |
| Successful try-weaning within 84 h, *n* (%) | | 4,114 (70.0) | |
| Successful try-weaning within 96 h, *n* (%) | | 4,281 (72.9) | |
| Successful try-weaning within 108 h, *n* (%) | | 4,373 (74.5) | |
| Successful try-weaning within 120 h, *n* (%) | | 4,506 (76.7) | |

*First: data of the first record in control mode. **Last: data of the last record in control mode. SD, Standard Deviation; IoT, Internet of Things; APACHE II, Acute Physiology and Chronic Health Evaluation II; TISS, Therapeutic intervention scoring system; FiO2, the fraction of inspired oxygen; PEEP, positive end-expiratory pressure; RR, respiratory rate; Mv, minute ventilation; Ppeak, peak inspiratory pressure; mPaw, mean airway pressure; SpO2, peripheral oxygen saturation; Vte, expiratory tidal volume; HR, heart rate; SBP, systolic blood pressure; DBP, diastolic blood pressure.

data. Due to the data imbalance problem (fewer cases in the minority class), we applied the Synthetic minority oversampling technique (SMOTE) method to process the training data (46). We performed a grid search for five-fold cross-validation on the training dataset to obtain the best hyperparameters for modeling. Finally, we used the testing dataset (also called hold-out dataset) for the final evaluation of the model quality. Four model quality indicators of accuracy, sensitivity, specificity, and AUC (area under the ROC) were applied to assess the model quality. However, the overall model performance is generally evaluated by AUC in many medical studies since both true/false positive and true/false negative are fairly considered. Thus, AUC was used in this study as the main indicator to determine the optimal model. We used the optimal models for subsequent implementation of the predictive system.

TABLE 2 Stage 2 demography.

| Feature | Overall |
|---|---|
| | ***N*= 4172** |
| Age, mean (SD) | 64.3 (15.3) |
| APACHE II score, mean (SD) | 18.9 (8.0) |
| TISS score, mean (SD) | 29.6 (7.7) |
| $FiO_2$, mean (SD) | 26.1 (2.1) |
| PEEP, mean (SD) | 5.2 (0.7) |
| RR, mean (SD) | 16.4 (5.0) |
| Mv, mean (SD) | 7.7 (2.4) |
| PSL, mean (SD) | 9.4 (2.0) |
| PSLvolume, mean (SD) | 484.4 (125.3) |
| Ppeak, mean (SD) | 15.4 (2.0) |
| mPaw, mean (SD) | 8.3 (1.8) |
| $SpO_2$, mean (SD) | 98.7 (1.6) |
| BT, mean (SD) | 36.6 (0.5) |
| HR, mean (SD) | 85.4 (16.7) |
| SBP, mean (SD) | 135.1 (23.8) |
| DBP, mean (SD) | 72.2 (14.9) |
| GCS_E, mean (SD) | 3.5 (0.7) |
| GCS_M, mean (SD) | 5.7 (0.7) |
| SBT times, mean (SD) | 1.4 (2.8) |
| Suction times, mean (SD) | 5.0 (4.4) |
| **Outcome** | |
| Successful weaning-MV within 24 h, *n* (%) | 1,807 (43.3) |
| Successful weaning-MV within 48 h, *n* (%) | 2,133 (51.1) |
| Successful weaning-MV within 72 h, *n* (%) | 2,451 (58.7) |
| Successful weaning-MV within 96 h, *n* (%) | 2,709 (64.9) |
| Successful weaning-MV within 120 h, *n* (%) | 2,910 (69.8) |
| Successful weaning-MV within 144 h, *n* (%) | 3,070 (73.6) |
| Successful weaning-MV within 168 h, *n* (%) | 3,198 (76.7) |
| Successful weaning-MV within 192 h, *n* (%) | 3,312 (79.4) |
| Successful weaning-MV within 216 h, *n* (%) | 3,402 (81.5) |
| Successful weaning-MV within 240 h, *n* (%) | 3,462 (83.0) |
| Successful weaning-MV within 264 h, *n* (%) | 3,518 (84.3) |

SD, Standard Deviation; APACHE II, Acute Physiology and Chronic Health Evaluation II; TISS, Therapeutic intervention scoring system; FiO2, the fraction of inspired oxygen; PEEP, positive end-expiratory pressure; RR, respiratory rate; Mv, minute ventilation; PSL, pressure support level; PSLvolume, tidal volume with pressure support; Ppeak, peak inspiratory pressure; mPaw, mean airway pressure; SpO2: BT, body temperature; HR: heart rate; SBP, systolic blood pressure; DBP, diastolic blood pressure; GCS_E: Glasgow Coma Scale eye-opening; GCS_M, Glasgow Coma Scale-motor response; SBT, spontaneous breathing trials.

Each outcome used a variety of ML algorithms to build models, including LR, RF, SVM, K Nearest Neighbor (KNN), lightGBM, XGBoost, and Multilayer Perception (MLP). The ML models were performed based on Sklearn library and related ML modules in Python.

The main purpose of this study was to predict the optimal timing to wean MV, not just successful weaning or not; thus, we divided each stage into 11 time periods based on clinical experience, and built 11 prediction models with the period data rather than building a single model with the end-point data of ICU patients with MV. That is, it is of great value to predict whether a patient can successfully wean or not over time. After all, most patients with MV in ICU will eventually be successfully weaned but we expect timely or even early safe weaning of MV to avoid overuse rather than just predicting success or not (in CMMC, the average extubation success rate exceeds 85%). However, hospitals can reduce or increase the predictive periods according to their needs while implementing.

Stage 1 of timing prediction for successful try-weaning involves the following: After the patient enters the ICU for intubation, we built 11 models for 11 prediction time points, namely: 8th hour, 12th hour, 24th hour, 36th hour, 48th hour, 60th hour, 72nd hour, 84th hour, 96th hour, 108th hour, and 120th hour. The first stage is considered a success if the MV is shifted from assist control to support mode for at least 48 h.

Stage 2 of timing prediction for successful weaning-MV involves the following: We built 11 models in days (after the patient completed the first stage successfully). The second stage is considered a success if the patient can last longer than 48 h after extubation from the support mode or leave the ICU safely for shorter than 48 h (47).

The data used in each model came from the data collected at this time point. For example, the 60th HR model used data of patients, which was collected at or nearer the 60th hour time point of using the respirator.

## Two-stage artificial intelligence prediction system development of the optimal models

We chose the optimal models for each stage to develop a two-stage prediction system in a digital dashboard style to assist the weaning decision of respiratory medical teams. The system was developed using Microsoft Visual Studio® with VB language.

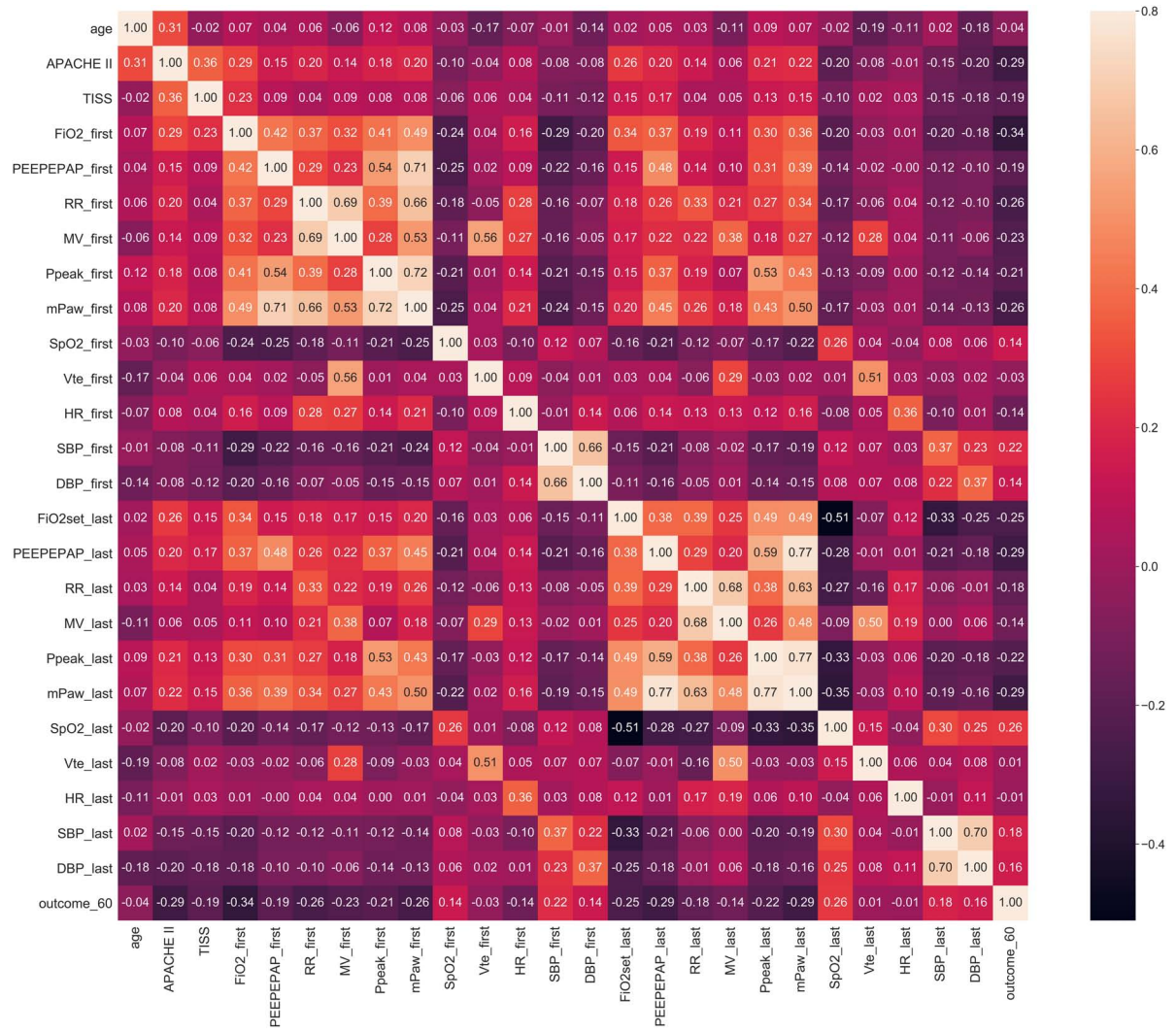



FIGURE 3
Stage 1 Spearman correlation (the 60th hour model). Note: outcome_60: Successful weaning or not before using 60 h of MV; FiO2_first, _last: the first/last value of FiO2 after using MV within 60 h, others are labeled similar.

The web-based dashboard is linked to the real-time database of the existing HIS which could retrieve the required feature values of a specific model. The clinical staff could obtain the related predictive data and figure out the best timing of MV weaning by just previewing the patient's data in the dashboard. The dashboard automatically retrieves the clinical data of the patient for AI prediction without the need for manual input and immediately displays the probabilities of successful MV weaning at each time from the beginning of ventilator use to the nearest future time point. The dashboard would automatically refresh the prediction for all patients every 60 min.

For example, if a patient has used MV for over 50 hours, the dashboard will show the probability of the 24th, 48th, and 72nd hour. The respiratory care team can further double-click on the targeted patient to prompt a new page to overview the detailed feature values at that predicting period. By monitoring the trend curve of the successful probabilities (in colored balls) and detailed feature values, the respiratory care team could evaluate whether each patient is eligible to start trying weaning or liberate the individual from MV more objectively and efficiently at this time point.

# Results

## Demographics

We retrospectively collected 6,184 cases of patients who used MV in CMMC ICU from 2016/1/1 to 2019/12/31. After excluding the cases with missing values, 5,873 cases were

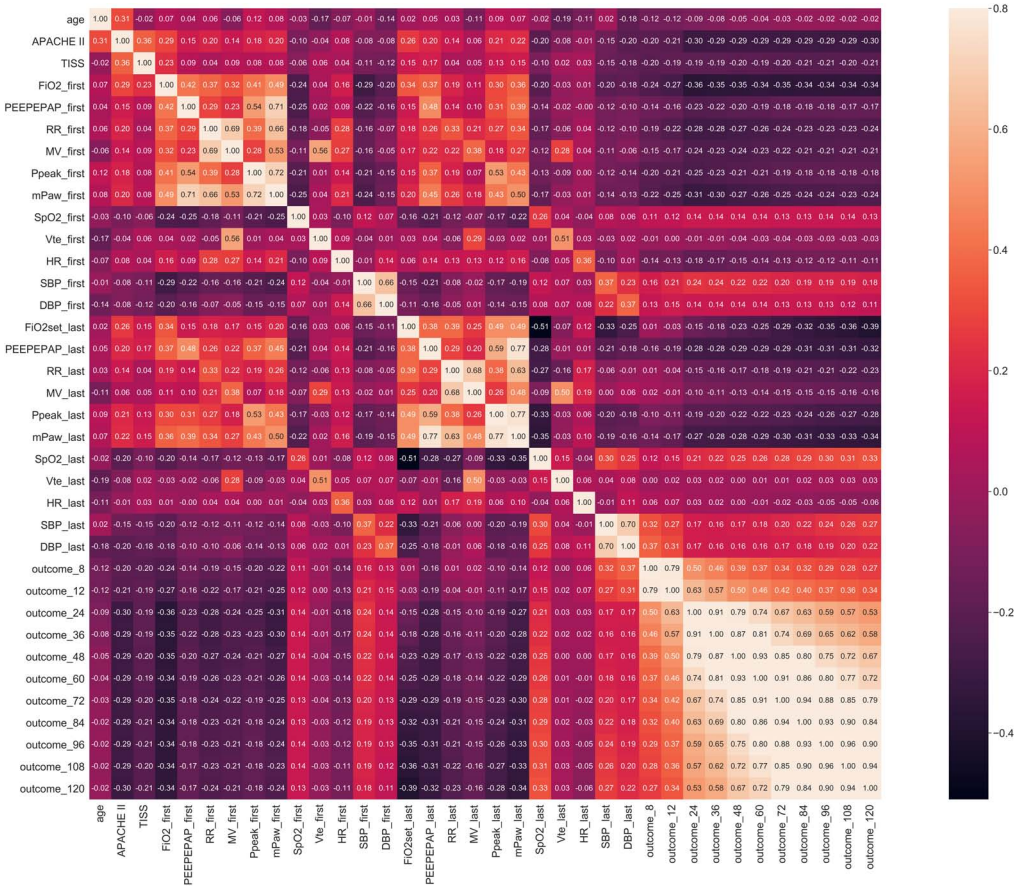



FIGURE 4
Stage 1 Spearman correlation.

included for modeling in Stage 1 and 4,172 cases were included in Stage 2. Tables 1, 2 show the patients' demographics in Stage 1 and Stage 2 respectively.

For example in Stage 1, Spearman correlation analysis for the 60th hour model (Figure 3) showed that the most relevant to the timing of successful try-weaning was the first $FiO_2$, followed by APACHE II score, and the last PEEP and mPaw. Spearman correlation analyses for all models in Stage 1 are shown in Figure 4. Moreover, for Stage 2, Spearman correlation analysis for the $120^{th}$ hour model (Figure 5) showed that the number of SBTs was most relevant to the timing of successful weaning-MV, followed by the number of Suctions. Spearman correlation analyses for all models in Stage 2 are shown in Figure 6.

## Modeling results

In this study, eleven models were established in each of the two stages. In Stage 1, the 60th-hour model was taken as an example. Each model used seven algorithms with optimal hyper-parameters. Models' performances with the seven ML algorithms are shown in Table 3 (Stage 1) (Supplementary Table 1 for other models in Stage 1). With the $60^{th}$-hour model as an example, according to the value of AUC model, the lightGBM model obtained the maximum value (AUC = 0.860) and was used as the basis for implementing the online prediction system. Besides, ROC curve is a performance measurement for a classification model at various thresholds. Figure 7 covers the ROC curves of the seven algorithms and the three highest AUCs (lightGBM, XGBoost and Random forest) ranged from 0.860 to 0.847 showing good model quality with smooth empirical ROC curves and AUCs near to 1.

In Stage 2, the 120th hour (5th day) model was taken as an example. The lightGBM model was selected for implementation based on the AUCs of the seven algorithms (AUC = 0.923) [Table 3 (Stage 2)] (Supplementary Table 2 for other models in Stage 2). Figure 8 shows the ROC curves of the seven algorithms and the three highest AUCs (lightGBM, Random forest and Logistic regression) ranged from 0.913 to 0.923. It also shows excellent models. Hyper-parameters used for building optimal model for each algorithm are listed in Supplementary Table 3.

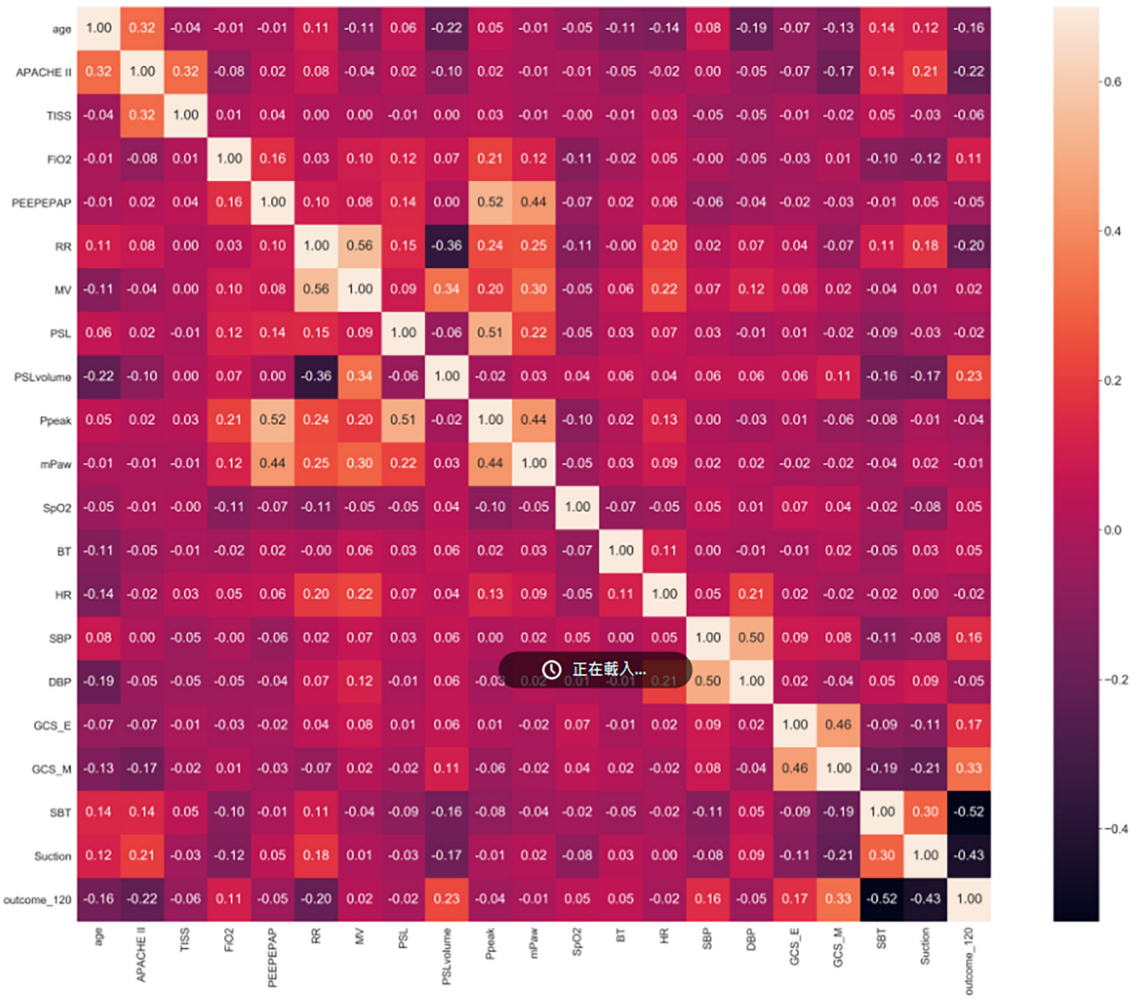



FIGURE 5
Stage 2 Spearman correlation (the 120th Hour model). Note: outcome_120: Successful weaning or not before using 120 h of MV, others are labeled similar.

Moreover, we randomly chose patients A, B, and C who successfully weaned from MV in 2021 (weaning time points at the 144th hr, 216th hr, and 242th hr, respectively) and observed them retrospectively. Taking the data at the 48th-hour ventilator use as features (the patients all failed to wean at the 48th hr), the probabilities predicted by our 48th-hr model were all <50%, which mean a tendency for unsuccessful weaning (probabilities were 32.58, 40.24, and 20.1%, respectively). These predictions were correct. We then fed the same data to a single model (usually the last model, represented here by our 264th-hr model) and all displayed a tendency for successful weaning (probabilities were 95.23, 79.06, and 61.38%). These predictions were incorrect. This proves that, adopting in practical, using multiple models is more appropriate to the prediction of weaning time than when using a single model only.

## Prediction system development and deployment

Using the optimal prediction models, the AI Center and the Department of Information Systems of CMMC jointly developed the timing prediction system (a dashboard) for try-weaning and weaning MV and integrated it with the existing hospital information system (respiratory care system). Such graphical presentation and drill-down interactive function help track the status of patients and enhance users' acceptance of the AI dashboard. Our results showed that this system could predict the optimal timing for try-weaning and weaning MV during the decision-making process of the clinicians. Moreover, the reference data from this system could be used effectively for communication with the patient's family.

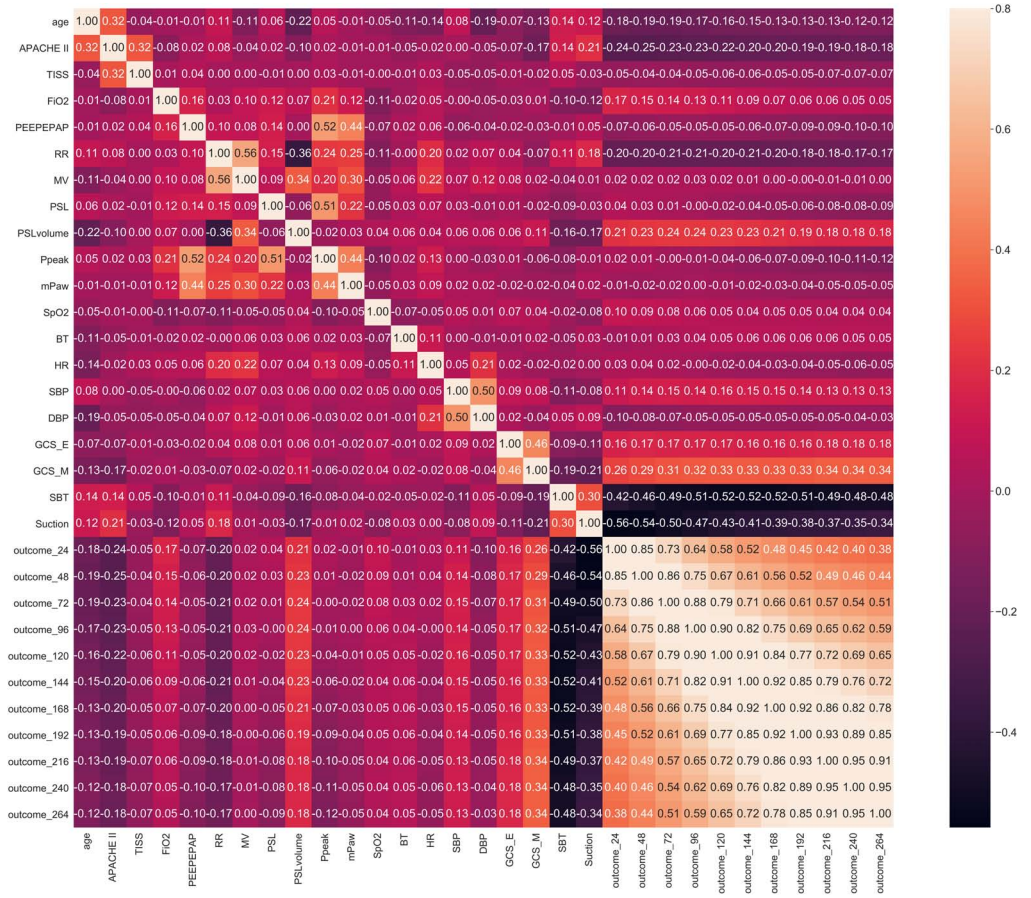



FIGURE 6
Stage 2 Spearman correlation.

## Use case scenario

The interface of the AI dashboard is shown in Figure 9. Stage 1 (try-weaning) displayed the patient's basic information (bed number, medical record number, name), the time when ventilator use was started, the current number of hours of use, and the probability of success in each period. For example, the first patient of Stage 1 had used the ventilator for 63 h; the system captured the patient's characteristic data and displayed the predictions for the nearest future. It could be seen that the probability of successful try-weaning within 72 h of this patient was 56.35%, which implies that the medical team may switch the mode of the patient's ventilator (start try-weaning) during this time. Stage 2 (weaning MV) presented content similar to Stage 1, which included basic information, starting time of support mode, current total hours of support mode, and the success probability of each period. For example, the first patient in Stage 2 had been in the support model for 51 h; the system predicted that the success probability of liberating the patient in MV within three days (72 h) was 33.36%, so it was not recommended to wean during this period.

TABLE 3 Testing results of the predictive models: Stage 1 try-weaning model of the 60th HR and Stage 2 MV-weaning model of the 120th HR.

| Algorithm | Accuracy | Sensitivity | Specificity | AUC |
|---|---|---|---|---|
| **Stage 1** | | | | |
| Logistic regression | 0.710 | 0.710 | 0.710 | 0.776 |
| Random forest | 0.760 | 0.760 | 0.760 | 0.847 |
| SVM | 0.716 | 0.778 | 0.609 | 0.759 |
| KNN | 0.686 | 0.749 | 0.578 | 0.730 |
| LightGBM | 0.768 | 0.788 | 0.733 | 0.860 |
| MLP | 0.732 | 0.746 | 0.709 | 0.815 |
| XGBoost | 0.774 | 0.806 | 0.718 | 0.853 |
| **Stage 2** | | | | |
| Logistic regression | 0.827 | 0.827 | 0.826 | 0.913 |
| Random forest | 0.824 | 0.822 | 0.829 | 0.918 |
| SVM | 0.713 | 0.714 | 0.712 | 0.797 |
| KNN | 0.649 | 0.679 | 0.580 | 0.683 |
| lightGBM | 0.842 | 0.842 | 0.842 | 0.923 |
| MLP | 0.805 | 0.804 | 0.807 | 0.905 |
| XGBoost | 0.810 | 0.810 | 0.810 | 0.908 |

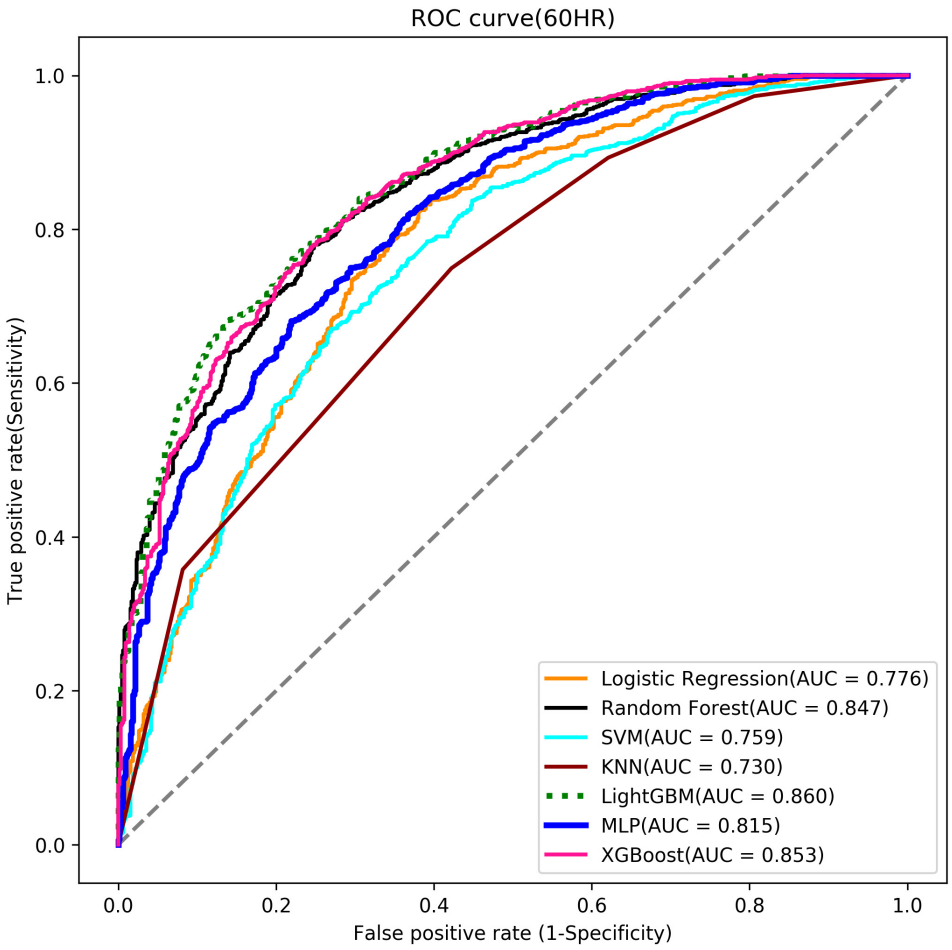



FIGURE 7
Stage 1 ROC curve (the 60th Hour model).

## User evaluation and impact analysis of artificial intelligence assistance

After the hospital launched the dashboard system and implemented it for one month, we interviewed some of the respiratory care members (3 physicians and 5 therapists) and gained high positive feedback. They thought that the dashboard was a very useful tool in helping them determine the optimal timing for trying to wean a patient from the ventilator. According to them, it was also a useful tool for shared decision-making (SDM) especially when communicating with patients or their families. Also, they raised expectations for improvement. For example, they hoped that the predicted value at each time point could be drawn as a polyline to easily see the trend of the predicted probabilities for a patient. These expectations were later realized.

So far, this AI dashboard has been online in ICU for nearly two years. Therefore, we conducted an anonymous 5-scaled questionnaire survey (with Google Form) for all 10 ICU physicians during September 30, 2022 and October 5, 2022, and received 8 valid questionnaires. Overall, they believe that the AI is easy to use (mean = 4.5), the prediction results provided by the AI are of reference value (mean = 4.0), and the AI is helpful to the MV weaning decision (mean = 4.25). However, 2 physicians answered "seldom use AI", 4 physicians answered "frequent use of AI", and the remaining 2 physicians answered "already use AI regularly". One of the physicians who answered " seldom use of AI" left a comment saying that physicians have had extensive experience in assessing MV weaning and AI assistance is not very necessary.

Moreover, we selected an ICU ward and recorded the successful extubation time and associated data. The collected data was then compared with that of the previous year. In other words, the parameters collected from July to November 2019 (without AI assistance) was contrasted with those of July to November 2020 (with AI assistance). Intubated adult patients weaned from MV successfully were enrolled in the study implementation. Patients with tracheostomy and transferred to

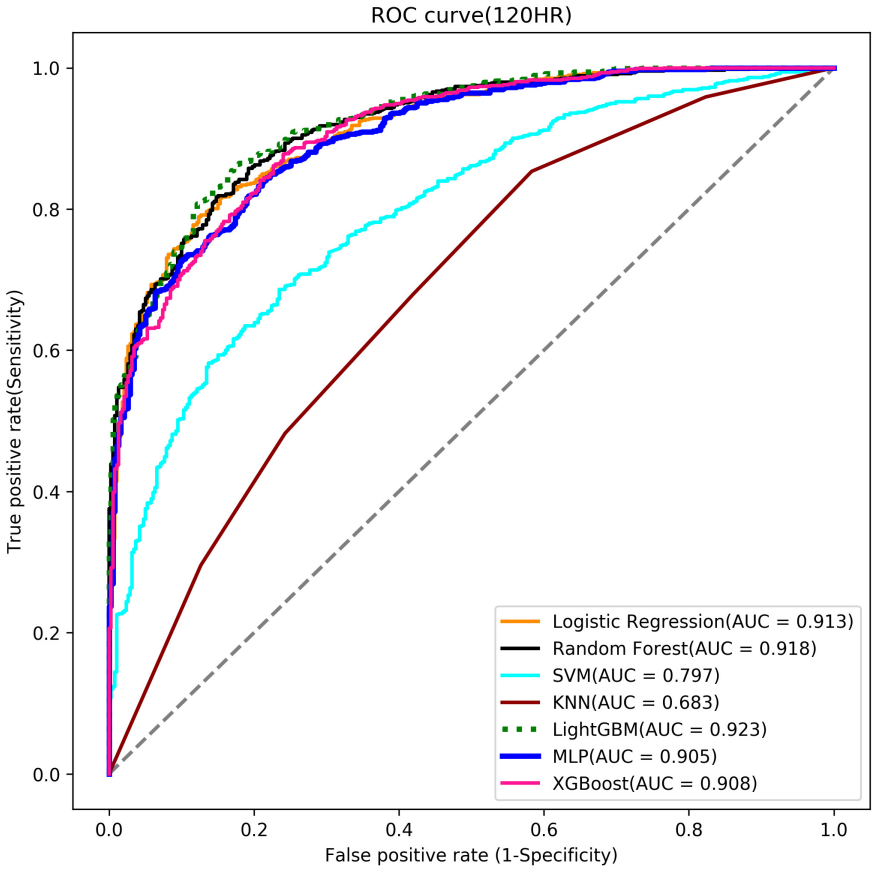



FIGURE 8
Stage 2 ROC curve (the 120th Hour model).

Optimal Timing Prediction for Weaning MV

try weaning

| Station | Bed | Patient ID | Name | Start Use | Hours | 8hr | 12hr | 24hr | 36hr | 48hr | 60hr | 72hr | 84hr | 96hr | 108hr | 120hr |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 3BI | | | | 2021-05-23 21:03 | 63 | | | 36.88% | 43.4% | 23.43% | 31.3% | 56.35% | | | | |
| 3CI | | | | 2021-05-24 14:51 | 45 | 5.96% | 11.23% | 16.41% | 22.52% | 25.38% | | | | | | |
| 4BI | | | | 2021-05-25 02:20 | 33 | | | 46.77% | 57.4% | | | | | | | |
| 5CI | | | | 2021-05-23 19:00 | 65 | | | 12.69% | 13.46% | 15.15% | 38.63% | 35.12% | | | | |
| 5CI | | | | 2021-05-22 20:05 | 88 | | | | | 17.25% | 16.88% | 21.64% | 23.1% | 26.9% | | |
| 5CI | | | | 2021-05-21 15:25 | 116 | 10.64% | | 10.14% | 6.9% | 10.15% | 2.69% | 11.27% | 7.92% | 8.77% | 7.07% | 8.61% |

weaning MV

| Station | Bed | Patient ID | Name | Start Use | Hours | 1 Day | 2 Days | 3 Days | 4 Days | 5 Days | 6 Days | 7 Days | 8 Days | 9 Days | 10 Days | 11 Days |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 3BI | | | | 2021-05-24 08:55 | 51 | 47.92% | 43.73% | 33.36% | | | | | | | | |
| 3BI | | | | 2021-05-25 09:15 | 26 | 56.69% | 56.57% | | | | | | | | | |
| 3BI | | | | 2021-05-24 09:35 | 50 | 20.99% | 57.54% | 72.48% | | | | | | | | |
| 3BI | | | | 2021-05-19 09:40 | 170 | 50.64% | 59.64% | 30.3% | 28.96% | 36.45% | 28% | 44.27% | 31.19% | | | |
| 3BI | | | | 2021-05-26 11:20 | 0 | 64.95% | | | | | | | | | | |
| 3BI | | | | 2021-05-22 09:05 | 99 | 34.62% | 41.61% | 50.98% | 38.43% | 21.2% | | | | | | |
| 3CI | | | | 2021-05-25 08:50 | 27 | 14.05% | 46.71% | | | | | | | | | |
| 4BI | | | | 2021-05-24 09:12 | 50 | 36.89% | 43.39% | 52.15% | | | | | | | | |
| 5CI | | | | 2021-05-24 08:50 | 51 | 33.53% | 51.81% | 66.49% | | | | | | | | |
| 6BI | | | | 2021-05-21 10:00 | 122 | 13.41% | 23.95% | 28.19% | 34.2% | 28.22% | 30.43% | | | | | |

FIGURE 9
A screenshot of the artificial intelligence (AI) prediction system interface.

TABLE 4 The results of clinical evaluation and comparison.

| Feature | Overall | 2019/07-11 (without AI) | 2020/07-11 (with AI) | *P*-Value |
|---|---|---|---|---|
| | *N*=171 | *N*=78 | *N*=93 | |
| **(A) Analysis of weaning rate** | | | | |
| Age, mean ± SD | 66.2 ± 15.7 | 65.7 ± 16.3 | 66.7 ± 15.2 | 0.695 |
| **Gender** | | | | |
| Female, *n* (%) | 58 (33.9) | 29 (37.2) | 29 (31.2) | 0.507 |
| Male, *n* (%) | 113 (66.1) | 49 (62.8) | 64 (68.8) | |
| APACHE II score, mean ± SD | 20.8 ± 8.5 | 20.3 ± 8.7 | 21.7 ± .5 | 0.262 |
| TISS, mean ± SD | 30.6 ± 6.5 | 30.1 ± 6.8 | 31.0 ± 6.4 | 0.243 |
| COMA scale, mean ± SD | 8.7 ± 3.5 | 9.1 ± 3.6 | 8.3 ± 3.4 | 0.068 |
| **Extubation** | | | | |
| Successful, *n* (%) | 167 (97.7) | 76 (97.4) | 91 (97.8) | 1.000 |
| Failure, *n* (%) | 4 (2.3) | 2 (2.6) | 2 (2.2) | |
| **Feature** | **Overall** | **2019/07-11 (without AI)** | **2020/07-11 (with AI)** | ***P*-Value** |
| | *N*=167 | *N*=76 | *N*=91 | |
| **(B) Analysis of successful weaning use-time** | | | | |
| Age, mean ± SD | 66.0 ± 15.8 | 65.7 ± 16.4 | 66.4 ± 15.1 | 0.814 |
| **Gender** | | | | |
| Female, *n* (%) | 56 (33.5) | 29 (38.2) | 27 (29.7) | 0.321 |
| Male, *n* (%) | 111 (66.5) | 47 (61.8) | 64 (70.3) | |
| APACHE II score, mean ± SD | 20.9 ± 8.5 | 20.4 ± 8.8 | 21.8 ± 8.5 | 0.278 |
| TISS, mean ± SD | 30.7 ± 6.5 | 30.3 ± 6.8 | 31.1 ± 6.4 | 0.318 |
| COMA scale, mean ± SD | 8.6 ± 3.5 | 9.0 ± 3.5 | 8.3 ± 3.5 | 0.099 |
| Intubation hours, mean ± SD | 170.9 ± 150.7 | 178.0 ± 147.7 | 156.6 ± 150.4 | 0.300 |
| ICU Days, mean ± SD | 9.3 ± 7.5 | 9.3 ± 8.0 | 8.8 ± 6.9 | 0.631 |

the respiratory care ward were excluded. The analysis results of Table 4A showed no statistically significant difference in the demographic distribution between these two groups, including the age, gender, and disease severity (Apache II, TISS, COMA) of patients. It provided a fair basis for AI intervention comparison. It also showed that there was no significant difference in successful extubation-rate, indicating that patient safety was not compromised (actually slightly improved) with AI. However, in Table 4B, we noticed the average intubation hours after AI intervention were about 21 hours shorter than that without AI intervention, and the average stay in ICU was reduced by 0.5 days, showing that our AI-assisted system does boost patients wean from ventilators earlier, which could improve the quality of care.

We also performed Kappa analysis ($P < 0.05$ for significance) on the patients with AI to estimate the consistency of AI prediction and regular care procedure. As shown in Table 5, all values of Kappa are above 0.61 indicating that all models have substantial consistency between these two decision-making strategies (48).

TABLE 5 Analysis of Kappa values in 11 models of Stage 2.

| Stage 2 model | Cohen Kappa |
|---|---|
| 24 HR model | 0.785 |
| 48 HR model | 0.710 |
| 72 HR model | 0.681 |
| 96 HR model | 0.841 |
| 120 HR model | 0.796 |
| 144 HR model | 0.677 |
| 168 HR model | 0.776 |
| 196 HR model | 0.752 |
| 216 HR model | 0.789 |
| 240 HR model | 0.711 |
| 264 HR model | 0.657 |

## Discussion

Most related studies in the past explored the factors that affect weaning from the ventilator or predicted the success of weaning. However, this study argues that precise weaning

decisions should consist of two phases, try-weaning and complete weaning MV, and that each should have a separate predictive model built. In addition, we believe that, clinically, deciding on the optimal timing for weaning is more crucial than predicting the final success, so we built 11 models at 11 time points for each stage. More importantly, we used the optimal models to build a prediction system (AI dashboard) for monitoring all patients with MV in ICU to validate the feasibility of our comprehensive AI approach. The impact study confirmed that the average intubation time was shortened by 21 h after AI intervention. Overall, this study has significant academic and practical values.

Mechanical ventilation use is a life-guarding technique providing critically ill respiratory support, and it is one of the most common interventions given to ICU patients (49). In this study, correlation analyses for all models in Stage 1 showed that FiO2, mPaw, APACHE II score, PEEP, SpO2 tend to be higher correlated to the predictive models. It implies that oxygenation, hemodynamics and disease severity have great influence on full support mode shift to partial support mode. Increasing FiO2, mPaw and PEEP is to improve the patient's oxygenation status, but too high mPaw and PEEP will cause lung overdistension and affect cardiac output. Similar, frequency of suction, numbers of SBT, GCS_M, APACHE II score, PSL volume, RR tend to higher correlate to the predictive models in Stage 2. It implies that cough strength, respiratory capacity and disease severity affected weaning success. This reminds clinical staff to assess the amount of sputum or secretions, the patient's mobility, ability to cough, and breathing patterns to ensure successful extubation. However, the biomedical etiology and pathophysiology of weaning failure are complex and often multifactorial, including airway and lung dysfunction, brain dysfunction, cardiac dysfunction, diaphragm dysfunction, and endocrine dysfunction. Accordingly, determining the reason and subsequently developing a treatment strategy require a dedicated clinician with in-depth knowledge of these parameters of weaning failure (50). Moreover, earlier recognition of the patient's capacity for some level of autonomous respiration is fundamental to progressively initiating the weaning of the patient from MV and finally gaining full independent respiratory function (51). Thus, our study provides a new AI-enabled solution to realize the expectation.

Ideally, the clinical weaning parameters collected in critical care need to be objective and easy to acquire, and the process would not impede patient management. The physiological mechanisms resulting in respiratory failure vary for different individuals, and diverse weaning parameters will contribute to one aspect of the pathophysiological mechanism. It has been proved that it is insufficient to improve the outcomes of ventilated patients by applying the weaning index only (52). Our AI models, which incorporate a full range of patients' basic parameters, physiological parameters, and respiratory parameters, and consider the dynamic changes in time series, were used to establish a two-stage, multi-time series prediction model, which significantly improves the success in predicting weaning and conforms to clinical experience.

There have been several studies in the past that explored the prediction of related ICU respirator use with machine learning methods or traditional statistical methods (e.g., regression analysis method). Our research found that ML methods roughly outperformed traditional statistics. In the studies of ML method, we also obtained more excellent results. Our models are not only of high quality (AUC $>$0.94) but also the two-stage design is closer to clinical experience of weaning decision-making than a single-stage design. Cheng (53) also proposed a two-stage decision-making approach; however, our prediction model quality was more superior to theirs since we also subdivided each stage into 11 time points which helped to precisely grasp the timing of weaning MV and even shorten the intubation time. More importantly, among these studies, only our research realized the ML models in practice. We summarized the comparison of our study with previous works (53–56) in **Table 6**.

Furthermore, LightGBM models were noted with the highest AUC values amidst the seven ML algorithms, consistent with Chen et al. (57). Moreover, LightGBM has been regarded as the most effective model to predict extubation success when compared with XGBoost, MLP, and SVM. LightGBM is a gradient boosting framework of tree-based learning algorithms with faster training speed and better accuracy, but with lower memory usage.

Further analysis indicated that our models had convinced predictability regarding the Swets classification ($0.5 \leq$ AUC $\leq 0.7$, lower predicted; $0.7 \leq$ AUC $\leq 0.9$, certain predictive ability; AUC $> 0.9$, high predictive ability) (58). However, it was found that the models over time have a tendency of decreasing AUC (0.953∼0.864, lightGBM models in Stage 1; 0.943∼916, lightGBM models in Stage 2), which may imply that the longer the patient uses the ventilator, the more complicated it becomes when considering whether the patient can undergo try-weaning. This finding can also support why we use multiple periods instead of single period to predict weaning MV.

Our AI system could allow the clinicians to grasp the appropriate weaning time precisely, which could prevent the worthless dangers due to delayed or premature weaning process. Thus, with our AI system, the risks of complications and medical costs related to ventilatory support for patients are expected to decrease. More importantly, our AI system could lessen the effect of inter-clinician variability and improve the overall ICU care quality.

The deficiency of thorough evidence and the difference of results between individuals and subpopulations demonstrates there is scanty consensus on the issue of the best weaning protocol in clinical literature (59, 60). Our research results could provide useful solution to this long-standing clinical difficulty.

TABLE 6 A comparison with related studies.

| Study | Patient group | Predictive outcome | ML algorithm (* best algorithm) | Sample size | Numbers of features | Model's performance (the highest AUC) | Real world implementation |
|---|---|---|---|---|---|---|---|
| **This Study** | Adult ICU patients with invasive mechanical ventilation | (1). Timing of full support shifting to partial support modes<br>(2). Timing of weaning MV | Seven ML algorithms:<br>LR, RF, SVM, KNN, LGBM, XGB, MLP.<br>11 models were established in each of the two stages.<br>*The best algorithm: LGBM. | Stage 1: 5,873<br>Stage 2: 4,172 | Stage 1: 25<br>Stage 2: 20 | Stage 1: 0.843-0.953<br>Stage 2: 0.889-0.944 | Yes.<br>A predictive dashboard with best AI models was implemented and integrated into the existing HIS |
| (52) | Adult ICU patients with invasive mechanical ventilation | (1). The success shifting from full to partial support ventilation<br>(2). Successful SBT | Seven ML algorithms:<br>LR, Ridge Regression, Elastic Net, RF, SVM, ANN, XGB.<br>1 model was established in each of the two stages.<br>*The best algorithm: XGB and RF. | First model: 2,153<br>Second model: 3,132 | First model: 16<br>Second model: 12 | First model: 0.76<br>Second model: 0.79 | No |
| (53) | Cardiac Surgery patients with invasive mechanical ventilation | Successful weaned within 24 h | Six ML algorithms:<br>LR, RF, SVM, DT, ANN, XGB.<br>*The best algorithm: SVM. | 1,439 | 28 | 0.88 | No |
| (54) | Adult ICU patients with invasive mechanical ventilation | Successful extubation | Three ML algorithm:<br>RF, LGBM, XGB.<br>*The best algorithms: LGBM. | 117 (Total number of labeled was 12,268) | 57 | 0.950 | No |
| (55) | Adult ICU patients with invasive mechanical ventilation | Successful extubation | Six ML algorithms:<br>CNN, ANN, LR, SVM, DT, RF<br>*The best algorithm: CNN. | 2,299 | 25 | 0.94 | No |

MV, Mechanical Ventilation; LR, Logistic Regression; RF, Random Forest; SVM, Support Vector Machines; KNN, K Nearest Neighbor; LGBM, lightGBM; XGB: XGBoost; MLP, Multilayer Perception; CNN, Convolutional Neural Network, ANN, Artificial Neural Network; DT: Decision Tree.



10.3389/fmed.2022.935366

AI applications should be aptly weighed parallel to other information sources and certified by well-designed prospective studies before comprehensive implementation. Although we noticed a substantial and even almost perfect consistency in the prediction of successful weaning from ventilators after AI intervention, we position our AI system as an auxiliary, not as a determiner for diagnosis.

Clinical decision assistance systems could aid clinicians in their decision-making (61) and provide individualized management protocols based on the patients' clinical data and updated knowledge (62). Besides, AI is a powerful instrument that lowers the medical error rate and improves healthcare consistency and efficacy (63). However, there has been a lot of concern about the demerits of AI model applications in the decision of MV weaning. First, deep learning lacks explanatory power and related potential bias is hard to identify (64). Moreover, new ethical issues have been presented such as issues of erroneous decisions by AI, legal responsibility, and private information security crisis are taken into consideration (65).

There are limitations to our study. First, it is a single-center study, and we do not have an external cohort to validate our obtained models despite using data routinely collected in a real-world setting. Thus, extra care in terms of research generality must be given when extrapolating the findings to other centers. Second, some weaning-relevant data, like rehabilitation program arrangement, were not assessed in our dataset. We consider the model's accuracy could be improved significantly after assessing this detailed information. Third, our enrolled patient number was relatively small, impacting the result. Fourth, our study failed to include essential features and modalities, like chest X-ray images, cuff-leak test, diaphragm ultrasonography, and fluid balance, which are widely assessed to predict successful extubation. Further, no information related to laryngeal edema after extubation was trained in our models. Therefore, it could be difficult for the developed model to forecast the extubation failure rate due to post-extubation laryngeal edema.

## Conclusion

Weaning timing assessment in ICU patients with MV is one of the most critical steps for respiratory care teams. We employed AI technology to develop a comprehensive system and embedded it into the existing HIS to predict the timing of weaning MV; this proves the clinical innovation of AI intervention in critical care. According to our knowledge, such a study with valuable academic and practical implications is rare.

Most studies only report the quality of predictive models; thus, it may be difficult to judge its actual clinical value. Our study established a predictive model and validated the model in the clinical field, which proved that it has better benefits than traditional ones. Therefore, our study supports that AI could be a promising approach in predicting MV weaning timing in ICU and is expected to advance clinical research in this field.

Although we can see that the AI prediction dashboard we proposed can be an effective tool to assist weaning decision-making, it should be noted that it cannot be regarded as the only dependence for final decision-making. That is, after referring to the AI's prediction, the medical team still need to conduct and discuss a professional and comprehensive observation and evaluation of the patient again before making the final weaning decision.

Our study showed that the use of ML approaches could obtain better predictive ability in ICU, however, some physicians also reported that AI assistance is not very necessary. Thus, how to increase physicians' willingness to accept AI is indeed a key research topic. Besides, AI algorithms are difficult to understand (so-called black-box), which may affect the trust of clinical staff. Therefore, follow-up research to improve the explainability of AI must be done. Furthermore, intensivists expect that AI can be applied to build a decision support tool for integrated consideration of a patient rather than simply providing predictions on an illness. This is a challenge that should be taken seriously. However, we still have a long way to go at this moment.

## Data availability statement

The original contributions presented in the study are included in the article/**Supplementary material**, further inquiries can be directed to the corresponding author/s.

## Ethics statement

The studies involving human participants were reviewed and approved by Chi Mei Medical Center (IRB Serial No.: 10912-016). Written informed consent for participation was not required for this study in accordance with the national legislation and the institutional requirements.

## Author contributions

C-MChen, K-CC, C-FL, and C-CC: conceptualization. C-MH, J-JW, C-CL, and C-CC: data curation. C-FL, M-IS, C-JC, and C-CC: formal analysis. C-MH, C-MChen, and C-CC: investigation. C-FL, M-IS, and C-JC: methodology. C-CL and C-MChen: project administration. S-CK, K-CC, M-IS, S-CH, and C-MChen: resources. C-FL and C-JC: software. C-CC: supervision. C-MH, S-CK, C-MChao, and C-CC: validation. C-MChao and C-MChen: visualization. C-FL and C-CC:

writing—original draft. C-CC: writing—review and editing. All authors read and approved the submitted version.

## Funding

This research was partially supported by the Ministry of Science and Technology of Taiwan (No. MOST111-2410-H-384-001-MY2).

## Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

## Publisher's note



## Supplementary material

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fmed.2022.935366/full#supplementary-material

## References

1. Wunsch H, Linde-Zwirble WT, Angus DC, Hartman ME, Milbrandt EB, Kahn JM. The epidemiology of mechanical ventilation use in the United States. *Crit Care Med.* (2010) 38:1947–53. doi: 10.1097/CCM.0b013e3181ef4460

2. Eskandar N, Apostolakos MJ. Weaning from mechanical ventilation. *Crit Care Clin.* (2007) 23:263–74. doi: 10.1016/j.ccc.2006.12.002

3. Navalesi P, Bruni A, Garofalo E, Biamonte E, Longhini F, Frigerio P. Weaning off mechanical ventilation: much less an art, but not yet a science. *Ann Transl Med.* (2019) 7:S353. doi: 10.21037/atm.2019.09.83

4. Schweickert WD, Gehlbach BK, Pohlman AS, Hall JB, Kress JP. Daily interruption of sedative infusions and complications of critical illness in mechanically ventilated patients. *Crit Care Med.* (2004) 32:1272–6. doi: 10.1097/01.ccm.0000127263.54807.79

5. Tsai TL, Huang MH, Lee CY, Lai WW. Data science for extubation prediction and value of information in surgical intensive care unit. *J Clin Med.* (2019) 8:1709. doi: 10.3390/jcm8101709

6. Asehnoune K, Seguin P, Lasocki S, Roquilly A, Delater A, Gros A, et al. Extubation success prediction in a multicentric cohort of patients with severe brain injury. *Anesthesiology.* (2017) 127:338–46. doi: 10.1097/ALN.0000000000001725

7. Strickland JH Jr., Hasson JH. A computer-controlled ventilator weaning system. A clinical trial. *Chest.* (1993) 103:1220–6. doi: 10.1378/chest.103.4.1220

8. Ely EW, Baker AM, Dunagan DP, Burke HL, Smith AC, Kelly PT, et al. Effect on the duration of mechanical ventilation of identifying patients capable of breathing spontaneously. *N Engl J Med.* (1996) 335:1864–9. doi: 10.1056/NEJM199612193352502

9. Kollef MH, Shapiro SD, Silver P, St John RE, Prentice D, Sauer S, et al. A randomized, controlled trial of protocol-directed versus physician-directed weaning from mechanical ventilation. *Crit Care Med.* (1997) 25:567–74. doi: 10.1097/00003246-199704000-00004

10. Marelich GP, Murin S, Battistella F, Inciardi J, Vierra T, Roby M. Protocol weaning of mechanical ventilation in medical and surgical patients by respiratory care practitioners and nurses: effect on weaning time and incidence of ventilator-associated pneumonia. *Chest.* (2000) 118:459–67. doi: 10.1378/chest.118.2.459

11. Krishnan JA, Moore D, Robeson C, Rand CS, Fessler HE. A prospective, controlled trial of a protocol-based strategy to discontinue mechanical ventilation. *Am J Respir Crit Care Med.* (2004) 169:673–8. doi: 10.1164/rccm.200306-761OC

12. Namen AM, Ely EW, Tatter SB, Case LD, Lucia MA, Smith A, et al. Predictors of successful extubation in neurosurgical patients. *Am J Respir Crit Care Med.* (2001) 163:658–64. doi: 10.1164/ajrccm.163.3.2003060

13. Navalesi P, Frigerio P, Moretti MP, Sommariva M, Vesconi S, Baiardi P, et al. Rate of reintubation in mechanically ventilated neurosurgical and neurologic patients: evaluation of a systematic approach to weaning and extubation. *Crit Care Med.* (2008) 36:2986–92. doi: 10.1097/CCM.0b013e31818b35f2

14. Rose L, Presneill JJ, Johnston L, Cade JF. A randomised, controlled trial of conventional versus automated weaning from mechanical ventilation using SmartCare/PS. *Intensive Care Med.* (2008) 34:1788–95. doi: 10.1007/s00134-008-1179-4

15. Simeone F, Biagioli B, Scolletta S, Marullo ACM, Marchet-Ti L, Caciorgna M, et al. Optimization of mechanical ventilation support following cardiac surgery. *J Cardiovasc Surg.* (2002) 43:633–41.

16. Piotto RF, Maia LN, de Nassau Machado M, Orrico SP. Effects of the use of mechanical ventilation weaning protocol in the coronary care unit: randomized study. *Rev Bras Cir Cardiovasc.* (2011) 26:213–21. doi: 10.1590/s0102-76382011000200011

17. Béduneau G, Pham T, Schortgen F, Piquilloud L, Zogheib E, Jonas M, et al. Epidemiology of Weaning Outcome according to a new definition. The WIND study. *Am J Respir Crit Care Med.* (2017) 195:772–83. doi: 10.1164/rccm.201602-0320OC

18. Jeong BH, Ko MG, Nam J, Yoo H, Chung CR, Suh GY, et al. Differences in clinical outcomes according to weaning classifications in medical intensive care units. *PLoS One.* (2015) 10:e0122810. doi: 10.1371/journal.pone.0122810

19. Girard TD, Alhazzani W, Kress JP, Ouellette DR, Schmidt GA, Truwit JD, et al. An official American thoracic society/American college of chest physicians clinical practice guideline: liberation from mechanical ventilation in critically ill adults. Rehabilitation protocols, ventilator liberation protocols, and cuff leak tests. *Am J Respir Crit Care Med.* (2017) 195:120–33. doi: 10.1164/rccm.201610-2075ST

20. Epstein SK, Ciubotaru RL, Wong JB. Effect of failed extubation on the outcome of mechanical ventilation. *Chest.* (1997) 112:186–92. doi: 10.1378/chest.112.1.186

21. DiRusso SM, Sullivan T, Holly C, Cuff SN, Savino J. An artificial neural network as a model for prediction of survival in trauma patients: validation for a regional trauma area. *J Trauma.* (2000) 49:212–20. doi: 10.1097/00005373-200008000-00006

22. LeCun Y, Bengio Y, Hinton G. Deep learning. *Nature.* (2015) 521:436–44. doi: 10.1038/nature14539

23. Hsieh MH, Hsieh MJ, Chen CM, Hsieh CC, Chao CM, Lai CC. An artificial neural network model for predicting successful extubation in intensive care units. *J Clin Med.* (2018) 7:240. doi: 10.3390/jcm7090240

24. Tu JV. Advantages and disadvantages of using artificial neural networks versus logistic regression for predicting medical outcomes. *J Clin Epidemiol.* (1996) 49:1225–31. doi: 10.1016/s0895-4356(96)00002-9

25. Yang F, Wang HZ, Mi H, Lin CD, Cai WW. Using random forest for reliable classification and cost-sensitive learning for medical diagnosis. *BMC Bioinformatics.* (2009) 10:S22. doi: 10.1186/1471-2105-10-S1-S22

26. Furey TS, Cristianini N, Duffy N, Bednarski DW, Schummer M, Haussler D. Support vector machine classification and validation of cancer tissue samples using microarray expression data. *Bioinformatics.* (2000) 16:906–14. doi: 10.1093/bioinformatics/16.10.906

27. Videtta W, Vallejos J, Roda G, Collazos H, Naccarelli N, Tamayo A, et al. Predictors of Successful Extubation in Neurocritical Care Patients. *Acta Neurochir Suppl.* (2021) 131:91–3.

28. Baptistella AR, Mantelli LM, Matte L, Carvalho MEDRU, Fortunatti JA, Costa IZ, et al. Prediction of extubation outcome in mechanically ventilated patients: development and validation of the Extubation Predictive Score (ExPreS). *PLoS One.* (2021) 16:e0248868. doi: 10.1371/journal.pone.0248868

29. Chung WC, Sheu CC, Hung JY, Hsu TJ, Yang SH, Tsai JR. Novel mechanical ventilator weaning predictive model. *Kaohsiung J Med Sci.* (2020) 36:841–9.

30. Fathy S, Hasanin AM, Raafat M, Mostafa MMA, Fetouh AM, Elsayed M, et al. Thoracic fluid content: a novel parameter for predicting failed weaning from mechanical ventilation. *J Intensive Care.* (2020) 8:20. doi: 10.1186/s40560-020-00439-2

31. Formenti P, Umbrello M, Dres M, Chiumello D. Ultrasonographic assessment of parasternal intercostal muscles during mechanical ventilation. *Ann Intensive Care.* (2020) 10:120. doi: 10.1186/s13613-020-00735-y

32. Mesquida J, Gruartmoner G, Espinal C, Masip J, Sabatier C, Villagrá A, et al. Thenar oxygen saturation (StO(2)) alterations during a spontaneous breathing trial predict extubation failure. *Ann Intensive Care.* (2020) 10:54. doi: 10.1186/s13613-020-00670-y

33. Chawla S, Natarajan G, Shankaran S, Carper B, Brion LP, Keszler M, et al. Markers of successful extubation in extremely preterm infants, and morbidity after failed extubation. *J Pediatr.* (2017) 189:113–9.e2. doi: 10.1016/j.jpeds.2017.04.050

34. Manley BJ, Doyle LW, Owen LS, Davis PG. Extubating extremely preterm infants: predictors of success and outcomes following failure. *J Pediatr.* (2016) 173:45–9. doi: 10.1016/j.jpeds.2016.02.016

35. Muralitharan S, Nelson W, Di S, McGillion M, Devereaux PJ, Barr NG, et al. Machine learning-based early warning systems for clinical deterioration: systematic scoping review. *J Med Internet Res.* (2021) 23:e25187. doi: 10.2196/25187

36. Fleuren LM, Klausch TLT, Zwager CL, Schoonmade LJ, Guo T, Roggeveen LF, et al. Machine learning for the prediction of sepsis: a systematic review and meta-analysis of diagnostic test accuracy. *Intensive Care Med.* (2020) 46:383–400.

37. Parreco J, Hidalgo A, Kozol R, Namias N, Rattan R. Predicting mortality in the surgical intensive care unit using artificial intelligence and natural language processing of physician documentation. *Am Surg.* (2018) 84:1190–4.

38. Zheng L, Lin F, Zhu C, Liu G, Wu X, Wu Z, et al. Machine learning algorithms identify pathogen-specific biomarkers of clinical and metabolomic characteristics in septic patients with bacterial infections. *Biomed Res Int.* (2020) 2020:6950576. doi: 10.1155/2020/6950576

39. Mueller M, Almeida JS, Stanislaus R, Wagner CL. Can machine learning methods predict extubation outcome in premature infants as well as clinicians?. *J Neonatal Biol.* (2013) 2:1000118. doi: 10.4172/2167-0897.1000118

40. Mueller M, Wagner CC, Stanislaus R, Almeida JS. Machine learning to predict extubation outcome in premature infants. *Proc Int Jt Conf Neural Netw.* (2013) 2013:1–6. doi: 10.1109/IJCNN.2013.6707058

41. 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) 8:17116. doi: 10.1038/s41598-018-35582-2

42. Hsieh MH, Hsieh MJ, Cheng AC, Chen CM, Hsieh CC, Chao CM, et al. Predicting weaning difficulty for planned extubation patients with an artificial neural network. *Medicine.* (2019) 98:e17392. doi: 10.1097/MD.0000000000017392

43. Chang YJ, Hung KC, Wang LK, Yu CH, Chen CK, Tay HT, et al. A real-time artificial intelligence-assisted system to predict weaning from ventilator immediately after lung resection surgery. *Int J Environ Res Public Health.* (2021) 18:2713. doi: 10.3390/ijerph18052713

44. Lai CC, Chen CM, Chiang SR, Liu WL, Weng SF, Sung MI, et al. Establishing predictors for successfully planned endotracheal extubation. *Medicine.* (2016) 95:41. doi: 10.1097/MD.0000000000004852

45. Meade M, Guyatt G, Cook D, Griffith L, Sinuff T, Kergl C, et al. Predicting success in weaning from mechanical ventilation. *Chest.* (2001) 120:400S–24S.

46. Chawla NV, Bowyer KW, Hall LO, Kegelmeyer WP. SMOTE: synthetic minority over-sampling technique. *J Artif Int Res.* (2002) 16:321–57.

47. Rose L, Adhikari NK, Leasa D, Fergusson DA, McKim D. Cough augmentation techniques for extubation or weaning critically ill patients from mechanical ventilation. *Cochrane Database Syst Rev.* (2017) 1:CD011833. doi: 10.1002/14651858.CD011833.pub2

48. Viera A, Garrett J. Understanding interobserver agreement: the kappa statistic. *Fam Med.* (2005) 37:360–3.

49. Lovejoy CA, Buch V, Maruthappu M. Artificial intelligence in the intensive care unit. *Crit Care.* (2019) 23:7. doi: 10.1186/s13054-018-2301-9

50. Heunks LM, van der Hoeven JG. Clinical review: the ABC of weaning failure–a structured approach. *Crit Care.* (2010) 14:245. doi: 10.1186/cc9296

51. Kwong MT, Colopy GW, Weber AM, Ercole A, Bergmann JHM. The efficacy and effectiveness of machine learning for weaning in mechanically ventilated patients at the intensive care unit: a systematic review. *Biodes Manuf.* (2019) 2:31–40. doi: 10.1186/s13054-016-1208-6

52. Huo Y, Guo S, Zhang K, Zhang T, Li B, Zhang Q, et al. A clinical study on the ability of the integrative weaning index to predict weaning from mechanical ventilation. *Ann Palliat Med.* (2020) 9:3162–9. doi: 10.21037/apm-20-1335

53. Cheng KH, Tan MC, Chang YJ, Lin CW, Lin YH, Chang TM, et al. The feasibility of a machine learning approach in predicting successful ventilator mode shifting for adult patients in the medical intensive care unit. *Medicina.* (2022) 58:360. doi: 10.3390/medicina58030360

54. Chen WT, Huang HL, Ko PS, Su W, Kao CC, Su SL. A simple algorithm using ventilator parameters to predict successfully rapid weaning program in cardiac intensive care unit patients. *J Pers Med.* (2022) 12:501. doi: 10.3390/jpm12030501

55. Otaguro T, Tanaka H, Igarashi Y, Tagami T, Masuno T, Yokobori S, et al. Machine learning for prediction of successful extubation of mechanical ventilated patients in an intensive care unit: a retrospective observational study. *J Nippon Med Sch.* (2021) 88:408–17. doi: 10.1272/jnms.JNMS.2021_88-508

56. Jia Y, Kaul C, Lawton T, Murray-Smith R, Habli I. Prediction of weaning from mechanical ventilation using convolutional neural networks. *Artif Intell Med.* (2021) 117:102087. doi: 10.1016/j.artmed.2021.102087

57. Chen T, Xu J, Ying H, Chen X, Feng R, Fang X, et al. Prediction of extubation failure for intensive care unit patients using light gradient boosting machine. *IEEE Access.* (2019) 7:150960–68. doi: 10.1186/s13040-022-00309-7

58. Swets JA. Measuring the accuracy of diagnostic systems. *Science.* (1988) 240:1285–93. doi: 10.1126/science.3287615

59. Conti G, Mantz J, Longrois D, Tonner P. Sedation and weaning from mechanical ventilation: time for 'best practice' to catch up with new realities?. *Multidiscip Respir Med.* (2014) 9:45. doi: 10.1186/2049-6958-9-45

60. Goldstone J. The pulmonary physician in critical care. 10: difficult weaning. *Thorax.* (2002) 57:986–91. doi: 10.1136/thorax.57.11.986

61. Montani S, Striani M. Artificial intelligence in clinical decision support: a focused literature survey. *Yearb Med Inform.* (2019) 28:120–7.

62. Reddy S, Fox J, Purohit MP. Artificial intelligence-enabled healthcare delivery. *J R Soc Med.* (2019) 112:22–8.

63. Hung CM, Shi HY, Lee PH, Chang CS, Rau KM, Lee HM, et al. Potential and role of artificial intelligence in current medical healthcare. *Artif Intell Cancer.* (2022) 3:1–10.

64. Tarassenko L, Watkinson P. Artificial intelligence in health care: enabling informed care. *Lancet.* (2018) 391:1260. doi: 10.1016/S0140-6736(18)30701-3

65. Gerke S, Minssen T, Cohen G. Chapter 12 - Ethical and legal challenges of artificial intelligence-driven healthcare. *Artif Intell Healthc.* (2020) 2020:295–336.

# Frontiers in Medicine

**Translating medical research and innovation into improved patient care**

A multidisciplinary journal which advances our medical knowledge. It supports the translation of scientific advances into new therapies and diagnostic tools that will improve patient care.

## Discover the latest Research Topics

See more →


**Frontiers**

Avenue du Tribunal-Fédéral 34
1005 Lausanne, Switzerland
frontiersin.org

**Contact us**

+41 (0)21 510 17 00
frontiersin.org/about/contact