frontiers | Research Topics

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

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

Published in
Frontiers in Medicine
Frontiers in Immunology
Frontiers in Artificial Intelligence

ISSN 1664-8714
ISBN 978-2-8325-4337-5
DOI 10.3389/978-2-8325-4337-5

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 IV

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. (2024). *Clinical application of artificial intelligence in emergency and critical care medicine, volume IV*. Lausanne: Frontiers Media SA. doi: 10.3389/978-2-8325-4337-5

Table of contents

frontiers | Frontiers in Medicine

TYPE Editorial
PUBLISHED 09 January 2024
DOI 10.3389/fmed.2023.1346070

OPEN ACCESS

EDITED BY
Gulzar H. Shah,
Georgia Southern University, United States

REVIEWED BY
Athanasios Chalkias,
University of Pennsylvania, United States

*CORRESPONDENCE
Rahul Kashyap
✉ kashyapmd@gmail.com

RECEIVED 28 November 2023
ACCEPTED 22 December 2023
PUBLISHED 09 January 2024

CITATION
Dhillon G, Zhang Z, Grewal H and Kashyap R (2024) Editorial: Clinical application of artificial intelligence in emergency and critical care medicine, volume IV. *Front. Med.* 10:1346070. doi: 10.3389/fmed.2023.1346070

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

Gagandeep Dhillon[1], Zhongheng Zhang[2], Harpreet Grewal[3] and Rahul Kashyap[4,5]*

[1]Department of Internal Medicine, University of Maryland Baltimore Washington Medical Center, Glen Burnie, MD, United States, [2]Department of Emergency Medicine, Sir Run Run Shaw Hospital, Zhejiang University School of Medicine, Hangzhou, China, [3]Department of Radiology, Florida State University College of Medicine, Pensacola, FL, United States, [4]Division of Pulmonary and Critical Care Medicine, Mayo Clinic, Rochester, MN, United States, [5]Department of Research, WellSpan Health, York, PA, United States

KEYWORDS

artificial intelligence, prediction, critical care, machine learning, intensive care unit

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

Integrating artificial intelligence (AI) into the realm of emergency and critical care medicine marks a transformative stage in healthcare delivery. In Volume IV of the research compilation titled "*Clinical application of artificial intelligence in emergency and critical care medicine*," a collection of 16 articles highlights the burgeoning intersection between advanced technology and acute medical interventions. This compilation looks at a spectrum of innovative applications, ranging from diagnostic support systems to predictive analytics, all poised to reshape the dynamics of emergency medical response and critical patient care.

Guiot et al. presented an interesting retrospective study on COVID-19 patients. Over the last few years, there have been millions of COVID-19 cases with many deaths (1). To help manage the load on radiologists, an artificial intelligence (AI) based analysis (CACOVID-CT) was implemented to evaluate the severity of the disease on the basis of CT chests performed on those patients. Progress in machine learning and artificial intelligence has led to the creation of tools that can augment the diagnostic skills of radiologists (2). The area of the lung affected by COVID-19 Affected Area (%AA) and CT severity score (total CT-SS) were quantified to help evaluate outcome and prognosis. It is interesting to note that both %AA and CT-SS had a high correlation with length of stay, risk for invasive ventilation, ICU admission, and death during hospital stay. It alleviated the workload of radiologists by measuring the severity of lung damage.

As the pandemic continued to grow, there was an increased number of COVID-19 patients with acute respiratory distress syndrome (ARDS) in the ICU (3). However, there was limited information about predictive studies of ARDS in those patients. Zhou et al. attempted to create predictive models to establish a correlation between ARDS and COVID-19. One hundred three critically ill COVID patients were included in the study, and the development of ARDS in patients admitted to ICU was the primary outcome. Based on convolutional neural network (CNN) and extreme gradient boosting (XGBoost), two

predictive models were established. Out of 104, 23 (22.3%) of patients developed ARDS. In critically ill COVID-19 patients, an integrated deep-learning model can be helpful to predict ARDS.

As the healthcare industry embarks on a paradigm shift toward a more data-driven and technologically enhanced future, this volume is a comprehensive exploration of AI's profound impact in optimizing clinical decision-making, resource allocation, and patient outcomes within the high-stakes environments of emergency and critical care settings. Saqib et al. carried out a comprehensive search in PubMed, Google Scholar, PLOS One, and Scopus to develop an understanding of AI applications in critical illness in a narrative review. They concluded that it is vital to ensure that AI systems are made robust and reliable in the care of critically ill patients. Also, there should be transparent and comprehensible reasoning behind recommendations generated by AI. Quality control measures must be in place to ensure safety and effectiveness.

Nonvariceal upper gastrointestinal bleeding (NVUGIB) in patients with decompensated cirrhosis can be critically ill and has been associated with a higher rate of readmissions and mortality (4). For patients with NVUGIB, Ungureanu et al. developed an artificial neural network with mortality as the primary outcome. Over 1000 NVUGIB patients hospitalized were divided into training and testing groups in this retrospective study. Glasgow Blatchford (GBS), AIM65, and admission Rockall (Rock) are non-endoscopic risk scores used in the past. In the study, four machine learning algorithms, Quadratic Discriminant Analysis (QDA), logistic regression (LR), Linear Discriminant Analysis (LDA), and K-Nearest Neighbor (K-NN) were used with GBS, Rock, AIM65, and others. It was noted that the machine learning models had more accuracy in identifying patients with a higher mortality risk than the current risk scores. An accuracy of 98% was seen with K-NN classifier, proving that there is scope for using machine learning in NVUGIB patients to predict mortality.

In hospitals, length of stay (LOS) indicates the efficiency of management (5). Zeleke et al. attempted to compare and develop various models to estimate LOS and prolonged LOS in patients admitted through the emergency room. Six algorithms (Random Forest (RF), Support Vector Machines (SVM), Gradient Boosting (GB), AdaBoost, K-Nearest Neighbors (KNN), and logistic regression (LR) were used, and they analyzed a total of 12,858 patients. Out of them, 61% had a prolonged LOS. The models were evaluated using the Brier score with the area under the curve, sensitivity, accuracy, precision, specificity, and F1 score.GB algorithm best predicted the accuracy of prolonged LOS, and there was tremendous potential seen in the machine learning-based methods to assess for LOS. They also give insights to help understand the risks behind increased LOS. If combined with provider expertise, they can be used to make informed decisions.

In conclusion, examining "*Clinical application of artificial intelligence in emergency and critical care medicine*," Volume IV, has provided a nuanced insight into the transformative potential of artificial intelligence within the critical domains of emergency and critical care medicine. It encapsulates a diverse array of AI applications, ranging from real-time diagnostics to prognostic modeling, each contributing to an evolving landscape where technology complements and enhances the capabilities of healthcare professionals. Although most topics have been covered in the collection of articles, we should be cognizant of the fact that technological advancement with AI does bring to light an acute need to address ethical considerations. Healthcare industry needs to protect the values of medicine and follow fundamental ethical principles. The future calls for papers for this special topic may consider including it. As we move forward, a collaborative effort between clinicians, technologists, and policymakers is crucial to harness the full potential of artificial intelligence for improving patient care in these critical settings.

Author contributions

GD: Conceptualization, Methodology, Visualization, Writing – original draft. ZZ: Conceptualization, Supervision, Writing – review & editing. HG: Conceptualization, Methodology, Visualization, Writing – review & editing. RK: Conceptualization, Project administration, Resources, Supervision, Writing – review & editing.

Funding

The author(s) declare that financial support was received for the research, authorship, and/or publication of this article. ZZ received funding from the China National Key Research and Development Program (Nos. 2023YFC3603104 and 2022YFC2504500), the Huadong Medicine Joint Funds of the Zhejiang Provincial Natural Science Foundation of China under Grant No. LHDMD24H150001, the Fundamental Research Funds for the Central Universities (226-2022-00148), National Natural Science Foundation of China (82272180), and the Project of Drug Clinical Evaluate Research of Chinese Pharmaceutical Association No. CPA-Z06-ZC-2021-004.

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.

The author(s) declared that they were an editorial board member of Frontiers, at the time of submission. This had no impact on the peer review process and the final decision.

Publisher's note

References

1. Naik R, Avula S, Palleti SK, Gummadi J, Ramachandran R, Chandramohan D, et al. From emergence to endemicity: a comprehensive review of COVID-19. *Cureus*. (2023) 15:e48046. doi: 10.7759/cureus.48046

2. Grewal H, Dhillon G, Monga V, Sharma P, Buddhavarapu VS, Sidhu G, et al. Radiology gets chatty: the ChatGPT saga unfolds. *Cureus*. (2023) 15:e40135. doi: 10.7759/cureus.40135

3. Fan E, Beitler JR, Brochard L, Calfee CS, Ferguson ND, Slutsky AS, et al. COVID-19-associated acute respiratory distress syndrome: is a different approach to management warranted? *Lancet Respir Med*. (2020) 8:816–21. doi: 10.1016/S2213-2600(20)30304-0

4. Kruger AJ, Abougergi MS, Jalil S, Sobotka LA, Wellner MR, Porter KM, et al. Outcomes of nonvariceal upper gastrointestinal bleeding in patients with cirrhosis: a national analysis. *J Clin Gastroenterol*. (2023) 57:848–53. doi: 10.1097/MCG.0000000000001746

5. Baek H, Cho M, Kim S, Hwang H, Song M, Yoo S. Analysis of length of hospital stay using electronic health records: a statistical and data mining approach. *PLoS ONE*. (2018) 13:e0195901. doi: 10.1371/journal.pone.0195901

frontiers | Frontiers in Medicine

TYPE Original Research
PUBLISHED 29 August 2022
DOI 10.3389/fmed.2022.930055

OPEN ACCESS

EDITED BY
Jin Woo Song,
Asan Medical Center, South Korea

REVIEWED BY
Jung Hwa Hwang,
Soonchunhyang University Seoul Hospital, South Korea
Ju Hyun Oh,
Inje University Sanggye Paik Hospital, South Korea

*CORRESPONDENCE
Julien Guiot
J.Guiot@chuliege.be

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

SPECIALTY SECTION
This article was submitted to Pulmonary Medicine, a section of the journal Frontiers in Medicine

RECEIVED 27 April 2022
ACCEPTED 25 July 2022
PUBLISHED 29 August 2022

CITATION
Guiot J, Maes N, Winandy M, Henket M, Ernst B, Thys M, Frix A-N, Morimont P, Rousseau A-F, Canivet P, Louis R, Misset B, Meunier P, Charbonnier J-P and Lambermont B (2022) Automatized lung disease quantification in patients with COVID-19 as a predictive tool to assess hospitalization severity. *Front. Med.* 9:930055. doi: 10.3389/fmed.2022.930055

Automatized lung disease quantification in patients with COVID-19 as a predictive tool to assess hospitalization severity

Julien Guiot[1]*, Nathalie Maes[2], Marie Winandy[1], Monique Henket[1], Benoit Ernst[1], Marie Thys[2], Anne-Noelle Frix[1], Philippe Morimont[3], Anne-Françoise Rousseau[3], Perrine Canivet[4], Renaud Louis[1], Benoît Misset[3], Paul Meunier[4], Jean-Paul Charbonnier[5†] and Bernard Lambermont[3†]

[1]Respiratory Department, University Hospital of Liège, Liège, Belgium, [2]Biostatistics and Medico-Economic Information Department, University Hospital of Liège, Liège, Belgium, [3]Intensive Care Department, University Hospital of Liège, Liège, Belgium, [4]Department of Radiology, University Hospital of Liège, Liège, Belgium, [5]Thirona B.v., Nijmegen, Netherlands

The pandemic of COVID-19 led to a dramatic situation in hospitals, where staff had to deal with a huge number of patients in respiratory distress. To alleviate the workload of radiologists, we implemented an artificial intelligence (AI) - based analysis named CACOVID-CT, to automatically assess disease severity on chest CT scans obtained from those patients. We retrospectively studied CT scans obtained from 476 patients admitted at the University Hospital of Liege with a COVID-19 disease. We quantified the percentage of COVID-19 affected lung area (% AA) and the CT severity score (total CT-SS). These quantitative measurements were used to investigate the overall prognosis and patient outcome: hospital length of stay (LOS), ICU admission, ICU LOS, mechanical ventilation, and in-hospital death. Both CT-SS and % AA were highly correlated with the hospital LOS, the risk of ICU admission, the risk of mechanical ventilation and the risk of in-hospital death. Thus, CAD4COVID-CT analysis proved to be a useful tool in detecting patients with higher hospitalization severity risk. It will help for management of the patients flow. The software measured the extent of lung damage with great efficiency, thus relieving the workload of radiologists.

KEYWORDS

SARS-CoV-2, CT scan analysis, artificial intelligence, mechanical ventilation risk, severity of hospital stay prediction, COVID-19, in-hospital death, ICU length of stay

Introduction

The rapid outbreak of coronavirus disease 2019 (COVID-19), originating from severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) infection, has become a public health emergency of international concern (1). During the first wave, a high proportion of infected patients required hospitalization in the intensive care unit (ICU) (2).

Since the onset of the COVID-19 pandemic, chest CT imaging has been widely used to help clinicians in the identification of patients infected with COVID-19 (3, 4). CT scans capture imaging features from the lung associated with COVID-19 since the earliest stages of the disease. CT scan could, thus, serve as an efficient and effective way to diagnose, and possibly prognosis patients with COVID-19 admitted to the hospital.

Visual assessment of disease severity by CT scoring usually includes ground-glass opacity, consolidation, air bronchogram, crazy paving, nodular opacities, and pleural effusion (5). This method is time-consuming, requires experienced radiologists, and could be error-prone when the workload is heavy. CT scoring methods may produce different severity levels, while some effort has been made to get a common lexicon and scoring (6, 7).

However, visual scoring of CT scans is known to be susceptible to high inter-reader variability (8, 9) and allows only for qualitative or semiquantitative assessment of the parenchymal involvement of the disease, even if a major effort has been made to standardize the description of CTs used for diagnostic purposes (6). Furthermore, in a context of a high burden on healthcare institutes during the COVID-19 pandemic, visual scoring of many CT scans could be highly challenging for radiologists (10).

The Dutch Radiological Society developed the coronavirus disease 2019 (COVID-19) Reporting and Data System (CO-RADS) as a categorical assessment scheme for pulmonary involvement of COVID-19 at unenhanced chest CT (9, 11–13). This method performs well in predicting COVID-19 in patients with moderate-to-severe symptoms and has a substantial interobserver agreement. Thus, it is helpful in COVID-19 diagnosis and the evaluation of disease severity and prognosis (14).

The CO-RADS artificial intelligence (AI) system consists of three deep learning algorithms that automatically segment the five pulmonary lobes, assign the CO-RADS score for the suspicion of COVID-19, and assign a CT severity score for the degree of parenchymal involvement per lobe. This algorithm proved to be in accordance with diagnoses obtained from experienced radiologists (10), exhibiting a high performance for diagnosis and disease prognosis.

Artificial intelligence-based models can aid the radiologist in assessing CT scans, providing rapid and quantitative information on disease-related parenchymal involvement (15). These models may help to provide precise and reproducible quantitative information on lung parenchyma affected by COVID-19, while relieving some of the burden on healthcare professionals (9, 11–13). CT scan analysis with deep learning methods even allowed for the diagnosis of COVID-19 disease earlier than reverse transcriptase-PCR (RT-PCR) (8).

Besides the chest imaging effectiveness in COVID-19 diagnosis, multiple prognosis models have been developed, without making it possible to establish a strong predictive model of the clinical evolution (3).

Computed tomography scores have been combined with other clinical or biological parameters, either directly related to lung function, or inflammation or infection [C-reactive protein (CRP), D-dimer, alkaline phosphatase, etc.,] (16–19) with a good correlation with disease severity and death risk. However, we must note a lack of consistency, as the considered laboratory data differ between studies.

One of the key questions when caring for hospitalized patients with COVID-19 infection remains to determine the risk of deterioration leading to ICU admission. It is even more important to predict the need for mechanical ventilation, given the limited number of ventilators and the need for specialized staff to monitor closely these patients.

In this study, CT quantification of COVID-19-related parenchymal abnormalities was performed using CAD4COVID-CT (Thirona, Nijmegen, Netherlands). CAD4COVID-CT is a CE (0344) class IIa certified AI-based software package that automatically quantifies the lobar extent of COVID-19 using state of the art deep learning techniques. The algorithm provides a quantitative assessment of the categorical CT severity score (CT-SS) such as the CO-RADS severity scoring system, and, in addition, quantifies the volume percentage of COVID-19-related affected areas (% AA) on the lobar level.

The main goal of this retrospective study was to explore the prognostic value of CAD4COVID-CT severity scores on hospitalization severity indicators of patients with COVID-19.

Materials and methods

Study design and participants

During the COVID-19 pandemic, our hospital expanded its total ICU capacity from 58 to 68 beds during wave 1 (W1) (from 10 March to 22 June 2020) and 71 beds during wave 2 (W2) (from 31 August to 12 October 2020), with 10–12 beds dedicated to non-COVID-19 critically ill patients. In the whole hospital, 196 beds were dedicated to patients with COVID-19 during the two waves, out of a total of 878 beds.

All the adult patients admitted to the University Hospital of Liege for acute respiratory failure related to SARS-CoV-2 pneumonia between 13 March 2020 and 18 April 2021 were

included, if they had undergone a chest CT scan at most 1 day before or after hospital admission. Patients primarily hospitalized for scheduled or urgent surgery with positive SARS-CoV-2 PCR were excluded.

Patients were diagnosed with a positive PCR for SARS-CoV-2 in nasal swabs or other respiratory samples during the 5 days of their admission to the hospital or the 14 days before admission. When performed in our hospital, the detection of SARS-CoV-2 was performed by reverse transcription PCR using the Cobas SARS-CoV-2 Assay (Roche, Switzerland) for the detection of the *ORF1ab* and *E* genes. The results were reported as cycle thresholds to have a semiquantitative measurement of the viral load.

Chest CT scans at maximum 1 day before or after hospital admission were considered for image analysis at admission. For some patients, we obtained multiple CT scan data. For these patients, we compared the data collected during hospitalization with those obtained after hospital discharge (1–7 CT scans/patient).

Imaging

All the CT images used in the study were acquired on one of our five multidetector CT scanners [Siemens Edge Plus (2), GE Revolution CT (1), and GE Brightspeed (2)]. Since CT images were collected retrospectively, no standardized scan protocol was available over the complete dataset.

All the acquired CT scans were analyzed using the CE 0344 certified Class IIa medical device CAD4COVID-CT (Thirona, Nijmegen, Netherlands). This AI-based software package analyses the lungs and each of the individual lobes for automatic quantification of COVID-19-related pulmonary parenchymal involvement. The software uses state of the art deep learning and image normalization techniques to provide robust and repeatable quantitative information in CT scans acquired with varying scanner parameters, typically found in a clinical setting. The analysis starts by identifying the lungs and each of the pulmonary lobes to provide their volumes. Within each of these areas of interest, emphysematous areas and COVID-19-related abnormality areas are identified and quantified. Emphysema is a pathological situation that worsens hypoxemia. Since hypoxemia is a critical factor in determining admission to the intensive care unit, it is important to distinguish whether the hypoxemia observed is due to emphysema or viral pneumonia.

This information is presented as the volume percentage of emphysema and volume percentage of the affected area for the whole lung and each of the pulmonary lobes. For each lobe, the percentage of the affected area (% AA) is used to calculate a severity score per lobe.

This lobar CT-SS is a categorization of the percentage of affected area defined as: 0 (affected area: 0%), 1 (affected area: 0.1–5.0%), 2 (affected area: 5.1–25.0%), 3 (affected area: 25.1–50.0%, 4 (affected area: 50.1–75.0%), and 5 (affected area: over 75.0%).

The total CT-SS is the accumulation of each of the individual lobar scores. Two examples of CAD4COVID-CT report with all the quantitative information and an example of a coronal section of the CT scan are shown in **Figure 1**.

Statistical methods

Quantitative variables were expressed by mean and SD, or median and P25 and P75 quartiles. Qualitative variables are presented using frequency tables (number and percentage).

The univariate linear regression models were used to study the relationship between the patient's characteristics and CT-severity score (total CT-SS) and% affected area (% AA) at hospital admission.

The impact of the CAD4COVID-CT scores on the hospital length of stay (LOS) and the ICU length of stay was studied using the multiple linear regression models on the log-transformed lengths of stay. Results were presented as adjusted estimated coefficients ± SEs and p-values. The impact of the CAD4COVID-CT scores on the risk of ICU stay, the risk of ventilation, and the risk of in-hospital death was studied using the multiple logistic regression models. Results were expressed as adjusted odds ratios (ORs) and 95% CI. All the multiple models were adjusted for age, gender, BMI, and wave.

Optimal total CT-SS cutoff points to predict ICU admission and the need for mechanical ventilation were calculated on the data of wave 2 using Youden's index method. The predictive models based on the data of W2 were recommended for optimal cutoff point estimation, since the risk of the considered outcomes had changed between W1 and W2. As the number of cases increased significantly during W2, it was also imperative to adapt the ICU admission procedure, to preserve resources for the most severe cases.

Generalized linear mixed models (GLMMs) were used to analyze the evolution of the CAD4COVID-CT scores with time during and after the hospital stay.

The results were considered significant at the 5% uncertainty level ($p < 0.05$). Statistical analyses were performed on all the available data and the missing data were not replaced. Calculations were done using SAS software (version 9.4) and graphics with R software (version 3.6.1).

Results

Patients characteristics

A total of 476 patients with COVID-19 hospitalized at the CHU Hospital of Liège between 13 March 2020 and 18 April

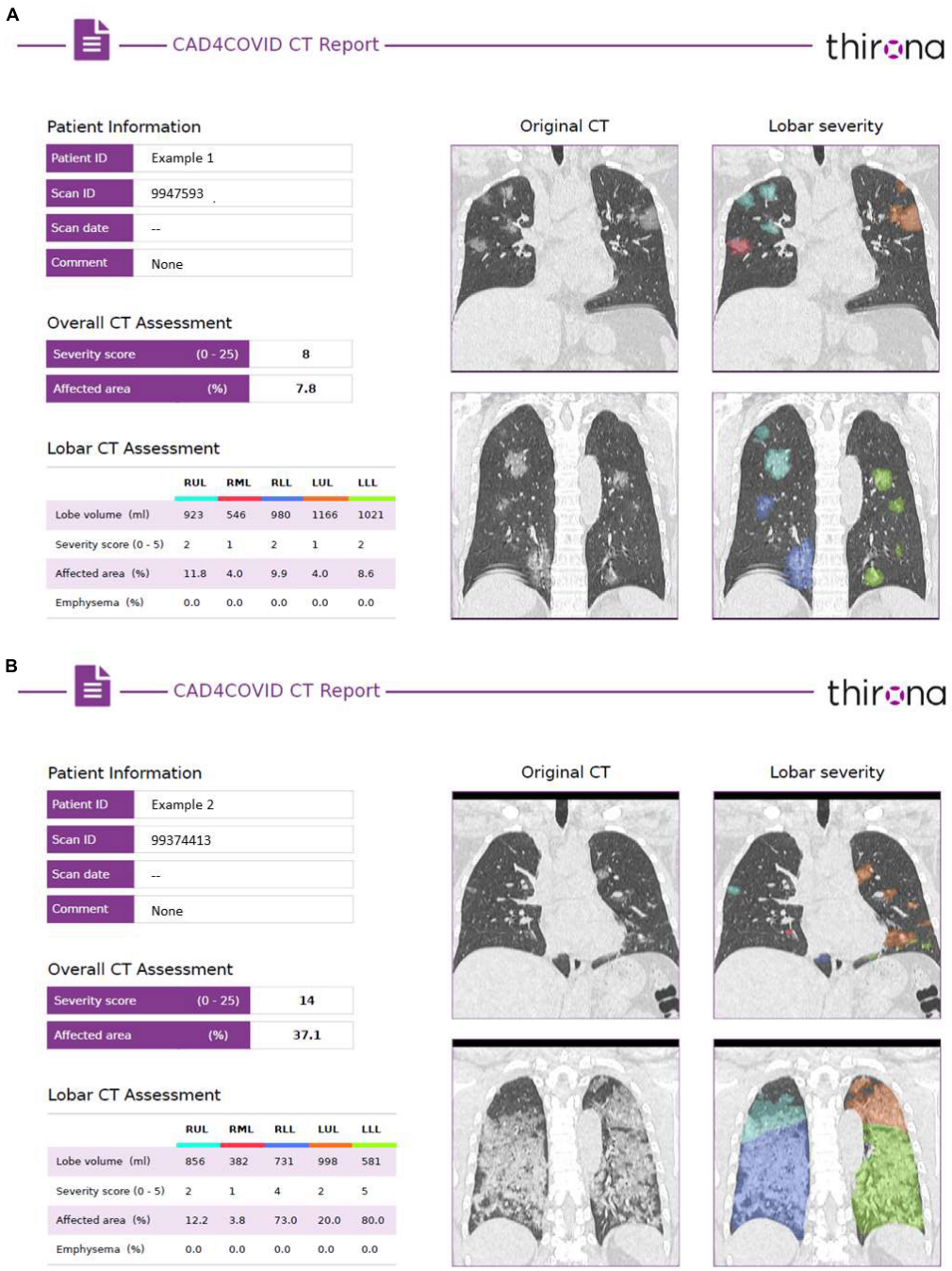

FIGURE 1
Example of the CAD4COVID-CT reports of two patients with COVID-19. **(A)** The left side of each report provides the quantitative assessment, including the lobar volume, the total and lobar CT severity scores and% affected area, and the lobar emphysema scores. **(B)** The right side of each report shows two coronal sections with a color-coated overlay of the identified affected areas, where each color represents a different lobe matching the colors indicated in the lobar CT assessment table.

2021 were included in the study. Patients' characteristics are given in Table 1.

Within our cohort, we identified that 37.1% of them were suffering from obesity, defined as a body mass index (BMI) above 30 kg/m^2.

The relationship between the characteristics of the patients and total CT-SS and% AA at hospital admission was calculated. They were significantly correlated with BMI and obesity ($p < 0.001$). We observed that older patients had lower% AA ($p < 0.0001$) and CT-SS ($p < 0.001$) at admission.

TABLE 1 Patients' characteristics.

Characteristics	Results Mean ± SD, Median (P25 – P75), or N (%)
Age (years)	67.3 ± 14.7
Gender, Men	311 (65.3)
BMI (kg/m^2)	28.5 ± 6.3
Chronic kidney disease	52 (14.2)
Diabetes	204 (48.7)
Arterial hypertension	293 (64.5)
Cardio-vascular disease	138 (38.3)
Chronic respiratory disease	114 (27.8)
Immunosuppressive therapy	22 (6.0)
Obesity	153 (37.1)
Oncological condition	63 (13.2)
Wave 1	229 (48.1)
Inter Wave 1-Wave 2	8 (1.7)
Wave 2	226 (47.5)
Wave 3	13 (2.7)
Hospital LOS (days)	11 (7 – 19)
ICU	240 (50.4)
Time between hospital admission and ICU admission (days)	2 (1 – 3)
ICU LOS (days)	8 (4 – 18)
Mechanical Ventilation	134 (28.1)
Dialysis	22 (4.6)
In-hospital death	144 (30.2)

Patients with the oncological conditions had lower% AA ($p < 0.01$) and CT-SS ($p < 0.05$) at admission. Neither CT-SS at admission nor% AA was related to gender (**Supplementary Data**, **Tables 1**, **2**).

We also analyzed the relationship between BMI, age, wave, and the severity of hospitalization indicators (**Supplementary Data**, **Table 3**). Patients with higher BMI had higher LOS ($p < 0.05$), higher risk of ICU admission ($p < 0.05$), higher ICU LOS ($p < 0.05$), and higher risk of mechanical ventilation ($p < 0.05$).

Hospital and ICU LOS were lower when diagnosis occurred after wave 1 ($p < 0.05$ and $p < 0.0001$, respectively).

Older patients have a higher risk of ICU admission ($p < 0.01$) and a higher risk of death during their hospital stay ($p < 0.0001$).

Gender was not related to the risk of hospitalization severity in this study for none of the severity parameters measured.

CT Scans at admission: CAD4COVID-CT analysis

The 476 CT scans at a maximum 1 day before or after hospital admission were analyzed. A corresponding severity score was assigned to each scan (**Table 2**), depending on% AA.

We also analyzed CT-SS and% AA for each lobe. We observed that the most affected lobes were the lower ones for both the analyses (**Supplementary Table 4**).

Relationship between CAD4COVID-CT analysis and hospitalization severity indicators

An increased total CT-SS at admission, as well as the% AA, was closely associated with a higher risk of prolonged hospital LOS, ICU admission, mechanical ventilation, or in-hospital death. Of note, there was no specific correlation with ICI LOS (**Table 3**).

When we considered the scores obtained from the individual lobes, we observed that the highest order of% AA and CT-SS was reached in the left and right lower lobes. However, the association with patients' outcomes was in the same range of statistical significance for LOS, risk of ICU admission, ICU LOS, and risk of mechanical ventilation, except for the risk of in-hospital death, where the p-values were > 0.05 (**Supplementary Table 5**).

Specificity and sensitivity of CAD4COVID-CT as a predictive tool of intensive care unit admission and mechanical ventilation risks

The predictive models based on the data of wave 2 (W2) were recommended for optimal cutoff point estimation since the risk of the considered outcomes had changed between waves 1 (W1) and W2. As the number of cases increased significantly

TABLE 2 CT scans analyses by Thirona (n = 476).

	Mean ± SD	Median (p25-p75)	Extremes
Volume (mL)	3411 ± 1184	3386 (2637 – 4105)	1340; 8578
Emphysema score (%)	0.74 ± 2.8	0.024 (0.0001 – 0.22)	0.00; 31.7
% AA	26.1 ± 22.4	19.0 (6.3 – 42.2)	0.00; 84.1
Total CT-SS	11.4 ± 6.0	11 (7 – 16)	0; 25

10.3389/fmed.2022.930055

TABLE 3 Relationship between CT scan analysis and hospitalization severity.

	Length of stay		Risk of ICU admission		ICU length of stay		Risk of mechanical ventilation		Risk of in-hospital death	
	Coef ± SE	*P*-value	OR (95% CI)	*P*-value	Coef ± SE	*P*-value	OR (95% CI)	*P*-value	OR (95% CI)	*P*-value
CT-SS	0.025 ± 0.0071	<0.001	1.3 [1.2; 1.3]	<0.0001	0.017 ± 0.012	0.16	1.2 [1.1; 1.2]	<0.0001	1.0 [1.002; 1.1]	<0.05
% AA	0.0060 ± 0.0019	<0.01	1.1 [1.1; 1.1]	<0.0001	0.0036 ± 0.0029	0.22	1.0 [1.03; 1.05]	<0.0001	1.0 [1.002; 1.02]	<0.05

All the models were adjusted for age, gender, BMI, and wave.

TABLE 4 The CT-SS cutoff points (wave 2, n = 226).

	ICU admission	Mechanical ventilation
AUC (95%CI)	0.84 (0.79; 0.90)	0.71 (0.63; 0.78)
Optimal cut-off point	14	16

during W2, it was also imperative to adapt the ICU admission procedure, to preserve resources for the most severe cases.

We calculated Youden's index to maximize specificity and sensitivity, for the CT-SS for the 226 patients who experienced COVID-19 during W2 (Table 4). Results showed a CT-SS cutoff of 14 to predict the risk of ICU admission and 16 for mechanical ventilation. Figure 2 shows the area under the curve (AUC) (95% CI) and optimal cutoff determination for CT-SS to predict the risk of ICU admission (A) and risk of mechanical ventilation (B).

External validation should be necessary for cutoff validation. Even if the W1 group was not representative of the current situation of the patients (most of the outcomes were improved between W1 and W2) to test the cutoff values, we nevertheless applied the cutoffs to the W1 group as internal validation (n patients = 229).

The CT-SS cutoff of 14 to predict the risk of ICU admission led to a sensitivity of 87% (95% CI: 81 to 92%) and a specificity of 58% (95% CI: 47 to 68%); the CT-SS cutoff of 16 to predict the risk of mechanical ventilation led to a sensitivity of 88% (95% CI: 83 to 93%) and a specificity of 48% (95%CI: 35 to 60%). These results are very similar to those obtained with W2 patients.

Computed tomography scan evolution over time

Of the 476 patients, 84 patients had repeated chest CT scans during hospitalization and follow-up period. Figures 3, 4 show the evolution of total CT-SS and% AA after hospital discharge (GLMM model). Value considered at hospital discharge (time = 0 in the figure) was the maximum CT-SS or% AA during the hospital stay (n = 20 patients with at least one CT scan after hospital discharge).

We observed that both the measured values are decreasing after discharge, with a significant evolution over time ($p < 0.001$ and $p < 0.01$, respectively).

Discussion

In our study, we showed that CAD4COVID-CT was able to help in the risk stratification of patients suffering from acute COVID-19 infection, based on chest CT images. Both the total CT-SS and% AA were able to predict hospital LOS, ICU admission risk, risk of mechanical ventilation, and in-hospital

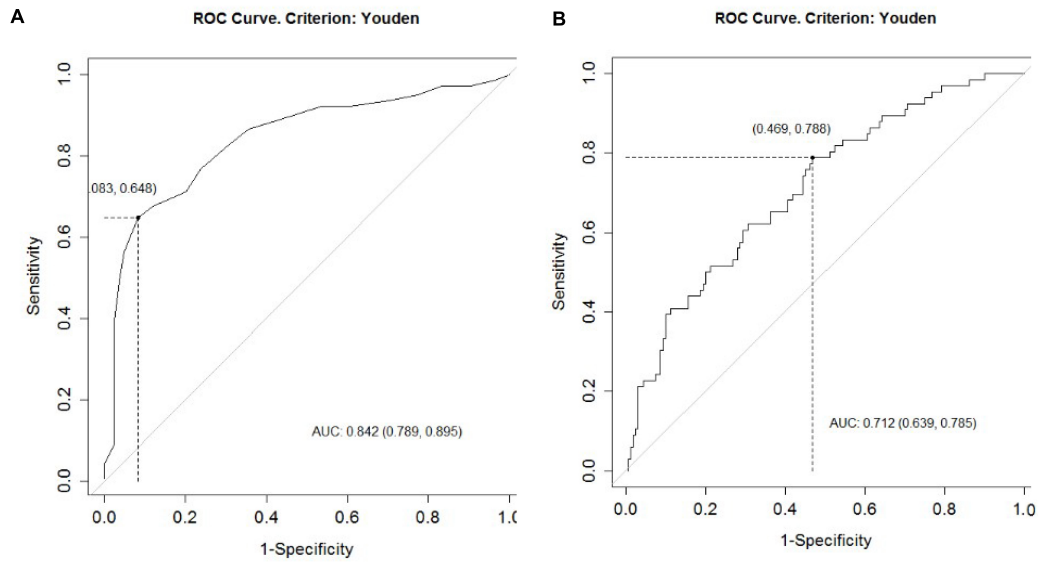

FIGURE 2
The ROC curve predicts ICU admission based on the initial CAD4COVID-CT evaluation. The AUC (95% CI) and optimal cutoff determination for the CT-SS to predict **(A)** risk of ICU admission and **(B)** risk of mechanical ventilation.

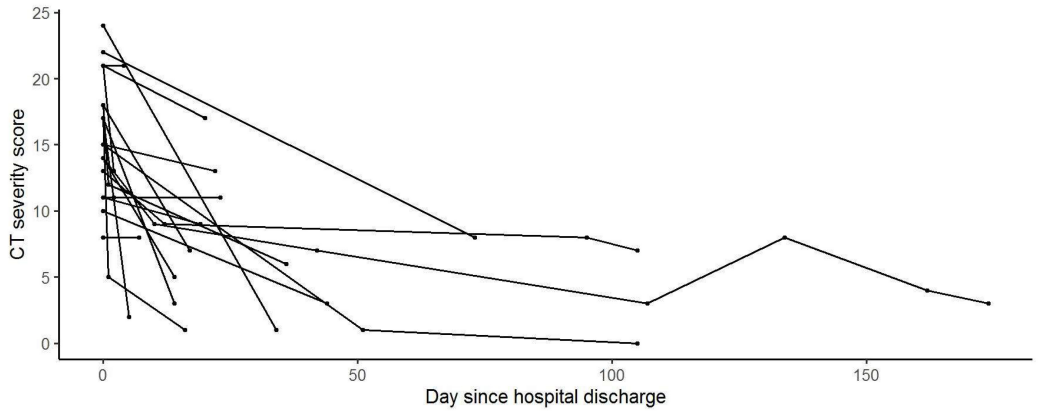

FIGURE 3
CT-SS evolution after hospital discharge. Generalized linear mixed models (GLMMS) were used to analyze the evolution of the CT's Thirona scores with time during and after the hospital stay.

death. This means that AI can be used with great efficiency to predict the risk of worsening the patient's condition, thus allowing for better management of patient flow in the hospital.

Our analysis showed that this CT scan image analysis tool can be of interest to better stratify the risk of ICU for patients acutely infected with COVID-19. This approach can help in the global management of patients in an in-hospital setup.

We also analyzed the relationship between patients' characteristics and the% AA or total CT-SS. Patients with higher BMI had higher% AA and CT-SS at admission, and the severity indicators were all significantly related to this condition. The mean BMI of our cohort (28.5) was higher than in the whole Belgian population (being 25.5) (20). About half of the Belgian population is overweight, making it one of the top five factors associated with mortality in our country. The Belgian adult population has an obesity rate of 15.9% (20), while our cohort showed a 37.1% obesity rate. This is consistent with much-published data associating age and BMI with the risk of severe COVID-19 (2, 21–24).

In a counterintuitive way, we observed that older patients had lower% AA and CT-SS at admission, while they died more frequently. This could be explained by the comorbidities found in this population, which can increase the global risk of experiencing complications associated with COVID-19 infection. Similarly, patients with oncological conditions had lower% AA and CT-SS at admission,

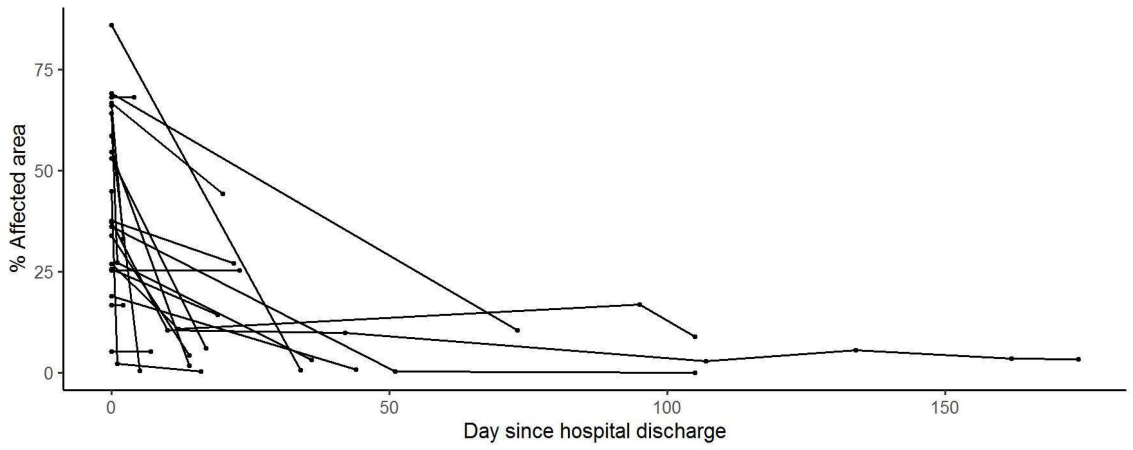

FIGURE 4
% Affected area evolution after hospital discharge. Generalized linear mixed models (GLMM) were used to analyze the evolution of the CT's Thirona scores with time during and after the hospital stay.

which can also be explained by non-COVID-19-related associated complications.

We showed that hospital LOS, ICU admission risk, and ICU LOS were lower when diagnosis occurred after W1. At this time, more effective care had been setup in our hospital, including early treatment with dexamethasone and remdesivir, as well as high-flow nasal oxygen therapy. This resulted in a decrease in ICU admission risk, ICU length of stay, and mechanical ventilation risk, among other indicators (25).

The secondary aim of this study was to determine whether CAD4COVID-CT was able to help in disease monitoring and follow-up. Some patients benefited from a long-term follow-up, with several CT scans after discharge from the hospital (5, 26). For those patients, we showed that the total CT-SS and% AA decreased over time. This finding was consecutively strongly correlated with health improvement as expected (27). Importantly, chest CT should not be considered as a follow-up tool for COVID-19 infection due to radiation-associated risk and the lack of rationale for systematic follow-up.

CAD4COVID-CT provides two main AI-based scores: (1) the percentage of the affected area (% AA) and (2) the categorical CT severity score derived from the affected area (total CT-SS). Although these two scores are highly related to each other, both provide a specific value for the CT assessment. The% AA is the most precise measure, which is calculated at the voxel level. This allows for an exact delineation and quantification of the percentage of affected lung tissue. The CT-SS score is derived from the% AA similar to the CO-RADS, in which certain cutoff points are used to make severity categories. Evidently, by categorizing a continuous variable, precision information is lost. However, categorical scores allow for a direct comparison to a visual assessment, as it mimics how a human would score disease severity, where true quantification for humans is virtually impossible. Therefore, using AI for precise quantification of COVID-19-affected lung areas gives reliable results, and a good correlation with disease severity, while scoring may reassure clinicians and radiologists, who are used to visually assigning severity scores when reviewing CT scans (13). An additional advantage of using AI-based quantification is in statistical analysis and risk stratification. Moreover, an AI-based algorithm provides consistent and objective output, while allowing avoiding potential discrepancies in inter- and intraobserver variability.

An important aspect of CAD4COVID-CT is that the entire analysis was designed to handle the considerable amount of CT scan variability, typically encountered in clinical practice and especially during the COVID-19 pandemic. Different sources of CT scan variability (such as differences in scanner manufacture, reconstruction kernels, and dose levels) can have a substantial impact on the quantitative score if not properly mitigated during algorithm development. This may lead to poor clinical correlations and conflicting longitudinal assessments. The design of CAD4COVID-CT allows the algorithm to deal with CT scan variability in two main ways. First, the AI-based algorithms were trained with a well-balanced set of CTs coming from various sources. This allowed the algorithms to learn from CT scan variations that they would likely encounter in a clinical setting. Second, CT scan normalization techniques are used to standardize each CT before the scan is presented to the AI algorithms. This procedure greatly reduces the inherent variability between scans and allows the AI algorithms to provide consistent results in a clinical setting.

A limitation of the CAD4COVID-CT analysis is the lack of separation of different textures within the identified affected areas. The AI-based algorithm was trained to identify COVID-19-related abnormalities as a single class, meaning that both the ground-glass opacities and consolidations are combined into the% AA and CT-SS scores. Although this

approach follows the severity scoring of the CO-RADS, separating ground-glass opacities and consolidation could provide additional clinically relevant information on disease severity and prognosis. Furthermore, the current quantification is lobar-based, allowing a quantitative assessment of the lobar disease distribution. With only the lobar information, quantification of the ventral vs. dorsal disease distribution is not possible, while this information may be clinically relevant in patients that require mechanical ventilation. However, since CAD4COVID-CT is identifying the abnormalities on a voxel level, and the relationship between the voxels and the lung and lobar boundaries is known, this information could be extracted and added as an additional feature.

Another limitation of this study comes from the fact that this is a retrospective study performed in a single center. It should also be validated on external data.

Conclusion

In conclusion, our study showed that the CAD4COVID-CT AI-based quantification of lung injury in COVID-19 infection was highly correlated with major clinical indicators and helped to predict ICU admission and the risk of mechanical ventilation. This method can be used as a clinical decision support system for patients' triage, to better manage the intrahospital flow, and to guide the indicated therapy promptly. Further clinical studies for validation are needed to confirm the added value of this model depending on the variant modification over time.

Data availability statement

The datasets generated or analyzed during this current study are not publicly available because these data are considered sensitive, but are available from the corresponding author on reasonable request.

Ethics statement

The studies involving human participants were reviewed and approved by the Ethics Committee of the University Hospital of Liege (Comité d'éthique hospitalo-universitaire de Liège) reviewed the study and approved it (Reference: 2022/21). Written informed consent for participation was not required for this study in accordance with the national legislation and the institutional requirements.

Author contributions

JG, RL, and BL conceived the study. RL, BM, and BL designed the experiments. MH and NM analyzed the results. JG, MT, A-NF, PM, A-FR, PC, and BM conducted the experiments and acquired the data. BE acquired the funding. J-PC provided the Thirona software. JG, NM, MW, A-FR, J-PC, and BL wrote the manuscript. All authors have read and approved the final version of the manuscript.

Funding

We acknowledge financial support from the following European Union's research and innovation programs. The DRAGON project has received funding from the Innovative Medicines Initiative 2 Joint Undertaking (JU) under grant agreement No. 101005122. The JU receives support from the European Union's Horizon 2020 research and innovation program and EFPIA. The iCOVID project has received funding from the European Union's Horizon 2020 research and innovation program under grant agreement No. 101016131.

Acknowledgments

We thank the study participants and the hospital staff for their participation in this study.

Conflict of interest

J-PC was employed by Thirona. JG reports personal fees for advisory board, work and lectures from Boehringer Ingelheim, Janssens, GSK, Roche, and Chiesi, non-financial support for meeting attendance from Chiesi, Roche, Boehringer Ingelheim, and Janssens. He is in the permanent SAB of Radiomics (Oncoradiomics SA) for the SALMON trial without any specific consultancy fee for this study. He is co-inventor of one issued patent on radiomics licensed to radiomics (Oncoradiomics SA). He confirms that none of the above entities of funding was involved in the study design, collection, analysis, interpretation of data, the writing of this article or the decision to submit it for publication.

The remaining authors declare that this study was conducted in the absence of any commercial or financial relationships that could be construed as the potential conflicts of interest.

Publisher's note

claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

Author disclaimer

This publication reflects the author's view. Neither IMI nor the European Union, EFPIA, or the DRAGON and iCOVID consortia is responsible for any use that may be made of the information contained therein.

Supplementary material

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fmed.2022.930055/full#supplementary-material

References

1. Hope M, Raptis CA, Shah A, Hammer MA, Henry TS. A role for CT in COVID-19? What data tell us really so far. *Lancet.* (2020) 395:1189–90. doi: 10.1016/S0140-6736(20)30728-5

2. Covid-Icu Group on behalf of the Reva Network and the Covid-Icu Investigators. Clinical characteristics and day-90 outcomes of 4244 critically ill adults with COVID-19: A prospective cohort study. *Intensive Care Med.* (2021) 47:60–73. doi: 10.1007/s00134-020-06294-x

3. Guiot J, Vaidyanathan A, Deprez L, Zerka F, Danthine D, Frix AN, et al. Development and validation of an automated radiomic CT signature for detecting COVID-19. *Diagnostics.* (2021) 11:41. doi: 10.3390/diagnostics11010041

4. Walsh SLF, Calandriello L, Silva M, Sverzellati N. Deep learning for classifying fibrotic lung disease on high-resolution computed tomography: A case-cohort study. *Lancet Respir Med.* (2018) 6:837–45. doi: 10.1016/S2213-2600(18)30286-8

5. Liu X, Zhou H, Zhou Y, Wu X, Zhao Y, Lu Y, et al. Temporal radiographic changes in COVID-19 patients: Relationship to disease severity and viral clearance. *Sci Rep.* (2020) 10:10263. doi: 10.1038/s41598-020-66895-w

6. Salehi S, Abedi A, Balakrishnan S, Gholamrezanezhad A. Coronavirus disease 2019 (COVID-19) imaging reporting and data system (COVID-RADS) and common lexicon: A proposal based on the imaging data of 37 studies. *Eur Radiol.* (2020) 30:4930–42. doi: 10.1007/s00330-020-06863-0

7. de Smet K, de Smet D, Ryckaert T, Laridon E, Heremans B, Vandenbulcke R, et al. Diagnostic performance of chest CT for SARS-CoV-2 infection in individuals with or without COVID-19 symptoms. *Radiology.* (2020) 298:E30–7.

8. Kardos AS, Simon J, Nardocci C, Szabó IV, Nagy N, Abdelrahman RH, et al. The diagnostic performance of deep-learning-based CT severity score to identify COVID-19 pneumonia. *Br J Radiol.* (2022) 95:20210759.

9. Bai HX, Hsieh B, Xiong Z, Halsey K, Choi JW, Tran TML, et al. Performance of radiologists in differentiating COVID-19 from Non-COVID-19 viral pneumonia at chest CT. *Radiology.* (2020) 296:E46–54.

10. Lessmann N, Sánchez CI, Beenen L, Boulogne LH, Brink M, Calli E, et al. Automated assessment of CO-RADS and chest CT severity scores in patients with suspected COVID-19 using artificial intelligence. *Radiology.* (2021) 298:E18–28. doi: 10.1148/radiol.2020202439

11. Lambin P, Rios-Velazquez E, Leijenaar R, Carvalho S, van Stiphout RGPM, Granton P, et al. Radiomics: Extracting more information from medical images using advanced feature analysis. *Eur J Cancer.* (2012) 48:441–6.

12. Kanne JP, Bai H, Bernheim A, Chung M, Haramati LB, Kallmes DF, et al. COVID-19 Imaging: What we know now and what remains unknown. *Radiology.* (2021) 299:E262–79. doi: 10.1148/radiol.2021204522

13. Xiong Z, Wang R, Bai HX, Halsey K, Mei J, Li YH, et al. Artificial intelligence augmentation of radiologist performance in distinguishing COVID-19 from pneumonia of other origin at chest CT. *Radiology.* (2020) 296:E156–65.

14. Prokop M, van Everdingen W, van Rees Vellinga T, van Ufford HQ, Stöger L, Beenen L, et al. CO-RADS: A categorical CT assessment scheme for patients suspected of having COVID-19-definition and evaluation. *Radiology.* (2020) 296:E97–104. doi: 10.1148/radiol.2020201473

15. Frix AN, Cousin F, Refaee T, Bottari F, Vaidyanathan A, Desir C, et al. Radiomics in lung diseases imaging: State-of-the-art for clinicians. *J Pers Med.* (2021) 11:602.

16. Francone M, Iafrate F, Masci GM, Coco S, Cilia F, Manganaro L, et al. Chest CT score in COVID-19 patients: Correlation with disease severity and short-term prognosis. *Eur Radiol.* (2020) 30:6808.

17. Sharifpour A, Safanavaei S, Tabaripour R, Taghizadeh F, Nakhaei M, Abadi A, et al. Alkaline phosphatase and score of HRCT as indicators for predicting the severity of COVID-19. *Ann Med Surg.* (2021) 67:102519. doi: 10.1016/j.amsu.2021.102519

18. Huespe I, Carboni Bisso I, di Stefano S, Terrasa S, Gemelli NA, las Heras M. COVID-19 severity index: A predictive score for hospitalized patients. *Med Intensiva.* (2022) 46:98–101.

19. Bennouar S, Bachir Cherif A, Kessira A, Bennouar DE, Abdi S. Development and validation of a laboratory risk score for the early prediction of COVID-19 severity and in-hospital mortality. *Intensive Crit Care Nurs.* (2021) 64:103012.

20. sciensano.be. *Overweight and obesity in Belgium – Numbers.* (2022). Available online at: https://www.sciensano.be/en/health-topics/obesity/numbers#overweight-and-obesity-in-belgium (accessed March 31, 2022)

21. Smati S, Tramunt B, Wargny M, Seret-Bégué D, Winiszewski P, Matthieu P, et al. Relationship between obesity and severe COVID-19 outcomes in patients with type 2 diabetes: Results from the CORONADO study. *Diabetes Obes Metab.* (2020) 23:391–403. doi: 10.1111/dom.14228

22. Drucker DJ. Diabetes, obesity, metabolism, and SARS-CoV-2 infection: The end of the beginning. *Cell Metab.* (2021) 33:479. doi: 10.1016/j.cmet.2021.01.016

23. Zhou F, Yu T, Du R, Fan G, Liu Y, Liu Z, et al. Clinical course and risk factors for mortality of adult inpatients with COVID-19 in Wuhan, China: A retrospective cohort study. *Lancet.* (2020) 395:1054–62. doi: 10.1016/S0140-6736(20)30566-3

24. Gao M, Piernas C, Astbury NM, Hippisley-Cox J, O'Rahilly S, Aveyard P, et al. Associations between body-mass index and COVID-19 severity in 6·9 million people in England: A prospective, community-based, cohort study. *Lancet Diabetes Endocrinol.* (2021) 9:350–9. doi: 10.1016/S2213-8587(21)00089-9

25. Lambermont B, Rousseau AF, Seidel L, Thys M, Cavalleri J, Delanaye P, et al. outcome improvement between the first two Waves of the coronavirus disease 2019 pandemic in a single tertiary-care hospital in Belgium. *Crit Care Explor.* (2021) 3:e0438. doi: 10.1097/CCE.0000000000000438

26. Darcis G, Bouquegneau A, Maes N, Thys M, Henket M, Labye F, et al. Long-term clinical follow-up of patients suffering from moderate-to-severe COVID-19 infection: A monocentric prospective observational cohort study. *Int J Infect Dis.* (2021) 109:209–16. doi: 10.1016/j.ijid.2021.07.016

27. Ojha V, Mani A, Pandey NN, Sharma S, Kumar S. CT in coronavirus disease 2019 (COVID-19): A systematic review of chest CT findings in 4410 adult patients. *Eur Radiol.* (2020) 30:6129–38. doi: 10.1007/s00330-020-06975-7

frontiers | Frontiers in Medicine

TYPE Original Research
PUBLISHED 17 February 2023
DOI 10.3389/fmed.2023.1134835

OPEN ACCESS

EDITED BY
Qinghe Meng,
State University of New York Upstate Medical University, United States

REVIEWED BY
Hasan Maulahela,
University of Indonesia, Indonesia
Ali Taha,
University Hospital Crosshouse, United Kingdom

*CORRESPONDENCE
Adina Turcu-Stiolica
✉ adina.turcu@umfcv.ro

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

RECEIVED 30 December 2022
ACCEPTED 06 February 2023
PUBLISHED 17 February 2023

CITATION
Ungureanu BS, Gheonea DI, Florescu DN, Iordache S, Cazacu SM, Iovanescu VF, Rogoveanu I and Turcu-Stiolica A (2023) Predicting mortality in patients with nonvariceal upper gastrointestinal bleeding using machine-learning.
Front. Med. 10:1134835.
doi: 10.3389/fmed.2023.1134835

Predicting mortality in patients with nonvariceal upper gastrointestinal bleeding using machine-learning

Bogdan Silviu Ungureanu[1], Dan Ionut Gheonea[1], Dan Nicolae Florescu[1], Sevastita Iordache[1], Sergiu Marian Cazacu[1], Vlad Florin Iovanescu[1], Ion Rogoveanu[1] and Adina Turcu-Stiolica[2]*

[1]Department of Gastroenterology, University of Medicine and Pharmacy of Craiova, Craiova, Romania, [2]Department of Pharmacoeconomics, University of Medicine and Pharmacy of Craiova, Craiova, Romania

Background: Non-endoscopic risk scores, Glasgow Blatchford (GBS) and admission Rockall (Rock), are limited by poor specificity. The aim of this study was to develop an Artificial Neural Network (ANN) for the non-endoscopic triage of nonvariceal upper gastrointestinal bleeding (NVUGIB), with mortality as a primary outcome.

Methods: Four machine learning algorithms, namely, Linear Discriminant Analysis (LDA), Quadratic Discriminant Analysis (QDA), logistic regression (LR), K-Nearest Neighbor (K-NN), were performed with GBS, Rock, Beylor Bleeding score (BBS), AIM65, and T-score.

Results: A total of 1,096 NVUGIB hospitalized in the Gastroenterology Department of the County Clinical Emergency Hospital of Craiova, Romania, randomly divided into training and testing groups, were included retrospectively in our study. The machine learning models were more accurate at identifying patients who met the endpoint of mortality than any of the existing risk scores. AIM65 was the most important score in the detection of whether a NVUGIB would die or not, whereas BBS had no influence on this. Also, the greater AIM65 and GBS, and the lower Rock and T-score, the higher mortality will be.

Conclusion: The best accuracy was obtained by the hyperparameter-tuned K-NN classifier (98%), giving the highest precision and recall on the training and testing datasets among all developed models, showing that machine learning can accurately predict mortality in patients with NVUGIB.

KEYWORDS

UGIB, Rockall score, Beylor Bleeding score, machine learning, Glasgow Blatchford score

Introduction

Upper gastrointestinal bleeding (UGIB) still represents a common cause of gastroenterological admission and usually requires risk stratification for the level of care determination as well as rapid decision management (1). In order to differentiate the high-risk groups of patients in the emergency department, multiple guidelines developed pre-endoscopic risk assessment scores which combine both clinical features and biological parameters (2). Both the American Society of Gastroenterology (ASGE) and the European Society of Gastrointestinal Endoscopy (ESGE) recommend Glasgow-Blatchford (GBS), Rockall admission score (Rock), and AIM65 as possible tools to assess UGIB patients on their first presentation (3, 4). However, some studies suggested that the most accurate score for patient risk differentiation is GBS with multiple outcomes such as necessary transfusions, endoscopic reintervention, and death (5, 6).

Postponing endoscopy is also recommended whenever low-risk patients are identified, however, delaying endoscopy could also lead to dramatic consequences if patient selections are not done well. Several studies have proposed a hierarchy of patients who present with UGIB by defining patients that might be delayed till endoscopy (7, 8). While many centers, still need to reschedule endoscopy until the next morning or over the weekend, new methods should be proposed for a better discerning of patient's evolution.

Artificial intelligence in gastroenterology is on continuous path-breaking development, especially on imaging recognition patterns with already proposed techniques for daily practice (9, 10). The term AI covers machine learning (ML) and specific techniques such as deep learning (DL) by using data sets for pattern recognition by combining several variables which will further allow transposing new data that uses the same variables. Available clinical models for UGIB allow patients' features and predictors to suggest the prognostic. By involving an artificial neural network (ANN), the data trained to determine the desired outcome may be used to predict the output on input data of newly identified cases that may be encountered. Thus, by doing a repetitive learning technique, the ANN will be able to foretell the outcomes of the patient's prognosis.

The development of new models of patient triage and follow-up should be promoted to reduce medical exposure, thus managing possible complications. Moreover, by using ANN the results might be even more effective since the human factor is bypassed. The patient's prognosis presented with non-variceal UGIB (NVUGIB) should be assessed as early as possible in order to determine the proper timing of endoscopy. The aim of our study was to provide a new ANN that sums up all available pre-endoscopic risk scores for patients with UGIB for predicting mortality, thus promoting patients for new endoscopic procedures or even surgery.

Abbreviations: GBS, Glasgow Blatchford; ANN, artificial neural network; QDA, quadratic discriminant analysis; LR, logistic regression; K-NN, K-nearest neighbor; Rock, Rockall; BBS, Beylor Bleeding score; UGIB, upper gastrointestinal bleeding; NVUGIB, non-variceal upper gastrointestinal bleeding; ASGE, American Society of Gastroenterology; ESGE, European Society of Gastrointestinal Endoscopy; DL, deep learning; ML, machine learning; LDA, linear discriminant analysis; SD, standard deviation; AUC, area under the curve; TN, true negative; FP, false positive; TP, true positive; FN, false negative.

Materials and methods

Patients

The Ethics Committee of the University of Medicine and Pharmacy of Craiova, Romania approved this retrospective study and informed consent from all patients were acquired in the County Hospital before patient enrolment in the study (11977/24.03.2020). We selected 1,096 patients who were admitted for UGIB from March 2018 to December 2021within the Gastroenterology Department of the Emergency County Hospital of Craiova, Romania. The selection was based on the criteria: (1) patients with NVUGIB, (2) age $\geq$ 18 years old, (3) existing information as mortality, GBS, Rock, Beylor Bleeding score (BBS), AIM65, and T-score. Furthermore, the following exclusion criteria were considered: (1) patients with variceal UGIB, (2) patients with any type of cancer, (3) patients with important missing data (for example, data for calculating the scores).

Machine learning analysis framework

We adopted multiple machine-learning (ML) models, including Linear Discriminant Analysis (LDA), Quadratic Discriminant Analysis (QDA), logistic regression (LR), K-Nearest Neighbor (KNN). We tried GridSearch, RandomSearch as model tuning techniques to see if that improves the model's final performance. Confusion matrix was used to check model performance with or without standardization and we recorded accuracy, precision, recall, and f1-score. These classifiers were compared in terms of predicting the likelihood of mortality. The data was split into 70% train and 30% test sets, using the stratified

TABLE 1 Socio-demographic and clinical characteristics of the study subjects.

Characteristics	Frequency in the population study (N = 1096)*
Age (years), mean ± SD, range	63.9 ± 14.6, 17–92
Gender, male	738 (67.34%)
Urban residence	530 (48.36%)
Hospital days	8 ± 7.2
Mortality	82 (7.48%)
Rebleeding	32 (2.92%)
Surgery	11 (1%)
Hematemesis	472 (43.07%)
Platelets (no/mcL)	211,096.1 ± 97,621.5
Creatinine (mg/dL)	1.2 ± 1.3
Cirrhosis, yes	121 (11.04%)
Comorbidities	
Cardiovascular diseases	65 (5.93%)
Chronic kidney diseases	34 (3.1%)

*Continuous variables are expressed in mean ± SD and discrete variables are expressed in frequency and percentages.

sampling technique to ensure that relative class frequencies are approximately preserved in each train and validation fold. We used descriptive statistics to summarize the patients' characteristics: counts (percentages) for categorical variables and mean ± standard deviation (SD) for continuous variables.

The models predicted whether a NVUGIB patient would experience mortality by learning a number of five clinical scoring systems: GBS, Rock, BBS, AIM65, and T-score. Covariance matrix was introduced in the equation to consider the variation among the independent variables (GBS, Rock, BBS, AIM65, T-score). The ROC curve (receiver operating characteristic curve) and the area under this curve (AUC) for every single scoring system were used to quantify the visual profile of the ability of a model that includes only one score.

The confusion matrix shows clockwise from top left: True Negative (TN, model predicts that a NVUGIB patient would live and the patient does not die), False Positive (FP, model predicts that a NVUGIB patient would die but the patient actually does not die), True Positive (TP, model predicts that a NVUGIB patient would die and the patient dies) and False Negative (FN, model predicts that a NVUGIB patient would live but the patient actually dies). The recall [the fraction of total actually positive cases that are predicted correct = TP/(TP+FN)] will predict the need for intervention without high westing of hospital resources. It is preferred to use recall because the healthcare system cannot afford to make false-negative errors. The greater the Recall, the higher the chances of minimizing FN. Precision is the fraction of total positive predictions that are actually correct [TP/(TP+FP)]. F1-score is used when both precision and recall seem to be important.

Linear Discriminant Analysis draws one hyperplane and projects the data onto this hyperplane in such a way as to maximize the separation of the patients who died, according to two criteria:

TABLE 2 Statistical characteristics of the two groups divided by the mortality in the study population.

Characteristics	Total (*n* = 1096)	Mortality		*p*-value
		No (*n* = 1014)	Yes (*n* = 82)	
Glasgow Blatchford	9.9 ± 3.6 10 (8–12)	9.76 ± 3.56 10 (8–12)	12.26 ± 3.2 12 (10–14.25)	<0.001
Rockall score	3.7 ± 1.9 4 (2–5)	3.64 ± 1.89 4 (2–5)	4.34 ± 1.74 4 (4–5)	0.001
Beylor Bleeding score	7.6 ± 4.1 8 (4–11)	7.5 ± 4.14 8 (3.75–11)	8.32 ± 3.93 8.5 (6–11)	0.099
AIM65	1.1 ± 0.9 1 (1–1)	1.02 ± 0.83 1 (0-1)	1.8 ± 1.05 1 (1–2.25)	<0.001
T-score	9.3 ± 2.0 9 (8–10)	9.34 ± 2.01 9 (8–11)	8.85 ± 1.91 9 (7–9.25)	0.024

Data are presented as mean ± SD and median (interquartile range).

FIGURE 1
The correlation heatmap between measured scores (colors range from dark blue for strong positive correlations, to bright yellow, for strong negative correlations).

maximizing the distance between the means of the two classes and minimizing the variation between each category (11).

Quadratic Discriminant Analysis is a probabilistic parametric classification technique that represents an evolution of LDA for nonlinear class separations. QDA, like LDA, is based on the hypothesis that the probability density distributions are multivariate normal but, in this case, the dispersion is not the same for all of the categories (12).

Logistic regression is a supervised learning algorithm where we used the sigmoid function to calculate the probability of dying given the five scores, using also Lasso regularization (13).

K-Nearest Neighbor is a non-parametric algorithm, it does not make any assumption on underlying data. Because all the variables are continuous, we can apply LDA, assuming normality assumption for P(X|Y = 1) and P(X|Y = 0), and homoscedasticity (the covariance matrices are equal among the 2 classes) and QDA if the class variance are not the same (14).

Statistical analysis

The models were implemented using an open-source program language (Python 3.7.1), using its packages (numpy, scikit-learn, matplotlib). Continuous numerical variables were expressed as means (± standard deviation) and median (interquartile range, 25% quantile–75% quantile) and categorical variables were expressed as percentages. We used the Mann–Whitney U test for continuous variables. The *p*-value less than 0.05 was significant.

Results

Patients characteristics

This study implied 1,096 patients with NVUGIB (738 men, 67.3%; mean age ± SD, 63.9 ± 14.6). Socio-demographic and clinical features of patients are shown in Table 1. A percentage of 11% of these patients had cirrhosis and 7.5% mortality was registered.

Performance of models and classifiers

The five scores for the class groups of mortality are summarized in Table 2 and no significant differences were observed only for BBS (*p*-value = 0.099). The values for GBS, Rock, and AIM65 were significantly higher, and T-score was significantly lower for patients that died.

Statistically significant correlations were found between the five scores, even if they are low or very low, as in Figure 1.

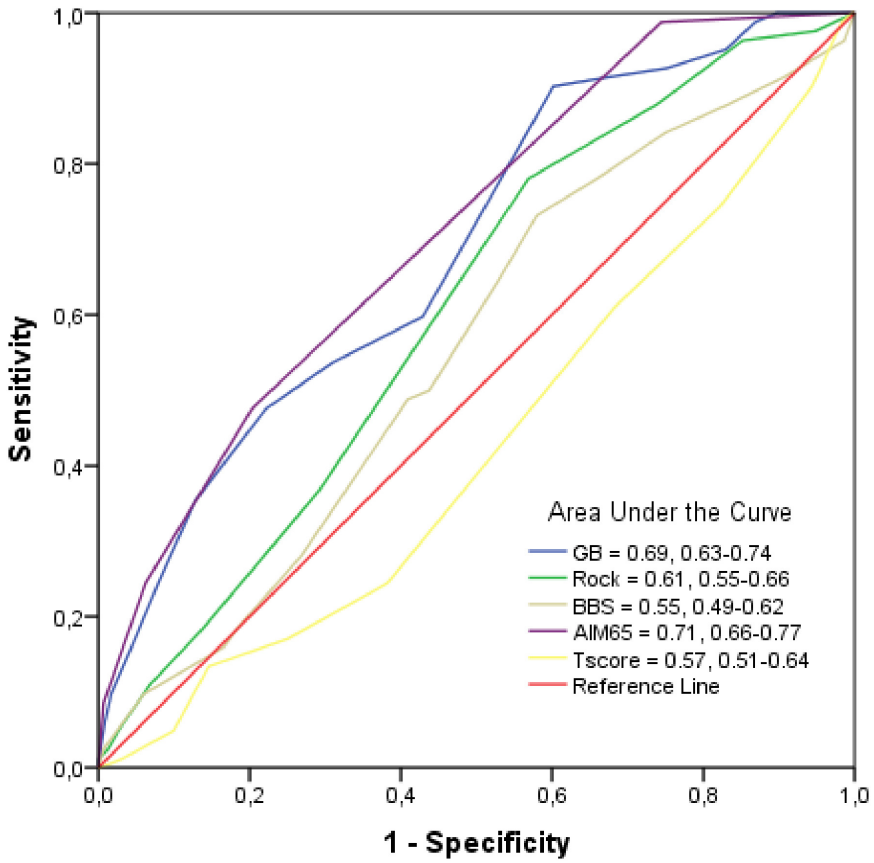

FIGURE 2
Receiver operating characteristic (ROC) Curve and Area under the curve (AUC). Mean AUC and its 95% confidence interval of the scores are shown in the legends of the subplots.

TABLE 3 Comparison of the confusion matrix and evaluation measures among prediction models.

	Precision	Recall	F1-score	Support
LDA				
Survival	0.95	0.95	0.95	709
Death	0.32	0.31	0.32	54
Accuracy			0.90	763
Macro avg	0.63	0.63	0.63	763
Weighted avg	0.90	0.90	0.90	763
QDA				
Survival	0.94	0.97	0.95	709
Death	0.32	0.19	0.24	54
Accuracy			0.91	763
Macro avg	0.63	0.58	0.60	763
Weighted avg	0.90	0.91	0.90	763
LG				
Survival	0.95	0.95	0.95	709
Death	0.32	0.28	0.30	54
Accuracy			0.91	763
Macro avg	0.63	0.62	0.62	763
Weighted avg	0.90	0.91	0.90	763
K-NN				
Survival	0.98	1.00	0.99	709
Death	1.00	0.78	0.88	54
Accuracy			0.98	763
Macro avg	0.99	0.89	0.93	763
Weighted avg	0.98	0.98	0.98	763

The reported average includes the macro average which averages the unweighted mean per label, and the weighted average which averages the support-weighted mean per label.

Positive correlation coefficients were observed, except by making the T-score where they were negative correlated (rho for T-score and GBS = −0.12, p-value < 0.001; rho for T-score and Rock = −0.30, p-value < 0.001; rho for T-score and AIM65 = −0.16, p-value < 0.001). the strongest correlation was observed between Rock and BBS (rho = 0.56, p-value < 0.001).

The AUC for GBS, Rock, BBS, AIM65, and T-score was low, the highest value was observed for AIM65 (AUC = 0.71, 95% CI: 0.66–0.77), as in Figure 2.

Classification accuracy of each machine-learning model (LDA, QDA, LR, and K-NN) was evaluated and summarized in Table 3.

The LDA model is performing well in terms of accuracy on the training data, as in Figure 3A1. The recall for death is quite low (0.31), which implies that this model will not perform well in differentiating the patients who have a high chance of survival, and hence this model would not help reduce the mortality rate. The model is giving a decent average recall when we balanced the precision and the recall for a threshold of about 0.25. A recall of 0.63 suggests that there is a 37% chance that the model will predict that a person is going to die even though he/she would not, and the health system would waste their time and money on these patients who are not at risk of mortality. We have built the LDA model. Furthermore, checking the coefficients, we found which variables are leading to mortality and which can help to reduce the mortality. The scores which positively affect the mortality are AIM65 (coefficient = 0.93) and GBS (coefficient = 0.57) and the ones that negatively affect it are T-score (coefficient = −0.31) and Rock (coefficient = −0.25). Based on LDA model, AIM65 is the most important feature in detecting whether a NVUGIB patient would die or not and BBS has almost no effect in predicting this (coefficient = 0.04). We checked the performance on the test data in Figure 3A2. The model was giving a similar performance on the test and train data, meaning the model has generalized well. The average recall, the precision and the accuracy are good, but we evaluated if we could get a better performance using other algorithms.

The QDA model did not obtained different outcomes from the LDA model (even worse recall), as in Figure 3B.

The LR model was giving a similar performance on the test and the train datasets (Figure 3C). The recall of the test data has increased while at the same time, the precision has decreased slightly, which was to be expected while adjusting the threshold at 0.18. The accuracy was of 0.91 on the train and of 0.90 on the test datasets. Checking the coefficients of the model, we observed the same variables that are leading to mortality rate: AIM65 (coefficient = 0.60) and GBS (coefficient = 0.57) and which can help to deduce the mortality rate: T-score (coefficient = −0.17) and Rock (coefficient = −0.21). The coefficients that positively and negatively affect the mortality rate were similar for LR and LDA. This means they capture the same pattern and give the same conclusions from the dataset.

Performing the KNN model from the Figure 3D, we selected the best value of k for which the error rate is the least in the validation data and k = 14 gave us the generalized model with very similar train and test errors, as in Figure 4.

We used GridSearchCV for hyperparameter tuning and we used them to find a better recall of the model. The recall and the precision have significantly increased by tuning the K-NN classifier. This is a high-performing model that a physician can use to control the mortality rate. There is a 98% chance that the model will detect NVUGIB patients who are likely to die, and the physician can take the appropriate action.

Discussion

Patient stratification in UGIB has been considered for prognosis assessment by differentiating high-risk patients (15). So far, available prediction scores use only some variables, both clinical or biological, and are based on conventional statistical analysis. While some of them are used for rebleeding or death prediction, a high precision rate has not been achieved. Ensuring a risk stratification at patient admission might be helpful in choosing the proper time for endoscopy, especially in small regional hospitals which do not provide a full-time endoscopy service. Moreover, probably a turning point in medicine in the last years, the COVID-19 pandemic almost changed patients' presentation in the emergency room, as well as patients' admission (16). A general decrease in patients' admission has been observed in the first months of the pandemic for all types of disease, and also for UGIB

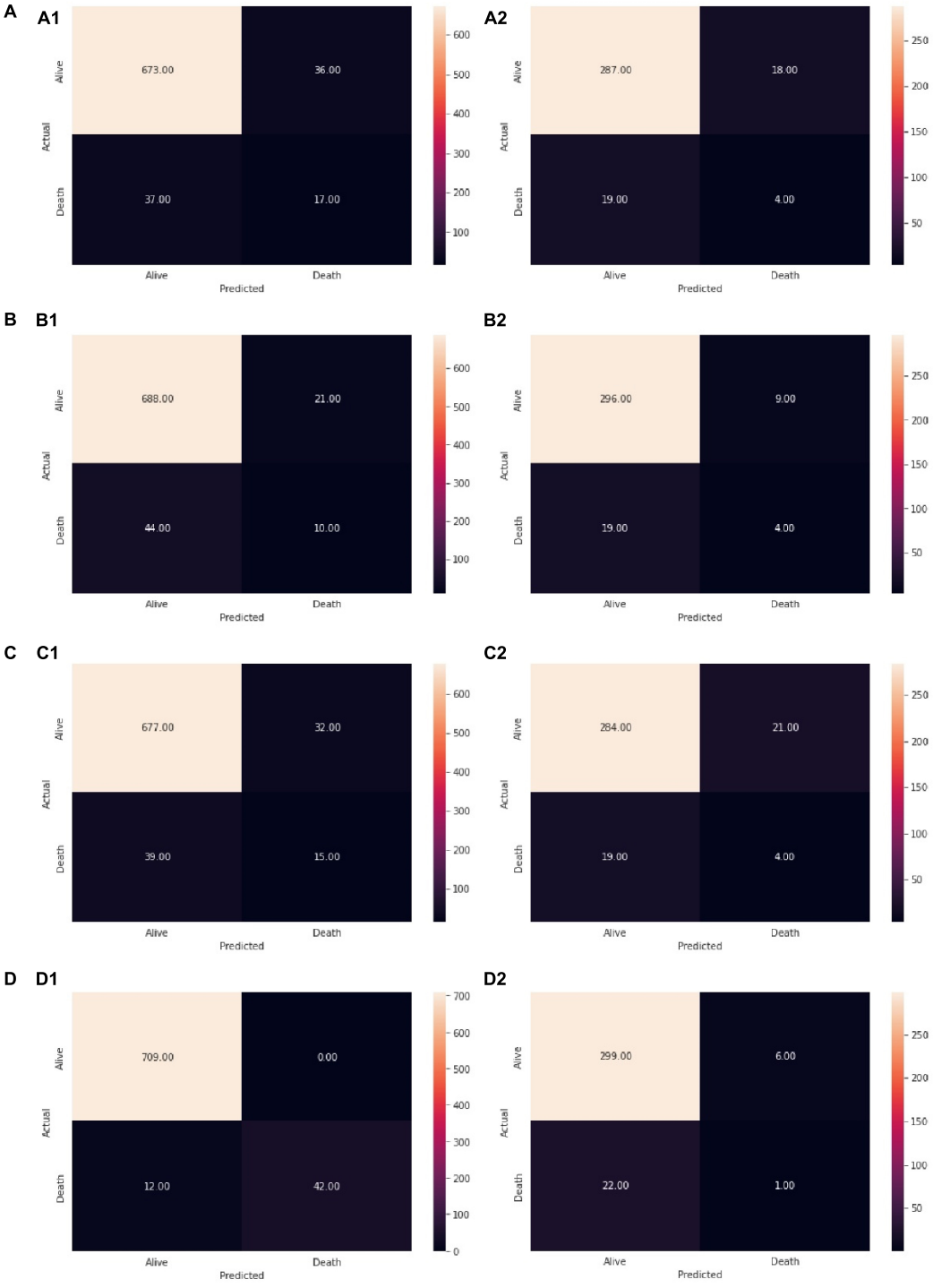

FIGURE 3
Checking model performance of **(A1)** LDA on training data, **(A2)** LDA on test data, **(B1)** QDA on training data, **(B2)** QDA on test data, **(C1)**. LR on training data, **(C2)** LR on test data, **(D1)** K-NN on training data, **(D2)** K-NN on test data. Reading the confusion matrix (clockwise from top left): True Negative (Actual = Alive, Predicted = Alive): Model predicts that the patient would live and the patients' lives, False Positive (Actual = Alive, Predicted = Death): Model predicts that the patient would die and the patients actually lives, True Positive (Actual = Death, Predicted = Death): Model predicts that the patient would die and the patients dies, False Negative (Actual = Death, Predicted = Alive): Model predicts that the patient would live and the patients actually dies.

patients. While the first consideration was that endoscopy was a high-risk procedure and should be performed only if patients required it, due to the lack of medical materials as well as the fear of contamination or hospital circuit reorganization, many patients still required rapid endoscopic assessment due to UGIB (17). Providing a tool to delay endoscopy or to predict the death secondary to UGIB might organize better the endoscopist decision-making process in choosing the right time for endoscopy.

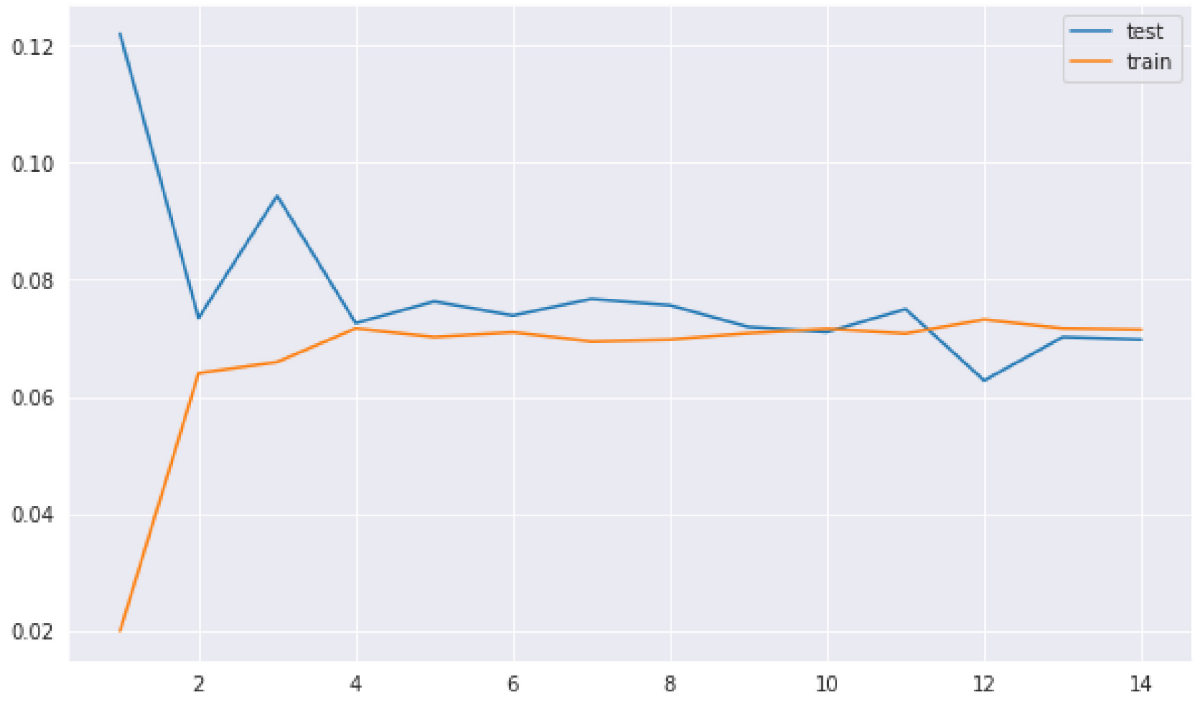

FIGURE 4
Test and train error for the K-NN model. As the number of neighbors increases, the test error and the train error are the same.

European Society of Gastrointestinal Endoscopy updated guidelines on NVUGIB recommend the use of GBS as the main risk stratification after patient admission (18). As stated, patients with a GBS ≤ 1 may be successfully managed as outpatients and may be discharged, however, patients require to be notified of the possibility of rebleeding, thus they should maintain contact with the discharging hospital. When discussing high-risk patients, there is a low probability of discharging and GBS score has shown a specificity of 12% for transfusions, hemostatic interventions as well as death (19). Also, when NVUGIB is associated with liver cirrhosis, mortality might increase due to the underlying disease complications such as hepatic encephalopathy and spontaneous bacterial peritonitis (20).

Our study provides a non-endoscopic ML model as an alternative tool to predict mortality in patients with NVUGIB at their admission to the emergency department. We obtained a high accuracy for death prediction and surpassed the available scores used for the initial assessment. There are some other studies that used AAN to predict mortality in UGIB, showing also better results than the current clinical scores (21–24). Available studies suggest that risk assessment tools have an AUC of 0.77 for mortality as mentioned in two multicenter studies (25, 26). However, our study points out that ANN might be more efficient in highlighting patients' prognoses related to mortality, with an AUC of 0.99. Moreover, the results are even more optimistic than the available ANN used so far for UGIB assessment by ANN or ML models.

A systematic review showed that ML models were more effective in predicting rebleeding, intervention, and mortality, with an AUC ranging from 0.80 to 0.90 (27). The ANN we propose focuses on five non-endoscopic scores used as an initial assessment to stratify the risk of UGIB. We combined GBS, AIMS65, Rock, T-score as well as BBS in a ML model, thus trying to better identify patients with a dismal prognosis. Our study end-point was mortality as we focused on exploring the potential of all five scores combined within a newly developed ML. Noteworthy is that taken separately all risk scores were definitely less accurate than our prediction model. Thus, our model might enable new opportunities for non-invasive tools to predict the NVUGIB mortality rate.

Risk assessment represents a cornerstone for the healthcare system, as it may provide high-quality care for patients and may also help save resources and direct them to more precise interventions. Even though there is a long distance to implementing this type of model in clinical practice, the potential of ML for UGIB assessment should not be downplayed (28, 29). We do acknowledge that it may be challenging to transfer an ML to a clinical setting, however, AI depicting background may attempt to integrate into clinical care and provide more reliable measures for UGIB assessment.

Nonetheless, our study has certain limitations. Firstly, this is a single-center experience study, thus we validated our AAN only on patients admitted to our Clinic. Secondly, we had a small sample size, but without missing data, and the Precision and Recall obtained in the validation dataset were not low. Finally, we prepared our dataset from the retrospective database, but the outcomes could not have changed over time due to the update of treatment guidelines in the last years. Testing the algorithm in a multicenter setting will surely help validate and improve our objective. On the other hand, we focused only on patients' mortality prediction and did not consider other important factors that might be encountered in day-to-day practice such as the rebleeding rate or surgical interventions.

The data we used were retrospectively collected from our registry which suggests heterogeneous information.

Conclusion

Our study suggests that a machine learning program based on the available pre-endoscopic bleeding scores might provide a more accurate prediction for patients' mortality rate after NVUGIB admission. By combining the results of the five scores in a ML algorithm, our tool might be considered useful, not only for endoscopists but also for emergency physicians to assess patients' prognosis at their presentation. While our single-center study may not be sufficient to validate and implement this tool, it may be a starting point for future integration in the healthcare system.

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 Ethics Committee of the University of Medicine and Pharmacy of Craiova, Romania approved this retrospective study and informed consent from all patients were acquired in the County Hospital before patient enrolment in the study (11977/24.03.2020). The patients/participants provided their written informed consent to participate in this study.

Author contributions

BSU and AT-S: project design and manuscript writing. BSU, DNF, SI, SMC, and VFI: data collection. AT-S: data analysis. BSU, DIG, and IR: revise the manuscript and interpretation of data. All authors read and approved the final manuscript.

Funding

The manuscript processing charges were funded by the University of Medicine and Pharmacy of Craiova, Romania.

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. Kate V, Sureshkumar S, Gurushankari B, Kalayarasan R. Acute upper non-variceal and lower gastrointestinal bleeding. *J Gastrointest Surg.* (2022) 26:932–49. doi: 10.1007/s11605-022-05258-4

2. Oakland K. Risk stratification in upper and upper and lower GI bleeding: which scores should we use? *Best Pract Res Clin Gastroenterol.* (2019) 42–3:101613. doi: 10.1016/j.bpg.2019.04.006

3. Hwang J, Fisher D, Ben-Menachem T, Chandrasekhara V, Chathadi K, Decker G, et al. Standards of practice committee of the american society for gastrointestinal endoscopy. The role of endoscopy in the management of acute non-variceal upper GI bleeding. *Gastrointest Endosc.* (2012) 75:1132–8. doi: 10.1016/j.gie.2012.02.033

4. Karstensen J, Ebigbo A, Aabakken L, Dinis-Ribeiro M, Gralnek I, Le Moine O, et al. Nonvariceal upper gastrointestinal hemorrhage: european society of gastrointestinal endoscopy (ESGE) cascade guideline. *Endosc Int Open.* (2018) 6:E1256–63. doi: 10.1055/a-0677-2084

5. Marmo R, Soncini M, Bucci C, Zullo A, Gised. Comparison of assessment tools in acute upper gastrointestinal bleeding: which one for which decision. *Scand J Gastroenterol.* (2022) 57:1–7. doi: 10.1080/00365521.2021.1976268

6. Chandnani S, Rathi P, Sonthalia N, Udgirkar S, Jain S, Contractor Q, et al. Comparison of risk scores in upper gastrointestinal bleeding in western India: a prospective analysis. *Indian J Gastroenterol.* (2019) 38:117–27. doi: 10.1007/s12664-019-00951-w

7. Gu L, Xu F, Yuan J. Comparison of AIMS65, Glasgow-Blatchford and Rockall scoring approaches in predicting the risk of in-hospital death among emergency hospitalized patients with upper gastrointestinal bleeding: a retrospective observational study in Nanjing, China. *BMC Gastroenterol.* (2018) 18:98. doi: 10.1186/s12876-018-0828-5

8. Iino C, Mikami T, Igarashi T, Aihara T, Ishii K, Sakamoto J, et al. Evaluation of scoring models for identifying the need for therapeutic intervention of upper gastrointestinal bleeding: a new prediction score model for Japanese patients. *Dig Endosc.* (2016) 28:714–21. doi: 10.1111/den.12666

9. Chadebecq F, Lovat L, Stoyanov D. Artificial intelligence and automation in endoscopy and surgery. *Nat Rev Gastroenterol Hepatol.* (2022). doi: 10.1038/s41575-022-00701-y

10. Udriștoiu A, Cazacu I, Gruionu L, Gruionu G, Iacob A, Burtea D, et al. Real-time computer-aided diagnosis of focal pancreatic masses from endoscopic ultrasound imaging based on a hybrid convolutional and long short-term memory neural network model. *PLoS One.* (2021) 16:e0251701. doi: 10.1371/journal.pone.0251701

11. McLachlan G. *Discriminant Analysis and Statistical Pattern Recognition.* Hoboken, NJ: Wiley (2004).

12. Deo RC. Machine learning in medicine. *Circulation.* (2018) 20:1920–30.

13. Wood SJ. Fast stable restricted maximum likelihood and marginal likelihood estimation of semiparametric generalized linear models. *Royal Stat Soc Ser B Stat Methodol.* (2011) 73:3–36.

14. Ali N, Neagu D, Trundle P. Evaluation of k-nearest neighbour classifier performance for heterogeneous data sets. *SN Appl Sci.* (2019) 1:1559.

15. Tham J, Stanley A. Clinical utility of pre-endoscopy risk scores in upper gastrointestinal bleeding. *Expert Rev Gastroenterol Hepatol.* (2019) 13:1161–7. doi: 10.1080/17474124.2019.1698292

16. Mauro A, De Grazia F, Anderloni A, Di Sabatino A. Upper gastrointestinal bleeding in coronavirus disease 2019 patients. *Curr Opin Gastroenterol.* (2022) 38:443–9. doi: 10.1097/MOG.0000000000000859

17. Khan R, Saha S, Gimpaya N, Bansal R, Scaffidi M, Razak F, et al. Outcomes for upper gastrointestinal bleeding during the first wave of the COVID-19 pandemic in the Toronto area. *J Gastroenterol Hepatol.* (2022) 37:878–82. doi: 10.1111/jgh.15804

18. Gralnek I, Stanley A, Morris A, Camus M, Lau J, Lanas A, et al. Endoscopic diagnosis and management of nonvariceal upper gastrointestinal hemorrhage (NVUGIH): european society of gastrointestinal endoscopy (ESGE) guideline – Update 2021. *Endoscopy.* (2021) 53:300–32. doi: 10.1055/a-1369-5274

19. Ramaekers R, Mukarram M, Smith C, Thiruganasambandamoorthy V. The predictive value of preendoscopic risk scores to predict adverse outcomes in emergency department patients with upper gastrointestinal bleeding: a systematic review. *Acad Emerg Med.* (2016) 23:1218–27. doi: 10.1111/acem.13101

20. Kalafateli M, Triantos C, Nikolopoulou V, Burroughs A. Non-variceal gastrointestinal bleeding in patients with liver cirrhosis: a review. *Dig Dis Sci.* (2012) 57:2743–54. doi: 10.1007/s10620-012-2229-x

21. Shung D, Au B, Taylor R, Tay J, Laursen S, Stanley A, et al. Validation of a machine learning model that outperforms clinical risk scoring systems for upper gastrointestinal bleeding. *Gastroenterology.* (2020) 158:160–7.

22. Rotondano G, Cipolletta L, Grossi E, Koch M, Intraligi M, Buscema M, et al. Artificial neural networks accurately predict mortality in patients with nonvariceal upper GI bleeding. *Gastrointest Endosc.* (2011) 73:218–26, 226.e1–2. doi: 10.1016/j.gie.2010.10.006

23. Rao V, Gupta N, Swei E, Wagner T, Aronsohn A, Reddy K, et al. Predictors of mortality and endoscopic intervention in patients with upper gastrointestinal bleeding in the intensive care unit. *Gastroenterol Rep (Oxf).* (2020) 8:299–305. doi: 10.1093/gastro/goaa009

24. Tan Q, Ma A, Deng H, Wong V, Tse Y, Yip T, et al. A hybrid residual network and long short-term memory method for peptic ulcer bleeding mortality prediction. *AMIA Annu Symp Proc.* (2018) 2018:998–1007.

25. Stanley A, Laine L, Dalton H, Ngu J, Schultz M, Abazi R, et al. International gastrointestinal bleeding C: comparison of risk scoring systems for patients presenting with upper gastrointestinal bleeding: international multicentre prospective study. *BMJ.* (2017) 356:i6432. doi: 10.1186/s13054-016-1208-6

26. Laursen S, Hansen J, Schaffalitzky de Muckadell, OB. The glasgow blatchford score is the most accurate assessment of patients with upper gastrointestinal hemorrhage. *Clin Gastroenterol Hepatol.* (2012) 10:e1131.

27. Shung D, Simonov M, Gentry M, Au B, Laine L. Machine learning to predict outcomes in patients with acute gastrointestinal bleeding: a systematic review. *Dig Dis Sci.* (2019) 64:2078–87. doi: 10.1007/s10620-019-05645-z

28. Das A, Ben-Menachem T, Farooq F, Cooper G, Chak A, Sivak M Jr., et al. Artificial neural network as a predictive instrument in patients with acute nonvariceal upper gastrointestinal hemorrhage. *Gastroenterology.* (2008) 134:65–74. doi: 10.1053/j.gastro.2007.10.037

29. Veisman I, Oppenheim A, Maman R, Kofman N, Edri I, Dar L, et al. Novel prediction tool for endoscopic intervention in patients with acute upper gastro-intestinal bleeding. *J Clin Med.* (2022) 11:5893. doi: 10.3390/jcm11195893

frontiers | Frontiers in Medicine

TYPE Original Research
PUBLISHED 15 March 2023
DOI 10.3389/fmed.2023.1150623

OPEN ACCESS

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

REVIEWED BY
Zhifeng Liu,
General Hospital of Southern Theater Command of PLA, China
Huasheng Tong,
General Hospital of Guangzhou Military Command, China

*CORRESPONDENCE
Jingchun Song
✉ songjingchun@126.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 24 January 2023
ACCEPTED 27 February 2023
PUBLISHED 15 March 2023

CITATION
Zeng Q, Zhong L, Zhang N, He L, Lin Q and Song J (2023) Nomogram for predicting disseminated intravascular coagulation in heatstroke patients: A 10 years retrospective study.
Front. Med. 10:1150623.
doi: 10.3389/fmed.2023.1150623

Nomogram for predicting disseminated intravascular coagulation in heatstroke patients: A 10 years retrospective study

Qingbo Zeng[1,2†], Lincui Zhong[1†], Nianqing Zhang[2], Longping He[1], Qingwei Lin[1] and Jingchun Song[1*]

[1]Intensive Care Unit, The 908th Hospital of Logistic Support Force, Nanchang, China, [2]Intensive Care Unit, Nanchang Hongdu Hospital of Traditional Chinese Medicine, Nanchang, China

Background: Disseminated intravascular coagulation (DIC) can lead to multiple organ failure and death in patients with heatstroke. This study aimed to identify independent risk factors of DIC and construct a predictive model for clinical application.

Methods: This retrospective study included 87 patients with heatstroke who were treated in the intensive care unit of our hospital from May 2012 to October 2022. Patients were divided into those with DIC (n = 23) or without DIC (n = 64). Clinical and hematological factors associated with DIC were identified using a random forest model, least absolute shrinkage and selection operator (LASSO) regression and support vector machine-recursive feature elimination (SVM-RFE). Overlapping factors were used to develop a nomogram model, which was diagnostically validated. Survival at 30 days after admission was compared between patients with or without DIC using Kaplan-Meier analysis.

Results: Random forest, LASSO, and SVM-RFE identified a low maximum amplitude, decreased albumin level, high creatinine level, increased total bilirubin, and aspartate transaminase (AST) level as risk factors for DIC. Principal component analysis confirmed that these independent variables differentiated between patients who experienced DIC or not, so they were used to construct a nomogram. The nomogram showed good predictive power, with an area under the receiver operating characteristic curve of 0.976 (95% CI 0.948–1.000) and 0.971 (95% CI, 0.914–0.989) in the internal validation. Decision curve analysis indicated clinical utility for the nomogram. DIC was associated with significantly lower 30 days survival for heatstroke patients.

Conclusion: A nomogram incorporating coagulation-related risk factors can predict DIC in patients with heatstroke and may be useful in clinical decision-making.

KEYWORDS

heatstroke, disseminated intravascular coagulation, predictor, nomogram, thromboelastography

1. Introduction

Heatstroke is a life-threatening illness manifesting as extreme hyperthermia (>40°C), dysfunction in the central nervous system, and multiple organ failure (1, 2). Although the treatment of heatstroke has improved, heatstroke-related deaths are increasing worldwide, which may worsen due to global warming (3, 4).

A substantial proportion of patients with heatstroke, from 22 to 45%, experience disseminated intravascular coagulation (DIC), which further increases risk of multiple organ dysfunction and mortality (5, 6). DIC is difficult to diagnose and challenging to treat. Reliable prediction of which heatstroke patients are at greater risk of DIC could help clinicians monitor such patients more closely and initiate preventive or therapeutic measures earlier. However, the conventional coagulation tests typically used to diagnose DIC are poor predictors of the complication (7).

Several studies have explored potentially better predictors of DIC, such as thromboelastography maximum amplitude, activated clotting time and clot rate as determined with a Sonoclot@ device (8–10). However, these biomarkers reflect primarily coagulation, so they may identify patients already in early stages of DIC rather than predict the complication before it occurs. Due to the complex pathogenesis of DIC, a more comprehensive panel of biomarkers may be needed to predict DIC in heatstroke patients.

The current study explored a range of potential risk factors of DIC and selected the best to create a predictive nomogram, which we validated using a 10 years retrospective dataset that included survival at 30 days after admission.

2. Materials and methods

2.1. Study design and patients

This retrospective study was approved the Ethics Committee of the 908th Hospital of Logistic Support Force (Nanchang,

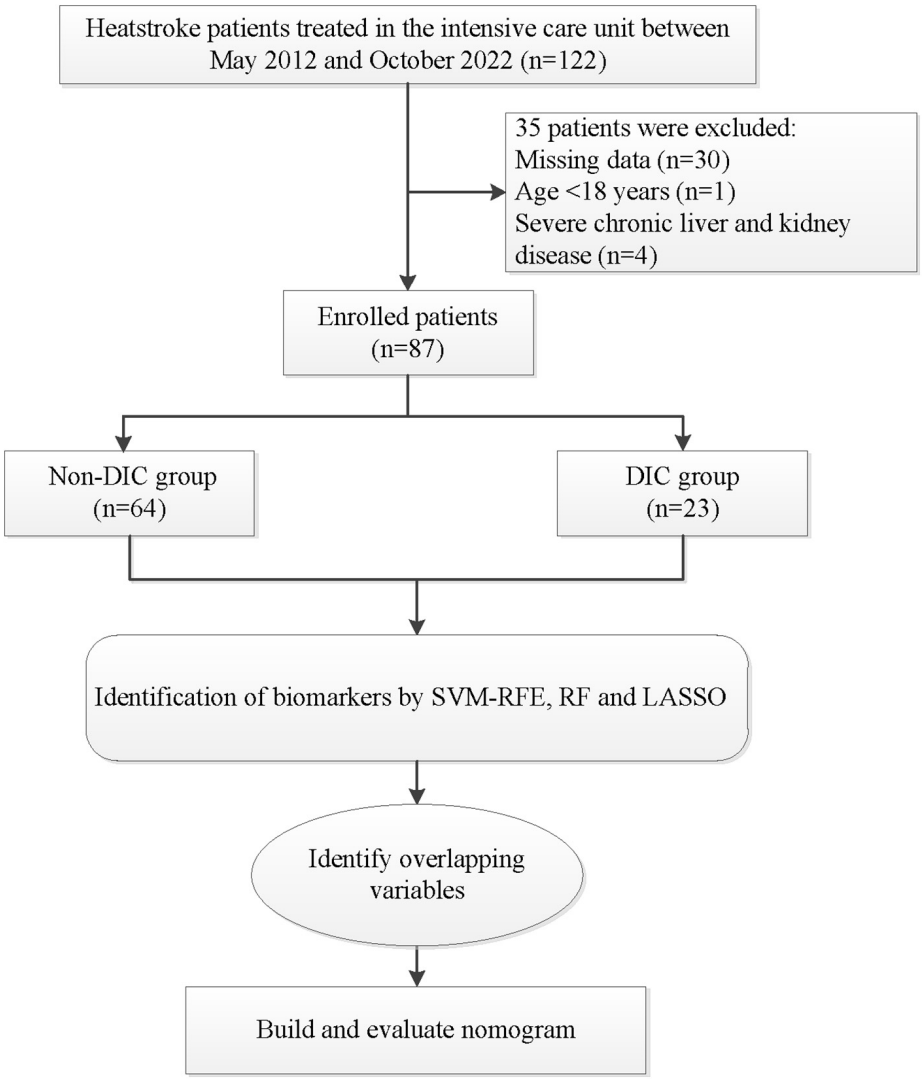

FIGURE 1
Flowchart of patient selection and analysis. DCA, decision curve analysis; LASSO, least absolute shrinkage and selection operator; RF, random forest; SVM-RFE, support vector machine-recursive feature elimination.

China), which waived the requirement for consent because all participants, at the time of treatment, signed written consent for their anonymized medical data to be analyzed and published for research purposes. All procedures involving human participants were performed in accordance with the 1975 Helsinki Declaration and its later amendments.

We screened for eligibility all patients with heatstroke who were admitted to the intensive care unit of the 908th Hospital from May 2012 to October 2022. Eligible patients had a history of exposure to hot and humid weather or high-intensity activity, and they met at least one of the following criteria based on the Chinese Expert Consensus on the Diagnosis and Treatment of Heatstroke (1): (1) neurological dysfunction, including coma, convulsions, delirium, or abnormal behavior; (2) core temperature $\geq 40°C$; (3) functional impairment of at least two organs; or (4) severe coagulopathy or DIC. Severe coagulopathy was defined as the presence of at least two of the following criteria: platelet count < 100,000 cells per μL, international normalized ratio > 1.5, fibrinogen level < 1.50 g/L,

TABLE 1 Clinicodemographic characteristics of patients with heatstroke at admission to the intensive care unit, stratified by disseminated intravascular coagulation (DIC) diagnosis.

Characteristics	Total (n = 87)	No DIC (n = 64)	DIC (n = 23)	P
Male	76 (87.4)	55 (85.9)	21 (91.3)	0.72
Age, yr	43 (22, 64)	28 (21.8, 64)	48 (40, 61)	0.211
Core temperature, °C	37.6 (36.8, 38.6)	37.4 (36.7, 38.4)	38.4 (37.6, 39)	0.006
R, min	7.4 (5.7, 12.4)	6.5 (5.2, 9.4)	14.9 (10.6, 28.6)	<0.001
K, min	3.2 (2.2, 6.2)	2.6 (2, 4.2)	10.6 (5.9, 22)	<0.001
Angle, °	50.3 (31.4, 60.5)	54.9 (43.8, 62.4)	22 (9.4, 35.1)	<0.001
MA, mm	49.8 (38.4, 56.5)	53.1 (46.3, 59.4)	30.2 (19.5, 40.1)	<0.001
PT, s	16.9 (13.7, 19.8)	14.8 (13.3, 17.4)	27.6 (21.8, 43.5)	<0.001
INR	1.4 (1.2, 1.6)	1.2 (1.1, 1.4)	2.2 (1.8, 3.2)	<0.001
APTT, s	36.1 (28.8, 48.4)	31.6 (28, 40.4)	46.3 (39.8, 97.8)	<0.001
Fibrinogen, g/L	2.0 (1.5, 2.7)	2.2 (1.8, 2.9)	1.4 (1.0, 2.1)	0.002
TT, s	16.8 (14.8, 20.5)	16.4 (14.6, 18.2)	22.0 (16.8, 26.2)	0.002
FDP, μg/L	8.1 (2.2, 26.4)	3.2 (1.2, 9.9)	34.0 (24.5, 65.9)	<0.001
D-dimer, μg/L	2.3 (0.5, 6.7)	1.0 (0.3, 2.7)	8.1 (4.1, 28.1)	<0.001
WBC, $\times 10^9$/L	11.7 (8, 16.6)	11.4 (7.8, 16.8)	12.5 (8.9, 14.6)	0.733
NLR	10.3 (5.0, 17.4)	10.2 (4.3, 14.7)	11.7 (6.9, 22.6)	0.075
CRP, μg/L	2.4 (0.6, 15.1)	2.2 (0.6, 10.8)	5.7 (1.0, 31.6)	0.192
RBC, $\times 10^{12}$/L	4.3 ± 0.7	4.4 ± 0.6	4.1 ± 0.7	0.058
HGB, g/L	133 (118, 143)	136 (121, 145)	119 (109, 133)	0.003
HCT, %	39.4 ± 5.9	40.4 ± 5.3	36.6 ± 6.5	0.018
Platelet, $\times 10^9$/L	115 (52, 202)	155 (101, 228)	35 (22, 58)	<0.001
AST, U/L	49.1 (21.0, 164.9)	30.4 (18.6, 62.8)	443.1 (98.6, 1105.0)	<0.001
ALT, U/L	75.7 (28.5, 306.0)	41.5 (23.7, 134.4)	687.8 (157.9, 1543.4)	<0.001
Tbil, μmol/L	16.4 (12.3, 25.3)	15.1 (11.3, 19.1)	28.0 (19.3, 78.2)	<0.001
Albumin, g/L	37.9 (34.3, 43.8)	40.0 (36.8, 44.5)	33.2 (25.6, 35.5)	<0.001
Cr, μmol/L	105.9 (77.4, 150.3)	91.0 (72.4, 123.9)	207.9 (113.7, 250.5)	<0.001
MYO, ng/mL	632.9 (118.6,926.9)	445.7 (48.0, 915.1)	834.8 (570.9, 944.5)	0.052
CK, U/L	696 (226, 1975)	402 (192, 1224)	2216 (702, 10179)	<0.001
HR, min^{-1}	100 (81, 110)	90 (73, 105)	116 (102, 134)	<0.001
PH	7.4 (7.3, 7.4)	7.4 (7.4, 7.5)	7.3 (7.3, 7.4)	<0.001
$PaCO_2$, mmHg	33.5 (28.0, 39.6)	33.8 (28.2, 39.0)	33 (28.1, 42.8)	0.535
PaO_2, mmHg	145 (86.5, 187)	143.5 (88.6, 182)	158 (82.2, 204.5)	0.765
Lac, mmol/L	2.4 (1.1, 4.9)	1.6 (1.0, 3.2)	4.9 (3.4, 7.6)	<0.001
GCS score	6 (4, 14)	10 (5, 15)	4 (3, 5)	<0.001
APACHE II score	21 (12, 26)	19 (11, 24)	28 (23, 34)	<0.001

Values are n (%), mean ± SD, or median (interquartile range), unless otherwise noted. R, reaction time; K, kinetics of clot development; MA, maximum amplitude; PT, prothrombin time; APTT, activated partial thrombin time; INR, international normalized ratio; TT, thrombin time; FDP, fibrinogen degradation product; WBC, white blood cell count; NLR, neutrophil to lymphocyte ratio; CRP, C-reaction protein; RBC, red blood cell count; HGB, hemoglobin; HCT, hematocrit; ALT, alanine transaminase; AST, aspartate transaminase; Tbil, total bilirubin; Cr, creatinine; MYO, myoglobin; CK, creatine kinase; HR, heart rate; Lac, lactate; GCS, glasgow coma scale; APACHE II, acute physiology and chronic health evaluation II.

10.3389/fmed.2023.1150623

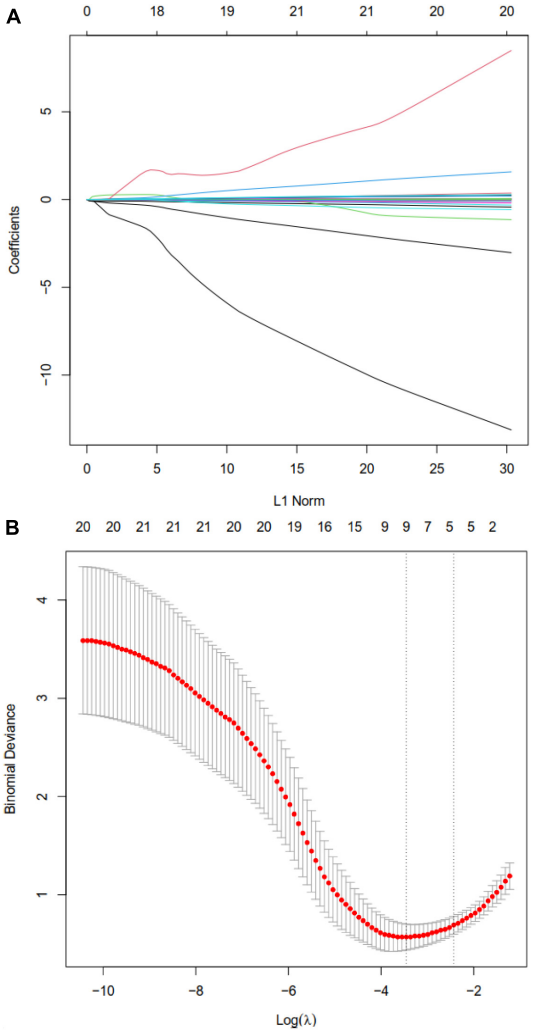

FIGURE 2
Least absolute shrinkage and selection operator (LASSO) regression to identify clinical features that may predict disseminated intravascular coagulation (DIC) in heatstroke patients. **(A)** LASSO coefficient profiles. **(B)** LASSO regression using 10-fold cross-validation and the "minimum plus one standard error" criterion to identify the optimal penalization coefficient lambda (λ).

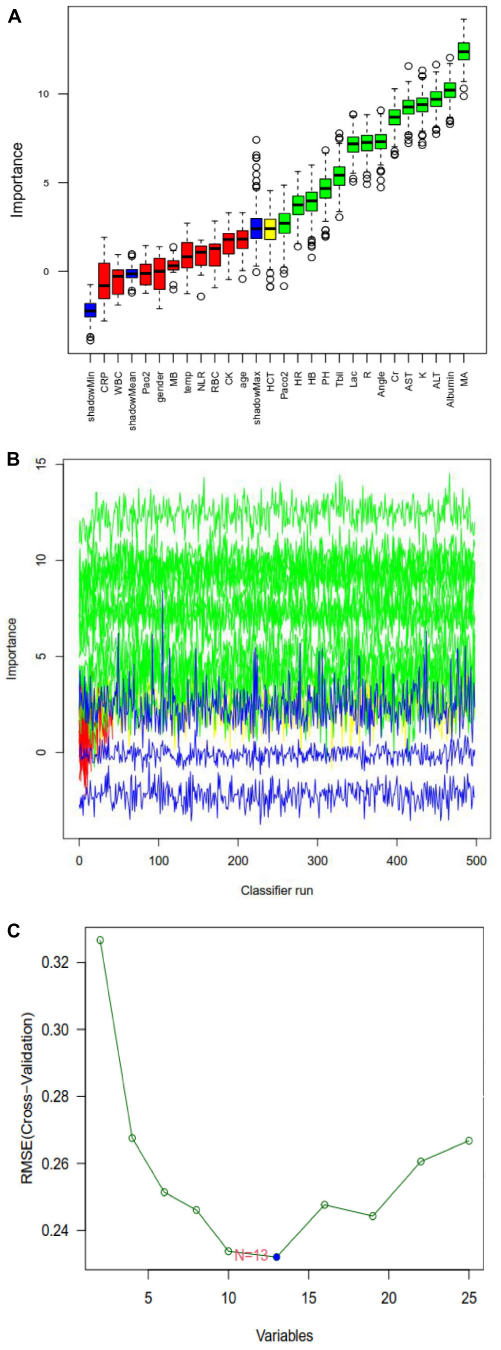

FIGURE 3
Random forest analysis to identify clinical features that may predict disseminated intravascular coagulation (DIC) in heatstroke patients. **(A)** Boxplot for all features in random forest analysis. Green indicates important variables; red, blue, or yellow, rejected variables. **(B)** Rejection or acceptance of factors during random forest classification runs. **(C)** Support vector machine-recursive feature elimination (SVM-RFE) to identify clinical predictors of DIC.

and D-dimer value above 10 times the upper limit of normal (11). Patients were excluded if they were younger than 18 years, if they had a congenital coagulopathy or severe chronic disease of the liver or kidney, or if they were using anticoagulant drugs at admission.

Included patients were divided into two groups based on presence or absence of DIC at admission, which was diagnosed based on the scoring system proposed by the International Society of Thrombosis and Hemostasis (12). DIC was diagnosed if a patient had a total score of at least five after summing the points for the following four parameters: platelet count, scored as one point if $<100 \times 10^9$/L or two points if $<50 \times 10^9$/L; prothrombin (PT) prolongation time, scored as one point if >3 s, or two points if >6 s; fibrinogen level, scored as one point if <1.0 g/L; fibrin degradation products or D-dimer level, scored as two points if $\geq$5-fold the upper limit of the normal range, or three points if $\geq$10-fold the upper limit (13).

2.2. Data collection

Baseline clinicodemographic data were extracted from electronic medical records, including age, sex, core temperature (rectal temperature), heart rate, Glasgow coma scale score, and the Acute Physiology and Chronic Health Evaluation II score. Data

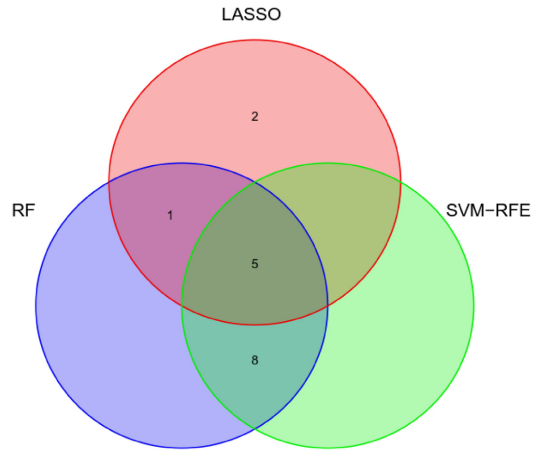

FIGURE 4
Venn diagram showing overlap of clinical factors that least absolute shrinkage and selection operator (LASSO), random forest, and support vector machine-recursive feature elimination (SVM-RFE) methods identified as predictors of disseminated intravascular coagulation (DIC) in heatstroke patients.

on the following routine coagulation indicators were collected: PT, activated partial thrombin time, fibrinogen, international normalized ratio, thrombin time, and levels of fibrin degradation product and D-dimer. Data on the following TEG indexes were collected: reaction time (R time), kinetics of clot development (K time), angle, maximum amplitude (MA). Data were collected on the following whole blood characteristics: counts of platelets and red and white blood cells, neutrophil-to-lymphocyte ratio, hemoglobin range and hematocrit percentage. In addition, data were collected on levels of C-reactive protein, alanine transaminase (ALT), AST, total bilirubin (Tbil), albumin, creatinine (Cr), myoglobin, and creatine kinase.

2.3. Statistical analysis

All statistical analyses were performed using R 4.2.1 software for windows (Chicago, IL), and all analyses were two-sided. Continuous variables with normal distribution were presented as mean ± standard deviation, while continuous data with a skewed distribution were expressed as median with interquartile range (IQR). Categorical variables were expressed as percentages (%). Pairwise comparisons were conducted using Student's *t*-test or the Mann-Whitney U test for continuous variables, while the chi-squared test or Fisher's exact test was used for categorical variables with normal or skewed distributions, as appropriate. Differences were considered significant if $P < 0.05$.

Potential risk factors of DIC were identified using three algorithms: least absolute shrinkage and selection operator (LASSO) regression, support vector machine-recursive feature elimination (SVM-RFE), and random forest. Risk factors identified by all three models were used to construct a nomogram using the *rms* package in R 4.2.1 software for windows. The discriminatory ability of the nomogram was evaluated in terms of areas under receiver operating characteristic curves (AUCs) and calibration curves. Principal component analysis was used to assess the ability of DIC biomarkers. The bootstrapping method (resampling = 500) was used for internal validation. The net benefit rate of the nomogram was assessed using decision curve analysis. Kaplan–Meier curve describing survival at 30 days after admission was compared between patients with or without DIC using the log-rank test.

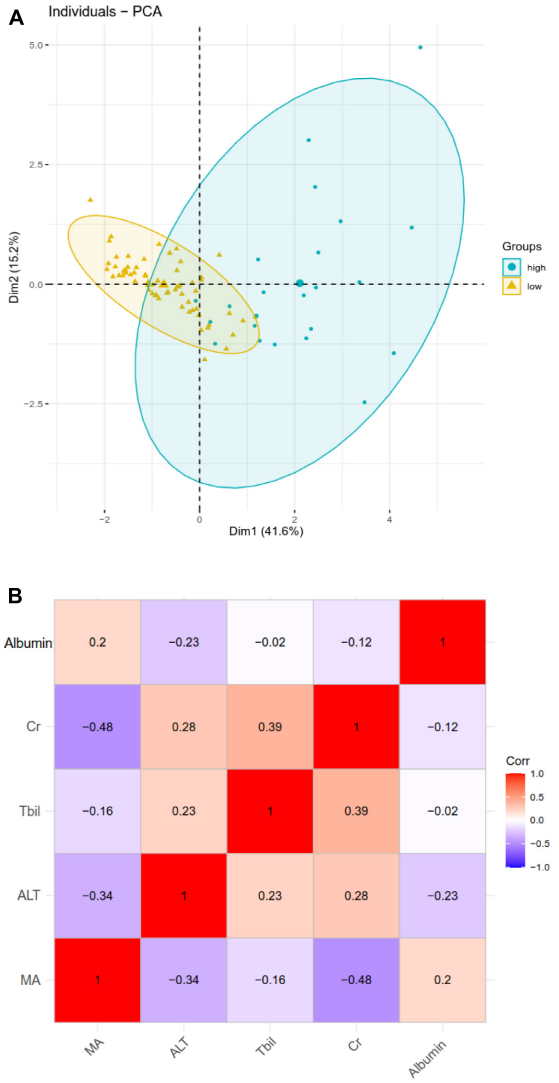

FIGURE 5
(A) Principal component analysis to assess the ability of the five selected variables to differentiate heatstroke patients with or without disseminated intravascular coagulation (DIC). **(B)** Heatmap of pairwise correlations among the five variables. ALT, alanine transaminase; Cr, creatinine; MA, maximum amplitude; Tbil, total bilirubin.

3. Results

3.1. Patient characteristics

Of the 122 patients considered for enrollment, 87 were included into the final analysis, of whom 23 had DIC, while 64 did not (Figure 1). There were no significant differences between the two

10.3389/fmed.2023.1150623

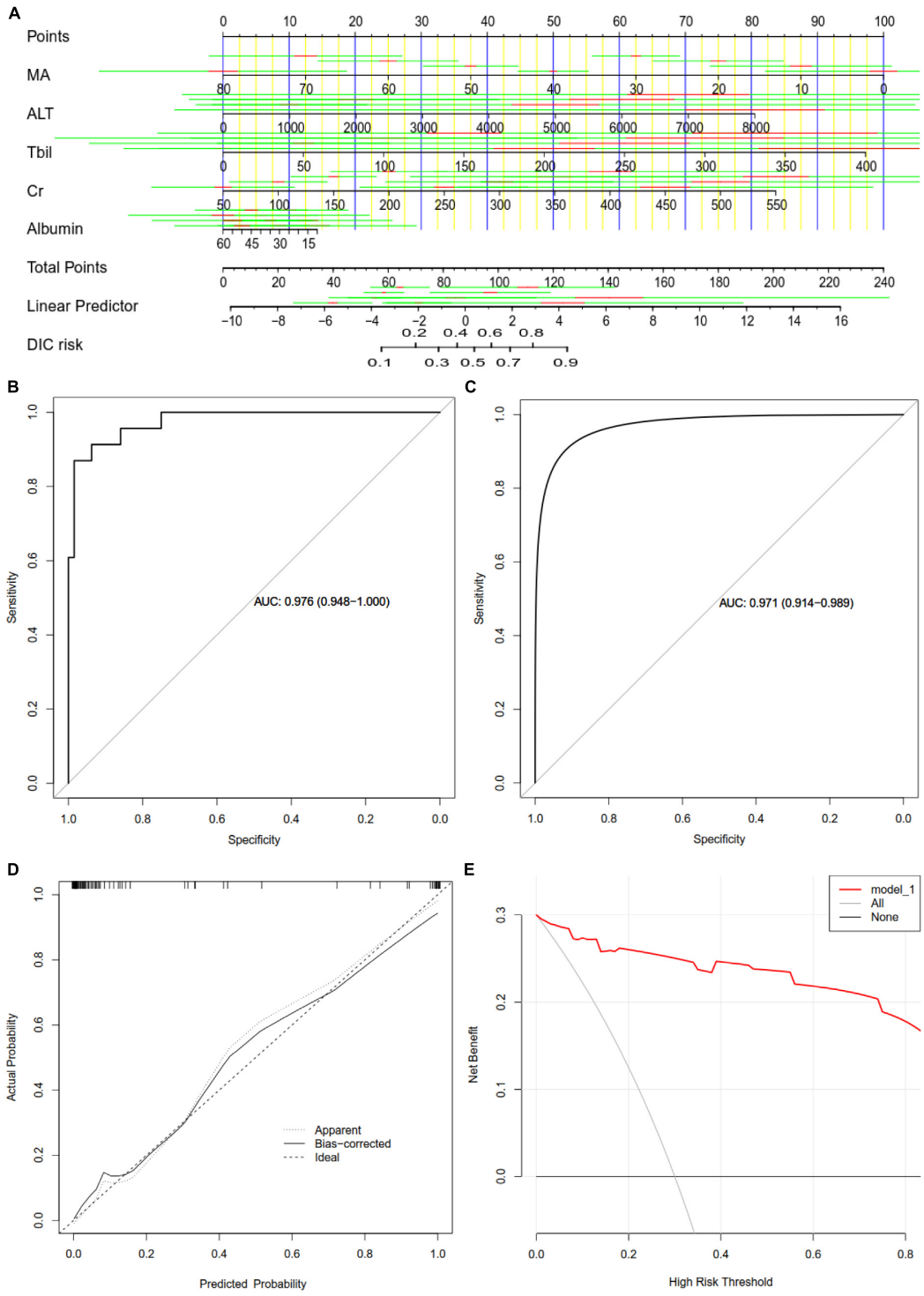

FIGURE 6
Assessment of a nomogram based on five predictors of disseminated intravascular coagulation (DIC) in patients with heatstroke. **(A)** Nomogram for predicting DIC in patients with heatstroke. **(B)** Receiver operating characteristic curves assessing the ability of the nomogram to predict DIC. **(C)** Internal validation using the bootstrap method (resampling = 500). **(D)** Calibration curve of the predictive model showing the degree of consistency between the predicted probability and actual probability (the Hosmer–Lemeshow test, $P > 0.05$, suggesting that it is of goodness-of-fit). **(E)** Decision curve analysis to assess the clinical benefit of the predictive nomogram. ALT, alanine transaminase; AUC, areas under receiver operating characteristic curves; Cr, creatinine; MA, maximum amplitude; Tbil, total bilirubin.

groups in terms of sex distribution, age, counts of white or red blood cells, or myoglobin levels (Table 1). Patients with DIC were more likely to have a higher core temperature and increased levels of the following: fibrinogen degradation product, D-dimer, AST, ALT, Tbil, Cr and creatine kinase. As expected, indicators of coagulation were also altered in patients with DIC, reflected by

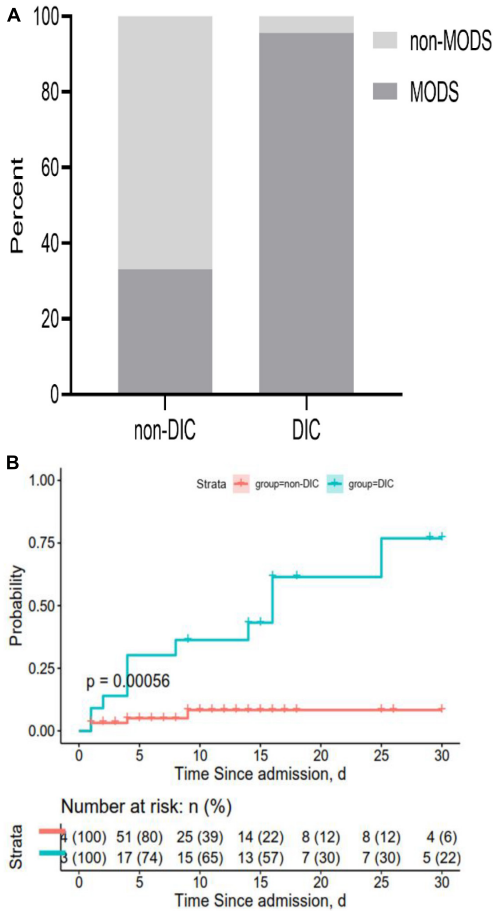

FIGURE 7
Rates of **(A)** multiple-organ dysfunction syndrome (MODS) and **(B)** overall survival at 30 days after admission. Both panels show data for all 87 patients in the sample.

longer R time and K time; longer PT, activated partial thrombin time and thrombin time; and a higher international normalized ratio. Conversely, compared to non-DIC patients, those with DIC had lower levels of angle, MA, fibrinogen, hemoglobin, hematocrit, albumin and platelet count. Overall, DIC patients showed more severe illness and injury to the central nervous system than patients without DIC.

3.2. Identification of DIC biomarkers

Least absolute shrinkage and selection operator regression analysis identified eight clinical features as potential predictors of DIC in patients with heatstroke: core temperature, ALT, maximum amplitude, hemoglobin, creatinine, albumin, total bilirubin, and creatine kinase (Figure 2). Random forest analysis identified the same six features as LASSO well as the following eight: AST, reaction time, kinetics of clot development, Angle, heart rate, lactate, $PaCO_2$, and pH (Figures 3A, B). SVM-RFE identified five of the same features as LASSO and random forest, as well as another eight: AST, reaction time, kinetics of clot development, Angle, heart rate, lactate, PaO_2, and pH (Figure 3C). The five factors overlapping across all three methods - MA, Cr, Tbil, albumin, and ALT - were considered potential predictors of DIC and used in further analyses (Figure 4).

3.3. Verification of DIC biomarkers

Principal component analysis was used to assess the ability of the five variables identified by LASSO, random forest, and SVM-RFE methods to differentiate patients with or without DIC (Figure 5A). There were no significant correlations among the five variables, suggesting that they had no function similarities (Figure 5B).

3.4. Development and validation of a predictive nomogram

We developed a nomogram to predict DIC based on the five verified factors (Figure 6A). Our nomogram showed good predictive power, with an AUC of 0.976 (Figure 6B), which was internally validated by bootstrapping, which gave an AUC of 0.971 (Figure 6C). A calibration curve of the predictive model showed a high degree of consistency between the predicted probability and actual probability and confirmed that the nomogram accurately predicted DIC (Figure 6D). Furthermore, decision curve analysis demonstrated that our nomogram had an extensive range of cutoff probabilities and excellent net benefits for threshold probabilities, which showed the potential clinical utility of the predictive model (Figure 6E).

3.5. Patient outcomes

Across all patients in our study, 95.6% in the DIC group experienced multiple-organ dysfunction by 30 days after admission, compared to only 33.3% in the non-DIC group ($P < 0.05$) (Figure 7A). DIC patients also showed a significantly lower overall survival rate at 30 days (47.8 vs. 6.3%; Figure 7B).

4. Discussion

To our knowledge, this article firstly reported a nomogram for prediction of heatstroke induced DIC. Patients suffering from heatstroke are at high risk of developing DIC, which remains a major cause of mortality (14). In this retrospective study, the incidence of DIC was 26.4% and the rate of mortality in patients with DIC was 47.8%. In an effort to predict DIC in order to improve management and timely treatment, we used three complementary methods including SVM-RFE, LASSO and random forest to screen for clinical factors that could reliably predict the complication, and we validated a nomogram for this purpose. SVM is a novel small sample method and a rather robust classification tool. Random forest and lasso can well deal with the high-dimensional data. The resulting model incorporates five routine clinical indexes that are easily acquired within 24 h of hospital admission and that capture complementary aspects of DIC pathophysiology, which may make our nomogram more reliable than other DIC predictors. Result from PCA analysis further showed these variables can clearly

distinguished DIC and non-DIC, which indicated that they may play important roles in the prediction of DIC.

Heatstroke directly affects platelet function and can induce organ function damage (15, 16). Previous studies reported that platelet abnormality and hypofibrinogenemia in heatstroke patients increases risk of multiple-organ dysfunction syndrome and heatstroke-induced coagulopathy, with the latter often progressing to DIC (17–19). Therefore, it was not completely surprising that we detected maximum amplitude, a measure of interaction between platelets and fibrinogen used in thromboelastography, as an independent risk factor for DIC (20, 21). We also found that low albumin level, elevated creatinine, high glutamic-pyruvic transaminase, and total bilirubin were positively related to the progression to DIC. Heatstroke patients suffer damage to the liver and kidney, and both organs produce hormones that affect coagulation homeostasis (22–24).

ROC analysis is a traditional method that evaluates the performance of a model (25). The predictive nomogram constructed in our study has a better ability for predicting DIC based on the value of AUC. However, an AUC alone is insufficient to determine that a model has good performance in improving decision-making. DCA and calibration curve were also introduced to estimate the clinical utility and predictive capacity of a nomogram, respectively (26, 27). Results showed the prediction model exhibited acceptable calibration and DCA gave the heatstroke population net benefit of nomogram at different threshold probabilities. Overall, the current predictive model exhibited good performance regarding DIC prediction.

Our model should be further developed and optimized in light of the fact that it is based only on the first 24 h in the intensive care unit, so it does not take into account dynamics in indicator levels. The model was developed with data from patients at a single medical center, so it should be validated in other patient populations.

This work establishes the feasibility of accurately predicting DIC in heatstroke patients on the basis of a few carefully selected clinical variables that are accessible to most medical centers. Our nomogram may become increasingly useful as the incidence of heatstroke increases worldwide.

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 the 908th Hospital of Logistic Support Force. Written informed consent for participation was not required for this study in accordance with the national legislation and the institutional requirements.

Author contributions

QZ, LZ, LH, QL, and JS: acquisition of data. QZ, LZ, and NZ: statistical analysis. QZ: drafting manuscript. JS: manuscript revision. All authors read and approved the final manuscript.

Funding

This study was funded by Chinese Medicine Education Association (No. 2022KTZ013) and Foundation Reinforce Project of Chinese PLA (No. 2022-JCJQ-ZD-097-11). The funders were not involved in research design, data collection, and manuscript preparation.

Acknowledgments

We thank the patients for participating in our study.

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. Liu SY, Song JC, Mao HD, Zhao JB, Song Q. Expert consensus on the diagnosis and treatment of heat stroke in China. *Mil Med Res.* (2020) 7:1. doi: 10.1186/s40779-020-00266-4

2. Hoover J. Heatstroke. *N Engl J Med.* (2019) 381:1186. doi: 10.1056/NEJMc1909690

3. Iba T, Connors JM, Levi M, Levy JH. Heatstroke-induced coagulopathy: biomarkers, mechanistic insights, and patient management. *EClinicalMedicine.* (2022) 44:101276. doi: 10.1016/j.eclinm.2022.101276

4. Allan RP, Barlow M, Byrne MP, Cherchi A, Douville H, Fowler HJ, et al. Advances in understanding large-scale responses of the water cycle to climate change. *Ann NY Acad Sci.* (2020) 1472:49–75. doi: 10.1111/nyas.14337

5. Hifumi T, Kondo Y, Shimazaki J, Oda Y, Shiraishi S, Wakasugi M, et al. Prognostic significance of disseminated intravascular coagulation in patients with heat stroke in a nationwide registry. *J Crit Care.* (2018) 44:306–11. doi: 10.1016/j.jcrc.2017.12.003

6. Huisse MG, Pease S, Hurtado-Nedelec M, Arnaud B, Malaquin C, Wolff M, et al. Leukocyte activation: the link between inflammation and coagulation during

heatstroke. A study of patients during the 2003 heat wave in Paris. *Crit Care Med.* (2008) 36:2288–95. doi: 10.1097/CCM.0b013e318180dd43

7. al-Mashhadani SA, Gader AG, al Harthi SS, Kangav D, Shaheen FA, Bogus F. The coagulopathy of heat stroke: alterations in coagulation and fibrinolysis in heat stroke patients during the pilgrimage (Haj) to Makkah. *Blood Coagul Fibrinolysis.* (1994) 5:731–6. doi: 10.1097/00001721-199410000-00009

8. He L, Lin Q, Zhong L, Zeng Q, Song J. Thromboelastography maximum amplitude as an early predictor of disseminated intravascular coagulation in patients with heatstroke. *Int J Hyperthermia.* (2022) 39:605–10. doi: 10.1080/02656736.2022.2066206

9. Min J, Wan P, Liu G, Yu M, Su L. Sonoclot signature analysis: a new point-of-care testing method for defining heat stroke-induced coagulopathy. *Int J Gen Med.* (2021) 14:6925–33. doi: 10.2147/IJGM.S321982

10. Wan P, Tong H, Zhang X, Duan P, Tang Y, Su L. Diagnosis of overt disseminated intravascular coagulation in critically Ill adults by Sonoclot coagulation analysis. *Int J Hematol.* (2014) 100:125–31. doi: 10.1007/s12185-014-1601-3

11. Faustino E, Karam O, Parker R, Hanson S, Brandão L, Monagle P, et al. Coagulation dysfunction criteria in critically Ill children: the PODIUM consensus conference. *Pediatrics.* (2022) 149:S79–83. doi: 10.1542/peds.2021-052888L

12. Taylor F Jr, Toh C, Hoots W, Wada H, Levi M. Scientific subcommittee on disseminated intravascular coagulation (DIC) of the international society on thrombosis and haemostasis (ISTH). Towards definition, clinical and laboratory criteria, and a scoring system for disseminated intravascular coagulation. *Thromb Haemost.* (2001) 86:1327–30.

13. Li WJ, Sha M, Ma W, Zhang ZP, Wu Y, Shi DM. Efficacy evaluation of D-dimer and modified criteria in overt and nonovert disseminated intravascular coagulation diagnosis. *Int J Lab Hematol.* (2016) 38:151–9. doi: 10.1111/ijlh.12467

14. Ohbe H, Isogai S, Jo T, Matsui H, Fushimi K, Yasunaga H. Treatment with antithrombin or thrombomodulin and mortality from heatstroke-induced disseminated intravascular coagulation: a nationwide observational study. *Semin Thromb Hemost.* (2019) 45:760–6. doi: 10.1055/s-0039-1700520

15. Iba T, Helms J, Levi M, Levy J. The role of platelets in heat-related illness and heat-induced coagulopathy. *Thromb Res.* (2022). doi: 10.1016/j.thromres.2022.08.009 [Epub ahead of print].

16. Pease S, Bouadma L, Kermarrec N, Schortgen F, Régnier B, Wolff M. Early organ dysfunction course, cooling time and outcome in classic heatstroke. *Intensive Care Med.* (2009) 35:1454–8. doi: 10.1007/s00134-009-1500-x

17. Zhong L, Wu M, Ji J, Wang C, Liu Z. Association between platelet levels on admission and 90-day mortality in patients with exertional heatstroke, a 10 years cohort study. *Front Med.* (2021) 8:716058. doi: 10.3389/fmed.2021.716058

18. Zhong L, Wu M, Wang C, Yu B, Liu Z, Liu Z. Clinical characteristics and outcomes of patients with severe heatstroke complicated with disseminated intravascular coagulation: a case-control study. *Thromb Res.* (2021) 197:120–3. doi: 10.1016/j.thromres.2020.11.009

19. Gader AM. Heat stroke and platelets. *Platelets.* (1992) 3:125–8.

20. Othman M, Kaur H. Thromboelastography (TEG). *Methods Mol Biol.* (2017) 1646:533–43. doi: 10.1007/978-1-4939-7196-1_39

21. Bowbrick VA, Mikhailidis DP, Stansby G. Influence of platelet count and activity on thromboelastography parameters. *Platelets.* (2003) 14:219–24. doi: 10.1080/0953710031000118849

22. Becker JA, Stewart LK. Heat-related illness. *Am Fam Physician.* (2011) 83:1325–30.

23. Wei D, Gu T, Yi C, Tang Y, Liu F. A nomogram for predicting patients with severe heatstroke. *Shock.* (2022) 58:95–102. doi: 10.1097/SHK.0000000000001962

24. Celep RB, Unlu BS. Retrospective platelet values measurement. Is it acceptable to discuss? *Eur Rev Med Pharmacol Sci.* (2014) 18:1108.

25. Iasonos A, Schrag D, Raj G, Panageas K. How to build and interpret a nomogram for cancer prognosis. *J Clin Oncol.* (2008) 26:1364–70. doi: 10.1200/JCO.2007.12.9791

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. Alba AC, Agoritsas T, Walsh M, Hanna S, Iorio A, Devereaux PJ, et al. Discrimination and calibration of clinical prediction models: users' guides to the medical literature. *JAMA.* (2017) 318:1377–84. doi: 10.1001/jama.2017.12126

frontiers | Frontiers in Medicine

TYPE Systematic Review
PUBLISHED 31 March 2023
DOI 10.3389/fmed.2023.1109411

OPEN ACCESS

EDITED BY
Hong Wang,
Central South University,
China

REVIEWED BY
Aljoscha Burchardt,
German Research Center for Artificial Intelligence (DFKI),
Germany
Prabhishek Singh,
Bennett University,
India
Xuewei Cheng,
Central South University,
China

*CORRESPONDENCE
Sobhan Moazemi
✉ sobhan.moazemi@med.uni-duesseldorf.de

†These authors share last authorship

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

RECEIVED 27 November 2022
ACCEPTED 10 March 2023
PUBLISHED 31 March 2023

CITATION
Moazemi S, Vahdati S, Li J, Kalkhoff S, Castano LJV, Dewitz B, Bibo R, Sabouniaghdam P, Tootooni MS, Bundschuh RA, Lichtenberg A, Aubin H and Schmid F (2023) Artificial intelligence for clinical decision support for monitoring patients in cardiovascular ICUs: A systematic review.
Front. Med. 10:1109411.
doi: 10.3389/fmed.2023.1109411

Artificial intelligence for clinical decision support for monitoring patients in cardiovascular ICUs: A systematic review

Sobhan Moazemi[1]*, Sahar Vahdati[2], Jason Li[2], Sebastian Kalkhoff[1], Luis J. V. Castano[1], Bastian Dewitz[1], Roman Bibo[1], Parisa Sabouniaghdam[3], Mohammad S. Tootooni[4], Ralph A. Bundschuh[5], Artur Lichtenberg[1], Hug Aubin[1†] and Falko Schmid[1†]

[1]Digital Health Lab Düsseldorf, Department of Cardiovascular Surgery, Medical Faculty and University Hospital Düsseldorf, Düsseldorf, Germany, [2]Institute for Applied Informatics (InfAI), Dresden, Germany, [3]Department of Computer Science, Heinrich-Hertz-Europakolleg, Bonn, Germany, [4]Department of Health Informatics and Data Science, Loyola University Chicago, Chicago, IL, United States, [5]Nuclear Medicine, Medical Faculty, University Augsburg, Augsburg, Germany

Background: Artificial intelligence (AI) and machine learning (ML) models continue to evolve the clinical decision support systems (CDSS). However, challenges arise when it comes to the integration of AI/ML into clinical scenarios. In this systematic review, we followed the Preferred Reporting Items for Systematic reviews and Meta-Analyses (PRISMA), the population, intervention, comparator, outcome, and study design (PICOS), and the medical AI life cycle guidelines to investigate studies and tools which address AI/ML-based approaches towards clinical decision support (CDS) for monitoring cardiovascular patients in intensive care units (ICUs). We further discuss recent advances, pitfalls, and future perspectives towards effective integration of AI into routine practices as were identified and elaborated over an extensive selection process for state-of-the-art manuscripts.

Methods: Studies with available English full text from PubMed and Google Scholar in the period from January 2018 to August 2022 were considered. The manuscripts were fetched through a combination of the search keywords including AI, ML, reinforcement learning (RL), deep learning, clinical decision support, and cardiovascular critical care and patients monitoring. The manuscripts were analyzed and filtered based on qualitative and quantitative criteria such as target population, proper study design, cross-validation, and risk of bias.

Results: More than 100 queries over two medical search engines and subjective literature research were developed which identified 89 studies. After extensive assessments of the studies both technically and medically, 21 studies were selected for the final qualitative assessment.

Discussion: Clinical time series and electronic health records (EHR) data were the most common input modalities, while methods such as gradient boosting, recurrent neural networks (RNNs) and RL were mostly used for the analysis. Seventy-five percent of the selected papers lacked validation against external datasets highlighting the generalizability issue. Also, interpretability of the AI decisions was identified as a central issue towards effective integration of AI in healthcare.

KEYWORDS

artificial intelligence (AI), machine learning (ML), clinical decision support (CDS), cardiovascular, intensive care unit (ICU), patient monitoring, explainable AI (XAI)

1. Introduction

Complications due to clinical deterioration and medical errors are often caused by human error, either due to forgetfulness, inattention, or inexperience and are far greater than technical failures (1, 2). Furthermore, intensive care units (ICUs) are prominent sources of large bulk of data collected from each patient. For the special case of cardiovascular ICU patients who mostly attribute higher complication rates and longer ICU stays (3, 4), it becomes even more challenging for the medical staff to spot certain complications or symptoms of patients. Considering the promising impact of artificial intelligence (AI) for clinical decision support (CDS) (5, 6), implementing AI into the cardiovascular ICUs could help minimize the number of medical errors by being able to guide the clinician to the correct diagnosis and ultimately to an appropriate therapy.

In the context of medical AI, the two major disciplines of Medicine and AI need to come together. Recent discoveries in medicine and medical technology as well as new advancement in AI modeling and computational power increased the application of ML-based methodologies in healthcare domains, such as disease diagnosis, prognosis and treatment planning (7–10), and overall/disease-free survival prediction (11–13). In particular, in intensive patient monitoring, AI methods have been used for different purposes such as prediction of readmission (3, 14–16) and sepsis (17–19) and mortality risk assessment (20, 21).

Despite the large body of evidence illustrating the promising relevance of AI methodologies in medical domains, there are some common challenges which limit the integration of AI-based methodologies in daily routines. For instance, trained classifiers may make biased predictions due to various sources of bias, such as gender bias, present in medical datasets (22, 23). Another challenge is the 'black box' nature of most of the modern deep and recurrent neural network models, which necessitates solutions to address explainability of these methods when applied to medical domains (24). Furthermore, ensuring consistency between the characteristics of open access data sets used for training and real clinical data is crucial for the successful integration of AI in intensive care routine practice (25). We aimed to draw attention to the limitations stemming from bias, interpretability, and data set shift issues, which expose a gap in the integration of AI in clinical decision making. This gap is mostly caused by medical staff's lack of trust in AI.

There are already a number of impactful articles which closely relate to the current systematic review. Fleuren et al. (26) conducted a systematic review and meta-analysis of AI models to predict sepsis onset in different wards including normal, emergency and ICU stations. Although their findings illustrate that ML models can achieve high accuracy in predicting sepsis in their corresponding experimental setups and might be considered as alternatives to some established scoring systems in clinical routines, they identify a lack of systematic reporting and clinical implementation studies in the domain which should be overcome in the future. Giordano et al. (27) argued that patient risk stratification and patient outcome optimization would be the first venues in which AI can practically contribute to routine practices. However, the mentioned work emphasizes the necessity for medical staff to receive extracurricular training on the mechanics of AI decision making and improved interpretability. This can ultimately lead to increased trust in AI in healthcare scenarios. Syed et al. (28) identified that predicting mortality, sepsis, acute kidney injury (AKI), and readmissions were the most common tasks for applied AI in patient monitoring in ICUs. Greco et al. (29) identified inconsistencies in diagnosis and treatment protocols between different health centers and countries as well as the lack of emotional intelligence to be the most critical aspects which confine the successful integration of AI driven approaches for patient monitoring. Antoniadi et al. (24) addressed interpretability as one of the most critical issues towards integration of ML-based approaches for CDS, identifying tabular data processing XAI-enabled systems and XAI-enabled CDS tools for text analysis as the most and the least common approaches in the literature, respectively. Also, Yang et al. (30) addressed the medical XAI aspects in multi-modal and multi-center scenarios in a mini-review study. They further showcased an XAI framework integrated for automated classification of corona virus disease (COVID)-19 patients and ventricle segmentation using computed tomography (CT) and magnetic resonance imaging (MRI) scans. Finally, Abdellatif et al. (31) reviewed the applications of reinforcement learning (RL) for intelligent healthcare (I-Health) systems, focusing on large networks of Internet of mobile things (IoMT) and software defined networks (SDNs) producing big data. In the realm of this evolving field, our work distinguishes itself by emphasizing the strategies and knowledge necessary to bridge the gap and successfully integrate AI for clinical decision support in daily intensive care routines, with a particular focus on cardiac diseases.

In this systematic review, following the PRISMA (32) and PICOS (33) guidelines, we designed the study in four steps including: identification of initial manuscripts through search engine queries and subjective searches, screening of original articles upon availability of full text in English, eligibility with regard to domain of interest and technical significance as well as medical relevance of the studies. We considered the most well-known publisher databases in the clinical and medical research domains to search and select high quality original research articles. We mainly focused on shortlisting the works that aimed at analyzing the applications of AI-assisted methodologies for automated patient monitoring in cardiovascular ICUs. We further analyzed most common data types as well as mostly applied AI algorithms for decision support in patient monitoring. The main contributions of this manuscript can be listed as following:

- Performing a systematic review over patient monitoring articles following PRISMA and PICOS guidelines
- Covering the technical foundations according to the medical AI life cycle (34)

- Providing an extensive factual and narrative analysis of the selected articles
- Providing expertise from both data science and medical science points of view
- Discussing limitations and insights for the successful integration of AI-driven methods for decision making in cardiac ICUs
- Recommending additional standardization and risk of bias criteria applicable to novel medical AI tools with regards to generalization and external validation aspects.

In the next sections, first, we discuss the basic concepts which are fundamental to be able to follow the reported findings from the selected articles. The Methods section provides the details on the screening and selection criteria of the papers followed by the Results and Discussion sections which provide a comprehensive outline of the findings from the selected contributions. Finally, a short conclusion of the findings is given.

2. Background and fundamental concepts

According to the best practices (34), the life cycle of medical AI includes (a) model development and evaluation, (b) data creation and collection, and (c) AI Safety. Therefore, we covered the current state of the methods used in major related work, the data used in the studies, and the recent advances in the interpretability and explainable AI for medicine. The rest of this section briefly describes some of the most important concepts in these three aspects which are critical for better understanding of the topics that are covered in the next sections. Note that, the choice of methods which are discussed in this section reflects the methodology implemented in the selected articles as a result of the systematic review process.

2.1. Common AI methods applied to clinical data for patient monitoring

From a high-level perspective, machine learning (ML) techniques can be categorized in three main groups: supervised, unsupervised and reinforcement learning. If ground truth labels are available and used to train and fit the model (e.g., binary classification using known classes), the model corresponds to supervised ML paradigm. Otherwise, if the model is trained without prior knowledge on the target variable (e.g., clustering), the model corresponds to unsupervised ML paradigm. Another ML paradigm that has been frequently used for clinical decision support is Reinforcement Learning (RL).

Reinforcement learning: In RL, a computational agent is trained to maximize the cumulative reward it receives over a series of time-steps by taking observations of the current state of the environment and by evaluating the feedback it receives after taking an action in that state (35). More formally, RL is founded on a Markov Decision Process (MDP) (36), where the RL agent is trained to learn an optimal policy pi^* that maximizes the cumulative reward by exploring the environment defined by $p(s, a, s')$ and r, and exploiting its knowledge of the environment represented by V_pi or Q_pi and y.

There is a long history of clinical decisions being formulated as an MDP. Initial efforts in this direction focused on dynamic programming solutions, while in recent work, variations of the Q-Learning algorithm have become more prominent, such as fitted-Q-iteration (FQI) (37) or deep Q-networks (DQN) (38). Areas where RL has been applied, that are relevant for cardiovascular monitoring include targeting of measurements during monitoring and choosing, timing and dosing of treatment steps. Many diagnostic and prognostic tasks in the healthcare domain are facilitated through the use of a variety of supervised ML models including logistic regression (LR), support vector machines (SVM), and ensemble methods such as random forest (RAF) and extra trees (39–42). This group of AI algorithms are often applied on time-independent tabular patient information. For textual, higher dimensional data, and grid like data types such as time series data and medical images, natural language processing (NLP), deep learning, convolutional neural networks (CNNs), and recurrent neural networks (RNNs) models are widely applied (25, 43, 44). It is quite common in this domain that basic classifiers such as LR and decision tree based methods are applied to simplified representations of datasets to provide baselines for comparison to more sophisticated methods (3, 14).

Logistic regression: As a supervised ML algorithm, logistic regression (LR) (45) is a predictive model leveraging the concept of probability to solve binary classification problems. Fundamentally, LR is a linear regression model with a special type of activation function, the so-called sigmoid function or logistic function which, based on a given decision boundary, quantifies the probability of belonging to each of the binary labels.

Support vector machines: Support vector machines (SVMs) (46) is a supervised ML algorithm that aims to find the optimal hyperplane which separates data points in one, two, or multi-dimensional space, depending on the complexity of the feature space. To maximize the probability of true classification of unseen data points, the chosen hyperplane has to expose the maximum possible distance, i.e., margin, between the data points of different classes, increasing the impact of the data points locating nearest to the hyperplane (i.e., support vectors).

Decision trees and ensemble algorithms: Decision trees employ tree-structured flowcharts of decisions based on the values of the input features to solve classification problems (47). At each node of such trees, a decision is made based on a single feature whether to make the final prediction or make another decision based on another feature. The leaves of a decision tree are the target labels. Ensemble algorithms such as random forest (RAF) (48) apply different randomized groups of decision trees, denoted as ensembles of trees, as well as different bootstrapping mechanisms to come up with the final decision on the target labels.

Gradient boosting and categorical boosting: Gradient boosting, which is used for classification and regression tasks, draws predictions as ensembles of some weak learners, mostly decision trees or random forests (49). When it comes to the analysis of categorical data, categorical boosting or CatBoost algorithm outperforms other gradient boosting methods (50).

Recurrent neural networks: In contrast to conventional feed-forward neural network models which are mostly used for processing time-independent datasets, RNNs are well-suited to extract non-linear interdependencies in temporal and longitudinal data as they are capable of processing sequential information, taking advantage of the notion of hidden states h. In such a model,

at each timestamp *t,* the input data is processed alongside the information which was processed in the previous timestamp *t-1* (51). Also, for patient monitoring, a variety of RNN-based models such as long short-term memory (LSTM) and gated recurrent unit (GRU) are commonly applied.

First introduced by Hochreiter and Schmidhuber in 1997, LSTM (52) aims at identifying both short-term and long-term dependencies in the sequential data such as clinical time series data. LSTMs consist of cells with input, output, and forget gates which regulate the flow of information to remember values over arbitrary time intervals.

Natural language processing: When it comes to automated processing of textual patient data, such as electronic health records (EHRs), natural language processing (NLP) comes into action. NLP [Allen2003] denotes the set of AI based approaches which are capable of identifying underlying patterns in the textual data, hence understanding human languages. Taking the examples of EHRs and temporal textual patient information stored in medical databases such as Medical Information Mart for Intensive Care (MIMIC) (53, 54), clinical and medical domains also take advantage of NLP (15).

2.2. Established conventional scoring systems used in critical care

Alongside continuous monitoring of patients by the intensivists and medical staff during patients stays at ICUs, several scoring systems are widely used in critical care units to monitor and manage patients states such as the acute physiologic and chronic health evaluation (APACHE), the sequential organ failure assessment (SOFA), and the mortality prediction model (MPM) (55–57). Such scoring systems become handy in studies which aim at analyzing emerging AI methods for clinical decision making as they provide established baselines for comparison.

Mortality Prediction Models (MPMs) (56) *and APACHEs* (55) are mathematical models that estimate the probability of death for critically ill patients in ICUs based on patient data such as demographics, diagnoses, and physiological measurements. Each of MPM and APACHE use a different set of variables and algorithms to predict mortality risk. These models are useful in guiding clinical decision-making, evaluating ICU performance, and identifying risk factors for mortality. However, they have limitations and should be used alongside clinical judgment as they are not designed to replace it or provide definitive prognoses. The accuracy of MPMs may vary depending on the patient population and the specific model used, and they should be validated and calibrated before use in clinical practice.

Sequential Organ Failure Assessment (SOFA) (57) is a scoring system used to track the progression of organ dysfunction in critically ill patients in the intensive care units. It is based on the evaluation of six organ systems: respiratory, cardiovascular, hepatic, renal, coagulation, and neurological, with the score ranging from 0 to 4 for each organ system, and higher scores indicating greater dysfunction. The total SOFA score is the sum of the scores for all six organ systems, ranging from 0 to 24, and is calculated daily for each patient in the ICU. SOFA score is often used in clinical research and quality improvement initiatives in ICUs, and it has been shown to be a useful predictor of mortality in critically ill patients.

2.3. Medical data modalities for intensive patient care

From a general perspective, one can subdivide medical data modalities into the following subgroups: structured data (with and without timestamp) and unstructured data such as medical image modalities and electronic health records (EHR). Like other fields of data science, numerical tabular information such as patient demographic information (e.g., age and weight) can be used to form feature vectors for AI- and ML-based methods. In case of time-dependent measurements such as lab values and vital signs, the dimension of time (i.e., timestamp) should be integrated in the corresponding analysis pipeline, hence the clinical time series data. This section provides a brief overview of different data modalities used in the scope of this systematic review.

Numerous kinds of data in diverse modalities are processed by medical experts and intelligent systems for patient monitoring in ICUs. Clinical time series and electrocardiograms (ECGs) are among the most common types of data applied in this domain. Furthermore, open access databases facilitate objective performance analyses of the implemented AI methods.

Clinical time series data: Continuous patient monitoring leads to a magnitude of measurements captured and stored at discrete timestamps. Regardless of the disease type, a variety of temporal datasets such as Electronic health records (EHR), lab values, vital signs, diagnoses and treatments records can be used for patient monitoring (58).

Electrocardiograms (ECGs): First invented by William Eindhoven in 1902, electrocardiograms (ECGs) (59) are recorded non-invasively from the patient's body surface and are used to represent the heart's electrical activity. ECGs are widely applied for diagnosing heart complications also in cardiac ICUs.

2.3.1. Open access datasets

Ensuring that methodology can be replicated is a key consideration in data science, which typically necessitates the sharing of data. However, in the medical and clinical field, there are often additional ethical limitations and considerations when it comes to sharing patient data, which is considered highly sensitive and confidential. These ethical concerns must be balanced with the need for reproducibility in research. This highlights the importance of open access datasets for medical and clinical research. This subsection briefly introduces some of the most applied publicly available datasets for intensive patient care.

One of the majorly used information platforms in biomedical research and education is PhysioNet which offers free access to large collections of physiological and clinical data and related open-source software, and educational tutorials (60). Among the recently published extensive clinical data collections that are present in PhysioNet, datasets of High time-Resolution ICU Dataset (HiRID) (61), Medical Information Mart for Intensive Care (MIMIC-II, MIMIC-III and MIMIC-IV) (53, 54, 62), and eiCU (63) are the ones majorly used for studies about intensive care units.

MIMIC is a public database of de-identified electronic health records of over 60,000 adult patients admitted to the intensive care units at the Beth Israel Deaconess Medical Center. It contains information on demographics, diagnoses, laboratory tests, medications, and clinical notes collected from various sources such as

bedside monitors, clinical documentation, and hospital information systems. The database has been widely used in clinical research and machine learning applications to develop predictive models, identify risk factors, and improve clinical outcomes. Access to the database requires an application process and approval from the Institutional Review Board at BIDMC, but it is publicly available through PhysioNet, a repository of physiological data and clinical information maintained by MIT.

Intensive care units (ICU) are a prominent source of time series data, as the nature of intensive care usually requires close and regular monitoring of patients and thereby produce a high density of measurements. Instances of time-dependent measurement data that can be found in publicly available ICU datasets include time-stamped nurse-verified physiological measurements such as hourly documentation of heart rate, arterial blood pressure, or respiratory rate. Other examples include documented progress notes by care providers, continuous intravenous drip medications, and fluid balances (53).

2.4. Interpretability and explainability of AI in healthcare

Usually, in intensive patient care, the mission of AI systems is to provide risk estimates and assist in decisions by providing predictions, which then need to be understood, interpreted and validated by clinicians. To assess the trustworthiness, the AI developers together with clinicians have different sorts of higher-order evidence at hand (64). Most importantly, as identified by related work (24) and discussed in some of the selected manuscripts (25, 65–67), before an AI system is being implemented in clinical settings, it is being technically and clinically validated. The validation yields evidence of a system's accuracy and reliability through a standard procedure. Besides these evaluations, it is important to transfer the knowledge about what the AI system has focused its attention on through some *post hoc* explanations. This AI transparency is crucial in medical AI, especially in the use case of patient monitoring (68). Transparency refers to algorithmic procedures that make the inner workings of a 'black box' algorithm interpretable to humans (69). Another factor is traceability that intersects with the concepts of method and results in reproducibility and replicability of underlying data analysis. Covering these aspects relates to providing sufficient detail about procedures and data so that the same procedures could be exactly repeated. Auditability of AI shapes itself more and more as a necessary tool in achieving innovation in a secure, transparent way.

To interpret decisions made by AI models with deep architectures and to cope with their 'black box' nature, recursive feature elimination (RFE) and SHapley Additive exPlanations (SHAP) methods are commonly applied also in the medical AI domain. RFE takes an ML classifier and the desired number of features as input and starts from the entire input feature set. Then at each recursion step, the features are ranked based on an importance metric and the least relevant variables are removed. This procedure continues until the desired number of features are chosen (70). Inspired by game theory, SHAP is used to explain the output of any machine learning model by connecting optimal credit allocation with local explanations, assigning each input feature an importance value for a particular prediction (71). Nevertheless, the explainability provided by most of conventional methods such as RFE and SHAP is rather located on model level and addresses understanding of how a model derives a certain result, lacking the semantic context which is required for providing human-understandable explanations. In medical applications, the quest for explainability is usually motivated by medical semantic understanding, thus explainability on e.g., syndrome level which is the language of physicians (72).

3. Methods

3.1. Search strategy and screening

We followed the preferred reporting items for systematic reviews and meta-analyses (PRISMA) (32) and the population, intervention, comparator, outcome, and study design (PICOS) (33) guidelines. However, as meta-analysis was not originally intended for this study, we only followed the parts of PRISMA that only apply to systematic reviews. As this had led to a group of studies covering a diverse selection of datasets and algorithms, a comprehensive meta-analysis was not feasible. From the PubMed and Google Scholar databases, the following keywords are searched: ("artificial intelligence" OR "AI" OR "machine learning" OR "ML") AND ("ICU" OR "intensive care" OR "intensive care unit" OR "intermediate care unit" OR "IMC" OR "IMU" OR "patient monitoring") AND ("cardiovascular" OR "cardiac"). Moreover, a subjective literature research according to most relevant related studies complement results of the search engine queries. The publications dated from January 2018 to August 2022.

In the screening phase, original studies focusing on clinical decision support for adult subjects (age $\geq$17 years) visiting cardiovascular ICUs were analyzed. Thus, studies focusing on pediatric cohorts and review articles were removed from the results of search in the screening process. The summary of PICOS scheme containing the inclusion as well as exclusion criteria is outlined in Table 1.

3.2. Quality assessment, selection criteria, and risk of bias assessment

All the papers collected as results of search engine queries were assessed whether they held enough significance and relevance from both data science and medical points of view. First, each of the papers underwent qualitative reviews by two independent reviewers which were selected randomly from a group of reviewers with data science and AI background. In case of agreement about selecting the manuscript between the two assigned reviewers, the manuscript would be short-listed or eliminated from the systematic review accordingly. On the contrary, in case of a mismatch between the assessments carried out by the first two reviewers, a third reviewer with higher qualification would decide whether to select or reject the manuscript. Consecutively, the selected papers underwent another assessment step by a group of medical experts whether they fit within the scope of this study: patient monitoring in cardiovascular ICUs. The technical criteria to assess the manuscripts qualitatively include proper research concept, representative train/test cohorts, and proper cross-validation either within the dataset or against external cohorts.

TABLE 1 Population, intervention, comparator, outcome, and study design (PICOS) criteria for the systematic review.

Parameter	Inclusion criteria	Exclusion criteria
Population	• Adults (age ≥ 17) • Patients admitted to cardiovascular ICU	• Age < 17 • No cardiovascular patients • No ICU patients
Intervention	Any	No restriction
Comparator	• At least one AI/ML algorithm • At least one control group	• No AI/ML algorithm • No control group
Outcomes	Any	No restriction
Study designs	• Retrospective, prospective, or ambispective data analysis • Hold enough data scientific significance • Hold enough medical relevance	• No proper statistical analysis significance • No proper cross-validation • No enough medical relevance

To visualize the risk of bias assessment results, the robvis package (73) is used. As the criteria for risk of bias, the following seven items have been considered: reasonable cohort size (D1), proper cross-validation (D2), external validation set (D3), blinding of participants and personnel (D4), blinding of outcome assessment (D5), incomplete outcome data (D6), and selective reporting (D7). To account for subjectivity, the bias assessment was conducted with the same approach as for the study selection, i.e., with random assignments to two reviewers followed by a final validation by a third expert.

4. Results

In this section, the results of the systematic review are elaborated. First, a summary of the screening step is given followed by narrative reviews of the selected papers. Afterwards, a comprehensive analysis of the papers is provided which comprise a risk of bias analysis and assessments of studies outcomes, used datasets, and applied algorithms. Furthermore, if existing, relevant discussions on the integration of AI in cardiovascular ICUs are reported.

4.1. Study selection

The search engine queries have resulted in 89 papers in total. Out of these papers, 60 were from PubMed database and 25 were from Google Scholar. Another four papers were selected from subjective literature research from most relevant related articles. In the screening phase, 12 papers were excluded due to not available full text and three studies were excluded because of being review articles. In the eligibility assessment step, 11 papers were eliminated as they analyzed non-adult cohorts, 27 studies were excluded as considered not to be of proper significance from data science point of view, and 15 papers eliminated because they did not particularly focus on cardiovascular ICU cohorts (see Figure 1). As a result, 21 papers have been selected for the qualitative and quantitative analyses.

4.2. Summary of the included studies

Table 2 provides a summary of the important contents of the 21 included papers. This subsection presents a narrative review of these studies.

Zhao et al. (65) integrated a categorical boosting ML model to predict extubation failure resulting in in-hospital or 90-day mortality in patients visiting ICUs. To train their model, they used clinical time series data from the MIMIC-IV database. For the test purposes, they applied an external data set. To identify the most important predictive factors, they applied RFE and SHAP methods. Their results suggest that critically ill patients might benefit from AI assisted mechanical ventilation. They also provide an UI for model validation which is freely accessible online. They mention interpretability and inconsistency in train and test datasets as the most critical challenges towards integrating AI in clinical practice.

Jentzer et al. (66) used multivariate logistic regression on numerical clinical variables extracted from ECGs from their own facilities to quantify mortality risk due to left ventricle systolic dysfunction in patients staying at ICUs. Their findings suggest the relevance of the AI-driven methodology for the quantification of cardiac patients' survival potential and identify lack of explainability as a challenge to be handled before it can be integrated in prognostic pipelines.

Gandin et al. (74) investigated the interpretability of an RNN model with long short-term memory (LSTM) architecture as used for survival prediction in a cohort of patients visiting cardiovascular ICUs. They analyzed the MIMIC-III dataset for both training and test purposes. The results of their study demonstrate that incorporating an attention layer into the LSTM model can enhance the interpretability of the AI model's decisions, leading to greater reliability in AI decision making.

Andersson et al. (67) took advantage of artificial neural networks (ANNs) to anticipate neurological outcomes due to out-of-hospital cardiac arrest (OHCA). They analyzed clinical variables and biomarkers from a cohort of patients from their own hospital and used SHAP method for identifying the most relevant factors. They showed that the clinical parameters captured in the first 3 days of ICU stay contribute to OHCA prognostication. Although their results suggest

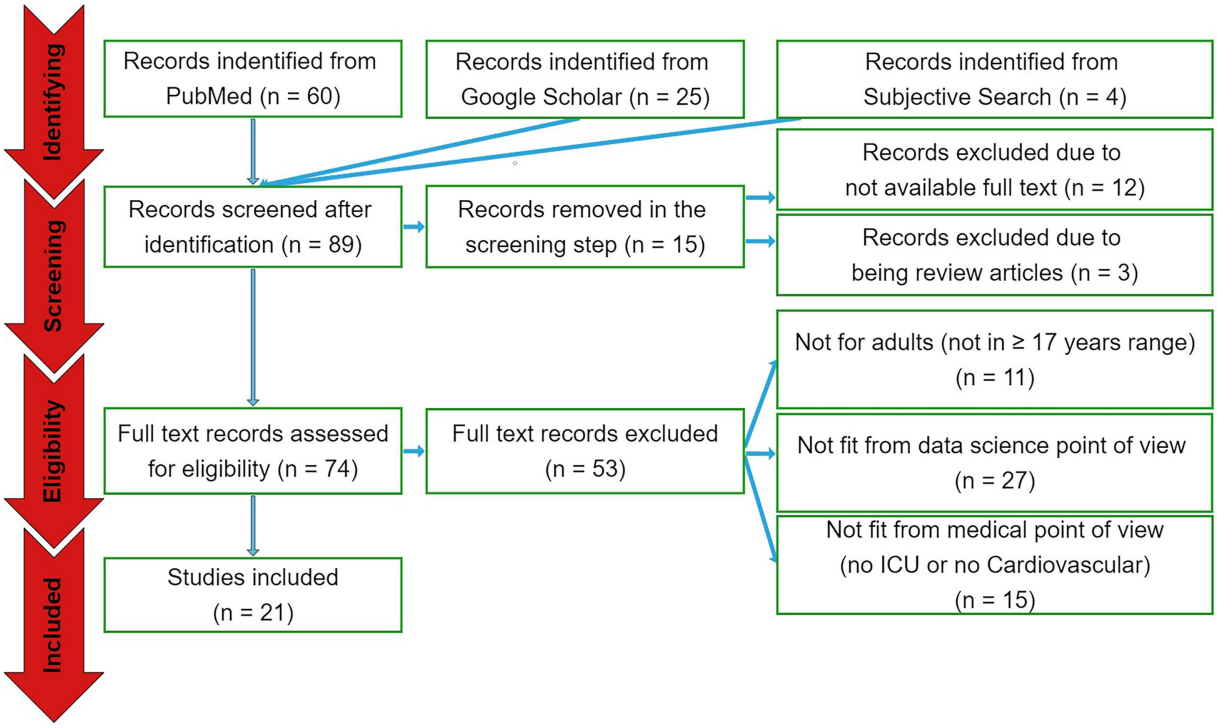

FIGURE 1
The PRISMA diagram. From a total of 89 papers identified by the search queries from the three sources, 15 and 53 papers were excluded in the screening and eligibility assessment phases, respectively. Accordingly, 21 papers were included to be reported.

reliable predictions, they insist on external validation with larger cohorts to assess generalizability of their methods.

Parsi et al. (39) took advantage of supervised machine learning methods such as support vector machines (SVM) to analyze data extracted from ECGs to predict paroxysmal atrial fibrillation in ICU patients with high accuracy. For their training and test, they applied open access data from the atrial fibrillation prediction database (AFPDB) of PhysioNet. Their primary contribution involves integrating an AI model with high performance onto implantable devices with low computational power.

Yu et al. (40) evaluated several ML models including logistic regression, random forest, and adaptive boosting (Ada) as applied to clinical time series data (from MIMIC-III database) for the prediction of long-term survival of patients after cardiac surgery, highlighting the significance of Ada model. As the generalizability plays an important role in integration of AI-assisted methods, they also provide a freely accessible online platform for the validation of their model against external sets of data.

To predict noninvasive ventilation (NIV) failure in cardiac ICU patients, Wang et al. (75) took advantage of categorical boosting alongside RFE and SHAP methods for analyzing most important factors among clinical time series data. They used open access data from the eICU-CRD database for training and data from their own hospital for test purposes. They have shown the relevance of the AI model and provide an online tool for model validation, while identifying lower specificity in predictions of AI as the most challenging issue which limits generalizability of their findings.

Chen et al. (41) analyzed different supervised ML classifiers (including logistic regression, SVM, random forest, artificial neural networks and XGBoost) for the task of predicting ventilator weaning in the next 24-h time windows, given non-time series clinical data corresponding to a cohort of cardiac ICU stays in their facilities. Their key finding is that ventilator weaning can be anticipated using a limited number of clinical factors such as expiratory minute ventilation, expiratory tidal volume, ventilation rate set, and heart rate. As they only applied data from their own center, generalizability of their findings remains in question.

Dutra et al. (76) applied a variety of statistical and ML methods including Cox and Kaplan–Meier estimators as well as ElasticNet (85) and survival trees to quantify mortality risks of ICU patients due to heart failure with mid-range ejection fraction (EF). Their findings suggest that there is no significant correlation between EF and survival probability of the patients. As they only analyzed data from a single center, their findings are subject to bias, hence the need for follow-up generalizability assessments.

Bodenes et al. (42) applied and compared AI classifiers such as k-NN, SVM, and decision trees to predict survival of the ICU patients due to heart rate variability (HRV). They analyzed clinical time series from a single center and proposed a low cost and efficient model for HRV analysis. However, their findings are subject to further assessments against external data cohorts. They also identified the lack of global standardization of HRV measurement methods and interpretability of AI models as limitations to overcome in the future.

Moazemi et al. (25) evaluated two alternative long short-term memory (LSTM)-based models to predict readmission risks in cohorts of cardiovascular ICU patients, analyzing clinical time series data as well as patient level information. They used a cohort of cardiac ICU stays from MIMIC-III as well as a dataset from their own hospital

TABLE 2 The summary of the included studies. The most important contents of the 21 studies are summarized.

Study	Population	Study designs	Predicted outcome(s)	Data type(s)	Method(s)	Main contribution(s)	Identified challenge(s) towards integration of AI in practice
Zhao et al. (65)	16,189 adult (age > 18) patients from MIMIC-IV	Retrospective training, prospective validation	Extubation failure	Clinical time series (MIMIC-IV and domestic)	Categorical boosting with SHAP and RFE	Well-performing AI model (up to 0.83 AUROC), increased interpretability, open access UI for model validation	Interpretability, dataset shift problem
Jentzer et al. (66)	11,266 adult (Mean age 68 ± 15 years) patients from Mayo Clinic ICU	Retrospective data analysis	Mortality risk	Numerical clinical data extracted from ECGs (domestic)	Multivariate logistic regression	Well-performing AI model (up to 0.83 AUROC)	Interpretability
Gandin et al. (74)	10,616 patients from MIMIC III	Retrospective data analysis	Mortality risk	EHR (MIMIC-III)	RNN (LSTM with attention layer)	Well-performing AI model (up to 0.79 AUROC), attention layer to increase the interpretability of LSTM	Interpretability and reliability
Andersson et al. (67)	932 adult (age ≥ 18) patients from 36 ICUs across Europe and Australia	Retrospective data analysis	Neurological outcome following out-of-hospital cardiac arrest (OHCA)	Clinical variables and biomarkers (domestic-multicenter)	ANN with SHAP	Reliable AI model (up to 0.94 AUROC) using cumulative clinical data from first 3 days of ICU stay	Generalizability, effect of outliers
Parsi et al. (39)	53 patients from PhysioNet	Retrospective data analysis	Paroxysmal atrial fibrillation	ECG (PhysioNet)	SVM, k-NN, RF, MLP	High performance AI (up to 0.79 accuracy) on implantable defibrillator with low computation power	Low computational power on wearable and implantable devices
Yu et al. (40)	7,368 adult (age > 18) patients from MIMIC-III	Retrospective data analysis	4-year mortality risk after cardiac surgery	Clinical time series (MIMIC-III)	LR, ANN, Ada, NB, RF, etc. with RFE	Well-performing AI model (up to 0.80 AUROC), open access UI for model validation	Generalizability
Wang et al. (75)	929 adult (age > 18) patients from eICU-CRD	Retrospective training, prospective validation	Noninvasive ventilation (NIV) failure	Clinical time series (eICU-CRD and domestic)	Categorical boosting with RFE and SHAP	Well-performing AI model (up to 0.87 AUROC) applied to easily available clinical variables, open access UI for model validation	Generalizability, low specificity of AI predictions
Chen et al. (41)	1,439 adult (mean age 65.05 ± 12.53 years) patients from Cheng Hsin General Hospital	Retrospective data analysis	Ventilator weaning time	Non-time series clinical data (domestic)	LR, SVM, RF, ANN, XGBoost	Well-performing AI model (up to 0.88 AUROC), identify most simplified key parameters	Generalizability
Dutra et al. (76)	519 adult (age > 18, mean age, 74.87 ± 13.56 years) patients admitted to a Brazilian cardiac ICU	Ambispective data analysis	Mortality risk from heart failure with mid-range ejection fraction (EF)	Non-time series clinical data (domestic)	Cox, Kaplan–Meier, ElasticNet, survival tree	EF is not significantly correlated with mortality	Generalizability

(Continued)

TABLE 2 (Continued)

Study	Population	Study designs	Predicted outcome(s)	Data type(s)	Method(s)	Main contribution(s)	Identified challenge(s) towards integration of AI in practice
Bodenes et al. (42)	540 adult patients admitted to Brest University Hospital's cardiac ICU	Prospective data analysis	Mortality risk and heart rate variability (HRV)	Clinical time series (domestic)	k-NN, SVM, LR, decision trees	Low cost and efficient AI model for HRV analysis	Generalizability, interpretability, lack of standardized HRV measurement methods
Moazemi et al. (25)	11,513 patients from MIMIC-III and 502 from University Hospital Düsseldorf's cardiac ICU (age ≥ 17)	Retrospective data analysis	ICU readmission	Clinical time series (MIMIC-III and domestic)	RNN (LSTM)	Well perforing AI (up to 0.82 AUROC), data-driven approach, validation with external cohort	Interpretability, dataset shift problem
Baral et al. (44)	7,611 patients (age > 15) from MIMIC-III cardiac ICUs	Retrospective data analysis	Cardiac arrest	Clinical time series (MIMIC-III)	Multi-layer perceptron (MLP), RNN (bidirectional LSTM)	Well-performing AI model (up to 0.94 AUROC) to reduce false alarm for cardiac arrest, improved model compared to normal LSTM	Generalizability
Qin et al. (43)	49,168 patients from MIMIC-III	Retrospective data analysis	Sepsis	Textual and structured clinical data (MIMIC-III)	NLP (BERT), Amazon Comprehend Medical for data processing, XGBoost (for classification)	Outperform PhysioNet's sepsis prediction challenge winner (up to 0.89 AUROC)	Generalizability
Nanayakkara et al. (77)	Adult (age ≥ 17) septic patients from MIMIC-III	Retrospective data analysis	Sepsis treatment planning	Clinical time series (MIMIC-III)	RL	Introducing a novel physiology-driven recurrent autoencoder, highly interpretable, uncertainty quantification	Lack of standardization, how/when AI is considered safe enough for clinical routine
Zheng et al. (78)	1,362 critically ill COVID patients (mean age 69.7) from New York University Langone Health	Retrospective data analysis	Managing oxygen flow rate to reduce mortality risk	EHR (domestic)	RL	AI model to identify optimal personalized oxygen flow rate to reduce mortality rate	Generalizability
Peine et al. (79)	61,532 and 200,859 ICU stays of adult patients from MIMIC-III and eICU datasets	Retrospective data analysis	Optimization of mechanical ventilation to reduce mortality risk	Clinical time series (MIMIC-III and eICU)	RL	Introduce VentAI to dynamically optimize mechanical ventilation for individual patients	Generalizability, algorithm bias, missing/false data
Akrivos et al. (80)	162 adult patients (18 < age < 90 on) from MIMIC-II	Retrospective data analysis	Cardiac arrest	Transformed clinical time series (MIMIC-II)	integrated model of sequential contrast patterns using Multichannel Hidden Markov Model	High sensitivity (with the average of 0.78) and specificity to identify high risk patients	False positive rate in classification results

(Continued)

TABLE 2 (Continued)

Study	Population	Study designs	Predicted outcome(s)	Data type(s)	Method(s)	Main contribution(s)	Identified challenge(s) towards integration of AI in practice
Aushev et al. (81)	75 adult (age > 18) patients from ShockOmics European database	Retrospective data analysis	Mortality due to septic and cardiogenic shock	ECG (ShockOmics Dataset)	SVM, Random Forest, RFE, Bayesian networks	Apply feature selection to identify the most relevant predictors of mortality due to septic and cardiogenic shock using ECG with high certainty (up to 0.84 AUROC)	–
Kim et al. (82)	29,181 adult (age > 18) ICU patients from Yonsei Health System (Severance and Gangnam Severance Hospitals)	Retrospective data analysis	Acute respiratory failure and cardiac arrest	Time series (domestic)	Deep Learning (LSTM)	Introduce FAST-PACE for preparing immediate intervention in emergency situations, outperforming some established scoring systems (e.g., SOFA) (up to 0.88 AUROC)	Lack of relevant input data to AI models, lack of external validation, imbalanced datasets, lack of real time measurements of vital signs
Meyer et al. (83)	11,492 ICU stays from 9,269 adult (age ≥ 18) patients from a German cardiovascular tertiary care center	Retrospective data analysis	Mortality, renal failure, postoperative bleeding leading to operative revision	Time series (domestic)	Deep learning (RNN)	Predict severe complications after cardiothoracic surgery with a higher certainty (up to 0.96 AUROC), validation against MIMIC-III dataset	Dataset shift, biased data, generalizability, transparency and interpretability of AI decision making
Yoon et al. (84)	2,809 Adult (age > 18) patients from MIMIC-II	Retrospective data analysis	Tachycardia as a surrogate for cardiorespiratory instability (CRI)	Vital signs time series (MIMIC-II)	Regularized logistic regression (LR), Random Forest	Developed a risk score for predicting tachycardia episodes, AI model with high accuracy (up to 0.86 AUROC)	Timestamp mismatching and data sparsity, specificity of predictions, lack of external validation

for train and test purposes, respectively. Their findings highlight the benefit of RNN models in general, and the need for consistency in train and validation cohorts in particular. They further highlight the dataset shift problem and interpretability of deep learning models as critical future challenges for AI in CDS.

Baral et al. (44) applied multi-layer perceptrons (MLP) and bidirectional LSTM models for the prediction of cardiac arrest and have shown the superiority of the enhanced bidirectional model to the normal LSTM. They analyzed a cohort of data from MIMIC-III for both training and test purposes. Their proposed RNN model showed reasonable performance in predicting cardiac arrest, reducing the false alarm rate significantly. As they did not validate their model with external data, their findings are subject to further generalizability assessments.

Qin et al. (43) applied Bidirectional Encoder Representations from Transformers (BERT) (86) and Amazon Comprehend Medical techniques (as natural language processing (NLP) approaches) to process textual data and XGBoost method to classify patients with high risk of sepsis. They leveraged open access and structured clinical data from the MIMIC-III database for training and test. Their proposed pipeline outperformed the winner of PhysioNet challenge for sepsis prediction in 2019 which had applied XGBoost and Bayesian optimization without processing textual data (87). However, their findings lack validation against independent external cohorts, hence the generalizability issue.

Nanayakkara et al. (77) took advantage of reinforcement learning approaches to introduce a novel recurrent autoencoder for the task of sepsis treatment planning. They used clinical time series data from the MIMIC-III database for their analysis which include interpretable uncertainty quantification of clinical factors. They further discussed the lack of globally agreed standards in the assessments of safeness of AI methodologies as one of the most critical challenges in the field.

Zeng et al. (78) also applied reinforcement learning methodologies to quantify the optimal personalized oxygen flow rate to minimize the risk of mortality in cardiac ICU patients. To this end, they analyzed electronic health record (EHR) data from cardiovascular patients' stays at their hospital in a single center study. Thus, their findings might be subject to future external validation.

In another study leveraging reinforcement learning methodologies, Peine et al. (79) introduced VentAI, an RL based pipeline for personalized optimization of mechanical ventilation in patients staying at cardiovascular ICUs. They analyzed clinical time series data from two open access databases (MIMIC-III and eICU) and identified generalizability, bias in AI algorithms, and

missing and false entries in the measured clinical parameters as the most important challenges towards integration of AI in clinical practice.

Applying regularized logistic regression and random forest algorithms to vital signs from MIMIC-II dataset, Yoon et al. (84) suggest that predicting tachycardia could increase clinical awareness of a higher risk of future hypotension and subsequently other forms of cardiorespiratory instability (CRI). But they did not directly compare their model to conventional scoring systems or conduct validation studies against independent sets of data.

Meyer et al. (83) applied a deep recurrent model to analyze time series data for the task of predicting severe complications in critical care units after cardiovascular surgery such as mortality, renal failure, and postoperative bleeding leading to operative revision. Their model outperforms clinical reference tools and is available to be integrated in EHR systems. They further validate the performance of their model which is trained using domestic data against external data from the MIMIC-III database and highlight the importance of generalizability and interpretability of AI methods in clinical practice.

Kim et al. (82) introduced Feasible Artificial Intelligence with Simple Trajectories for Predicting Adverse Catastrophic Events (FAST-PACE), an LSTM model to process clinical time series data, to predict events of acute respiratory failure and cardiac arrest. They fit their model using a domestic cohort of data and show the superiority of their model compared to some established scoring systems such as sequential organ failure assessment (SOFA) and mortality prediction model (MPM). Their findings further identify lack of external validation and inconsistencies in real time measurement schemes in critical care units as some limitations of data-driven approaches towards clinical decision making.

Aushev et al. (81) applied different feature selection techniques such as recursive feature elimination (RFE) in combination with SVM and random forest classifiers to identify most relevant features that could predict mortality due to shock in the intensive care unit. To this end, they analysed ECG data from ShockOmics dataset as part of an Europe funded project. As their patient cohort with 75 subjects is relatively small, their findings might be subject to further assessment.

Akrivos et al. (80) took advantage of the MIMIC-II dataset to integrate a model of sequential contrast patterns using the Multichannel Hidden Markov Model which is able to predict cardiac arrest in cardiovascular ICUs. Their approach takes advantage of clinical time series data after transforming them to sequential patterns. Their model achieves high performance, while suffering from a relatively low false positive rate in classifier predictions. This identifies rooms for follow-up studies including data from independent databases.

4.3. Risk of bias assessment

Figure 2 provides an overview of the risk of bias analysis results. Most of the studies conducted proper cross-validation methods. However, only five studies used independent external datasets for the validation of their models (Figure 3), which identifies lack of generalizability as a common issue towards integration of AI methodologies across different research groups and medical centers.

4.4. Studies' outcomes

As illustrated in Figure 4, mortality as well as cardiac, sepsis and respiratory complications rank amongst the most common clinical outcomes analyzed by the selected literature. This is justified as most of the patients visiting cardiovascular ICUs have had cardiac surgeries beforehand or are subject to higher cardiac and respiratory complications.

4.5. Analyzed data types

Figure 5 shows an overview of the data modalities analyzed in the selected papers. Clinical time series is the most common group, while EHR and textual data are the least common groups. Moreover, as presented in Table 2, 13 studies out of 21 selected studies utilized open access datasets with 10 studies using different versions of MIMIC database either for training or validation purposes.

4.6. AI algorithms and models

Figure 6 outlines the AI methods for model development and interpretation of the models' decisions as utilized by the included

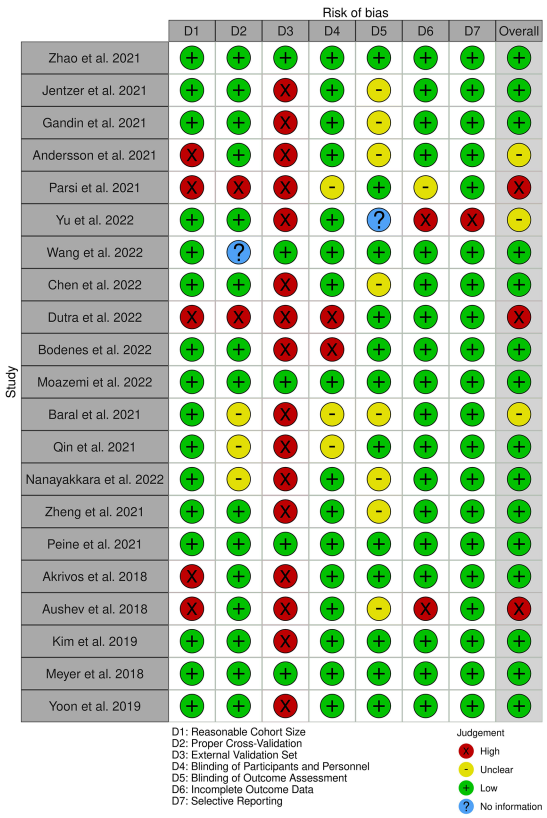

FIGURE 2
The risk of bias diagram for the selected studies. Each row corresponds to a selected study. The columns D1–D7 correspond to different risk criteria. The subjective judgements are color-coded as explained in the legend. The final column represents the overall judgement for the corresponding study.

studies. The most common group of algorithms are linear or decision tree-based methods, followed by recurrent models. Only five studies included feature selection or explainable AI methods. Although the high level of diversity in the datasets and algorithms which are utilized in the selected papers hinders us from conducting comprehensive performance meta-analysis, as outlined in Table 2, area under the receiver operating characteristics curve (AUROC) ranging from 79 to 96% throughout the entire cohort of papers, is the most commonly reported metrics item.

4.7. Concerns towards integration of AI in clinical routine

Figure 7 provides an overview of the concerns and limitations for the integration of AI for CDS in cardiac ICUs as discussed in the included papers, highlighting generalizability, interpretability, and dataset shift as the most central issues.

5. Discussion

Conventionally, patients visiting different care units undergo continuous examinations and interventions during their stays at the corresponding units. Thus, the physicians and medical staff are required to proactively monitor all the patients' critical signs and examination results regardless of their types and frequencies. In particular, for cardiovascular patients who are subject to higher complication rates and longer stays at intensive care units (ICUs) (4, 14), the increasing amounts of propagated and interconnected health-related factors captured along the patients' stays expose challenges towards taking appropriate and timely decisive actions for the physicians. These challenges are signified as many of the sources of multimodal temporal data used to make diagnostic or prognostic decisions, such as EHR extracted laboratory variables and vital signs, might be non-linearly correlated. Therefore, to assist the physicians and to complement their decision-making routines, there is an evolving need for appropriate clinical decision support systems (CDSS) leveraging modern AI-driven methodologies which are capable of investigating and identifying non-linear correlations in the multimodal patient data.

Advancements in AI are taking place continuously. Their presence in medicine is ever growing, and they could soon be present in cardiac ICUs. AI has the ability to assist clinicians in diagnosing arrhythmias, as shown in Parsi et al. where they were able to detect atrial fibrillation with a sensitivity and specificity >96% (39). Atrial fibrillation is a very common complication post cardiac surgery, which if not recognized, can have a significant negative impact on a patient's health. The sooner

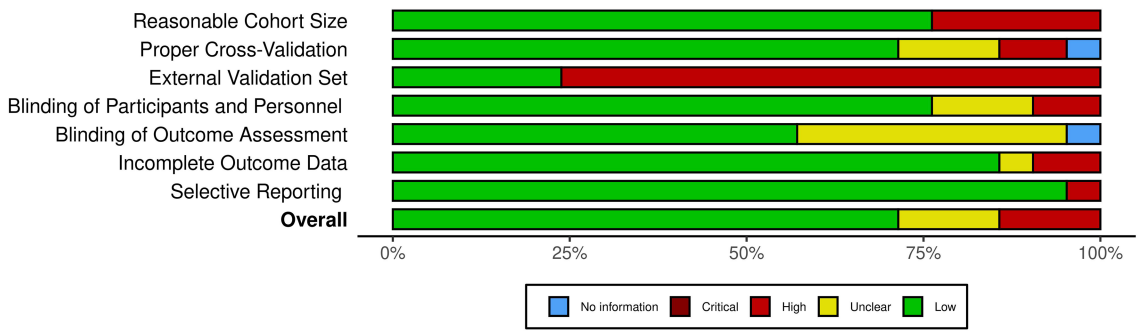

FIGURE 3
The summary of the risk of bias analysis. Each bar chart corresponds to one criteria of bias, stacked along the Y axis. The X axis quantifies the percentage of the studies with the corresponding color-coded subjective assessment as explained in the legend.

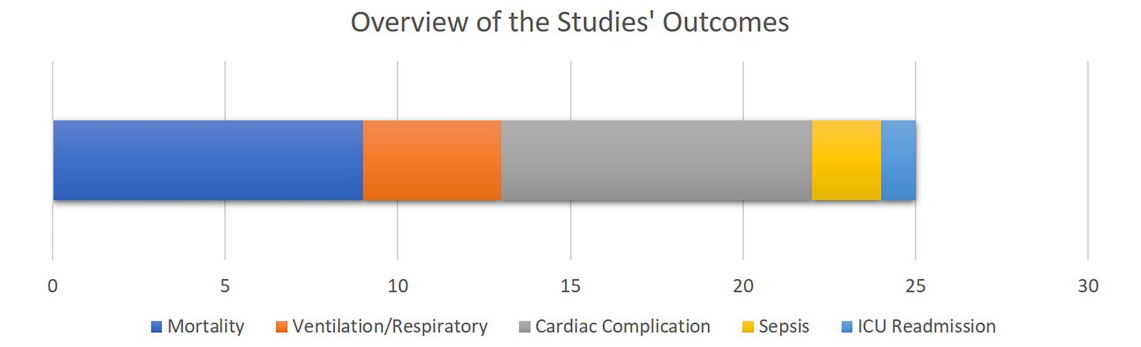

FIGURE 4
The overview of the outcomes of the selected studies. The bar chart shows how frequent each study outcome has been, with the X axis quantifying the number of studies. Note that some studies analyzed multiple outcomes.

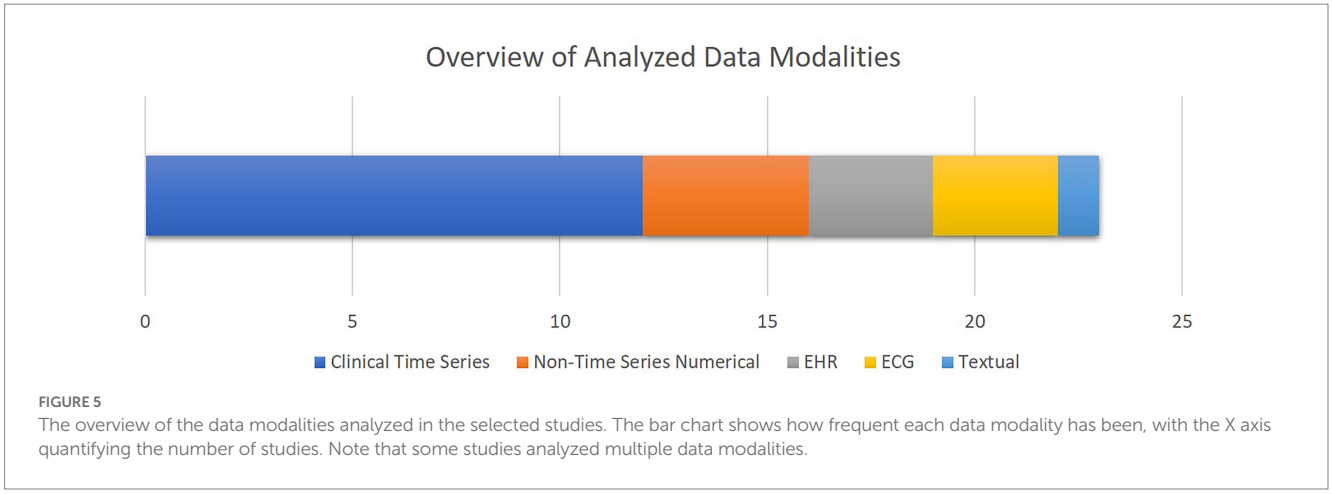

FIGURE 5
The overview of the data modalities analyzed in the selected studies. The bar chart shows how frequent each data modality has been, with the X axis quantifying the number of studies. Note that some studies analyzed multiple data modalities.

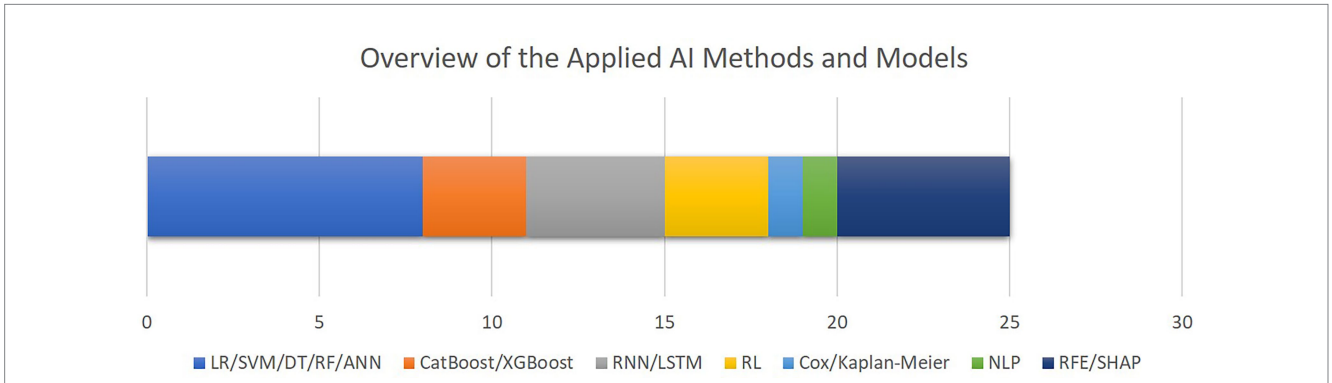

FIGURE 6
The overview of the AI methods and models applied for outcome prediction or interpretability. The bar chart shows how frequent each AI method has been, with the X axis quantifying the number of studies. Note that some studies applied multiple algorithms or methods (LR, logistic regression; SVM, support vector machine; DT, decision trees; RF, random forest; ANN, artificial neural networks; CatBoos, categorical boosting; XGBoost, extreme gradient boosting; RNN, recurrent neural networks; LSTM, long short-term memory; RL, reinforcement learning; NLP, natural language processing; RFE, recursive feature elimination; SHAP, SHapley Additive exPlanations).

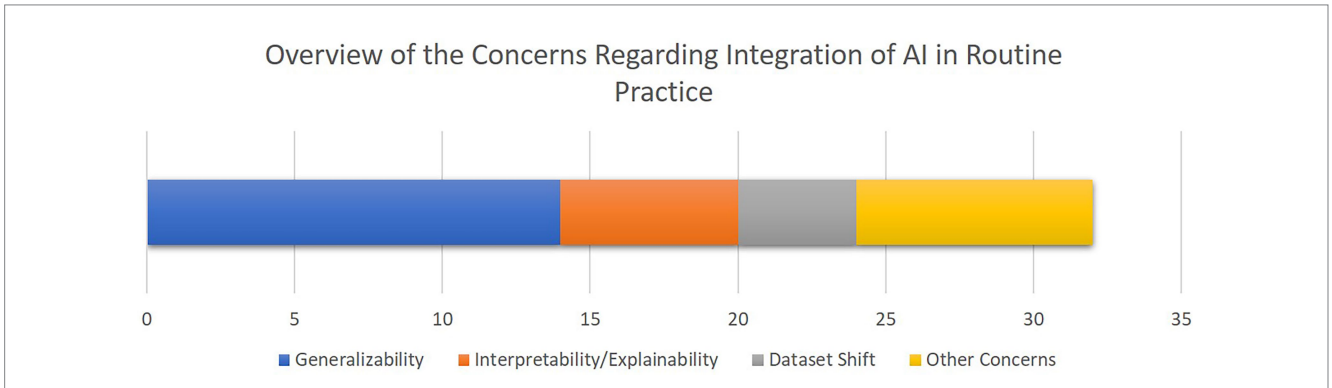

FIGURE 7
The overview of the concerns towards integration of AI-driven decision support tools in clinical routines as discussed in the selected studies. The bar chart shows how frequent each concern has been, with the X axis quantifying the number of studies. Note that some studies mentioned multiple concerns.

atrial fibrillation is detected and treated, the higher are the chances of conversion into sinus rhythm. Another role AI can play is predicting therapeutic outcomes and thereby helping plan for further treatment. In the paper by Andersson et al. the authors showed their ANN provided good prognostic accuracy in predicting neurological outcomes in comatose patients post out-of-hospital cardiac arrest (67). By having the capability to predict neurological outcomes, AI can help physicians decide whether further treatment would be beneficial for

patients with neurological complications post cardiac arrest in the form of neurological rehabilitation for instance. Thus, it could help improve patient quality of life in those who would benefit, as well as filtering those who would not, thus ideally lowering the demand for neurological rehabilitation spots in clinics, which are already oversaturated with patients on waiting lists. Finally, AI is capable of optimizing and fine tuning therapies, as shown in Peine et al. concluding AI was capable of delivering high performance optimization of mechanical ventilation in critical care, sometimes even exceeding physicians in comparison (79), and in Zheng et al. where AI was able to calculate the optimal oxygen therapy in COVID-19 patients, which was shown to be less on average than the amount recommended by physicians (78). This goes to show how AI is capable of improving general treatment and patient outcomes in ICUs while at the same time reducing the usage of costly materials, resources and services.

As illustrated in Figures 2, 3, our risk of bias analysis shows that most of the studies pass the criteria regarding blinding of the assessments and reporting bias. However, the findings revealed rooms for further consideration of universal validation guidelines, highlighting the lack of validation against external data cohorts. Thus, compared to conventional risk of bias criteria, we included three extra criteria (D1–D3) which address data-driven aspects of bias considering cohort size, proper cross-validation, and external datasets for validation purposes. We believe, integrating these extra bias assessment criteria should be followed in systematic reviews in the medical AI field.

To provide an overview of the results of this systematic review, most of the selected studies focused on critical cardiac and respiratory complications resulting in mortality of patients visiting cardiac ICUs (Figure 4). To this end, as illustrated in Figure 5, numerical measurements (either singular or time-dependent) captured during patients' stays at ICUs are extensively used for model training and evaluation in most of the studies, while textual data are the least used modality in this regard. Consecutively, depending on the input data, suitable AI-methods are utilized for model development. As shown in Figure 6, supervised ML classifiers such as SVM and random forest alongside XGBoost and CatBoost and reinforcement learning (RL) are the most common methods. Moreover, when it comes to analyzing clinical time series data and textual data, recurrent neural networks (RNNs) and natural language processing (NLP) come to action, respectively. For the special case of integrating NLP for processing textual health records, the lack of systematic guidelines for reporting EHRs becomes critical when no persistent vocabulary exists, especially for the non-English speaking centers for which less data is available for training and validation purposes.

Our findings further highlight the importance of utilizing open access datasets to provide AI-assisted clinical decision support in cardiovascular ICUs. While there are clear benefits to using open access datasets such as MIMIC in the field of critical care, it is important to consider the potential limitations of such datasets. Open access datasets may not fully capture the nuances of specific healthcare systems or populations in certain regions, which may impact the generalizability of the AI models trained on them. Therefore, researchers and clinicians should carefully evaluate the suitability of open access datasets for their particular use case and consider supplementing them with domestic datasets if necessary. Nonetheless, open access datasets can facilitate collaboration and knowledge sharing, which are essential for advancing the field of AI-assisted clinical decision making. Also, open access datasets are often rigorously curated and annotated by experts, ensuring the data is of high quality and can be used reliably. On the other hand, domestic datasets may not have the same level of diversity and may be limited in size, leading to suboptimal AI models. Nevertheless, regardless of the fact that which kind of data is used to fit AI agents, a proper cross-validation scheme should be applied to account for generalizability.

As illustrated in the analysis results, logistic regression (LR), SVM, decision trees, random forests, neural networks, and recurrent deep learning models are all popular machine learning algorithms used for various tasks in the field. Each of these algorithms has its own strengths and weaknesses, and the choice of algorithm depends on the specific task at hand and the available data. Most of the time, LR, SVM, and often tree-based methods are used as baseline methods to complement other more complex methodologies such as deep or recurrent neural networks (RNNs). Furthermore, decision trees and random forests are good choices when dealing with small to medium-sized datasets that have both categorical and numerical features. They work well when the data has a clear and interpretable structure, and when the decision-making process can be represented as a sequence of simple if-then-else rules. Decision trees are also good when there is a need to explain the reasoning behind a model's decision-making process. Neural networks, including deep learning models, are ideal for large and complex datasets with many features, such as image, speech, and text data. They are especially powerful when the relationships between input and output data are highly nonlinear and difficult to capture with simple models. However, neural networks can be computationally expensive to train and require a lot of data to generalize well. Recurrent deep learning models are a type of neural network that are well-suited for sequential and longitudinal data, such as time series, speech, and text data. They can capture long-term dependencies and patterns in the data and are especially useful when the output depends on past inputs. However, they can be more difficult to train than linear or tree-based models and require more specialized expertise. In summary, it's important to evaluate the strengths and weaknesses of each machine learning algorithm carefully and select the one that is best suited to the specific needs.

The findings from the selected articles have shown the predictive potential of different AI approaches including RNNs and RL. While many of the included studies integrated supervised ML classifiers like SVMs or RNNs for continuous patient monitoring in cardiac ICUs, one general advantage reinforcement learning provides over other paradigms of ML is that this way of defining the problem allows RL to take into account long-term rewards. This characteristic makes it especially appealing for clinical applications since, in numerous healthcare issues, the response to treatment decisions is frequently delayed (88). Additionally, the exploration-exploitation approach shares similarities with the actual clinical setting, where treatment responses can be heterogeneous (89) and finding the optimal treatment regime can also be characterized by trade-offs between exploration and exploitation.

Based on the findings of the included literature, the most critical limitations towards integration of AI-driven methods in routine clinical decision making are generalizability and explainability issues. As illustrated in Figure 3, more than 75% of the studies lack validation against external datasets which highlights the lack of generalizability associated with their findings. Nevertheless, as presented in Table 2, only three of the 21 included studies provided open access web-based user interfaces to facilitate validating their models with external datasets. Although providing freely accessible tools for external validation should be marked as a benefit for novel AI tools, the lack of standardization of external validation schemes considering the high levels of privacy and confidentiality associated with medical data cohorts rank amongst the most important limitations towards integration of AI in clinical routines, especially in multicentric and federated scenarios (90).

Furthermore, despite all the promising achievements of AI in the medical domain, the medical experts are still responsible for patients' lives. Therefore, to reduce the burden of responsibility and to provide further support, it is of critical importance to build trust in decisions made by the AI-assisted agents. As discussed in the related work (24), interpretability facilitated by explainable AI (XAI) best practices plays an important role to build further trust in AI in the medical domain. Although the authors of most of the reported articles recognize interpretability as a central issue in this domain, only five studies integrated methods such as RFE and SHAP to provide a level of transparency to complement their proposed models' decisions (Figure 6). In a related work, Asan et al. (91) identified transparency, robustness, and fairness as the most important criteria to enhance trust when it comes to human-AI collaboration in the healthcare domain which is confirmed by our risk of bias analysis as well. This emphasizes the evolving need for extra efforts to identify and mitigate different sources of bias since the early stages of designing and developing AI models for the clinical and medical domains.

Another concern which affects the effective integration of AI methodologies in the healthcare domain is the certification of the established models and products upon proper evaluations and clinical trials. Although an increasing number of approved AI/ML products has been traceable since 2015 in the united states and Europe in domains such as radiology, related works urge for more transparency on the criteria for the approval of AI/ML-based products facilitated through publicly accessible databases from authorities such as the food and drug association (FDA) of united states of America (United States) and Conformité Européene (CE) of Europe (92). As an insightful example, Zanca et al. suggest some practical guidelines for the medical physicists (MPs) who conventionally act as responsible authorities to ensure safety and quality of emerging diagnostic and therapeutic technologies in healthcare. They empathize that MPs need to acquire enough knowledge about AI tools and how they conceptually differ from traditional medical software and hardware devices, because they often attribute higher levels of autonomy compared to traditional medical products (93).

The current study presents a comprehensive overview of the most widely used AI-related methodologies as reported in recent literature, which were selected in a systematic and objective manner. As a result, the majority of the methodology employed is based on modern machine learning solutions. However, as per some other studies such as Roller et al. (94), there is a suggestion to begin with simpler systems which make the use of explicit, structured knowledge such as guidelines, decision-making procedures, and thresholds which are commonly found in clinical environments. As our comprehensive analysis outlined, these often simpler "rule-based" processes have been mostly overlooked in the selected articles. This is an important concern which needs to be further addressed in follow-up studies.

As a limitation of current study, due to diverse datasets and algorithms used in the selected cohort of studies, it was not feasible to conduct comprehensive meta-analysis covering comparison of all the methods across all the databases. Nonetheless, we reported performance results from all the articles in Table 2. Although the results are not directly comparable with each other, area under the receiver operating characteristics curve (AUROC), ranging from 0.79 to 0.96, was the most universal performance metric across all the selected studies.

In this study, we included studies from PubMed and Google Scholar databases alongside additional papers chosen from subjective search queries within impactful related works. Also, we focused on the studies written in the English language. Thus, our findings might be biased with regard to the choices of search engines and text language and might not be fully comprehensive. However, we covered the application oriented, model-driven, and data-driven aspects of AI-assisted methodologies utilized for patient monitoring and medical intervention in cardiovascular ICUs, following the PRISMA (32) and medical AI life cycle (34) paradigms.

6. Conclusion and future work

Technical conclusion: Recent advancements in AI-driven methodologies in intensive patient monitoring open up new horizons for the integration of clinical decision support in practice. However, regardless of being totally automated or requiring an expert's input or annotation, AI assisted methodologies for clinical decision support are meant to operate as a complementary aid to physicians and intensivists' subjective decisions rather than acting in complete autonomy. To achieve this, certain limitations should be mitigated. Most importantly, to address the generalizability issue which has been highlighted by our findings to be a common source of bias, proper validation against independent unseen sets of data should be taken care of. This becomes more critical as the medical datasets attribute high levels of confidentiality, affecting multicentric and federated learning scenarios.

Medical conclusion: AI has the potential to simplify part of the decision making in intensive patient monitoring by reducing the burden of processing huge amounts of information available from different sources of vital signs and critical patient parameters. However, still efforts need to be made to enhance interpretability of state-of-the-art AI methods for clinicians. In addition, proper training and understandable insights should be provided for the medical staff to enhance the level of trust in AI decisions. Moreover, AI algorithms should be tested in prospective clinical trials similar to other new

medical devices under observation of legal instances such as FDA in the United States and CE in Europe.

Future work: For the future, we plan to conduct studies on the integration of eXplainable AI (XAI) best practices for patient monitoring in cardiac ICUs, focusing on federated learning scenarios in which data from multiple hospitals are processed.

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.

Author contributions

SM, SV, MT, HA, and FS: conceptualization and study design. SM, SV, JL, PS, SK, RB, and BD: identification of papers through search engine queries. SM, SV, SK, and PS: technical review of searched papers. LC, HA, RAB, and AL: medical review of searched papers. JL, BD, RB, and PS: conducting visual analysis and preparing figures. SM, SV, PS, and LC: risk of bias analysis. SM, SV, PS, SK, BD, and RB: summarizing selected papers' contents. SM, SV, MT, RAB, HA, and FS: narrative discussion on the findings. All authors contributed to the article and approved the submitted version.

Funding

The project is funded by the German Federal Ministry of Education and Research (BMBF) under the grant number 16SV8601.

Acknowledgments

We would like to express special gratitude to Mrs. Susanne Bunnenberg for her generous donation that made this project possible in the first place.

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. Oyebode F. Clinical errors and medical negligence. *Med Princ Pract*. (2013) 22:323–33. doi: 10.1159/000346296

2. Reason J. Understanding adverse events: human factors. *BMJ Qual Saf*. (1995) 4:80–9. doi: 10.1136/qshc.4.2.80

3. 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

4. Almashrafi A, Elmontsri M, Aylin P. Systematic review of factors influencing length of stay in ICU after adult cardiac surgery. *BMC Health Serv Res*. (2016) 16:318. doi: 10.1186/s12913-016-1591-3

5. Sutton RT, Pincock D, Baumgart DC, Sadowski DC, Fedorak RN, Kroeker KI. An overview of clinical decision support systems: benefits, risks, and strategies for success. *NPJ Digit Med*. (2020) 3:17. doi: 10.1038/s41746-020-0221-y

6. Michel JJ, Flores EJ, Dutcher L, Mull NK, Tsou AY. Translating an evidence-based clinical pathway into shareable CDS: developing a systematic process using publicly available tools. *J Am Med Inform Assoc*. (2020) 28:52–61. doi: 10.1093/jamia/ocaa257

7. Gorgi Zadeh S, MWM W, Wiens V, Thiele S, Holz FG, Finger RP, et al. CNNs enable accurate and Fast segmentation of Drusen in optical coherence tomography In: MJ Cardoso, T Arbel, G Carneiro, T Syeda-Mahmood, T JMRS and M Moradiet al, editors. *Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support*. Cham: Springer International Publishing (2017). 65–73.

8. Andrearczyk V, Fontaine P, Oreiller V, Castelli J, Jreige M, Prior JO, et al. Multi-task deep segmentation and Radiomics for automatic prognosis in head and neck cancer In: I Rekik, E Adeli, SH Park and J Schnabel, editors. *Predictive Intelligence in Medicine*. Cham: Springer International Publishing (2021). 147–56.

9. Moazemi S, Khurshid Z, Erle A, Lütje S, Essler M, Schultz T, et al. Machine learning facilitates hotspot classification in PSMA-PET/CT with nuclear medicine specialist accuracy. *Diagnostics*. (2020) 10:622. doi: 10.3390/diagnostics10090622

10. Moazemi S, Essler M, Schultz T, Bundschuh R. Predicting Treatment Response in Prostate Cancer Patients Based on Multimodal PET/CT for Clinical Decision Support. In: *Multimodal Learning for Clinical Decision Support*. eds. T Syeda-Mahmood, X Li, A Madabhushi, H Greenspan, Q Li and R Leahyet al. (Cham: Springer International Publishing) (2021). 22–35.

11. Lao J, Chen Y, Li ZC, Li Q, Zhang J, Liu J, et al. A deep learning-based Radiomics model for prediction of survival in glioblastoma Multiforme. *Sci Rep*. (2017) 7:10353. doi: 10.1038/s41598-017-10649-8

12. Meng X, Zhao B, Xi R, Guo B, Huang B, Li S, et al. The radiomic signature derived from pre-treatment PET and CT images: A predictor of overall survival in non-small cell lung cancer. *J Nucl Med*. (2019) 60:1333–3.

13. Moazemi S, Erle A, Lütje S, Gaertner F, Essler M, Bundschuh R. Estimating the potential of radiomics features and radiomics signature from pretherapeutic PSMA-PET-CT scans and clinical data for prediction of overall survival when treated with 177Lu-PSMA. *Diagnostics (Basel)*. (2021) 11:186. doi: 10.3390/diagnostics11020186

14. Lin WT, Chen WL, Chao CM, Lai CC. The outcomes and prognostic factors of the patients with unplanned intensive care unit readmissions. *Medicine (Baltimore)*. (2018) 97:e11124. doi: 10.1097/MD.0000000000011124

15. Huang K, Altosaar J, Ranganath R. ClinicalBERT: modeling clinical notes and predicting hospital readmission. *arXiv*. (2019) doi: 10.48550/arXiv.1904.05342

16. Kessler S, Schroeder D, Korlakov S, Hettlich V, Kalkhoff S, Moazemi S, et al. Predicting readmission to the cardiovascular intensive care unit using recurrent neural networks. *Digit Health*. (2023) 9:205520762211495. doi: 10.1177/20552076221149529

17. Wang D, Li J, Sun Y, Ding X, Zhang X, Liu S, et al. A machine learning model for accurate prediction of sepsis in ICU patients. *Front Public Health*. (2021) 9:754348. doi: 10.3389/fpubh.2021.754348

18. Ghalati PF, Samal SS, Bhat JS, Deisz R, Marx G, Schuppert A. Critical transitions in intensive care units: a sepsis case study. *Sci Rep*. (2019) 9:12888. doi: 10.1038/s41598-019-49006-2

19. Scherpf M, Gräßer F, Malberg H, Zaunseder S. Predicting sepsis with a recurrent neural network using the MIMIC III database. *Comput Biol Med*. (2019) 113:103395. doi: 10.1016/j.compbiomed.2019.103395

20. Harerimana G, Kim JW, Jang B. A deep attention model to forecast the length of stay and the in-hospital mortality right on admission from ICD codes and demographic data. *J Biomed Inform*. (2021) 118:103778. doi: 10.1016/j.jbi.2021.103778

21. Ye J, Yao L, Shen J, Janarthanam R, Luo Y. Predicting mortality in critically ill patients with diabetes using machine learning and clinical notes. *BMC Med Inform Decis Mak*. (2020) 20:295. doi: 10.1186/s12911-020-01318-4

22. Parikh RB, Teeple S, Navathe AS. Addressing bias in artificial intelligence in health care. *JAMA*. (2019) 322:2377–8. doi: 10.1001/jama.2019.18058

23. Larrazabal AJ, Nieto N, Peterson V, Milone DH, Ferrante E. Gender imbalance in medical imaging datasets produces biased classifiers for computer-aided diagnosis. *Proc Natl Acad Sci*. (2020) 117:12592–4. doi: 10.1073/pnas.1919012117

24. Antoniadi AM, Du Y, Guendouz Y, Wei L, Mazo C, Becker BA, et al. Current challenges and future opportunities for XAI in machine learning-based clinical decision support systems: a systematic review. *Appl Sci*. (2021) 11:5088. doi: 10.3390/app11115088

25. Moazemi S, Kalkhoff S, Kessler S, Boztoprak Z, Hettlich V, Liebrecht A, et al. Evaluating a recurrent neural network model for predicting readmission to cardiovascular ICUs based on clinical time series data. *Eng Proc*. (2022) 18:1. doi: 10.3390/engproc2022018001

26. 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. doi: 10.1007/s00134-019-05872-y

27. Giordano C, Brennan M, Mohamed B, Rashidi P, Modave F, Tighe P. Accessing artificial intelligence for clinical decision-making. *Front Digit Health*. (2021) 3:645232.

28. Syed M, Syed S, Sexton K, Syeda HB, Garza M, Zozus M, et al. Application of machine learning in intensive care unit (ICU) settings using MIMIC dataset: systematic review. *Inform MDPI*. (2021) 8:16. doi: 10.3390/informatics8010016

29. Greco M, Caruso PF, Cecconi M. Artificial intelligence in the intensive care unit. *Semin Respir Crit Care Med*. (2021) 42:2–9. doi: 10.1055/s-0040-1719037

30. Yang G, Ye Q, Xia J. Unbox the black-box for the medical explainable AI via multi-modal and multi-Centre data fusion: A mini-review, two showcases and beyond. *Inf Fusion*. (2022) 77:29–52. doi: 10.1016/j.inffus.2021.07.016

31. Abdellatif AA, Mhaisen N, Chkirbene Z, Mohamed A, Erbad A, Guizani M. Reinforcement learning for intelligent healthcare systems: A comprehensive survey. *arXiv*. (2021) doi: 10.48550/arXiv.2108.04087

32. Page MJ, McKenzie JE, Bossuyt PM, Boutron I, Hoffmann TC, Mulrow CD, et al. The PRISMA 2020 statement: an updated guideline for reporting systematic reviews. *BMJ*. (2021) 372:n71. doi: 10.1136/bmj.n71

33. Amir-Behghadami M, Janati A. Population, intervention, comparison, outcomes and study (PICOS) design as a framework to formulate eligibility criteria in systematic reviews. *Emerg Med J*. (2020) 37:387–7. doi: 10.1136/emermed-2020-209567

34. Ng MY, Kapur S, Blizinsky KD, Hernandez-Boussard T. The AI life cycle: a holistic approach to creating ethical AI for health decisions. *Nat Med*. (2022) 28:2247–9. doi: 10.1038/s41591-022-01993-y

35. Sutton RS. Introduction: The challenge of reinforcement learning In: RS Sutton, editor. *Reinforcement Learning. The Springer International Series in Engineering and Computer Science*, vol. *173*. Boston, MA: Springer

36. Lippman SA. Dynamic programming and Markov decision processes In: . *The New Palgrave Dictionary of Economics*. London: Palgrave Macmillan UK (2016). 1–7. doi: 10.1057/978-1-349-95121-5_80-1

37. Riedmiller M. Neural fitted Q iteration – first experiences with a data efficient neural reinforcement learning method In: J Gama, R Camacho, PB Brazdil, AM Jorge and L Torgo, editors. *Machine Learning: ECML 2005*. Berlin, Heidelberg: Springer Berlin Heidelberg (2005). 317–28.

38. Mnih V, Kavukcuoglu K, Silver D, Graves A, Antonoglou I, Wierstra D, et al. Playing Atari with deep reinforcement learning. *arXiv*. (2013) doi: 10.48550/arXiv:1312.5602

39. Parsi A, Glavin M, Jones E, Byrne D. Prediction of paroxysmal atrial fibrillation using new heart rate variability features. *Comput Biol Med*. (2021) 133:104367. doi: 10.1016/j.compbiomed.2021.104367

40. Yu Y, Peng C, Zhang Z, Shen K, Zhang Y, Xiao J, et al. Machine learning methods for predicting long-term mortality in patients after cardiac surgery. *Front Cardiovasc Med*. (2022) 9:831390. doi: 10.3389/fcvm.2022.831390

41. 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

42. Bodenes L, N'Guyen QT, Le Mao R, Ferrière N, Pateau V, Lellouche F, et al. Early heart rate variability evaluation enables to predict ICU patients' outcome. *Sci Rep*. (2022) 12:2498. doi: 10.1038/s41598-022-06301-9

43. Qin F, Madan V, Ratan U, Karnin Z, Kapoor V, Bhatia P, et al. Improving early sepsis prediction with multi modal learning. *arXiv*. (2021)

44. Baral S, Alsadoon A, Prasad PWC, Al Aloussi S, Alsadoon OH. A novel solution of using deep learning for early prediction cardiac arrest in sepsis patient: enhanced bidirectional long short-term memory (LSTM). *Multimed Tools Appl*. (2021) 80:32639–64. doi: 10.1007/s11042-021-11176-5

45. Cox DR. The regression analysis of binary sequences. *J R Stat Soc Ser B Methodol*. (1958) 20:215–32. doi: 10.1111/j.2517-6161.1958.tb00292.x

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

47. Quinlan JR. Induction of decision trees. *Mach Learn*. (1986) 1:81–106. doi: 10.1007/BF00116251

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

49. Friedman JH. Greedy function approximation: A gradient boosting machine. *Ann Stat*. (2001) 29:1189–232. doi: 10.1214/aos/1013203451

50. Dorogush AV, Ershov V, Gulin A. CatBoost: Gradient boosting with categorical features support. *arXiv*. (2018)

51. Sherstinsky A. Fundamentals of recurrent neural network (RNN) and long short-term memory (LSTM) network. *Phys Nonlinear Phenom*. (2020) 404:132306. doi: 10.1016/j.physd.2019.132306

52. Hochreiter S, Schmidhuber J. Long short-term memory. *Neural Comput*. (1997) 9:1735–80. doi: 10.1162/neco.1997.9.8.1735

53. Johnson AEW, Pollard TJ, Shen L, Lehman L, Wei H, Feng M, et al. MIMIC-III, a freely accessible critical care database. *Sci Data*. (2016) 3:160035. doi: 10.1038/sdata.2016.35

54. Johnson A, Bulgarelli L, Shen L, Gayles A, Shammout A, Horng S, et al. MIMIC-IV, a freely accessible electronic health record dataset. *Scientific data*. (2023) 10:1.

55. Ho KM, Dobb GJ, Knuiman M, Finn J, Lee KY, Webb SA. A comparison of admission and worst 24-hour acute physiology and chronic health evaluation II scores in predicting hospital mortality: a retrospective cohort study. *Crit Care*. (2005) 10:R4.

56. Higgins TL, Teres D, Copes WS, Nathanson BH, Stark M, Kramer AA. Assessing contemporary intensive care unit outcome: an updated mortality probability admission model (MPM0-III). *Crit Care Med*. (2007) 35:827–35. doi: 10.1097/01.CCM.0000257337.63529.9F

57. Vincent JL, de Mendonca A, Cantraine F, Moreno R, Takala J, Suter PM, et al. Use of the SOFA score to assess the incidence of organ dysfunction/failure in intensive care units: results of a multicenter, prospective study. *Crit Care Med*. (1998) 26:1793–800. doi: 10.1097/00003246-199811000-00016

58. Jabali AK, Waris A, Khan DI, Ahmed S, Hourani RJ. Electronic health records: three decades of bibliometric research productivity analysis and some insights. *Inform Med Unlocked*. (2022) 29:100872. doi: 10.1016/j.imu.2022.100872

59. Sattar Y, Chhabra L. Electrocardiogram In: . *StatPearls*. Treasure Island (FL): StatPearls Publishing (2022)

60. 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.

61. Yèche H, Kuznetsova R, Zimmermann M, Hüser M, Lyu X, Faltys M, et al. HiRID-ICU-benchmark - A comprehensive machine learning benchmark on high-resolution ICU data. *arXiv*. (2022) doi: 10.48550/arXiv.2111.08536

62. Saeed M, Villarroel M, Reisner AT, Clifford G, Lehman LW, Moody G, et al. Multiparameter intelligent monitoring in intensive care II: A public-access intensive care unit database*. *Crit Care Med*. (2011) 39:952–60. doi: 10.1097/CCM.0b013e31820a92c6

63. 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

64. Shaban-Nejad A, Michalowski M, Buckeridge DL. Explainability and interpretability: AAAI international workshop on health intelligence, W3PHIAI 2020 In: A Shaban-Nejad, M Michalowski and DL Buckeridge, editors. *Explainable AI In Healthcare and Medicine - Building A Culture of Transparency and Accountability*: Springer Science and Business Media Deutschland GmbH (2020). 1–10.

65. 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

66. Jentzer JC, Kashou AH, Lopez-Jimenez F, Attia ZI, Kapa S, Friedman PA, et al. Mortality risk stratification using artificial intelligence-augmented electrocardiogram in cardiac intensive care unit patients. *Eur Heart J Acute Cardiovasc Care*. (2021) 10:532–41. doi: 10.1093/ehjacc/zuaa021

67. Andersson P, Johnsson J, Björnsson O, Cronberg T, Hassager C, Zetterberg H, et al. Predicting neurological outcome after out-of-hospital cardiac arrest with cumulative information; development and internal validation of an artificial neural network algorithm. *Crit Care Lond Engl*. (2021) 25:83. doi: 10.1186/s13054-021-03505-9

68. Kiseleva A, Kotzinos D, De Hert P. Transparency of AI in healthcare as a multilayered system of accountabilities: between legal requirements and technical limitations. *Front Artif Intell*. (2022) 5:879603. doi: 10.3389/frai.2022.879603

69. Srinivasu PN, Sandhya N, Jhaveri RH, Raut R. From Blackbox to explainable AI in healthcare: existing tools and case studies. *Mob Inf Syst*. (2022) 2022:1–20. doi: 10.1155/2022/8167821

70. Guyon I, Weston J, Barnhill S, Vapnik V. Gene selection for cancer classification using support vector machines. *Mach Learn*. (2002) 46:389–422. doi: 10.1023/A:1012487302797

71. Lundberg S, Lee SI. A unified approach to interpreting model predictions. *arXiv*. (2017)

72. Pesquita C. "Towards semantic integration for explainable artificial intelligence in the biomedical domain," in *In HEALTHINF 2021 Feb 11* (2021). 747–753.

73. McGuinness LA, Higgins JPT. Risk-of-bias VISualization (robvis): an R package and shiny web app for visualizing risk-of-bias assessments. *Res Synth Methods*. (2021) 12:55–61. doi: 10.1002/jrsm.1411

74. 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

75. Wang H, Zhao QY, Luo JC, Liu K, Yu SJ, Ma JF, et al. Early prediction of noninvasive ventilation failure after extubation: development and validation of a machine-learning model. *BMC Pulm Med*. (2022) 22:304. doi: 10.1186/s12890-022-02096-7

76. Dutra GP, Gomes Bf De O, Do CJP, JLF P, Nascimento EM, Pereira De B, et al. Mortality from heart failure with mid-range ejection fraction. *Arq Bras Cardiol*. (2022) 118:694–700. doi: 10.36660/abc.20210050

77. Nanayakkara T, Clermont G, Langmead CJ, Swigon D. Unifying cardiovascular modelling with deep reinforcement learning for uncertainty aware control of sepsis treatment. *PLOS Digit Health*. (2022) 1:e0000012. doi: 10.1371/journal.pdig.0000012

78. Zheng H, Zhu J, Xie W, Zhong J. Reinforcement learning assisted oxygen therapy for COVID-19 patients under intensive care. *BMC Med Inform Decis Mak*. (2021) 21:350. doi: 10.1186/s12911-021-01712-6

79. Peine A, Hallawa A, Bickenbach J, Dartmann G, Fazlic LB, Schmeink A, et al. Development and validation of a reinforcement learning algorithm to dynamically optimize mechanical ventilation in critical care. *Npj Digit Med*. (2021) 4:1–12. doi: 10.1038/s41746-021-00388-6

80. Akrivos E, Papaioannou V, Maglaveras N, Chouvarda I. Prediction of Cardiac Arrest in Intensive Care Patients Through Machine Learning. (2018). p. 25–29.

81. Aushev A, Ripoll VR, Vellido A, Aletti F, Pinto BB, Herpain A, et al. Feature selection for the accurate prediction of septic and cardiogenic shock ICU mortality in the acute phase. *PLoS One*. (2018) 13:e0199089. doi: 10.1371/journal.pone.0199089

82. Kim J, Chae M, Chang HJ, Kim YA, Park E. Predicting cardiac arrest and respiratory failure using feasible artificial intelligence with simple trajectories of patient data. *J Clin Med*. (2019) 8:1336. doi: 10.3390/jcm8091336

83. 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

84. Yoon JH, Mu L, Chen L, Dubrawski A, Hravnak M, Pinsky MR, et al. Predicting tachycardia as a surrogate for instability in the intensive care unit. *J Clin Monit Comput*. (2019) 33:973–85. doi: 10.1007/s10877-019-00277-0

85. Zou H, Hastie T. Regularization and variable selection via the elastic net. *J R Stat Soc Ser B Stat Methodol (Statistical Methodology)*. (2005) 67:301–20. doi: 10.1111/j.1467-9868.2005.00503.x

86. Devlin J, Chang MW, Lee K, Toutanova K. BERT: pre-training of deep bidirectional transformers for language understanding. *arXiv*. (2019)

87. Yang M, Wang X, Gao H, Li Y, Liu X, Li J, et al. "Early Prediction of Sepsis Using Multi-Feature Fusion Based XGBoost Learning and Bayesian Optimization," in *The IEEE Conference on Computing in Cardiology (CinC)* (2019) 46:1–4. Available at: http://www.cinc.org/archives/2019/pdf/CinC2019-020.pdf (Accessed October 31, 2022)

88. Gottesman O, Johansson F, Komorowski M, Faisal A, Sontag D, Doshi-Velez F, et al. Guidelines for reinforcement learning in healthcare. *Nat Med*. (2019) 25:16–8. doi: 10.1038/s41591-018-0310-5

89. Qian Y, Zhang C, Krishnamachari B, Tambe M. Restless poachers: Handling exploration-exploitation tradeoffs in security domains. (2016). p. 9.

90. Aouedi O, Sacco A, Piamrat K, Marchetto G. Handling privacy-sensitive medical data with federated learning: challenges and future directions. *IEEE J Biomed Health Inform*. (2022) 27:1–14.

91. Asan O, Bayrak AE, Choudhury A. Artificial intelligence and human Trust in Healthcare: focus on clinicians. *J Med Internet Res*. (2020) 22:e15154. doi: 10.2196/15154

92. Muehlematter UJ, Daniore P, Vokinger KN. Approval of artificial intelligence and machine learning-based medical devices in the USA and Europe (2015–20): a comparative analysis. *Lancet Digit Health*. (2021) 3:e195–203. doi: 10.1016/S2589-7500(20)30292-2

93. Zanca F, Brusasco C, Pesapane F, Kwade Z, Beckers R, Avanzo M. Regulatory aspects of the use of artificial intelligence medical software. *Semin Radiat Oncol*. (2022) 32:432–41. doi: 10.1016/j.semradonc.2022.06.012

94. Roller R, Budde K, Burchardt A, Dabrock P, Möller S, Osmanodja B, et al. When performance is not enough - A multidisciplinary view on clinical decision support. *arXiv*. (2022)

frontiers | Frontiers in Immunology

TYPE Original Research
PUBLISHED 03 April 2023
DOI 10.3389/fimmu.2023.1140755

OPEN ACCESS

EDITED BY
Rahul Kashyap,
WellSpan Health, United States

REVIEWED BY
Pratikkumar Vekaria,
University of South Carolina, United States
Pranjal Sharma,
Northeast Ohio Medical University,
United States

*CORRESPONDENCE
Fang Liu
xyliufang@csu.edu.cn
Li Li
llicu@qq.com
Qiongjing Yuan
yuanqiongjing@csu.edu.cn

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

SPECIALTY SECTION
This article was submitted to Systems Immunology, a section of the journal Frontiers in Immunology

RECEIVED 09 January 2023
ACCEPTED 17 March 2023
PUBLISHED 03 April 2023

CITATION
Zhou H, Liu L, Zhao Q, Jin X, Peng Z, Wang W, Huang L, Xie Y, Xu H, Tao L, Xiao X, Nie W, Liu F, Li L and Yuan Q (2023) Machine learning for the prediction of all-cause mortality in patients with sepsis-associated acute kidney injury during hospitalization. *Front. Immunol.* 14:1140755. doi: 10.3389/fimmu.2023.1140755

Machine learning for the prediction of all-cause mortality in patients with sepsis-associated acute kidney injury during hospitalization

Hongshan Zhou[1†], Leping Liu[2†], Qinyu Zhao[3†], Xin Jin[4], Zhangzhe Peng[1,5,6], Wei Wang[1,5,6], Ling Huang[1,5,6], Yanyun Xie[1,5,6], Hui Xu[1], Lijian Tao[1,5,6], Xiangcheng Xiao[1], Wannian Nie[1], Fang Liu[7*], Li Li[8*] and Qiongjing Yuan[1,5,6,9*]

[1]Department of Nephrology, Xiangya Hospital of Central South University, Changsha, Hunan, China, [2]Department of Pediatrics, The Third Xiangya Hospital, Central South University, Changsha, China, [3]College of Engineering and Computer Science, Australian National University, Canberra, ACT, Australia, [4]Critical Care Medicine, The Third Xiangya Hospital, Central South University, Changsha, Hunan, China, [5]Organ Fibrosis Key Lab of Hunan Province, Central South University, Changsha, Hunan, China, [6]National International Joint Research Center for Medical Metabolomics, Xiangya Hospital, Central South University, Changsha, Hunan, China, [7]Health Management Center, Xiangya Hospital of Central South University, Changsha, Hunan, China, [8]Critical Care Medicine, Xiangya Hospital of Central South University, Changsha, Hunan, China, [9]National Clinical Medical Research Center for Geriatric Diseases, Xiangya Hospital of Central South University, Changsha, Hunan, China

Background: Sepsis-associated acute kidney injury (S-AKI) is considered to be associated with high morbidity and mortality, a commonly accepted model to predict mortality is urged consequently. This study used a machine learning model to identify vital variables associated with mortality in S-AKI patients in the hospital and predict the risk of death in the hospital. We hope that this model can help identify high-risk patients early and reasonably allocate medical resources in the intensive care unit (ICU).

Methods: A total of 16,154 S-AKI patients from the Medical Information Mart for Intensive Care IV database were examined as the training set (80%) and the validation set (20%). Variables (129 in total) were collected, including basic patient information, diagnosis, clinical data, and medication records. We developed and validated machine learning models using 11 different algorithms and selected the one that performed the best. Afterward, recursive feature elimination was used to select key variables. Different indicators were used to compare the prediction performance of each model. The SHapley Additive exPlanations package was applied to interpret the best machine learning model in a web tool for clinicians to use. Finally, we collected clinical data of S-AKI patients from two hospitals for external validation.

Results: In this study, 15 critical variables were finally selected, namely, urine output, maximum blood urea nitrogen, rate of injection of norepinephrine, maximum anion gap, maximum creatinine, maximum red blood cell volume distribution width, minimum international normalized ratio, maximum heart rate,

maximum temperature, maximum respiratory rate, minimum fraction of inspired O_2, minimum creatinine, minimum Glasgow Coma Scale, and diagnosis of diabetes and stroke. The categorical boosting algorithm model presented significantly better predictive performance [receiver operating characteristic (ROC): 0.83] than other models [accuracy (ACC): 75%, Youden index: 50%, sensitivity: 75%, specificity: 75%, F1 score: 0.56, positive predictive value (PPV): 44%, and negative predictive value (NPV): 92%]. External validation data from two hospitals in China were also well validated (ROC: 0.75).

Conclusions: After selecting 15 crucial variables, a machine learning-based model for predicting the mortality of S-AKI patients was successfully established and the CatBoost model demonstrated best predictive performance.

KEYWORDS

sepsis, acute kidney injury, mortality, predictive model, machine learning

Introduction

Sepsis, which is one of the principal causes of mortality worldwide and affects more than 19 million people every year (1–3), is defined as a sequential fatal organ dysfunction after infection with a dysregulated host response by the Third International Consensus Definitions for Sepsis and Septic Shock (Sepsis-3). Similarly, the Kidney Disease: Improving Global Outcomes (KDIGO) group integrated previous diagnostic criteria and proposed an international consensus for acute kidney injury (AKI) to be defined as (i) an increase in SCr level by more than 26.5 μmol/L (0.3 mg/dl) within 48 h; (ii) an increase in SCr level by more than 1.5 times the baseline (confirmed or presumed to occur within 7 days); and (iii) urine volume <0.5 ml/(kg·h) lasting for more than 6 h (4). In critically ill patients, the main cause of AKI has been considered to be sepsis for a long time, and 45%–70% of AKI patients are considered to have sepsis (5). Thus, sepsis-associated acute kidney injury (S-AKI) should be defined as a syndrome that meets the Sepsis-3 and KDIGO criteria simultaneously (6).

The epidemiology of S-AKI has not been fully clarified probably because of uncoordinated epidemiology of sepsis and AKI criteria, but the global incidence is estimated to be 6 million cases annually (6). The mortality of S-AKI was reported to be 45.99% in the intensive care unit (ICU) (7), and a retrospective cohort study discovered that S-AKI was correlated with a significantly higher mortality rate compared to sepsis without AKI (71.7% vs. 21.3%) (8). At present, many studies have shown that S-AKI imposed a heavy burden on patients. In a review, Hoste et al. summarized that the occurrence of AKI was related to the severity of sepsis and that S-AKI was responsible for the increase in disease acuity and burden of organ dysfunction (9). Bagshaw et al. conducted an observational cohort study spanning multiple nations and centers, which reported that S-AKI was associated with a high-crude in-hospital case fatality rate (51.8%) (5). Furthermore, a multicenter retrospective cohort study in China concluded that sepsis resulted in 32.0% of hospital-acquired AKI and 15.2% of community-acquired AKI. In addition, AKI was correlated with high mortality, longer length of stay, and heavier daily expenses while in the hospital (10). Additionally, an observational study of 618 ICU patients with AKI, the Program to Improve Care in Acute Renal Disease (PICARD), revealed that the in-hospital mortality rate of S-AKI was noticeably high, regardless of sepsis occurring before AKI (48%) or after AKI (44%) (11).

Considering that S-AKI patients experience high morbidity and mortality, the precise prediction of their prognosis is necessary. Novel biomarkers like tissue inhibitor of metalloproteinases-2 (TIMP-2), neutrophil gelatinase-associated lipocalin (NGAL), and insulin-like growth factor binding protein-7 (IGFBP-7) have been evaluated to forecast the prognosis of S-AKI; however, their sensitivity has not been verified in large multicenter studies (12). Conventional scoring systems of severity, such as Sequential Organ Failure Assessment (SOFA) and Acute Physiology and Chronic Health Evaluation II (APACHE II), have been widely used in the ICU to predict outcomes. Regrettably, they lack discrimination and prediction accuracy, and external validation is required before application to S-AKI cohorts (13). Consequently, it is essential to establish a new model that efficiently and accurately predicts the outcomes of S-AKI.

As a novel technology, machine learning has been utilized in various medical fields owing to its ability to develop robust risk models and improve prediction power (14, 15). The accuracy of predicting the occurrence of S-AKI utilizing machine learning has been confirmed (16–18). However, this radical new technology has not been applied to predict the mortality of patients with S-AKI, which is equally noteworthy. Gradient boosted decision trees (GBDTs) are powerful machine learning ensemble techniques, particularly when massive amounts of data are involved in classification and regression tasks. As one of the GBDT families, CatBoost is perfectly suited to processing categorical, heterogeneous data (19). Since its debut, CatBoost has been used in some medical studies and demonstrated its excellent predictive ability.

This study aimed to identify the risk factors associated with mortality in patients with S-AKI and develop a machine learning

model to predict death in hospitals on the basis of primary research emphasizing the prediction of occurrence. The performance of this machine learning model was compared with 10 other machine learning models to validate the superiority of the proposed model.

Materials and methods

Study subjects

The Medical Information Mart for Intensive Care IV (MIMIC-IV) is a database containing patient data from all ICU and emergency departments at Beth Israel Deaconess Medical Center from 2008 to 2019. The contents of the database include basic patient information, diagnosis, clinical data, and medication records, among others. We extracted the data of patients with sepsis and AKI after admission from the MIMIC-IV database as training and validation sets. Then, we collected the data of patients with sepsis and AKI in the ICU of Xiangya Hospital (from 2015 to 2022) and Third Xiangya Hospital (from 2022) of Central South University, Changsha, China as an external validation set (Figure 1).

According to the KDIGO guidelines, AKI is characterized by one or more of the following: (i) an increase in SCr level by more than 26.5 μmol/L (0.3 mg/dl) within 48 h; (ii) an increase in SCr level by more than 1.5 times the baseline (confirmed or presumed to occur within 7 days); and (iii) urine volume <0.5 ml/(kg·h) lasting for more than 6 h. According to the Third International Consensus Definitions for Sepsis and Septic Shock (Sepsis-3), sepsis is characterized by life-threatening organ dysfunction as a result of infection coupled with an impaired host response. According to the SOFA, organ dysfunction is a change in the total SOFA score of 2 points caused by infection. As part of this study, patients who were younger than 18 years of age, had stayed in the ICU for less than 24 h, and missed essential data were excluded. We used multiple imputations to supplement the missing values of patients. The death group is composed of patients who died in the hospital, and the alive group consists of patients who did not die during hospitalization.

According to the ethical standards of the responsible committee on human experimentation in China and to the Helsinki Declaration of 1975, all procedures in this study were conducted in accordance with the ethical standards of the responsible committee. The study was initiated under the guidance of Transparent Reporting of a Multivariable Prediction Model for Individual Prognosis or Diagnosis (TRIPOD) (Supplementary Figure 1). The Xiangya Hospital of Central South University

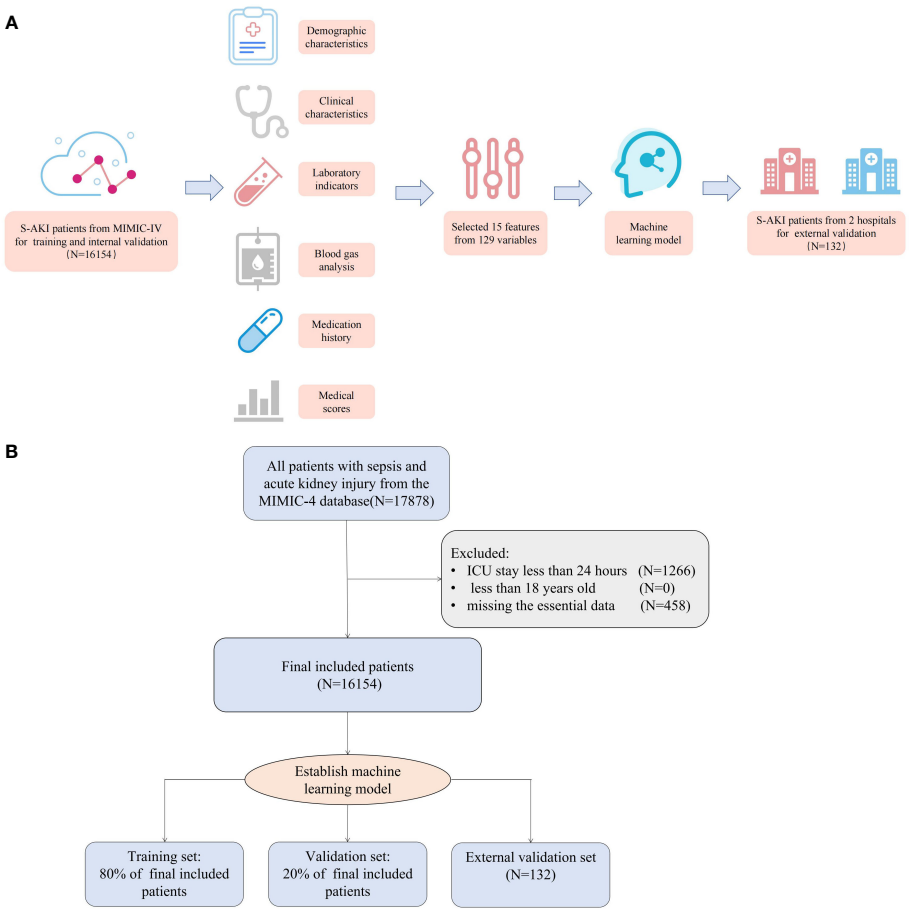

FIGURE 1
(A) The workflow of the study. **(B)** The algorithm chart of the study.

Ethics Committee reviewed and approved this study on 27 April 2022 (protocol number 202204101), which used machine learning to predict all-cause mortality among patients with S-AKI while hospitalized.

Study design and data collection

We collected 129 variables within 24 h of admission. The collected variables included patients' basic information, diagnosis, medication records, clinical data such as temperature, blood pressure, concomitant disease, laboratory indicators, urine output (24-h urine volume after diagnosis of S-AKI), injection rate of norepinephrine (initial concentration of norepinephrine after diagnosis of S-AKI), and commonly used scores such as Simplified Acute Physiology Score II (SAPS-II), SOFA score, and Glasgow Coma Scale (GCS). The external validation set was derived from the electronic health record systems of Xiangya Hospital and Third Xiangya Hospital. The data were collected by two authors (LL and HZ). Data collected by different hospitals were converted and unified. As an example, the injection rate of norepinephrine at 1 mcg/kg/min equaled 1 μg/kg/min. The concentration of creatinine in the blood was 88.4 μmol/L per mg/dl.

Statistical analysis

As appropriate, continuous variables were compared between the death and alive groups using either Student's *t*-test or the rank-sum test. A chi-square test or Fisher's exact test was used to compare categorical variables.

Then, the data were standardized such that the mean value was 0 and the standard deviation was 1. The K-nearest neighbor (KNN) algorithm was used to impute missing values. Next, the dataset was randomly split into a training set (80%) and a validation set (20%). On the training set, the recursive feature elimination (RFE) algorithm was utilized to identify crucial variables, and we developed a machine learning model based on categorical boosting (CatBoost) (20). Basically, RFE is a way of selecting features that recursively fit a model derived from smaller feature sets until a given termination criterion is reached. A feature's importance in the trained model is graded in each loop. In an RFE model, dependencies and collinearities are eliminated by recursively eliminating the lowest-priority feature. As a final step, the most important features were screened out, and the CatBoost model was developed based on the final set of features. Other features were not included because they only brought a small increment in the area under the receiver operating characteristic (AUROC) curve but significantly increased the difficulty of model applications. The trained model was validated on the validation set, and the AUROC curve was calculated correspondingly.

This study compared 10 other machine learning models to the proposed one, namely, KNN, AdaBoost, multilayer perceptron (MLP), support vector machine (SVM), logistic regression (LR), NaiveBayes, gradient boosting decision tree (GBDT), random forest, light gradient boosting (LightGBM), and extreme gradient boosting (XGBoost). These models were also developed on the training set and validated on the validation set. AUROC curves were compared between these models and our CatBoost model. Additionally, other performance measures were examined, such as accuracy (ACC), Youden index, sensitivity, specificity, F1 score, positive predictive value (PPV), and negative predictive value (NPV).

To explain the model, the SHapley Additive exPlanations (SHAP) package in Python was used. A game-theoretic approach is used by the SHAP package to interpret the output of the machine learning model (21). The model was able to connect optimal credit allocation to local explanations for each prediction sample. Two cases were analyzed by using SHAP values to examine model interpretability. The statistical analyses that were carried out in the present study were performed using Python (version 3.7.6); a significance level of $p < 0.05$ was considered to be statistically significant.

Results

Study population

There were 16,154 patients included in the MIMIC-IV set, and relevant information of the cohort can be viewed in Table 1. The average age of the patients was 67.7 years, men accounted for 42.3%, and the average body mass index (BMI) was 30.9. In the cohort, 20.5% of the patients died in the hospital, and their length of stay in the ICU was 3.7 days, longer than that of patients in the alive group. Information of external validation cohort is shown in Supplementary Table 1 and overall workflow and algorithm chart are shown in Figure 1.

Key variables

After utilizing the RFE algorithm, 15 essential variables were selected, namely, urine output, maximum blood urea nitrogen (BUN), rate of injection of norepinephrine, maximum anion gap, maximum creatinine, maximum red blood cell volume distribution width (RDW), minimum international normalized ratio (INR), maximum heart rate, maximum temperature, maximum respiratory rate, minimum fraction of inspired O_2 (FiO_2), minimum creatinine, minimum GCS score, and diagnosis of diabetes and stroke (Figure 2).

Then, machine learning was used for predicting hospital death of patients. The AUC of the proposed CatBoost model was 0.827, which is shown in Figure 3. The CatBoost model markedly outperformed conventional LR (AUC: 0.788) and nine other machine learning models. As described in Table 2, the ACC, best cutoff, Youden index, sensitivity, specificity, F1 score, PPV, and NPV of the CatBoost model were 75%, 19.5%, 50%, 75%, 75%, 56%, 44%, and 92%, respectively. These indicators of LR were 73%, 20.1%, 44%, 71%, 74%, 52%, 41%, and 90%, respectively. In addition, the ROC curve of the validation set reached 0.75, indicating the good applicability of our model (Supplementary Figure 2). To compare with the conventional scoring system, a

10.3389/fimmu.2023.1140755

TABLE 1 Most of the variables that differ between the two groups in the MIMIC-IV set.

Variable		All (*n* = 16,154)	Alive group (*n* = 12,836)	Death group (*n* = 3,318)	*p*-Value
N		16,154	12,836	3,318	
Charlson Index, median [Q1,Q3]		6.0 [4.0,8.0]	6.0 [4.0,8.0]	7.0 [5.0,9.0]	<0.001
Age, mean (SD)		67.7 (15.2)	67.1 (15.2)	70.3 (14.8)	<0.001
Gender, *n* (%)	F	6,836 (42.3)	5,362 (41.8)	1,474 (44.4)	0.006
	M	9,318 (57.7)	7,474 (58.2)	1,844 (55.6)	
Ethnicity, *n* (%)	Asian	377 (2.3)	294 (2.3)	83 (2.5)	<0.001
	Black	1,733 (10.7)	1,421 (11.1)	312 (9.4)	
	Hispanic	538 (3.3)	443 (3.5)	95 (2.9)	
	Other	2,586 (16.0)	1,839 (14.3)	747 (22.5)	
	White	10,920 (67.6)	8,839 (68.9)	2,081 (62.7)	
Liver disease, *n* (%)		3,253 (20.1)	2,096 (16.3)	1,157 (34.9)	<0.001
Stroke, *n* (%)		1,014 (6.3)	666 (5.2)	348 (10.5)	<0.001
BMI, mean (SD)		30.9 (8.8)	31.2 (8.7)	29.5 (8.7)	<0.001
SAPS-II, median [Q1,Q3]		42.0 [34.0,52.0]	40.0 [32.0,49.0]	54.0 [44.0,66.0]	<0.001
SOFA, median [Q1,Q3]		6.0 [4.0,9.0]	5.0 [4.0,8.0]	9.0 [6.0,12.0]	<0.001
GCS, median [Q1,Q3]		15.0 [13.0,15.0]	15.0 [13.0,15.0]	15.0 [12.0,15.0]	<0.001
Heart rate max, mean (SD)		106.1 (21.6)	104.3 (20.5)	113.0 (23.9)	<0.001
Heart rate min, mean (SD)		72.0 (15.9)	71.3 (15.0)	74.6 (18.9)	<0.001
Respiratory rate max, mean (SD)		28.7 (6.7)	28.1 (6.4)	30.7 (7.1)	<0.001
Respiratory rate min, mean (SD)		12.5 (3.9)	12.2 (3.7)	13.4 (4.5)	<0.001
MBP max, mean (SD)		105.0 (28.7)	104.7 (27.3)	106.3 (33.6)	0.016
MBP min, mean (SD)		54.6 (13.4)	55.9 (12.5)	49.7 (15.3)	<0.001
SBP max, mean (SD)		146.4 (23.9)	147.2 (23.3)	143.4 (25.7)	<0.001
SBP min, mean (SD)		85.9 (16.4)	87.6 (15.4)	79.2 (18.3)	<0.001
PaO_2 max, median [Q1,Q3]		174.0 [104.0,321.0]	188.0 [109.0,343.0]	144.0 [94.0,227.0]	<0.001
PaO_2 min, median [Q1,Q3]		84.0 [65.0,111.0]	87.0 [68.0,115.0]	73.0 [56.0,96.0]	<0.001
SpO_2 max, median [Q1,Q3]		100.0 [100.0,100.0]	100.0 [100.0,100.0]	100.0 [100.0,100.0]	<0.001
SpO_2 min, median [Q1,Q3]		92.0 [90.0,95.0]	93.0 [90.0,95.0]	91.0 [86.0,94.0]	<0.001
Temperature max, mean (SD)		37.5 (0.8)	37.5 (0.8)	37.4 (1.0)	<0.001
Temperature min, mean (SD)		36.2 (0.8)	36.3 (0.7)	36.0 (1.1)	<0.001
AST max, median [Q1,Q3]		54.0 [28.0,161.0]	47.0 [26.0,121.0]	91.0 [37.0,345.0]	<0.001
AST min, median [Q1,Q3]		48.0 [26.0,123.0]	42.0 [24.0,97.0]	72.0 [32.0,217.0]	<0.001
PTT max, median [Q1,Q3]		34.4 [29.3,46.4]	33.4 [28.8,42.7]	40.9 [31.7,64.5]	<0.001
PTT min, median [Q1,Q3]		30.7 [27.1,36.7]	30.0 [26.8,35.1]	34.4 [28.8,43.6]	<0.001
Platelet max, median [Q1,Q3]		189.0 [135.0,257.0]	190.0 [139.0,255.0]	183.0 [114.0,268.0]	<0.001
Platelet min, median [Q1,Q3]		160.0 [107.0,226.0]	162.0 [112.0,226.0]	151.0 [84.0,228.0]	<0.001
RBC max, mean (SD)		3.6 (0.7)	3.6 (0.7)	3.5 (0.8)	<0.001
RBC min, mean (SD)		3.2 (0.7)	3.3 (0.7)	3.2 (0.8)	<0.001
WBC max, median [Q1,Q3]		13.5 [9.5,18.6]	13.2 [9.4,18.0]	15.0 [10.0,21.2]	<0.001

(Continued)

10.3389/fimmu.2023.1140755

TABLE 1 Continued

Variable		All (n = 16,154)	Alive group (n = 12,836)	Death group (n = 3,318)	p-Value
WBC min, median [Q1,Q3]		10.4 [7.3,14.5]	10.2 [7.2,13.8]	11.7 [7.4,17.0]	<0.001
RDW max, mean (SD)		15.9 (2.5)	15.6 (2.4)	16.9 (2.8)	<0.001
RDW min, mean (SD)		15.5 (2.4)	15.3 (2.3)	16.5 (2.7)	<0.001
Glucose max, median [Q1,Q3]		143.0 [115.0,194.0]	140.0 [114.0,186.0]	162.0 [122.0,225.2]	<0.001
Glucose min, median [Q1,Q3]		115.0 [95.0,141.0]	115.0 [96.0,139.0]	114.0 [88.0,148.0]	0.017
Lactate max, median [Q1,Q3]		2.3 [1.5,3.8]	2.2 [1.5,3.3]	3.5 [1.9,7.2]	<0.001
Lactate min, median [Q1,Q3]		1.6 [1.2,2.3]	1.5 [1.1,2.1]	2.2 [1.4,3.8]	<0.001
BUN max, median [Q1,Q3]		27.0 [18.0,45.0]	25.0 [17.0,41.0]	38.0 [25.0,58.0]	<0.001
BUN min, median [Q1,Q3]		23.0 [15.0,39.0]	22.0 [15.0,35.0]	32.0 [21.0,52.0]	<0.001
Creatinine max, median [Q1,Q3]		1.4 [0.9,2.5]	1.3 [0.9,2.2]	1.9 [1.2,3.1]	<0.001
Creatinine min, median [Q1,Q3]		1.2 [0.8,2.1]	1.1 [0.8,1.9]	1.6 [1.0,2.6]	<0.001
Urine output, median [Q1,Q3]		1,040.0 [537.0,1,665.0]	1,150.0 [675.0,1,760.0]	605.0 [186.0,1,110.0]	<0.001
RRT, n (%)		1,633 (10.1)	1,135 (8.8)	498 (15.0)	<0.001
IMV, n (%)		9,518 (58.9)	7,398 (57.6)	2,120 (63.9)	<0.001
Vasopressor support, n (%)		5,912 (36.6)	3,942 (30.7)	1,970 (59.4)	<0.001
Rate of norepinephrine, median [Q1,Q3]		0.0 [0.0,0.1]	0.0 [0.0,0.0]	0.1 [0.0,0.4]	<0.001
IMV durations, median [Q1,Q3]		0.4 [0.0,2.6]	0.2 [0.0,1.9]	1.6 [0.0,5.1]	<0.001
Hospital mortality, n (%)		3,318 (20.5)	0(0.0)	3,318 (100.0)	<0.001
Length of ICU stay, median [Q1,Q3]		3.0 [1.7,6.0]	2.9 [1.7,5.6]	3.7 [1.7,7.5]	<0.001
Length of hospital stay, median [Q1,Q3]		8.7 [5.2,15.1]	9.1 [5.8,15.7]	6.3 [2.4,13.3]	<0.001

SD, standard deviation; BMI, body mass index; SAPS-II, Simplified Acute Physiology Score II; SOFA, Sequential Organ Failure Assessment; GCS, Glasgow Coma Scale; MBP, mean blood pressure; SBP, systolic blood pressure; PaO_2, partial pressure of oxygen; SpO_2, saturation of pulse oxygen; AST, aspartate aminotransferase; PTT, partial thromboplastin time; RBC, red blood cell; WBC, white blood cell; RDW, red blood cell volume distribution width; BUN, blood urea nitrogen; RRT, renal replacement therapy; IMV, intermittent mandatory ventilation; ICU, intensive care unit.

CatBoost model for the SOFA score was made, and the results show that the prediction ability of SOFA is inferior to the proposed model in the training and validation set (Supplementary Figure 3). As AST was almost double in the death group, and in the raw data, the number of patients with AST greater than 45 U/L was almost equal to the number of patients with liver disease. Therefore, a CatBoost model was also established to conduct a liver disease subgroup analysis that also demonstrates a good prediction power on the mortality of S-AKI among these subgroup patients (Supplementary Figure 4).

Application of the model

Analyzing the integral cohort by the SHAP package showed the crucial variables for predicting death (Figure 4). Input the information of a patient into the model: history of stroke, minimum GCS score of 15, maximum heart rate of 121 beats per minute, maximum temperature of 36.56°C, maximum respiratory rate of 68 breaths per minute, maximum BUN level of 73 mg/dl, minimum INR of 2.9, maximum creatinine level of 3 mg/dl, minimum creatinine level of 2.1 mg/dl, maximum RDW of 16.8%, minimum FiO_2 of 100%, maximum anion gap of 31 mEq/L, urine output of 405 ml/day, and a rate of injection of norepinephrine of 0.499 mcg/kg/min. The model showed that the risk of hospital mortality was 28.9% (higher than the best cutoff), suggesting that the patient had a high risk of death (Example 1, Figure 4). Input the information of another patient into the model: no history of stroke or diabetes, minimum GCS score of 15, maximum heart rate of 86 beats per minute, maximum temperature of 36.94°C, maximum respiratory rate of 28 breaths per minute, maximum BUN level of 74 mg/dl, minimum INR of 1.1, maximum creatinine level of 4.1 mg/dl, minimum creatinine level of 3.5 mg/dl, maximum RDW of 14.9%, minimum FiO_2 of 70%, maximum anion gap of 18 mEq/L, urine output of 1,060 ml, and a rate of injection of norepinephrine of 0 mcg/kg/min. The probability of hospital mortality was predicted to be 18.37%, suggesting a good prognosis (Example 2, Figure 4).

Discussion

Machine learning has been widely applied to solve medical and clinical problems, by which it has become a popular research topic.

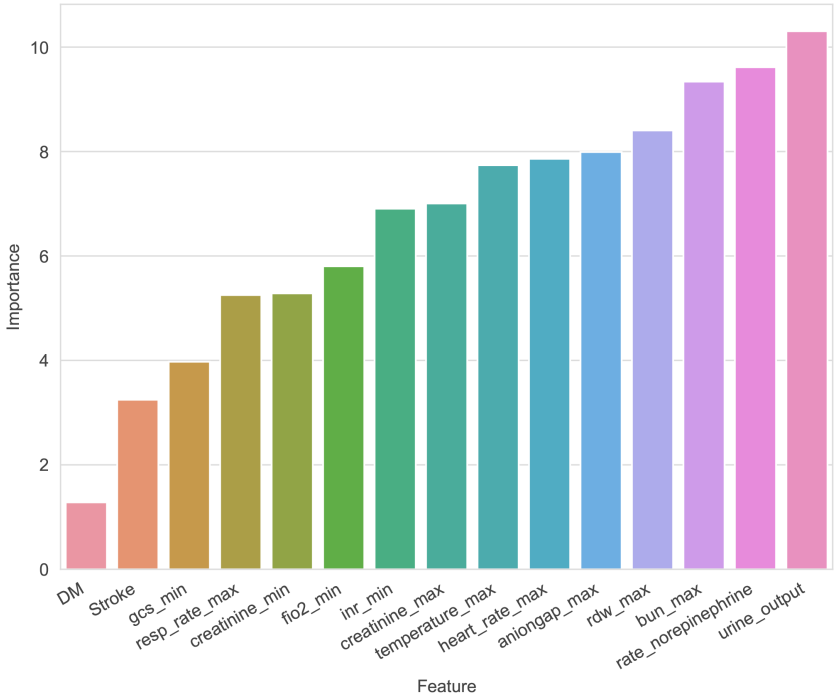

FIGURE 2
The importance of each feature to the machine learning model.

Based on their shortcomings, novel biomarkers and conventional scoring systems lack enough power to estimate the mortality of S-AKI patients (12, 13). In this article, we discussed whether machine learning improves the mortality prediction of S-AKI patients and then selected the model with the strongest prediction ability.

From the MIMIC-IV database used as a training set, 15 crucial variables were selected using the RFE algorithm. These variables are common in various clinical settings, which means information on them can be easily obtained, and the application of machine learning models will not be limited to a variable that is difficult to

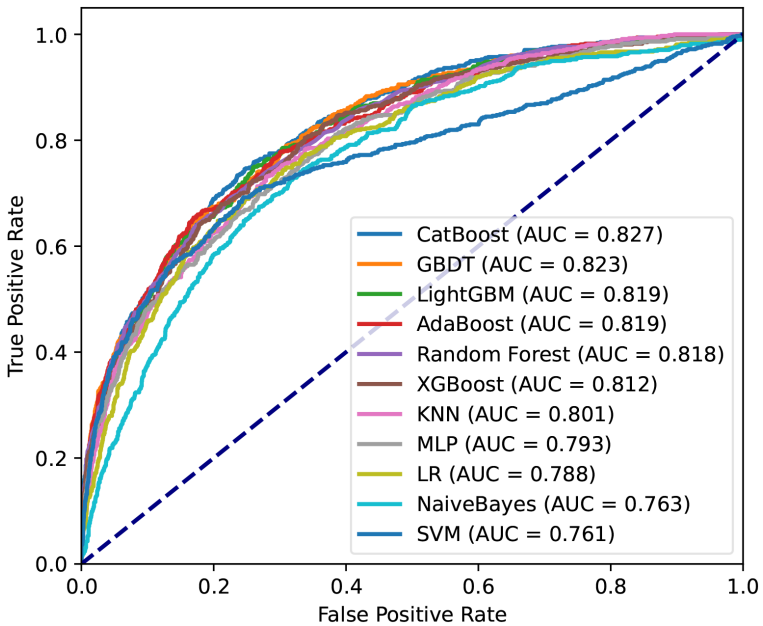

FIGURE 3
Receiver operating characteristic curves for the machine learning model and logistic regression in the training set. CatBoost, categorical boosting; GBDT, gradient boosting decision tree; LightGBM, light gradient boosting; AdaBoost, adaptive boosting; XGBoost, extremely gradient boosting; KNN, K-nearest neighbor; MLP, multilayer perceptron; LR, logistic regression. SVM, support vector machine.

10.3389/fimmu.2023.1140755

TABLE 2 Performance of machine learning models.

Model	AUC	ACC (%)	Best cutoff	Youden index (%)	Sensitivity (%)	Specificity (%)	F1 score	PPV (%)	NPV (%)
CatBoost	0.83	75	0.195	50	75	75	0.56	44	92
GBDT	0.82	71	0.16	48	79	69	0.53	40	93
LightGBM	0.82	74	0.183	49	75	74	0.55	43	92
AdaBoost	0.82	79	0.494	48	65	83	0.57	51	90
Random Forest	0.82	78	0.28	47	66	81	0.55	48	90
XGBoost	0.81	77	0.204	47	68	79	0.55	46	90
KNN	0.8	72	0.176	45	73	72	0.52	41	91
MLP	0.79	73	0.162	43	70	73	0.52	41	90
LR	0.79	73	0.201	44	71	74	0.52	41	90
NaiveBayes	0.76	68	0.092	41	74	67	0.49	37	91
SVM	0.76	74	0.149	45	69	75	0.53	43	90

CatBoost, categorical boosting; GBDT, gradient boosting decision tree; LightGBM, light gradient boosting; AdaBoost, adaptive boosting; XGBooST, extremely gradient boosting; KNN, K-nearest neighbor; MLP, multilayer perceptron; LR, logistic regression. SVM, support vector machine; ACC, accuracy, PPV, positive predictive value; NPV, negative predictive value.

FIGURE 4
Two examples of website tool usage. Enter the values of 15 key variables to predict the risk of death and show the contribution of each value to the outcome. Example 1 has a higher risk of death, and example 2 may have a better prognosis.

detect. Some studies have focused on the relative importance of each variable in predicting prognosis. For example, a retrospective study from a prospective cohort conducted by Sukmark et al. suggested that a lower GCS score was associated with in-ICU mortality with an adjusted odds ratio of 4.16 (3.10, 5.60) (22). Serum creatinine has been extensively utilized as a predictor in severity scores that assess renal function and adverse effects of renal dysfunction, such as SOFA and APACHE II. In addition, it has been reported that BUN is associated with multiorgan failure of ICU patients regardless of admission diagnosis, including kidney failure and long-term mortality (23). Sukmark also elaborated that BUN possibly reflected multiorgan failure better than serum creatinine (22). As mentioned before, some variables were found to be correlated with prognosis. However, few have put them into one prediction model and successfully quantified their ability to predict mortality.

After identifying these 15 variables, machine learning was applied to predict the mortality of patients during hospitalization. CatBoost is an open-source package and a new GBDT algorithm announced in 2017. Compared to other GBDT algorithms, it outperformed in handling categorical variables and reducing overfitting (24). To prove the efficiency of the CatBoost model, it was compared with 10 other machine learning models and SOFA. Satisfactorily, the proposed model significantly outperformed the others with an AUC of 0.827. Furthermore, we collected data from Xiangya Hospital and Third Xiangya Hospital, Central South University, China, to use as an external validation set. The ROC curve of the validation set was also as high as 0.754.

Compared with several other S-AKI-related clinical model studies (16–18), the innovation of this study is that the fourth edition of the MIMIC database used includes more patients from 2017 to 2019 than the third edition, with a larger amount of data and more recent data. In addition, in contrast to the related studies, emphasis was placed on predicting the mortality of S-AKI patients for the first time. Second, this study not only utilized data from the database but also collected data from other hospitals for validation, making the model more reliable. In addition, our training set is from Western countries, while the validation set is from China, indicating that the model has applicability among different populations. Moreover, instead of just using one machine learning algorithm to build the model, we compared multiple machine learning algorithms and selected the one that performed the best. Finally, since the chosen variables are easily accessible, the prediction model has a wide range of applications in areas with different medical levels.

However, our study has some limitations. First, the training set data originated from only one database, while the validation set data came from two hospitals in one region; thus, selection bias may have occurred. Even in view of this, the proposed model constructed by the MIMIC-IV database still passed the validation set from China, which, in turn, proved the superiority of our model. However, we must admit that more external validations are needed. Second, the variables were selected by the RFE algorithm, but the underlying mechanism was not discussed in our study.

As found in previous studies, S-AKI patients were treated with mechanical ventilation and vasoactive therapy with greater possibility (9), so was dialysis (70%) (11), which was simultaneously associated with a longer hospital stay (5). Prolonging hospital stays and expensive treatments mean an increasingly larger economic burden on patients and medical insurance. Meanwhile, it is sometimes challenging for clinicians to decide the priority treatment in the next step when condition deteriorates rapidly. Consequently, applying the CatBoost-based model to discern high-risk S-AKI patients and predict prognoses in a timely and accurate manner and providing clinicians with optimal treatment decision-making suggestions may help reduce these burdens. In conclusion, we hope that the proposed model will assist clinicians with better decision-making and allocating medical resources reasonably.

Conclusions

This study demonstrates that predicting the mortality of S-AKI patients in the ICU is critical and that the CatBoost-based model we proposed outperformed conventional LR and nine other machine learning models. Further validations across diverse study centers will help verify the reliability and improve the validation efficiency of this model.

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

This study was reviewed and approved by the Ethics Committee of the Xiangya Hospital of Central South University on 27 April 2022 (protocol number 202204101).

Author contributions

HZ: resources and writing—original draft; LPL: methodology, resources, validation, visualization, and writing—original draft; QZ: formal analysis, methodology, and validation; XJ: resources; ZP: investigation; WW: investigation; LH: investigation; YX: investigation; HX: supervision; LT: supervision; XX: supervision; WN: investigation; FL: review and editing; LL: review and editing; QY: review and editing and supervision. All authors contributed to the article and approved the submitted version.

Funding

This work was supported by the Natural Science Foundation of Hunan province China (Grant Nos. 2020JJ5942, 2019JJ40515, and 2019JJ20035), the Major Program of the National Natural Science Foundation of China (Grant No. 82090024), the General Programs

of the National Natural Science Foundation of China (Grant No. 82173877), and the Key Research and Development Program of Hunan Province (Grant No. 2021SK2015).

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/fimmu.2023.1140755/full#supplementary-material

SUPPLEMENTARY FIGURE 1
TRIPOD Checklist: Prediction Model Development.

SUPPLEMENTARY FIGURE 2
Receiver operating characteristic curves for the machine learning model and logistic regression in the validation set.

SUPPLEMENTARY FIGURE 3
Receiver operating characteristic curves for the machine learning model of SOFA score. **(A)** ROC of the training set. **(B)** ROC of the validation set.

SUPPLEMENTARY FIGURE 4
Receiver operating characteristic curves for the machine learning model of liver disease subgroup. **(A)** ROC of the training set. **(B)** ROC of the validation set.

SUPPLEMENTARY TABLE 1
Demographic and comorbidity information on the external validation cohort.

References

1. Cecconi M, Evans L, Levy M, Rhodes A. Sepsis and septic shock. *Lancet* (2018) 392:75–87. doi: 10.1016/S0140-6736(18)30696-2

2. Prescott HC, Angus DC. Enhancing recovery from sepsis: A review. *JAMA* (2018) 319:62–75. doi: 10.1001/jama.2017.17687

3. 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

4. Palevsky PM, Liu KD, Brophy PD, Chawla LS, Parikh CR, Thakar CV, et al. KDOQI US commentary on the 2012 KDIGO clinical practice guideline for acute kidney injury. *Am J Kidney Dis* (2013) 61:649–72. doi: 10.1053/j.ajkd.2013.02.349

5. Bagshaw SM, Uchino S, Bellomo R, Morimatsu H, Morgera S, Schetz M, et al. Septic acute kidney injury in critically ill patients: clinical characteristics and outcomes. *Clin J Am Soc Nephrol* (2007) 2:431–9. doi: 10.2215/CJN.03681106

6. Peerapornratana S, Manrique-Caballero CL, Gómez H, Kellum JA. Acute kidney injury from sepsis: current concepts, epidemiology, pathophysiology, prevention and treatment. *Kidney Int* (2019) 96:1083–99. doi: 10.1016/j.kint.2019.05.026

7. Liu J, Xie H, Ye Z, Li F, Wang L. Rates, predictors, and mortality of sepsis-associated acute kidney injury: a systematic review and meta-analysis. *BMC Nephrol* (2020) 21:318. doi: 10.1186/s12882-020-01974-8

8. Bouchard J, Acharya A, Cerda J, Maccariello ER, Madarasu RC, Tolwani AJ, et al. A prospective international multicenter study of AKI in the intensive care unit. *Clin J Am Soc Nephrol.* (2015) 10:1324–31. doi: 10.2215/CJN.04360514

9. 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

10. Xu X, Nie S, Liu Z, Chen C, Xu G, Zha Y, et al. Epidemiology and clinical correlates of AKI in Chinese hospitalized adults. *Clin J Am Soc Nephrol* (2015) 10:1510–8. doi: 10.2215/CJN.02140215

11. Honoré PM, Jacobs R, Boer W, Joannes-Boyau O. Sepsis and AKI: more complex than just a simple question of chicken and egg. *Intensive Care Med* (2011) 37:186–9. doi: 10.1007/s00134-010-2097-9

12. Bellomo R, Kellum JA, Ronco C, Wald R, Martensson J, Maiden M, et al. Acute kidney injury in sepsis. *Intensive Care Med* (2017) 43:816–28. doi: 10.1007/s00134-017-4755-7

13. 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

14. Handelman GS, Kok HK, Chandra RV, Razavi AH, Lee MJ, Asadi H. eDoctor: machine learning and the future of medicine. *J Intern Med* (2018) 284:603–19. doi: 10.1111/joim.12822

15. Deo RC. Machine learning in medicine. *Circulation* (2015) 132:1920–30. doi: 10.1161/CIRCULATIONAHA.115.001593

16. Chaudhary K, Vaid A, Duffy Á, 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

17. He J, Lin J, Duan M. Application of machine learning to predict acute kidney disease in patients with sepsis associated acute kidney injury. *Front Med (Lausanne)* (2021) 8:792974. doi: 10.3389/fmed.2021.792974

18. Luo XQ, Yan P, Zhang NY, Luo B, Wang M, Deng YH, 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

19. Hancock JT, Khoshgoftaar TM. CatBoost for big data: an interdisciplinary review. *J Big Data.* (2020) 7:94. doi: 10.1186/s40537-020-00369-8

20. Zhao QY, Liu LP, Luo JC, Luo YW, Wang H, Zhang YJ, et al. A machine-learning approach for dynamic prediction of sepsis-induced coagulopathy in critically ill patients with sepsis. *Front Med (Lausanne)* (2021) 7:637434. doi: 10.3389/fmed.2020.637434

21. Lundberg SM, Erion G, Chen H, DeGrave A, Prutkin JM, Nair B, et al. From local explanations to global understanding with explainable AI for trees. *Nat Mach Intell* (2020) 2:56–67. doi: 10.1038/s42256-019-0138-9

22. Sukmark T, Lumlertgul N, Praditpornsilpa K, Tungsanga K, Eiam-Ong S, Srisawat N. THAI-ICU score as a simplified severity score for critically ill patients in a resource limited setting: Result from SEA-AKI study group. *J Crit Care* (2020) 55:56–63. doi: 10.1016/j.jcrc.2019.10.010

23. Arihan O, Wernly B, Lichtenauer M, Franz M, Kabisch B, Muessig J, et al. Blood urea nitrogen (BUN) is independently associated with mortality in critically ill patients admitted to ICU. *PloS One* (2018) 13:e0191697. doi: 10.1371/journal.pone.0191697

24. Dorogush AV, et al. CatBoost: gradient boosting with categorical features support. Available at: https://arxiv.org/abs/1810.11363 (Accessed September 1, 2020).

frontiers | Frontiers in Medicine

TYPE Original Research
PUBLISHED 17 April 2023
DOI 10.3389/fmed.2023.1154041

OPEN ACCESS

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

REVIEWED BY
Bogdan Silviu Ungureanu,
University of Medicine and Pharmacy of Craiova, Romania
Toshiyo Tamura,
Waseda University, Japan

*CORRESPONDENCE
Sondre Heimark
✉ sondhe@ous-hf.no

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

RECEIVED 30 January 2023
ACCEPTED 14 March 2023
PUBLISHED 17 April 2023

CITATION
Heimark S, Bøtker-Rasmussen KG, Stepanov A, Haga ØG, Gonzalez V, Seeberg TM, Fadl Elmula FEM and Waldum-Grevbo B (2023) Accuracy of non-invasive cuffless blood pressure in the intensive care unit: Promises and challenges.
Front. Med. 10:1154041.
doi: 10.3389/fmed.2023.1154041

Accuracy of non-invasive cuffless blood pressure in the intensive care unit: Promises and challenges

Sondre Heimark[1,2]*, Kasper Gade Bøtker-Rasmussen[3,4], Alexey Stepanov[3], Øyvind Gløersen Haga[4], Victor Gonzalez[4], Trine M. Seeberg[3,4], Fadl Elmula M. Fadl Elmula[5] and Bård Waldum-Grevbo[1]

[1]Department of Nephrology, Oslo University Hospital, Ullevål, Oslo, Norway, [2]Institute of Clinical Medicine, University of Oslo, Oslo, Norway, [3]Aidee Health AS, Oslo, Norway, [4]Department of Smart Sensors and Microsystems, SINTEF Digital, Oslo, Norway, [5]Cardiorenal Research Centre, Oslo University Hospital, Ullevål, Oslo, Norway

Objective: Continuous non-invasive cuffless blood pressure (BP) monitoring may reduce adverse outcomes in hospitalized patients if accuracy is approved. We aimed to investigate accuracy of two different BP prediction models in critically ill intensive care unit (ICU) patients, using a prototype cuffless BP device based on electrocardiogram and photoplethysmography signals. We compared a pulse arrival time (PAT)-based BP model (generalized PAT-based model) derived from a general population cohort to more complex and individualized models (complex individualized models) utilizing other features of the BP sensor signals.

Methods: Patients admitted to an ICU with indication of invasive BP monitoring were included. The first half of each patient's data was used to train a subject-specific machine learning model (complex individualized models). The second half was used to estimate BP and test accuracy of both the generalized PAT-based model and the complex individualized models. A total of 7,327 measurements of 15s epochs were included in pairwise comparisons across 25 patients.

Results: The generalized PAT-based model achieved a mean absolute error (SD of errors) of 7.6 (7.2) mmHg, 3.3 (3.1) mmHg and 4.6 (4.4) mmHg for systolic BP, diastolic BP and mean arterial pressure (MAP) respectively. Corresponding results for the complex individualized model were 6.5 (6.7) mmHg, 3.1 (3.0) mmHg and 4.0 (4.0) mmHg. Percentage of absolute errors within 10mmHg for the generalized model were 77.6, 96.2, and 89.6% for systolic BP, diastolic BP and MAP, respectively. Corresponding results for the individualized model were 83.8, 96.2, and 94.2%. Accuracy was significantly improved when comparing the complex individualized models to the generalized PAT-based model in systolic BP and MAP, but not diastolic BP.

Conclusion: A generalized PAT-based model, developed from a different population was not able to accurately track BP changes in critically ill ICU patients. Individually fitted models utilizing other cuffless BP sensor signals significantly improved accuracy, indicating that cuffless BP can be measured non-invasively, but the challenge toward generalizable models remains for future research to resolve.

KEYWORDS

cuffless, blood pressure, pulse arrival time, machine learning, intensive care unit

1. Introduction

At present, blood pressure (BP) monitoring in hospitalized patients is limited to either intermittent cuff-based measurements or invasive arterial catheterization. Invasive arterial BP monitoring is the only method capable of accurate in-hospital continuous BP monitoring and is considered the gold standard given correct operating conditions. However, it is only available during surgery, post-operatively or in intensive care units (ICU) and requires specialized personnel. In addition, arterial catheterization carries risk such as bleeding, arterial occlusion and infection. For the remainder of hospitalized patients, BP is taken intermittently at varying intervals. Undetected hypotensive episodes may lead to organ damage such as acute kidney injury, and undetected clinical deterioration may delay adequate treatment and lead to adverse outcomes (1, 2). Studies indicate that adverse events are related to the intermittent nature of vital signs monitoring on hospital wards (3, 4). Thus, there is a clear need for non-invasive continuous cuffless BP monitoring in hospitalized patients to bridge the gap between intermittent cuff-based measurements and invasive arterial catheterization.

Despite substantial research on methods to enable non-invasive cuffless BP monitoring, its general accuracy remains uncertain, and few studies have investigated accuracy in critically ill patients. In addition, non-invasive cuffless BP methods use different approaches such as pulse wave propagation-based measurements (such as pulse arrival time (PAT)) and photo-plethysmography (PPG) waveform features. Studies, including research performed by our multidisciplinary team, have shown strong correlations between PAT and BP, particularly during various exercise methods (5–9) but its accuracy across differing populations and hemodynamic conditions are uncertain (6). New advances in non-invasive cuffless BP indicate that complex modeling by machine learning methods of sensor-based measurements are key toward improved results (6). In the present study, we aimed to investigate accuracy of two different BP-prediction models using the signals from a prototype chest belt BP sensor in critically ill patients. Specifically, we investigated a PAT-based model, derived from a general population cohort (generalized PAT-based model) compared to continuous invasive BP measurements and compared it with accuracy of individually fitted machine learning models (complex individualized models) that utilized other features of the signals obtained by the cuffless BP sensor.

2. Materials and methods

2.1. Subjects

Patients older than 18 years admitted to the general medical ICU at Oslo University Hospital, Ullevål were considered for inclusion. Inclusion criteria were signed consent and an inserted arterial line. Exclusion criteria were ongoing arrythmias generating irregular R-R intervals, failure to obtain adequate signals from the cuffless device or any medical contraindication to having a chest belt mounted. Each patient was monitored for a duration of 1–12 h, depending on length of stay, discontinuation of the intra-arterial catheter or other clinical interruptions.

2.2. Reference blood pressure

Reference BP was measured continuously with a radial artery catheter connected by a fluid filled tube to a pressure transducer (Xtrans; Codan, Forstinning, Germany). The pressure transducer was leveled at the phlebostatic axis and had a saline flush connected with a counterpressure of approximately 300 mmHg. The system was connected to a Philips IntelliVue MX 800 patient monitor (Philips, Böblingen, Germany). Zeroing was performed every 8-h according to the ICUs procedures. All vital signs, including the raw arterial waveform and the monitor-generated absolute BP values sampled every 5 s, were recorded directly to a laptop *via* an RS-232 connection using the Vital Recorder software (10).

2.3. Cuffless blood pressure device

A prototype cuffless BP sensor (cuffless BP device) was used in this study (7–9). It consists of a one-channel electrocardiogram (ECG) sensor, a photo-plethysmography (PPG) sensor and an inertial measurement unit (3D accelerometer and 3D gyroscope) integrated in a wearable chest belt. Raw signals from the ECG and PPG sensors were sampled at 1,000 Hz, while accelerometer data was sampled at 208 Hz and gyroscope data that were sampled at 26 Hz. The gyroscope data was not used. The cuffless BP device was fitted as illustrated in Figure 1. The generalized PAT-based model was developed from BP changes during isometric exercise in a general population cohort (9), using PAT and HR as cuffless surrogates but not any demographic information. A linear best fit equation with a coefficient for PAT, a coefficient for interaction between PAT and HR (this term was negligible) and a coefficient for HR was used. Additionally, we computed a best fit linear model using only PAT. The complex individualized models, utilizing other signal features, were trained using the first half of each patient's data. Thus, the test period for both models were defined as the second half of each patient's data. The cuffless BP device was calibrated against the first three minutes of reference BP at the start of each test period. This was a simple static calibration to correct the offset between average reference BP and cuffless BP across the initial three minutes. Since the pressure transducer was mounted on a bracket next to the patient bed, temporary periods occurred of which the pressure transducer moved relative to the phlebostatic axis. To reliably exclude such periods, an investigator continuously observed all data collections. In addition, if the pressure transducer moved significantly during such a period and was relevelled by the ICU staff, the cuffless BP device was re-calibrated against reference BP during the test period. Recalibration occurred in 14 patients (once in seven patients, twice in four patients and three times in two patients). Reasons for recalibration were related to nursing care, changing from supine bed rest to seated position or temporary detachment from the invasive monitoring system because of imaging studies or bathroom visits. Recalibration was decided necessary to avoid systematic biases introduced during relevelling. For example, if the pressure transducer was relevelled one time during a patient's data collection with an offset of 5 cm relative to the previous leveling, a systematic bias of 3.7 mmHg would be introduced for the remaining observation time.

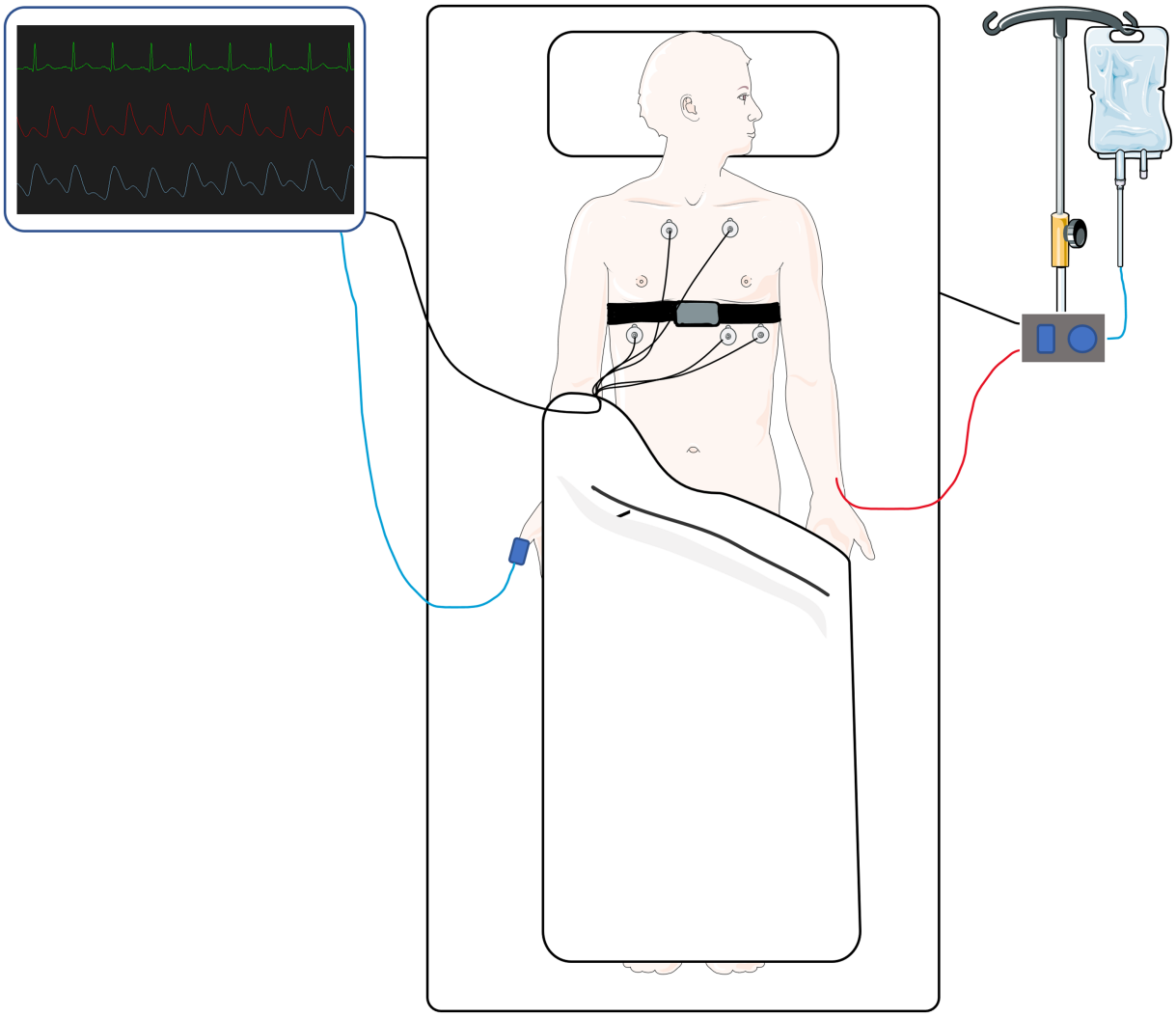

FIGURE 1
A simplified illustration of the chest belt device (cuffless device) fitted on a patient in the intensive care unit alongside basic monitoring equipment. Parts of the figure were created by using pictures from Servier Medical Art. Servier Medical Art by Servier is licensed under a Creative Commons Attribution 3.0 Unported License (https://creativecommons.org/licenses/by/3.0/).

2.4. Data analysis

2.4.1. Patient selection

Of 44 patients, 25 were available for the present study after exclusions (Figure 2). Prior to data analysis six patients were excluded due to the following reasons: (1) excessive movement causing the transducer to move relative to the leveling set point and excessive noise ($n = 2$), (2) arterial catheter failure ($n = 2$), (3) irregular RR intervals from pacemaker ($n = 1$) and (4) erroneous vital recorder data capture ($n = 1$). Thus, 38 patients were included in the formal data analysis. Next, the cuffless BP device data was processed to allow for proper training of the complex individualized models and 13 of the 38 patients were excluded because one or more of three criteria were met: (1) Ratio of valid device signals to reference data above 0.6 ($n = 9$), (2) short recordings (total number of reference and cuffless datapoints below 200) ($n = 11$) and (3) to ensure that adequate BP variation was available for the machine learning algorithm, the standard deviation of reference BP in the first half had to be at least 50% of the standard deviation of the reference BP for the whole duration of each individuals data ($n = 3$). Most patients met the criteria related to signal quality and number of reference and device measurement pairs.

2.4.2. Data filtering and processing

Filtering and processing of the data was performed post-hoc in a custom-made database using the Python programming language. Reference BP values were extracted from the raw arterial waveforms. The raw arterial waveform signals were filtered both manually and automatically to reliably remove artefacts from around arterial blood sampling, detachments and re-attachments to the arterial monitoring system, compression of waveforms from wrist flexion, cuff measurements taken at the same arm and high frequency noise. After filtering, reference BP and cuffless BP estimations from the two models were averaged on 15 s epochs. To allow for direct comparison between the two cuffless models, pairwise comparisons between cuffless BP and reference BP were made on the same data in each

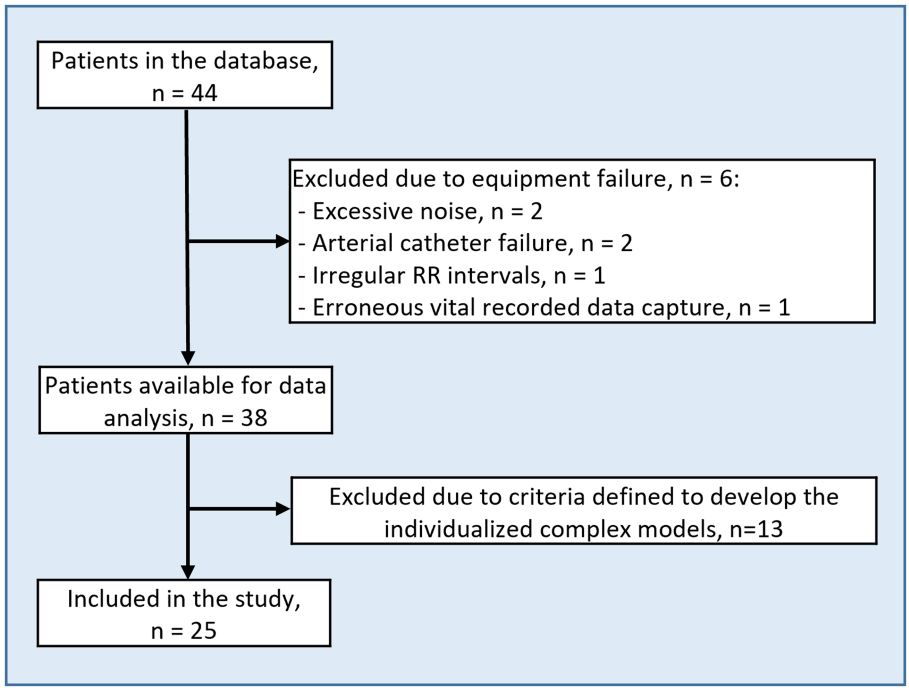

FIGURE 2
Flow chart of patient selection.

patient, i.e., the test period defined as the last 50% of data for each patient.

2.4.3. Statistical analyses

Statistical analyses were performed using Stata (StataCorp. 2019. Stata Statistical Software: Release 16. College Station, TX: StataCorp LLC). Data is presented as mean (standard deviation (SD)) or median (interquartile range) if non-normal distribution. We computed mean errors, mean absolute errors (MAE), SD of errors and Bland–Altman plots with bias and 95% limits of agreement (LOA). We are aware that pooling all measurement pairs across all patients may violate the Bland–Altman assumption of independent measurements (11). However, all comparable studies have pooled all measurements in Bland–Altman analyses (12–15). Thus, we chose same methodology for comparative purposes. We also computed Bland–Altman bias and LOA using a proposed method for repeated measures (16) which resulted in bias and LOA (not reported) with negligible differences from the pooled analyses. Correlation analysis was performed using repeated measures correlation as proposed by Bland and Altman (17). In this way the dependency of repeated within subjects are correctly handled. To be able to compare with similar studies, Pearson's correlation coefficients were also calculated for all measurements across all subjects pooled together.

Comparison of model performance was analyzed in three steps. First, we compared error estimations to determine if they were different from each other. The absolute errors of all measurement pairs ($n = 7{,}327$) were compared by a non-parametric test for equality of means. Equality of the standard deviation of the errors were compared using a variance comparison test. Second, aggregated BP means per subject from reference BP, the generalized PAT-based model, and the complex individualized model were computed. These means were fitted with the corresponding reference values in a linear regression model for the two models. As these models are not nested, they could not be directly compared by any statistical test. Thus, they were compared numerically on the coefficient of determination (R^2), root mean squared error and Akaike's and the Bayesian information criterion. Finally, the predictive accuracy of the two models were tested using the Diebold-Mariano predictive accuracy test. The stationary assumption was tested using the augmented Dickey-Fuller test. Sensitivity of the predictive accuracy test, as the stationary assumption may not hold regardless of the result of the augmented Dickey-Fuller test because the data is comprised of different subjects, were tested by performing the Diebold-Mariano test in each subject separately. The overall significance was tested using Fisher's method of combining p values. To test the influence of HR as an additional parameter in the PAT-based model, we also predicted BP using a PAT-only model derived from the data as the PAT and HR-based model. A value of p below 0.05 was considered statistically significant.

3. Results

Patient characteristics are presented in Table 1 and distribution of reference BP across all patients are presented in Table 2. The average number of pairwise comparisons (SD) between reference and the cuffless BP device per subject were 293.2 (161.2), ranging from 124 to 754 with a total of 7,327. Median (Interquartile range) observation time was 4.0 (3.1–4.6) hours with a range from 1.4–8.0 h. Performance of the generalized PAT-based model compared to the complex individualized models are presented in Table 3. The complex individualized models were numerically superior to the generalized PAT-based model across all parameters. Particularly when comparing

the repeated measures correlation, more covariation was captured by the complex individualized models compared to the generalized PAT-based model for SBP and MAP where repeated measures correlation coefficients were 0.23 vs. 0.39 and 0.25 vs. 0.37. Results were more similar for DBP compared to SBP and MAP with correlation coefficients of 0.29 (generalized PAT-based model) vs. 0.33 (complex individualized models). Bland–Altman plots with bias and LOA are presented in Figure 3. Bias was close to zero for all BP parameters in both models; −0.2 mmHg vs. −1.4 mmHg, −0.2 vs. 0.0 mmHg and 0.1 mmHg vs. −0.9 mmHg for the generalized PAT-based model vs. the complex individualized models regarding SBP, DBP, and MAP, respectively. LOA favored the complex individualized models for SBP [−21.5, 21.1 mmHg] vs. [−19.2, 16.2 mmHg] and MAP [−13.4, 13.5 mmHg] vs. [−13.9, 11.4 mmHg] but were similar for DBP [−9.8, 9.8 mmHg] vs. [−9.6, 9.6 mmHg]. Percentages of absolute errors within 15, 10 and 5 mmHg (Table 4) also favored the complex individualized models where all percentages were numerically higher for the complex individualized models except for within 15 mmHg regarding DBP. The complex individualized models were significantly different from and outperformed the generalized PAT-based model for SBP and MAP. To the contrary, for DBP, the SD of the errors were not significantly different, and the Diebold-Mariano test of predictive accuracy was not significant. Comparison of the PAT and HR-based model to a PAT-only model showed negligible differences. Pearson's correlation coefficient and R^2 between the two models were 0.999 and 0.997, respectively.

An important difference between the generalized PAT-based model and the complex individualized models appeared during the detailed data inspection The generalized PAT-based model performed inadequately in cases of decreasing BP with corresponding heart rate (HR) increase. Therefore, we plotted four different timeseries plots (Figure 4) of four different patients where reduction in BP was coupled with a rise in HR. In the first case (upper left panel) both models were unable to predict the BP reduction, while for the remaining cases, only the complex individualized models correctly predicted the direction of change in BP. Importantly, regarding periods of reduction in BP coupled with a rise in HR, the generalized PAT-based model compared to the PAT-only model showed negligible differences.

TABLE 1 Patient characteristics.

Sex, male no (%)	18 (72)
Age, years (SD), range	62.0 (15.4), 27–89
Body mass index, Kg/m^2(SD)	27.1 (6.4)
Cardiovascular Disease, no (%)	10 (40)
Hypertension, no (%)	17 (68)
Diabetes mellitus type I or II, no (%)	9 (36)
Ongoing intravenous vasopressor treatment, no (%)	2 (8)
Ongoing intravenous vasodilator treatment, no (%)	4 (16)
Ongoing non-invasive continuous or bi-level positive airway pressure, no (%)	2 (8)

4. Discussion

Continuous and cuffless non-invasive BP monitoring may improve in-hospital patient monitoring by early detection of clinical deterioration and reduction of adverse outcomes (18). The present study investigated the accuracy of two different predictive BP models using sensor data from a prototype cuffless BP chest belt against intra-arterial measurements in a critically ill ICU cohort. Specifically, we compared a PAT-based model derived from a general population cohort to complex individualized models. The present study had two main findings. First, the generalized PAT-based model did not achieve high accuracy results, indicating that PAT-based BP monitoring in critically ill patients may not be possible, particularly when considering the inability to detect periods of hypotension and tachycardia. Second, the complex individualized models significantly improved accuracy of the cuffless BP device for SBP and MAP, but not DBP, and were able to better track BP changes during hypotension and tachycardia.

The significantly improved accuracy by the complex individualized models sheds light on important challenges regarding non-invasive cuffless BP devices. PAT is frequently cited as a potential non-invasive cuffless surrogate feature in recent years (5). Our results, however, suggests that PAT may not be adequate as cuffless surrogate measurement alone to achieve high accuracy non-invasive BP measurement in critically ill patients. An underlying assumption for general accuracy is stability of the relationship between changes in PAT and changes in BP across individuals, populations and across differing hemodynamic conditions. One or more of these factors likely affect generalizability of PAT as a cuffless surrogate measurement. Several studies have shown that varying between-individuals relationships between PAT and BP are a major limitation (9, 18, 19). The improved accuracy of the complex individualized models indicates that features extracted from ECG and PPG sensors can enable non-invasive cuffless BP monitoring, but these models are patient-specific (and potentially cannot be generalized for all subjects) and rely on machine learning without any *a priori* physiological knowledge. In addition to improved errors, an important finding was the ability of the complex individualized models to better track BP fluctuations, reflected by correlations corrected for repeated within subjects' measurements (0.23 for the generalized PAT-based model vs. 0.39 for the complex individualized models regarding SBP). It should

TABLE 2 Blood pressure distribution.

	Systolic blood pressure	Diastolic blood pressure	Mean arterial pressure
Mean (SD), mmHg	131.0 (25.7)	61.2 (14.6)	83.9 (18.1)
Range, min-max, mmHg	70.6–194.3	34–100.3	50.9–136.3
Within subject change, median (IQR), mmHg	29.3 (25.0–42.1)	13.4 (12.0–17.0)	18.6 (25.8–27.7)

TABLE 3 Performance of the generalized PAT-based model, the complex individualized models and comparison of the two.

	Generalized PAT-based model	Complex individualized models	*p* value for comparison
Systolic blood pressure			
Mean error, mmHg	−0.2	−1.4	
Mean absolute error (SD), mmHg	7.6 (5.3)	6.5 (4.8)	<0.001*
SD of errors, mmHg	7.2	6.7	<0.001**
Median of absolute errors (IQR), mmHg	5.3 (4.5–10.7)	5.8 (4.7–7.3)	
Repeated measures correlation coefficient	0.23	0.39	
Correlation coefficient, all subjects pooled	0.91	0.94	
Linear regression of aggregated data between model and reference***, R^2	0.91	0.96	
Akaike's information criterion***	173	154	
Bayesian information criterion***	175	156	
Diebold-Mariano comparison of predictive accuracy	Individualized model is significantly better		0.001
Diastolic blood pressure			
Mean error, mmHg	0.2	0.0	
Mean absolute error, mean (SD), mmHg	3.3 (3.3)	3.1 (2.2)	<0.001*
SD of errors, mmHg	−3.1	3.0	0.56**
Median of absolute errors (IQR), mmHg	2.7 (1.8–4.1)	2.2 (1.7–3.5)	
Repeated measures correlation coefficient	0.29	0.33	
Correlation coefficient, all subjects pooled.	0.94	0.94	
Linear regression of aggregated data between model and reference***, R^2	0.94	0.94	
Akaike's information criterion***	131	130	
Bayesian information criterion***	134	133	
Diebold-Mariano comparison of predictive accuracy	Individualized model is non-significantly better		0.14
Mean arterial pressure			
Mean error, mmHg	0.1	−0.1	
Mean absolute error, mean (SD), mmHg	4.6 (3.2)	4.0 (2.9)	<0.001*
SD of errors, mmHg	4.4	4.0	<0.001**
Median of absolute errors (IQR), mmHg	3.3 (2.4–6.4)	3.3 (2.5–4.5)	
Repeated measures correlation coefficient	0.25	0.37	
Correlation coefficient, all subjects pooled.	0.93	0.95	
Linear regression of aggregated data between model and reference***, R^2	0.93	0.95	
Akaike's information criterion***	146	138	
Bayesian information criterion***	149	140	
Diebold-Mariano comparison of predictive accuracy	Individualized model is significantly better		0.006

*Compared using non-parametric test of difference in means of all absolute errors between the two models. **Compared using variance comparison test of equality of standard deviations. ***Means of predicted BP from each model for each subject fitted in a linear regression model against reference BP.

be kept in mind that correlation across all the data is suppressed by the fact that there were stable periods where BP had low variation.

A concerning finding in our analyses was the inability of the generalized PAT-based model to predict BP changes during some periods of BP reductions coupled with elevation in HR (Figure 4). In our data, the complex individualized models estimated BP better in these situations. In the first scenario in Figure 4 (upper left panel) all models fail, whereas for the next three scenarios the complex individualized models predict the correct direction of BP change while the generalized PAT-based model and the PAT-only model predicts an increase in BP during reduction of reference BP and increases of HR. Our findings suggest that PAT is dependent on HR; an increase in HR causes PAT to decrease independently of the underlying change in BP (a decrease in PAT should always indicate an increase in BP according to the theory). Although conflicting results exists, HR has been shown to affect pulse wave propagation independently of BP similarly to our observations (20, 21). It is also possible that elevated HR is an indication of elevated sympathetic tone, which is shown to increase pulse wave propagation speed independently of central aortic BP (22). This can mask the true BP change in cases were HR and BP change in opposite directions. It should be noted that this was not a pre-specified analysis nor tested in

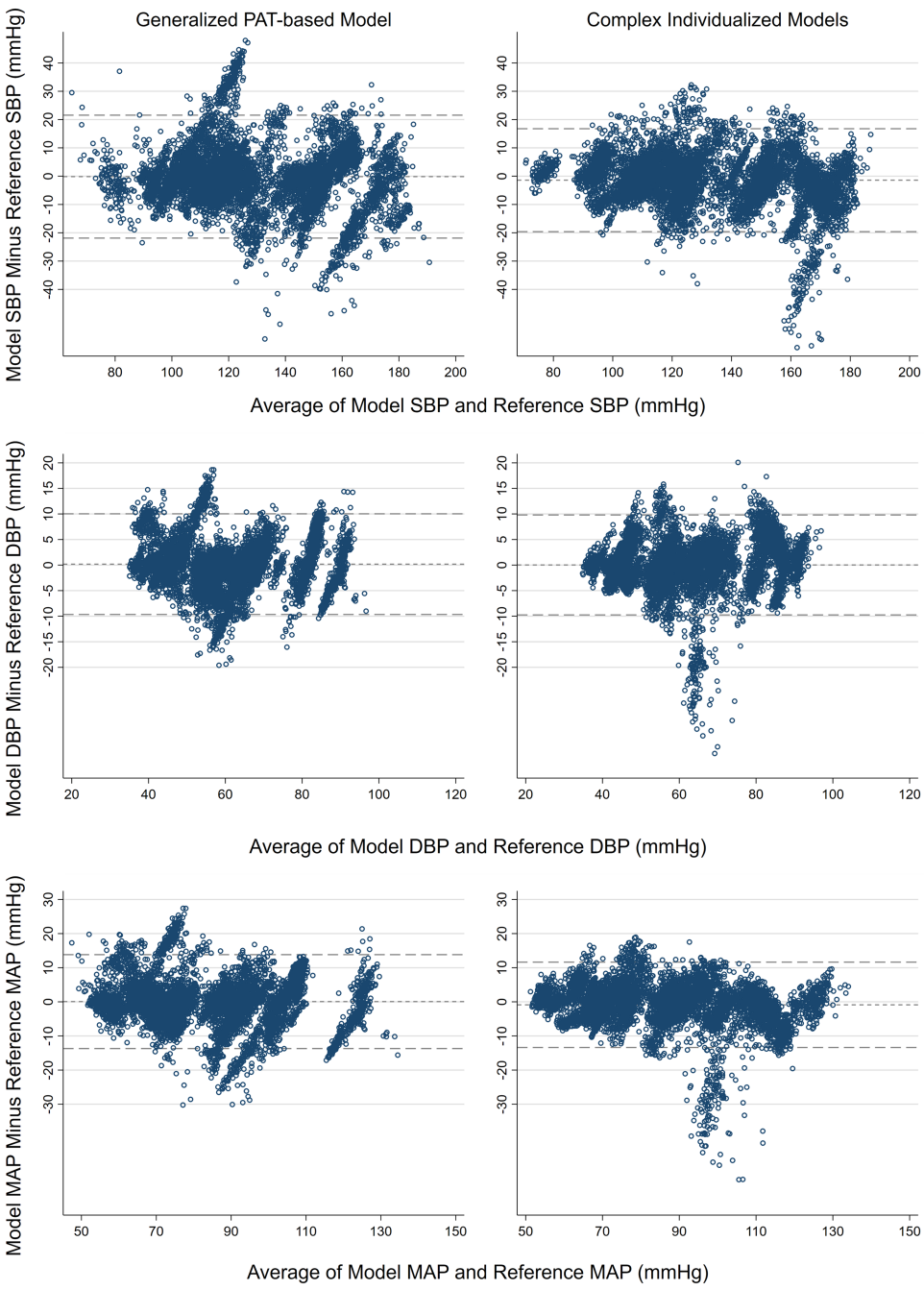

FIGURE 3
Bland–Altman plots. Mean of reference and model (*x*-axis) plotted against the difference between reference and model (*y*-axis). Horizontal lines indicate bias and upper and lower 95% limits of agreement. SBP, systolic blood pressure. DBP, diastolic blood pressure. MAP, mean arterial pressure.

any statistical model, merely, an indication of a potential serious limitation of cuff-based BP monitoring. We interpret this as a need for more data to develop robust models that can accurately estimate BP across differing hemodynamic conditions.

The generalized PAT-based model and complex individualized models achieved LOA of [−21.5, 21.1 mmHg] vs. [−19.2, 16.2 mmHg] regarding SBP and [−13.4, 13.5 mmHg] vs. [−13.9, 11.4 mmHg] regarding MAP. Corresponding results of MAE (SD of errors) were 7.6 (7.2) vs. 6.5 (6.7) and 4.6 (4.4) vs. 4.0 (4.0) regarding SBP and MAP, respectively. These results fall short of accuracy demands required in potentially unstable ICU patients. Particularly when considering the inability of the generalized PAT-based model to predict BP reductions coupled with elevated HR, which is critical in hospitalized patients as such circulatory changes may suggest onset of shock. On the other hand, considering more stable patients and that 78% (generalized PAT-based) and 84% (complex individualized models) of the absolute differences were below 10 mmHg regarding SBP, one may argue that our results are acceptable. It should also be kept in mind that the accuracy of the "gold standard" itself is dependent on appropriate damping as well as leveling and zeroing of the pressure transducer. In

TABLE 4 Percentage of absolute errors within 15, 10, and 5mmHg.

	Model	Systolic blood pressure	Diastolic blood pressure	Mean arterial pressure
≤5 mmHg	Generalized PAT-based model, %	53.1	78.9	69.2
	Complex individualized models, %	59.2	85.3	78.8
≤10 mmHg	Generalized PAT-based model, %	77.6	96.2	89.6
	Complex individualized models, %	83.8	97	94.2
≤15 mmHg	Generalized PAT-based model, %	87.9	99.7	95.9
	Complex individualized models, %	92.9	98.5	97.8

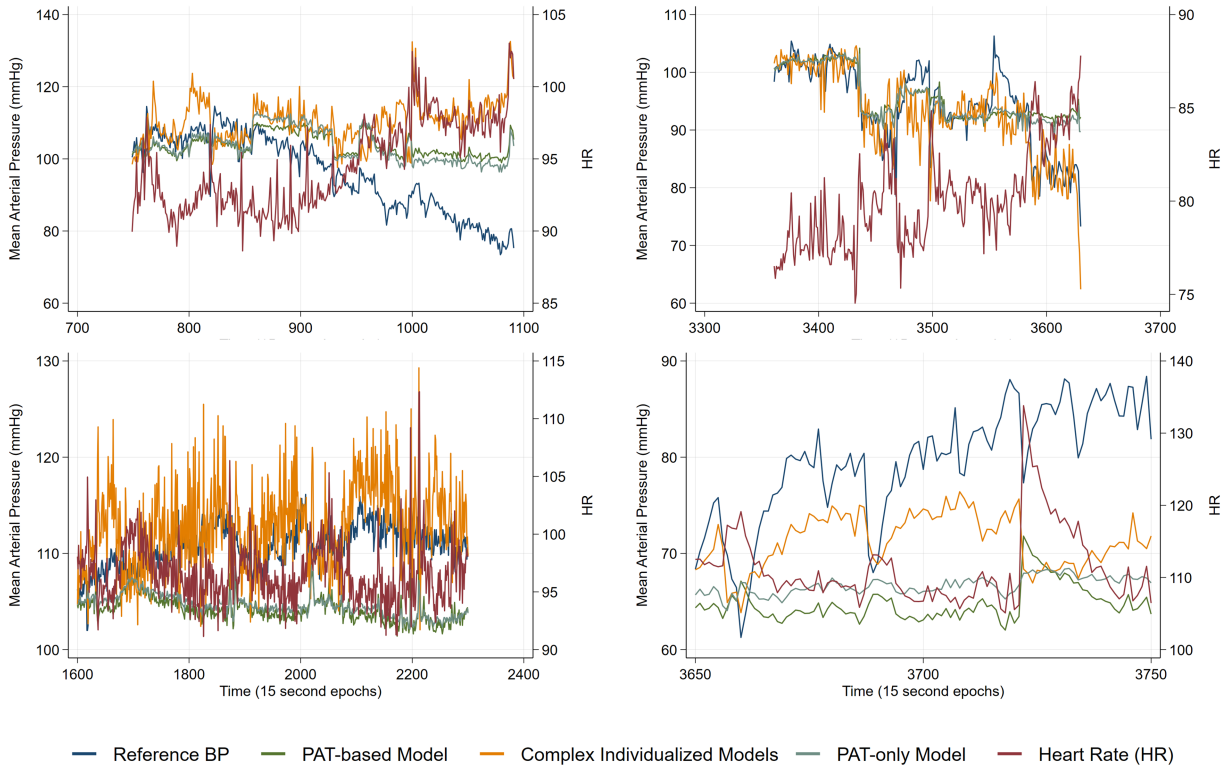

FIGURE 4
Time series plots from four different patients of reference mean arterial pressure (MAP), heart rate (HR) and predicted MAP from the two models in addition to predicted MAP from a PAT-only model.

everyday management of patients in the ICU, brachial oscillometric cuff BPs are taken regularly. Our LOA were considerably narrower compared to SBP LOA of [−30.2, 31.7 mmHg] revealed in a retrospective analysis comparing oscillometric cuff measurements to invasive measurements in 736 ICU patients (23).

We did not pre-specify any cut-off error statistic because we were evaluating a prototype of the cuffless BP device and the anticipated ISO 81060-3 validation standard applicable to cuffless BP devices was not completed at the time of study planning and data analysis. Acceptance criteria from validation standards aimed at cuff-based devices are not appropriate (24). As a consequence of lack of appropriate validation requirements regarding cuffless BP devices, many have compared against the Association for the Advancement of Medical Instrumentation/European Society of Hypertension/International Organization for Standardization (AAMI/ESH/ISO) criterion; mean error less than 5 mmHg and SD of errors less than 8 mmHg regarding SBP (12, 14, 15). Both our models satisfy this criterion as all mean errors were close to zero. This criterion is, however, intended for standardized cuff measurements seated at rest. Thus, it is difficult to specify clinically accepted accuracy in the study setting. Validation of novel cuffless BP devices dependent on calibration, of which all are at present, should be performed according to the new AAMI/ESH/ISO consensus validation protocol (24). Cuffless BP devices that pass the cuff-intended AAMI/ESH/ISO criterion may not be interpreted as accurate until also passing the new protocol intended to validate initial stability, accuracy during BP changes and reproducibility of stability within the time window of intended use.

Our device performances were comparable to the few similar studies that have investigated accuracy in a cuffless BP device, based on either ECG and PPG or PPG alone, against invasive measurements (12–15). Three of these devices are available on the market (12–14)

and one is a prototype (15). It is however difficult to compare results from those directly due to heterogenicity. Our results demonstrated the least narrow LOA compared to SBP LOA of [−10, 10 mmHg] in 10 post cardiac surgery patients (Biobeat wrist watch) (13), [−11.9, 12.2 mmHg] in 23 ICU patients (Aktiia wrist band, PPG) (12), [−11, 16 mmHg] during cardiac catheterization in 17 patients (Senbiosys prototype finger ring, PPG) (15) and [−7.4, 12.8 mmHg] in 20 cardiac ICU patients during controlled short-term supine and in bed measurements (Vitaliti continuous vital signs monitor, ECG and PPG) (14). However, while not achieving as narrow LOA, our study had the most subjects, 25 vs. 10 (Biobeat, ECG and PPG), 23 (Aktiia), 17 (Senbiosys) and 20 (Vitaliti) and by far the largest number of pairwise comparisons of 7,327 compared to 4,000 (Biobeat), 326 (Aktiia), 708 (Senbiosys) and 120 (Vitaliti). Sampling rate also varied between studies from 10 s epochs by Senbiosys to 1-min epochs by Biobeat. All studies excluded a large proportion of patients of which the majority were related to signal selection by algorithms or noise. A particularly important factor regarding cuffless BP devices is the degree of BP change within each patient during data collection. As all devices are dependent on initial calibration, a low change in BP within subjects may result in narrow LOA but the actual ability of these devices to track changes in BP remains unknown. Vitaliti reported measurements only from a stable period immediately following calibration, and Biobeat reported that their subjects were relatively stable as a limitation (within subject ranges not reported). Our subjects had reasonable within subject variations in BP with median SBP (IQR) of 29.3 (25.0–42.1) mmHg with a maximum of 63.2 mmHg. A related issue is reporting of Pearson's correlation coefficients which are pooled across all subjects, particularly when the devices are calibration dependent and there are repeated measurements within individuals. For comparative purposes we also computed Pearson's correlation coefficients from all measurements pooled and achieved 0.91 (generalized PAT-based model) and 0.94 (complex individualized models) for SBP compared to 0.94 (Biobeat), 0.87 (Aktiia) and 0.93 (Senbiosys). However, Pearson's correlation coefficients in this setting does not reflect device accuracy. In contrast, one study found a cuffless BP device using ECG and PPG inaccurate during coronary angiography with SBP LOA of [−2, 70 mmHg] (25). The study was, however, criticized by the manufacturer for incorrect calibration (26).

5. Strengths and limitations

A strength in our study is that neither model used any demographic information. The use of demographic information in cuff less research is criticized (27) because demographics itself are known to correlate with BP. Thus, when evaluating accuracy, it is not known how much is related merely to demographics as input in a model. We also provided, to the best of our knowledge, the most datapoints to date in a study evaluating accuracy of a cuffless BP device against invasive arterial measurements. Testing on critically ill patients admitted to an ICU enabled us to reveal the weaknesses of a PAT-based model and the strengths of complex individually fitted models.

We excluded many subjects (43%). However, the majority were related to criteria for developing the complex individualized models and we had comparable proportions and reasons for exclusion to similar studies. Algorithm selection imposes potential limitations on which patients may benefit from cuffless BP in the future. Re-calibration during the data collection in 14 patients may have introduced some overestimation of accuracy. If the device estimation of BP had drifted from reference BP, recalibration would artificially improve error estimates. However, as stated in the methods section, not recalibrating could introduce systemic errors and since the majority only had one recalibration it was decided to recalibrate if the transducer was relevelled. We did not formally test quality of the arterial line by for example the square wave test and calculation of damping coefficients. Since the transducer is levelled on a bracket next to the patient, arterial line BP accuracy is vulnerable to patient movement. We cannot exclude that some variations in reference BP were introduced in this manner. To reliably exclude all periods of which the pressure transducer was out of system, all data collection were observed by an investigator. The critically ill cohort is heterogenous. With a limited number of subjects, we cannot determine which, if any, clinical parameters affected accuracy. PAT can be measured at various places and we are limited to infer our findings to PAT measured at chest level.

6. Conclusion

Cuffless BP monitoring is promising, but challenges remain. In the present study, we demonstrated that a generalized PAT-based model measured on the chest did not achieve high accuracy results in critically ill ICU patients and failed to detect clinically important situations. We further demonstrated that more complex and individually fitted models, utilizing more information from the ECG and PPG signals, significantly outperformed the generalized PAT-based model. More data is needed to build robust general models based on machine learning to enable cuffless BP in hospitalized patients.

Data availability statement

The datasets presented in this article are not readily available because raw signals and data regarding model development may not be disclosed. BP predictions from both models together with reference measurements can be made available upon a formal request. Requests to access the datasets should be directed to sondhe@ous-hf.no.

Ethics statement

The studies involving human participants were reviewed and approved by REK sør-øst (REC south-east), Oslo, Norway. The patients/participants provided their written informed consent to participate in this study.

Author contributions

SH, TS, FF, and BW-G contributed to conception and design of the study. SH performed the data collection. KB-R, AS, ØH, and VG organized the database. SH, KB-R, AS, ØH, and VG performed the

data analysis and statistical analysis. SH wrote the first draft of the manuscript. All authors contributed to the manuscript revision, read, and approved the submitted version.

Funding

The research project (Hypersension) was funded by BIA program of the Norwegian research council (project number 332371).

Acknowledgments

The study appreciates patients for their willingness to participate and the intensive care unit at Oslo University Hospital, Ullevål for allowing us to conduct the study.

Conflict of interest

KB-R, AS, and TS were employed by company Aidee Health AS.

The remaining 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. Maheshwari K, Nathanson BH, Munson SH, Khangulov V, Stevens M, Badani H, et al. The relationship between ICU hypotension and in-hospital mortality and morbidity in septic patients. *Intensive Care Med*. (2018) 44:857–67. doi: 10.1007/s00134-018-5218-5

2. Turan A, Chang C, Cohen B, Saasouh W, Essber H, Yang D, et al. Incidence, severity, and detection of blood pressure perturbations after abdominal surgery: a prospective blinded observational study. *Anesthesiology*. (2019) 130:550–9. doi: 10.1097/ALN.0000000000002626

3. Khanna AK, Hoppe P, Saugel B. Automated continuous noninvasive ward monitoring: future directions and challenges. *Crit Care*. (2019) 23:194. doi: 10.1186/s13054-019-2485-7

4. Sessler DI, Saugel B. Beyond 'failure to rescue': the time has come for continuous ward monitoring. *Br J Anaesth*. (2019) 122:304–6. doi: 10.1016/j.bja.2018.12.003

5. Welykholowa K, Hosanee M, Chan G, Cooper R, Kyriacou PA, Zheng D, et al. Multimodal Photoplethysmography-based approaches for improved detection of hypertension. *J Clin Med*. (2020) 9:41203. doi: 10.3390/jcm9041203

6. Pilz N, Patzak A, Bothe TL. Continuous cuffless and non-invasive measurement of arterial blood pressure-concepts and future perspectives. *Blood Press*. (2022) 31:254–69. doi: 10.1080/08037051.2022.2128716

7. Seeberg TM, Orr JG, Opsahl H, Austad HO, Roed MH, Dalgard SH, et al. A novel method for continuous, noninvasive, cuff-less measurement of blood pressure: evaluation in patients with nonalcoholic fatty liver disease. *IEEE Trans Biomed Eng*. (2017) 64:1469–78. doi: 10.1109/TBME.2016.2606538

8. Heimark S, Eitzen I, Vianello I, Bøtker-Rasmussen KG, Mamen A, Hoel Rindal OM, et al. Blood pressure response and pulse arrival time during exercise testing in well-trained individuals. *Front Physiol*. (2022) 13:863855. doi: 10.3389/fphys.2022.863855

9. Heimark S, Rindal OMH, Seeberg TM, Stepanov A, Boysen ES, Bøtker-Rasmussen KG, et al. Blood pressure altering method affects correlation with pulse arrival time. *Blood Press Monit*. (2022) 27:139–46. doi: 10.1097/MBP.0000000000000577

10. Lee HC, Jung CW. Vital recorder-a free research tool for automatic recording of high-resolution time-synchronised physiological data from multiple anaesthesia devices. *Sci Rep*. (2018) 8:1527. doi: 10.1038/s41598-018-20062-4

11. Myles PS, Cui J. Using the bland-Altman method to measure agreement with repeated measures. *Br J Anaesth*. (2007) 99:309–11. doi: 10.1093/bja/aem214

12. Pellaton C, Vybornova A, Fallet S, Marques L, Grossenbacher O, de Marco B, et al. Accuracy testing of a new optical device for noninvasive estimation of systolic and diastolic blood pressure compared to intra-arterial measurements. *Blood Press Monit*. (2020) 25:105–9. doi: 10.1097/MBP.0000000000000421

13. Kachel E, Constantini K, Nachman D, Carasso S, Littman R, Eisenkraft A, et al. A pilot study of blood pressure monitoring after cardiac surgery using a wearable, non-invasive sensor. *Front Med*. (2021) 8:693926. doi: 10.3389/fmed.2021.693926

14. McGillion MH, Dvirnik N, Yang S, Belley-Côté E, Lamy A, Whitlock R, et al. Continuous noninvasive remote automated blood pressure monitoring with novel wearable technology: a preliminary validation study. *JMIR Mhealth Uhealth*. (2022) 10:e24916. doi: 10.2196/24916

15. Schukraft S, Haddad S, Faucherre Y, Arroyo D, Togni M, Barison A, et al. Remote blood pressure monitoring with a wearable photoplethysmographic device in patients undergoing coronary angiography: the senbiosys substudy. *Blood Press Monit*. (2022) 27:402–7. doi: 10.1097/MBP.0000000000000615

16. Linden A. RMLOA: Stata module to compute limits of agreement for data with repeated measures," Statistical Software Components S458980, Boston College Department of Economics.

17. Linden A. RMCORR: Stata module to compute a correlation for data with repeated measures," Statistical Software Components S458971, Boston College Department of Economics, revised 17 Nov 2021.

18. Mukkamala R, Stergiou GS, Avolio AP. Cuffless blood pressure measurement. *Annu Rev Biomed Eng*. (2022) 24:203–30. doi: 10.1146/annurev-bioeng-110220-014644

19. Finnegan E, Davidson S, Harford M, Jorge J, Watkinson P, Young D, et al. Pulse arrival time as a surrogate of blood pressure. *Sci Rep*. (2021) 11:22767. doi: 10.1038/s41598-021-01358-4

20. Haesler E, Lyon X, Pruvot E, Kappenberger L, Hayoz D. Confounding effects of heart rate on pulse wave velocity in paced patients with a low degree of atherosclerosis. *J Hypertens*. (2004) 22:1317–22. doi: 10.1097/01.hjh.0000125447.28861.18

21. Tan I, Butlin M, Spronck B, Xiao H, Avolio A. Effect of heart rate on arterial stiffness as assessed by pulse wave velocity. *Curr Hypertens Rev*. (2018) 14:107–22. doi: 10.2174/1573402113666170724100418

22. Cox J, Avolio AP, Louka K, Shirbani F, Tan I, Butlin M. Blood pressure-independent neurogenic effect on conductance and resistance vessels: a consideration for cuffless blood pressure measurement? *Ann Int Conf IEEE Eng Med Biol Soc*. (2021) 2021:7485–8. doi: 10.1109/EMBC46164.2021.9629936

23. Kaufmann T, Cox EGM, Wiersema R, Hiemstra B, Eck RJ, Koster G, et al. Non-invasive oscillometric versus invasive arterial blood pressure measurements in critically ill patients: a post hoc analysis of a prospective observational study. *J Crit Care*. (2020) 57:118–23. doi: 10.1016/j.jcrc.2020.02.013

24. Stergiou GS, Mukkamala R, Avolio A, Kyriakoulis KG, Mieke S, Murray A, et al. Cuffless blood pressure measuring devices: review and statement by the European Society of Hypertension Working Group on blood pressure monitoring and cardiovascular variability. *J Hypertens*. (2022) 40:1449–60. doi: 10.1097/HJH.0000000000003224

25. Moharram MA, Wilson LC, Williams MJ, Coffey S. Beat-to-beat blood pressure measurement using a cuffless device does not accurately reflect invasive blood pressure. *Int J Cardiol Hypertens*. (2020) 5:100030. doi: 10.1016/j.ijchy.2020.100030

26. Patzak A. Measuring blood pressure by a cuffless device using the pulse transit time. *Int J Cardiol Hypertens*. (2021) 8:100072. doi: 10.1016/j.ijchy.2020.100072

27. Mukkamala R, Yavarimanesh M, Natarajan K, Hahn JO, Kyriakoulis KG, Avolio AP, et al. Evaluation of the accuracy of Cuffless blood pressure measurement devices: challenges and proposals. *Hypertension*. (2021) 78:1161–7. doi: 10.1161/HYPERTENSIONAHA.121.17747

frontiers | Frontiers in Medicine

TYPE Review
PUBLISHED 20 April 2023
DOI 10.3389/fmed.2023.1176192

OPEN ACCESS

EDITED BY
Qinghe Meng,
Upstate Medical University, United States

REVIEWED BY
Yu-wen Chen,
Chinese Academy of Sciences (CAS), China
Bin Yi,
Army Medical University, China

*CORRESPONDENCE
Hassan Mumtaz
✉ hassanmumtaz.dr@gmail.com

RECEIVED 28 February 2023
ACCEPTED 04 April 2023
PUBLISHED 20 April 2023

CITATION
Saqib M, Iftikhar M, Neha F, Karishma F and Mumtaz H (2023) Artificial intelligence in critical illness and its impact on patient care: a comprehensive review.
Front. Med. 10:1176192.
doi: 10.3389/fmed.2023.1176192

Artificial intelligence in critical illness and its impact on patient care: a comprehensive review

Muhammad Saqib[1], Muhammad Iftikhar[1], Fnu Neha[2], Fnu Karishma[3] and Hassan Mumtaz[4]*

[1]Khyber Medical College, Peshawar, Khyber Pakhtunkhwa, Pakistan, [2]Ghulam Muhammad Mahar Medical College, Sukkur, Sindh, Pakistan, [3]Jinnah Sindh Medical University, Karachi, Sindh, Pakistan, [4]Health Services Academy, Islamabad, Pakistan

Artificial intelligence (AI) has great potential to improve the field of critical care and enhance patient outcomes. This paper provides an overview of current and future applications of AI in critical illness and its impact on patient care, including its use in perceiving disease, predicting changes in pathological processes, and assisting in clinical decision-making. To achieve this, it is important to ensure that the reasoning behind AI-generated recommendations is comprehensible and transparent and that AI systems are designed to be reliable and robust in the care of critically ill patients. These challenges must be addressed through research and the development of quality control measures to ensure that AI is used in a safe and effective manner. In conclusion, this paper highlights the numerous opportunities and potential applications of AI in critical care and provides guidance for future research and development in this field. By enabling the perception of disease, predicting changes in pathological processes, and assisting in the resolution of clinical decisions, AI has the potential to revolutionize patient care for critically ill patients and improve the efficiency of health systems.

KEYWORDS

artificial intelligence, intensive care units, critical illness, risk assessment, decision making

1. Introduction

The word Artificial Intelligence (AI) describes the methods by which a system may imitate human cognitive functions, such as reasoning capacity, decision-making, generalization, or learning from past experiences, to achieve goals without being expressly programmed for specific activities. AI is characterized as intelligent machines, as opposed to the intelligence of individuals or other living things (1). The areas of learning algorithms, processing natural languages, and robotics may thus fall under the umbrella of artificial intelligence (AI), which has the potential to advance biomedical research, primary care, and health systems. These fields can be adapted to almost any area of medicine.

One of the most hotly contested uses of artificial intelligence (AI) in the healthcare industry has been the development of technology. The use of software, algorithms for machine learning, or artificial intelligence (AI) to simulate mental abilities in the interpretation, evaluation, and comprehension of healthcare data is referred to as AI in healthcare. For instance, AI-based medical algorithms used in mammograms help radiologists by providing a second opinion while aiding in the diagnosis of breast cancer (1).

AI was used in the healthcare industry to produce well-performing medicine. For instance, Insilico Medical has created AI algorithms that can halt viral infection. By providing nutritional guidance to expectant mothers based on their health state and algorithm estimates, another proposal seeks to safeguard them. Epileptic seizure detection, another excellent use of AI, assisted in lessening the severity of epileptic convulsions. With AI and the creation of a cutting-edge movement-detecting device, early stroke might also be accurately predicted.

Although using AI in medical healthcare seems to have the potential to drastically increase the effectiveness of clinical diagnosis and biomedicine in general, it has also raised some ethical questions. One of the main obstacles for medical AI is safety. IBM Watson for oncology is a very good example. It uses AI algorithms to analyse data from patient records and assist physicians in exploring cancer options for treatment for their patient populations. However, it has since come under fire for allegedly making risky and unreliable cancer therapy recommendations.

The quality of medical treatment for critically ill patients has greatly improved due to advancements in care standards (1). Despite this progress, traditional critical care has limitations in fully understanding and addressing the complexities of patients' health, predicting deterioration, and providing timely treatment. The advent of advanced monitoring systems and non-invasive and invasive treatments has improved bedside care, but it is yet to be determined if these advancements represent the next step in critical care medicine. Artificial intelligence (AI) aims to help computers identify patterns in complex and diverse data, which was once only possible in limited fields like physics or astronomy due to limited computing resources. However, with the recent growth in computing power, AI can now be applied to other fields, including critical care medicine, where there is an abundance of complex data (2). According to a recent study (3), the number of articles about AI in the field of critical care medicine (CCM) has been increasing rapidly, particularly from 2018 to 2020. The majority of these articles are of high quality and come from top-ranked journals. Research into artificial intelligence (AI) has shown promise in terms of predicting disease outcomes and improving patient care (3).

While there are increasing numbers of studies using AI-powered models in the intensive care unit (ICU), our understanding of AI's potential in critical care is still limited. Additionally, there are challenges that AI must overcome before becoming a routine part of clinical practice. Using the most recent literature, this review aims to improve understanding of the applications of AI in critical illness and its impact on patient care, and it makes recommendations for the future.

2. Methods

A comprehensive search was carried out in PubMed, Google Scholar, PLOS One, and Scopus for all relevant literature using the following terms: "critical care," "intensive care medicine," "ICU medicine," "artificial intelligence," "AI," "machine learning," and "critical illness" from January 2018 through February 2023 in the English language. Similar articles were also reviewed using the suggested articles for each paper, and gray literature was also searched using relevant terms. All papers were imported into reference management software, and duplicates were removed. Older versions of the same papers were not included if newer versions were available. All relevant papers were read, and corresponding authors were contacted using email if the full text of a paper was not available. No unpublished papers were included in our review.

Abbreviations: AI, Artificial intelligence; ML, Machine learning; RL, Reinforcement learning; ICU, Intensive care unit; CCM, Critical care medicine; RF, Random forest; SHAP, Shapley additive explanations; HAPrI, Hospital-acquired pressure-related injury; COVID-19, Coronavirus disease – 2019; HDF5, Hierarchical data format-version 5; AUC, Area under the curve; REMAP-CAP, Randomized embedded multifactorial adaptive platform for community-acquired pneumonia.

3. Applications of AI in critical care patient management

AI has a multitude of diverse applications for the care of critically ill patients. The Figure 1 includes the recognition of disease, the prediction of disease progression, and the recognition of unique patterns in complex patient data. AI can also significantly aid caregivers in complex decision-making, as shown in Figure 1.

3.1. Recognition of disease

Diagnosing the source of a critically ill patient's clinical decline can be a complex task due to the subtle onset of the disease or the presence of other conditions that obscure the main issue. Properly understanding the underlying context can be a challenging feat. For example, the presence of pulmonary infiltrates does not always indicate an excessive accumulation of fluid in the air sacs; it could be a sign of cardiac-related pulmonary edema, fluid in the pleural cavity, inflammation- or infection-related fluid buildup, or blood collections from trauma. Without proper clinical context and additional testing, appropriate and prompt treatment may be hindered. Artificial

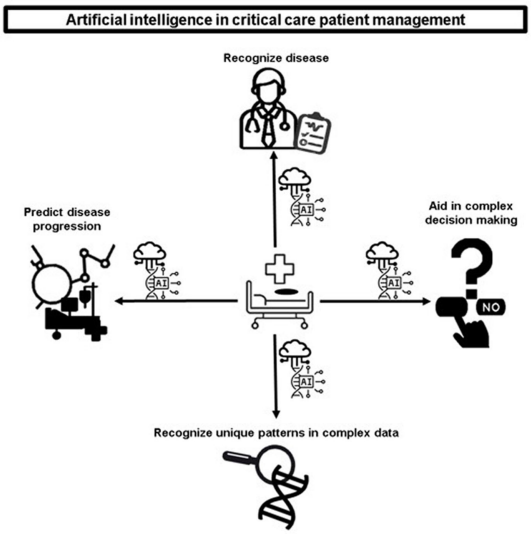

FIGURE 1
Artificial intelligence in critical care patient management.

intelligence (AI) can aid in the medical diagnosis of critically ill patients by utilizing its advanced text and image processing abilities (4). A machine learning model, for instance, can differentiate congestive heart failure from other lung diseases and quantify pulmonary edema using a technique that provides a probabilistic manner for describing an observation (5). Furthermore, recent advancements in image analysis using convolutional neural networks have enabled the evaluation of traumatic brain injury with more accuracy than manual methods when viewed on head computed tomography scans (6). In a retrospective analysis by Prasad et al. (7), a reinforcement learning (RL) approach was used to develop a treatment protocol for electrolyte replacements in an ICU setting. This system provides recommendations for patient care that can be continuously updated based on the patient's specific needs. The RL algorithm used available data from electronic health records, including vital signs, lab test results, and information about administered drugs and procedures, to estimate a patient-specific protocol for electrolyte repletion at six-hour intervals. The recommendations were presented by the AI algorithm in an interpretable and hierarchical manner, with the system first suggesting whether electrolyte replacement is needed and the best route for it, followed by the most appropriate dosage if the clinician chose to administer it. The RL system provided a more controlled and data-driven approach to electrolyte repletion as compared to traditional provider- or protocol-driven methods, which are often prone to error and deviation. This system also allows for greater flexibility and adaptability, considering patient context and clinical priorities. Optimal RL policy is reported to be able to recommend electrolyte replacements in a more targeted manner, potentially reducing the number of repletion events and the cost and time associated with unnecessary or repeat orders. Additionally, the system uses a reward and punishment system, reducing the costs and risks associated with intravenous delivery (7). This is not to underscore the value and significance of care-givers in the critical care setting; instead, it is a remarkable example of how new technologies such as AI can have a significant impact on the care of critically ill patients.

3.2. Prediction of disease progression using random forest models

Predicting disease progression is crucial for critically ill patients, as a delay in detecting clinical instability can result in harm or death (4). A dynamic random forest model is a type of machine learning algorithm that can be used to predict outcomes in the critical care setting. It works by using an ensemble of decision trees that can adapt and update in real-time as new data becomes available. A study by Yoon et al. (2) found that a dynamic model using random forest classification could predict cardiorespiratory instability, defined as a combination of hypotension, tachycardia, respiratory distress, or decreased oxygen saturation 90 min before it occurred in reality (2, 4). The use of AI and machine learning has expanded across various fields such as public health, disease prediction, and drug development, including the ability to predict viral mutations before they arise (4). The power of AI approaches continues to be utilized in a wide range of disease prediction and drug development applications (8). In a study by Davoudi et al. (9), tachycardia, which frequently precedes shock, was predicted 75 min before its onset using a random forest model (9). Although not in the critical care setting, hypotension was also predicted prior to its occurrence in the operating room and confirmed by a randomized controlled trial, reducing the rate of intraoperative hypotension to 1.2% (10, 11). In the critical care space, the prediction of hypotension events in the ICU has already been achieved using a random forest model that analyzed electronic health records and vital signs data, with 92.7% sensitivity, 15 min before the event even occurred (12). Another area where machine learning, a subset technology of artificial intelligence is in the assessment of pain in critically ill patients. In a study by Kobayashi et al. (13) which focused on using machine learning to assess the pain experienced by ICU patients, reported that vital signs, which are measured continuously in the ICU, can be used to predict pain with an accuracy upwards of 85% using a random forest (RF) model. This shows that machine learning can be used to continuously evaluate pain, which is important for pain management and the use of pain medication in ICUs. Their study also suggests that the use of an automated and continuous pain assessment algorithm may help relieve pain in patients who cannot communicate which could improve their life expectancy (13). All these examples show how the utilization of such models can prove significantly useful for management of critically ill patients.

3.3. Recognition of unique patterns in complex data

Critical illness is a complex condition that presents itself in various and unpredictable ways, leading to organ dysfunction and complicating the disease and recovery processes. To effectively manage these critical states, a careful consideration of underlying etiologies and clinical conditions is necessary. AI can help by recognizing unique patterns within complex data and identifying specific phenotypes or endotypes that reflect the individual's critical state, leading to more personalized treatment plans (14). This relies heavily on access to large amounts of training data and phenotypic information. The complexity of medical care is highlighted by the fact that the same symptoms can be caused by different underlying conditions, making it difficult to provide personalized treatment. Diseases such as brain disorders, cardiovascular issues, and digestive problems are examples of this complexity. Innovative techniques and tools have been used to achieve personalized phenotyping in patients, combining practical experiences and scientific knowledge to realize the potential for using AI in a systems medicine approach to personalize medical care (15). The advancement of AI techniques has enabled researchers to uncover the underlying causes of various phenotypes, including genetic variations and cancer diseases, and by utilizing these tools and combining them with other methods, the biomedical field will be able to advance their knowledge and understanding of the relationship between genomics and expressions in diseases, promising faster and more accurate discoveries (16). This exemplifies how AI can serve as an aid to personalized patient care for critically ill patients.

3.4. Aid to complex decision-making in critical care

AI has the potential to assist doctors in the complex process of assessing patient risk levels for treatments, determining those who are most likely to experience a sudden deterioration, and analyzing

multiple small outcomes to enhance overall patient outcomes. However, the complexity of AI techniques can affect physician comprehension and interpretation of results (17). To overcome this challenge, it is important for medical education to involve physicians in model creation and educate them in this field. AI platforms have the potential to be more efficient in some aspects compared to caregivers. For example, when compared to senior consultants, an AI platform such as Childhood Cataract Cruiser has proven to be more efficient and time-efficient for diagnoses, with high patient satisfaction rates (18). Such platforms can also be tested in the critical care setting. If they prove successful, it could significantly increase the efficiency of care delivery in the ICU. One-size-fits-all solutions are not effective in dealing with complex problems, as evidenced by the lack of improvement in septic shock outcomes in recent years despite various treatment guidelines (19, 20). Utilizing the concept of reinforcement learning has the potential to offer individualized solutions to the diverse nature of septic shock and varying host responses. A study by Komorowski et al. (21) used reinforcement learning on time series data with 44 features collected from mechanically ventilated patients, which resulted in improved outcomes compared to standard clinical care, reducing 90 day and ICU mortality rates (21). AI can also perform real-time electrocardiogram analysis to detect myocardial infarctions. A study by Chen et al. (22) reported using AI-assisted real-time analysis of electrocardiograms in the prehospital setting and found that it was feasible and had the potential to reduce delays in treatment times for patients requiring percutaneous coronary interventions (22). These examples demonstrate the use of AI for therapeutic guidance in medical decision-making for critically ill patients with good efficacy.

3.5. Intelligent decision making intervention in critical illness

By assisting in decision-making and enabling healthcare professionals to concentrate their efforts on investing more time with patients, artificial intelligence can help to promote shared decision-making (SDM) (22). AI technologies offer a wide range of information and have the capacity to evaluate enormous amounts of data and find correlations that scholars and healthcare professionals would have overlooked (23). The bioethics of employing AI for health decision-making, the challenges involved, patients' and healthcare practitioners' perspectives of AI-based decision aids, and how it should be included to provide patient-centered healthcare are all topics of developing study. Nevertheless, little is known about the actual application of AI in SDM or how it may help with the decision-making phase of SDM.

4. Challenges and obstacles to AI in critical care patient management

Despite the potential benefits of AI in healthcare, particularly in the critical care setting, it is important to be aware of the potential challenges and obstacles that may arise when implementing AI models for critically ill patients. These roadblocks should not be ignored or overlooked, as they can have significant consequences for patient care and outcomes. The Figure 2 includes interpretability, data privacy and sharing, decreased clinical readiness and sub-optimal adherence to standard. A figure depicting the challenges and ethical concerns of AI in critical care patient management is shown in Figure 2.

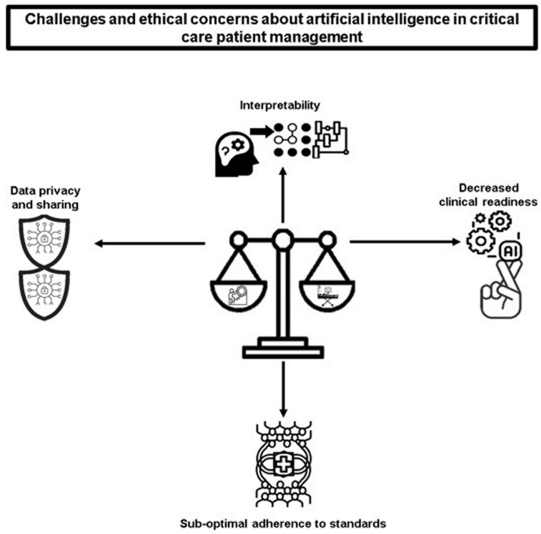

FIGURE 2
Challenges and ethical concerns of AI in critical care patient management.

4.1. Interpretability of AI in the intensive care unit

The deployment of AI in a healthcare setting, specifically at the bedside, requires careful planning and consideration of key factors such as usability and trustworthiness. The involvement of all relevant stakeholders, including patients, clinicians, researchers, and hospital administrators, is crucial for the success of the deployment. To ensure that the AI systems are effective and well-received, the implementation strategy should focus on creating models with a manageable amount of information that is presented in an understandable and visually appealing manner. This can be achieved through the use of interpretable logic and a user-friendly graphic interface. One of the key challenges in deploying AI systems at the bedside is ensuring that the AI-generated alerts are accurate and not overwhelming, so as to prevent alarm fatigue (23). In recent research on predicting hypotension in the ICU, the use of a stacked random forest model was found to reduce the number of alerts tenfold while still maintaining accuracy (12). To build trust and acceptance of AI systems among end users, it is important to understand the AI-generated predictions and recommendations. Despite the complex nature of many AI models, researchers are working to enhance their interpretability. The creation of a graphic user interface is essential for the effective deployment of AI systems at the bedside, as it helps to improve hospital workflow and reduce the burden on healthcare workers. Additionally, the use of deep learning in the analysis of patient behavior and environmental stimuli can provide useful information for detecting delirium in ICU patients (9). Care-takers are keen to understand how machines arrive at predictions that involve patient care. There are different software technologies that can help caregivers to understand how machines

arrive at these predictions. One such example is the Shapley additive explanation (SHAP), a method that explains how machines arrive at individual predictions. In a study by Alderden et al. (24), the risk of developing hospital-acquired pressure-related injury (HAPrI) was analyzed in COVID-19 patients who were hospitalized in the ICU. The study aimed to utilize machine learning algorithms to create a predictive model for HAPrI risk and ensure that the model was transparent and understandable for medical professionals. The best-performing model was an ensemble SuperLearner, which showed good discrimination in HAPrI risk assessment. The use of explainable AI methods such as SHAP plots was a novel approach in this study and provided a way to visualize the relationships between the patient's characteristics and the predictions made by the model. This study found that COVID-19 positive critical care patients have a higher risk of HAPrI compared to non-COVID patients. The use of machine learning algorithms to evaluate HAPrI risk in COVID-19 patients in the ICU is reported to be a feasible approach, and explainable AI methods such as SHAP plots provide a means of ensuring that the model is understandable and trustworthy for medical professionals. Medical professionals need to understand how the model reached its decisions for each individual patient to decide whether the model is trustworthy for that patient (24). Care-takers generally have a positive attitude towards the adoption of AI. Mlodzinski et al. (25) set out to examine the perspectives of both healthcare providers and non-providers regarding the use of machine learning (ML) in critical care. The study found that both groups generally have positive attitudes towards the use of ML in healthcare; however, non-providers with more knowledge about ML and AI are more likely to feel favorable towards its use. The study also found that there were no major differences in the level of comfort or knowledge among providers, regardless of their level of experience. Furthermore, the study identified common concerns such as systemic bias in data, patient safety, negative effects on the doctor-patient relationship, and data privacy an security. Among providers, workflow interruptions were also identified as a major concern, while limited knowledge of ML and AI was a concern among non-providers. It provided important insights into provider and non-provider perspectives on ML-based tools and will play a crucial role in optimizing their clinical utility (25). In the future, it will be important to design ICU systems that embrace the capabilities of AI and address caregiver concerns in order to enable early detection of patient deterioration and improve the accuracy and trustworthiness of AI-generated predictions. The complex nature of many AI models often makes it difficult to understand the rationale behind the computation and output, leading to resistance among healthcare professionals to adopting these models in daily practice. The fear of performing unnecessary interventions or changing treatment strategies without scientific evidence can have serious consequences, especially in critical care where patient outcomes are directly linked to such decisions (26, 27). However, there are efforts underway to address the issue of complex AI models. ML techniques are being used to determine what kinds of strategies caregivers use to make their decisions. For example, using game theory to measure the importance of features in predicting near-term hypoxic events during surgery has helped explain the contribution of various features to the AI model's output. This approach has been shown to provide consistent results with prior knowledge and literature, leading to improved clinical decision-making and preventing hypoxia during surgery (28). This can also be extrapolated to the critical care setting to explain the contribution of different features in the output of AI models. Additionally, providing detailed methodologies for model validation, robustness of analysis, and expert knowledge can help alleviate concerns and increase the reliability and trust in AI models (4).

In this study, in contrast to SHAP, we will concentrate on two more example post-hoc model accuracy techniques that have gained minimal attention in the physical scientific world, namely breakDown (BD) research and Ceteris-Paribus (CP) analyses. The BD technique, like the SHAP method, is founded on the variety attribution principle, which divides each observation's estimate into its individual variable components (29, 30). The BD values offer action descriptions of the impacts of variables in a clever way, in contrast to the SHAP values. The independence and non-interaction of the input characteristics (factors or descriptors) constitutes a component of the BD method's presumptions (31). For BD evaluation, there are two algorithms: step-up and step-down. The step-down approach begins with a complete collection of input characteristics.

Finally, in order to minimize the proximity to the prediction models, each selected feature contribution is determined by successively eliminating one characteristic from a set accompanied by variable relaxation. In contrast to the step-down approach, the step-up method begins with a null set and proceeds in the other manner. In feature contributions, both techniques have been proved to deliver consistent results.

On the contrary hand, the CP profiles, also known as individual conditional expectancies (ICE) plots, assess the impact of a variable from a learned ML model while assuming that the levels of all other variables remain constant (akin to what-if analysis).

Using CP profiles, one can quickly see how the source and responses are connected and how the projected response depends on a characteristic (e.g., in a non-linear, linear or complex). In this approach, the CP analysis aids in quantifying the influence of a particular variable on the conclusions drawn from a black version and offers a brief, visual description of the functional form linking an input with an output. From either the SHAP or BD analysis, it is difficult to draw conclusions regarding this type of functional reliance. Hence, adding CP profile plots to SHAP and BD studies is of great utility.

4.2. Reproducibility issues of AI systems during application

Frequently, determining the causative factors of deteriorating patients from the complete list of differential diagnoses is tough, because of the subtle feature of early illness progression or the existence of co-existing disorders disguising the underlying problem. Above all, it is important to accurately interpret the underlying context, which is sometimes difficult to do. For instance, it is not enough to infer that pulmonary infiltrates are caused by an excessive amount of alveolar fluid. These may signify pleural effusion, pulmonary embolism fluid from an infection or inflammation, pulmonary edema with a cardiac origin, collections of blood due to trauma, or any of these conditions. Lacking clinical context and additional testing, proper and prompt care might be delayed. AI might aid in such circumstances by obtaining a more exact diagnosis, given enhanced text and picture processing power. Using a machine learning algorithm, congestive heart failure (CHF) could

be distinguished from other cause of lung illness (32), and the quantity of pulmonary edema brought on by the CHF could be measured using semi-supervised machine learning and a finite difference autoencoder. An AI model was used to evaluate imaging data from hospitalized patients in during acute pulmonary syndrome coronavirus 2 (SARS-CoV-2) epidemic in order to identify coronavirus disease 2019 (COVID-19) (33).

The application of AI in the clinical setting is hindered by a lack of sufficient clinical trials and experiments, leading to a low rate of reproducibility and future analysis. A review of 172 AI solutions created from chart data revealed that the clinical readiness level of AI was low, with 93% of the analyzed solutions falling below stage 4 for real-world application and only 2% undergoing prospective validation (29). The reproducibility of AI solutions is uncertain due to limitations in data openness and algorithmic complexity, and there are no clear protocols in place to examine this thoroughly. A study showed that attempting to reproduce mortality prediction projects resulted in large sample size differences in half of the experiments, highlighting the importance of accurate labeling, clinical context, and precise reporting methods (30). Adherence to reporting standards and the risk of bias are also sub-optimal, as a study of 81 non-randomized and 10 randomized trials using deep learning showed that only 6 of the 81 non-randomized studies had been tested in a real-world clinical setting, and 72% of the studies had high risks of bias (31). Even more complex AI models, such as reinforcement learning, face challenges as they require significant computational resources and are difficult to test on patients in a clinical environment. However, new approaches such as inverse reinforcement learning may offer a solution by inferring information about rewards, potentially making decision-assisting engines more robust and reliable with varying input data, which is crucial in critical care data science where data is vast and extremely diverse (34).

5. Ethical concerns

The use of AI in critical care is a new and developing field, and the ethical issues that may arise from its use are not fully understood. However, there are a few aspects that can be discussed to anticipate potential ethical dilemmas. One issue is data privacy and sharing. The process of collecting and manipulating data to find patterns could lead to the leakage of confidential information, particularly during the pre-processing stage and external validation. De-identification and novel models such as federated learning might help to minimize data leakage and increase the speed of the validation process (4). Another ethical concern is the safety of AI models in patient care. The maturity metric used for self-driving cars has been used to describe the safety of AI models, with 6 levels ranging from no automation to full automation (35). Based on this scale, most AI-driven solutions would currently fall into the categories of partial or no automation, meaning that human oversight and decision-making are still required. This also raises questions about patient autonomy and informed consent, as AI recommendations may not always align with a patient's preferences. In order to address these ethical issues and overcome the limitations of AI, researchers and clinicians need to be aware of the potential problems and develop solutions to mitigate them. This also includes understanding patient perspectives and incorporating them into the development of practical and ethical AI solutions (4).

6. Guidance for future

The Figure 3 includes efficient data transfer, data de-identification, rapid processing, quality control, and decentralized federated learning. The field of AI has the potential to greatly impact critical care, but there are several steps that must be taken in order to make this happen, as highlighted in Figure 3. Many of the recommendations have either already been implemented or are in the process of being implemented.

One of the most important is ensuring that the data used for training AI models is properly de-identified and standardized. This is important for both privacy and data quality, as data from different hospitals may be structured differently and contain different amounts of personal information. The Society of Critical Care Medicine and the European Society of Intensive Care Medicine have developed a process for de-identifying data, that involves separating personal data from anonymous data, conducting a risk-based process to de-identify the personal data, and conducting an external review to ensure that all privacy and legal considerations are met (36). Another important step is standardizing the data in order to facilitate efficient exchange between different hospitals. This requires developing a standard format for storing and exchanging clinical and physiological data. One such format, the Hierarchical Data Format, Version 5 (HDF5), allows for the storage, compression, and real-time streaming of multiparameter data. This would allow for the integration of other types of large-scale datasets, such as those in imaging or genomics (37). Another solution is the use of federated learning, where models can be trained locally at different hospitals rather than having the data sent to a central location for training. This helps to preserve privacy and can be particularly useful when the data distribution is imbalanced or skewed. A successful example of this approach was seen during the COVID-19 pandemic, where 20 academic centers collaborated to predict clinical outcomes

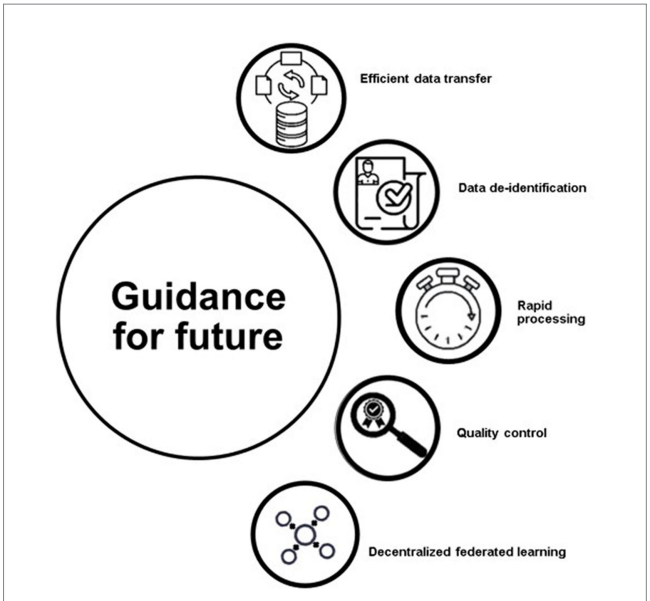

FIGURE 3
Guidance for future.

from COVID-19 using a federated learning approach. The AI model was trained on chest X-ray data, and achieved an average area under the curve (AUC), of 0.92 for predicting 24–72 h outcomes (4). The task of labeling events for AI models can be labor-intensive and resource-intensive, but novel AI models, such as weakly supervised learning, are being developed to make the process more efficient. This type of learning can build desired labels with only partial participation of domain experts, which preserves resources. Additionally, clinical trials can also be designed with AI models to maximize benefits and minimize risks to participants, as well as to make the best use of limited resources. One example of this is the Randomized Embedded Multifactorial Adaptive Platform for Community-Acquired Pneumonia (REMAP-CAP) trial, which uses a Bayesian inference model, to identify the optimal treatment for community-acquired pneumonia and has contributed to improved survival among critically ill COVID-19 patients (38). The labeling process for AI models can be a difficult and resource-intensive task. To make this process more efficient, new AI models such as weakly supervised learning have been developed. This method of learning allows for the partial involvement of domain experts and can reduce resource usage. For example, in the case of COVID-19 patients visiting the emergency department, weakly supervised learning was used in conjunction with medical ontologies and expert-driven rules to classify patients with related symptoms. This combination of weakly supervised learning and pretrained language models improved performance compared to a majority vote classifier, reducing the cost of creating classifiers in a short period of time, especially during a pandemic when experts may not be available for labeling. Innovative trial designs can also be developed with AI models to make the best use of resources and minimize risks to participants (39). This platform, initially developed for community-acquired pneumonia, has continued to enroll patients during the COVID-19 pandemic and has contributed to improved survival among critically ill patients (4, 40–42). For an AI model to be useful in real-life settings, it needs to provide important information in a timely manner, especially for critically ill patients who require quick feedback. The AI model should have a fast data pre-processing platform, parsimoniously feature input data, and deliver output rapidly. To date, no such model has been developed that can successfully do the above-mentioned tasks in such a quick manner. Although true real-time prediction is a challenging task, the application of a real-time AI model in the critical care environment could offer significant benefits without delay. Once the AI model is deemed useful in a clinical setting, quality assessment efforts should follow to ensure its maturity and integration with healthcare. The National Academy of Medicine of the United States has published a white paper on AI use in healthcare, emphasizing the development of guidelines and legal terms for safer, more effective, and personalized medicine (43).

7. Conclusion

The utilization of artificial intelligence (AI) in critical care presents numerous opportunities for enhancing outcomes in critically ill patients by enabling the perception of disease, predicting changes in pathological processes, recognizing unique patterns in disease presentations, and assisting in the process of clinical decision-making in a symbiotic fashion with care-givers. Moreover, AI can facilitate the understanding of medical processes by presenting recommendations for patient care in an interpretable and hierarchical manner through techniques such as reinforcement learning. The technology has the potential to improve understanding of the diverse clinical needs of critically ill patients, risk assessment for treatments, and the analysis of patient outcomes.

Author contributions

MS and MI: data curation, formal analysis, investigation, conceptualization, supervision, visualization, writing – original draft, and writing – review and editing. HM: data curation, formal analysis, investigation, methodology, writing – original draft, writing – review and editing, and supervision. All authors reviewed the final version of the paper and approved it for submission and publication.

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. Zimmerman JE, Kramer AA, Knaus WA. Changes in hospital mortality for United States intensive care unit admissions from 1988 to 2012. *Crit Care*. (2013) 17:R81. doi: 10.1186/cc12695

2. Yoon JH, Pinsky MR. Predicting adverse hemodynamic events in critically ill patients. *Curr Opin Crit Care*. (2018) 24:196–203. doi: 10.1097/MCC.0000000000000496

3. Cui X, Chang Y, Yang C, Cong Z, Wang B, Leng Y. Development and trends in artificial intelligence in critical care medicine: a bibliometric analysis of related research over the period of 2010–2021. *J Pers Med*. (2023) 13:50. doi: 10.3390/jpm13010050

4. Yoon JH, Pinsky MR, Clermont G. Artificial intelligence in critical care medicine. *Crit Care*. (2022) 26:75. doi: 10.1186/s13054-022-03915-3

5. Seah JCY, Tang JSN, Kitchen A, Gaillard F, Dixon AF. Chest radiographs in congestive heart failure: visualizing neural network learning. *Radiology*. (2019) 290:514–22. doi: 10.1148/radiol.2018180887

6. Monteiro M, Newcombe VFJ, Mathieu F, Adatia K, Kamnitsas K, Ferrante E, et al. Multiclass semantic segmentation and quantification of traumatic brain injury lesions on head CT using deep learning: an algorithm development and multicentre validation study. *Lancet Digital Health*. (2020) 2:e314–22. doi: 10.1016/S2589-7500(20)30085-6

7. Prasad N, Mandyam A, Chivers C, Draugelis M, Hanson CW 3rd, Engelhardt BE, et al. Guiding efficient, effective, and patient-oriented electrolyte replacement in critical care: an artificial intelligence reinforcement learning approach. *J Personalized Med*. (2022) 12:661. doi: 10.3390/jpm12050661

8. Zhang Q, Xie Q, GJCToIT W. A survey on rough set theory and its applications. *CAAI Trans Intel Technol*. (2016) 1:323–33. doi: 10.1016/j.trit.2016.11.001

9. 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

10. Wijnberge M, Geerts BF, Hol L, Lemmers N, Mulder MP, Berge P, et al. Effect of a machine learning-derived early warning system for intraoperative hypotension vs standard care on depth and duration of intraoperative hypotension during elective noncardiac surgery: the HYPE randomized clinical trial. *JAMA*. (2020) 323:1052–60. doi: 10.1001/jama.2020.0592

11. Joosten A, Rinehart J, van der Linden P, Alexander B, Penna C, de Montblanc J, et al. Computer-assisted individualized hemodynamic management reduces intraoperative hypotension in intermediate- and high-risk surgery: a randomized controlled trial. *Anesthesiology*. (2021) 135:258–72. doi: 10.1097/ALN.0000000000003807

12. Yoon JH, Jeanselme V, Dubrawski A, Hravnak M, Pinsky MR, Clermont G. Prediction of hypotension events with physiologic vital sign signatures in the intensive care unit. *Crit Care*. (2020) 24:661. doi: 10.1186/s13054-020-03379-3

13. Kobayashi N, Shiga T, Ikumi S, Watanabe K, Murakami H, Yamauchi M. Semi-automated tracking of pain in critical care patients using artificial intelligence: a retrospective observational study. *Sci Rep*. (2021) 11:5229. doi: 10.1038/s41598-021-84714-8

14. Yoon JH, Pinsky MR, Clermont G. Artificial Intelligence in critical care medicine. *Crit Care*. (2022) 26:75. doi: 10.1186/s13054-022-03915-3

15. Cesario A, D'Oria M, Bove F, Privitera G, Boškoski I, Pedicino D, et al. Personalized clinical phenotyping through systems medicine and artificial intelligence. *J Pers Med*. (2021) 11:265. doi: 10.3390/jpm11040265

16. Frey LJ. Artificial intelligence and integrated genotype–phenotype identification. *Genes*. (2019) 10:18. doi: 10.3390/genes10010018

17. Azar AT, El-Metwally SM. Decision tree classifiers for automated medical diagnosis. *Neural Comput Appl*. (2013) 23:2387–403. doi: 10.1007/s00521-012-1196-7

18. Lin H, Li R, Liu Z, Chen J, Yang Y, Chen H, et al. Diagnostic efficacy and therapeutic decision-making capacity of an artificial intelligence platform for childhood cataracts in eye clinics: a multicentre randomized controlled trial. *eClinicalMedicine*. (2019) 9:52–9. doi: 10.1016/j.eclinm.2019.03.001

19. Rivers E, Nguyen B, Havstad S, Ressler J, Muzzin A, Knoblich B, et al. Early goal-directed therapy in the treatment of severe sepsis and septic shock. *New England J Med*. (2001) 345:1368–77. doi: 10.1056/NEJMoa010307

20. PRISM InvestigatorsRowan KM, Angus DC, Bailey M, Barnato AE, Bellomo R, et al. Early, goal-directed therapy for septic shock—a patient-level meta-analysis. *New England J Med*. (2017) 376:2223–34. doi: 10.1056/NEJMoa1701380

21. Komorowski M, Celi LA, Badawi O, Gordon AC, Faisal AA. The artificial intelligence clinician learns optimal treatment strategies for sepsis in intensive care. *Nat Med*. (2018) 24:1716–20. doi: 10.1038/s41591-018-0213-5

22. Chen K-W, Wang Y-C, Liu M-H, Tsai B-Y, Wu M-Y, Hsieh P-H, et al. Artificial intelligence-assisted remote detection of ST-elevation myocardial infarction using a mini-12-lead electrocardiogram device in prehospital ambulance care. *Front Cardiovasc Med*. (2022) 9:1001982. doi: 10.3389/fcvm.2022.1001982

23. Hravnak M, Pellathy T, Chen L, Dubrawski A, Wertz A, Clermont G, et al. A call to alarms: current state and future directions in the battle against alarm fatigue. *J Electrocardiol*. (2018) 51:S44–8. doi: 10.1016/j.jelectrocard.2018.07.024

24. Alderden J, Kennerly SM, Wilson A, Dimas J, McFarland C, Yap DY, et al. Explainable artificial intelligence for predicting hospital-acquired pressure injuries in COVID-19-positive critical care patients. *Comput Inf Nurs*. (2022) 40:659–65. doi: 10.1097/CIN.0000000000000943

25. Mlodzinski E, Wardi G, Viglione C, Nemati S, Crotty Alexander L, Malhotra A. Assessing barriers to implementation of machine learning and artificial intelligence-based tools in critical care: web-based survey study. *JMIR Perioperative Med*. (2023) 6:e41056. doi: 10.2196/41056

26. 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

27. Durán JM, Jongsma KR. Who is afraid of black box algorithms? On the epistemological and ethical basis of trust in medical AI. *J Med Ethics*. (2021):medethics-2020-106820. doi: 10.1136/medethics-2020-106820

28. Lundberg SM, Nair B, Vavilala MS, Horibe M, Eisses MJ, Adams T, et al. Explainable machine-learning predictions for the prevention of hypoxaemia during surgery. *Nat Biomed Eng*. (2018) 2:749–60. doi: 10.1038/s41551-018-0304-0

29. Fleuren LM, Thoral P, Shillan D, Ercole A, Elbers PWG. Machine learning in intensive care medicine: ready for take-off? *Intensive Care Med*. (2020) 46:1486–8. doi: 10.1007/s00134-020-06045-y

30. Johnson AE, Pollard TJ, Mark RG, Reproducibility in critical care: a mortality prediction case study. Machine learning for healthcare conference PMLR; (2017).

31. Nagendran M, Chen Y, Lovejoy CA, Gordon AC, Komorowski M, Harvey H, et al. Artificial intelligence versus clinicians: systematic review of design, reporting standards, and claims of deep learning studies. *BMJ*. (2020) 368:m689. doi: 10.1136/bmj.m689

32. Seah JCY, Tang JSN, Kitchen A, Gaillard F, Dixon AF. Chest radiographs in congestive heart failure: visualizing neural network learning. *Radiology*. (2019) 290:514–22. doi: 10.1148/radiol.2018180887

33. Horng S, Liao R, Wang X, Dalal S, Golland P, Berkowitz SJ. Deep learning to quantify pulmonary edema in chest radiographs. *Radiol Artif Intell*. (2021) 3:e190228. doi: 10.1148/ryai.2021190228

34. Fu J, Luo K, S Levine. Learning robust rewards with adversarial inverse reinforcement learning (2017).

35. Topol EJ. High-performance medicine: the convergence of human and artificial intelligence. *Nat Med*. (2019) 25:44–56. doi: 10.1038/s41591-018-0300-7

36. Thoral PJ, Peppink JM, Driessen RH, Sijbrands EJG, Kompanje EJO, Kaplan L, et al. Sharing ICU patient data responsibly under the Society of Critical Care Medicine/European Society of Intensive Care Medicine Joint Data Science Collaboration: the Amsterdam University Medical Centers database (AmsterdamUMCdb) example. *Crit Care Med*. (2021) 49:e563–77. doi: 10.1097/CCM.0000000000004916

37. Laird P, Wertz A, Welter G, Maslove D, Hamilton A, Heung Yoon J, et al. The critical care data exchange format: a proposed flexible data standard for combining clinical and high-frequency physiologic data in critical care. *Physiol Meas*. (2021) 42:065002. doi: 10.1088/1361-6579/abfc9b

38. Angus DC, Berry S, Lewis RJ, Al-Beidh F, Arabi Y, van Bentum-Puijk W, et al. The REMAP-CAP (randomized embedded multifactorial adaptive platform for community-acquired pneumonia) study. Rationale and design. *Ann Am Thorac Soc*. (2020) 17:879–91. doi: 10.1513/AnnalsATS.202003-192SD

39. Fries JA, Steinberg E, Khattar S, Fleming SL, Posada J, Callahan A, et al. Ontology-driven weak supervision for clinical entity classification in electronic health records. *Nat Commun*. (2021) 12:2017. doi: 10.1038/s41467-021-22328-4

40. Angus DC, Derde L, Al-Beidh F, Annane D, Arabi Y, Beane A, et al. Effect of hydrocortisone on mortality and organ support in patients with severe COVID-19: the REMAP-CAP COVID-19 corticosteroid domain randomized clinical trial. *JAMA*. (2020) 324:1317–29. doi: 10.1001/jama.2020.17022

41. Gordon AC, Mouncey PR, Al-Beidh F, Rowan KM, Nichol AD, Arabi YM, et al. Interleukin-6 receptor antagonists in critically ill patients with Covid-19. *N Engl J Med*. (2021) 384:1491–502. doi: 10.1056/NEJMoa2100433

42. Patel BN, Rosenberg L, Willcox G, Baltaxe D, Lyons M, Irvin J, et al. Human-machine partnership with artificial intelligence for chest radiograph diagnosis. *NPJ Digital Medicine*. (2019) 2:111. doi: 10.1038/s41746-019-0189-7

43. Matheny ME, Whicher D, Thadaney Israni S. Artificial intelligence in health care: a report from the National Academy of medicine. *JAMA*. (2020) 323:509–10. doi: 10.1001/jama.2019.21579

frontiers | Frontiers in Medicine

TYPE Original Research
PUBLISHED 27 April 2023
DOI 10.3389/fmed.2023.1166896

OPEN ACCESS

EDITED BY
Rahul Kashyap,
WellSpan Health, United States

REVIEWED BY
Saraswathi Lakkasani,
Saint Michael's Medical Center, United States
Mack Sheraton,
Trinity Health System, United States

*CORRESPONDENCE
Xuandong Jiang
✉ lxqjiang@hotmail.com

RECEIVED 15 February 2023
ACCEPTED 10 April 2023
PUBLISHED 27 April 2023

CITATION
Jiang X, Zhang W, Pan Y and Cheng X (2023) Identification of subphenotypes in critically ill thrombocytopenic patients with different responses to therapeutic interventions: a retrospective study.
Front. Med. 10:1166896.
doi: 10.3389/fmed.2023.1166896

Identification of subphenotypes in critically ill thrombocytopenic patients with different responses to therapeutic interventions: a retrospective study

Xuandong Jiang*, Weimin Zhang, Yuting Pan and Xuping Cheng

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

Introduction: The causes of thrombocytopenia (TP) in critically ill patients are numerous and heterogeneous. Currently, subphenotype identification is a popular approach to address this problem. Therefore, this study aimed to identify subphenotypes that respond differently to therapeutic interventions in patients with TP using routine clinical data and to improve individualized management of TP.

Methods: This retrospective study included patients with TP admitted to the intensive care unit (ICU) of Dongyang People's Hospital during 2010–2020. Subphenotypes were identified using latent profile analysis of 15 clinical variables. The Kaplan–Meier method was used to assess the risk of 30-day mortality for different subphenotypes. Multifactorial Cox regression analysis was used to analyze the relationship between therapeutic interventions and in-hospital mortality for different subphenotypes.

Results: This study included a total of 1,666 participants. Four subphenotypes were identified by latent profile analysis, with subphenotype 1 being the most abundant and having a low mortality rate. Subphenotype 2 was characterized by respiratory dysfunction, subphenotype 3 by renal insufficiency, and subphenotype 4 by shock-like features. Kaplan–Meier analysis revealed that the four subphenotypes had different in-30-day mortality rates. The multivariate Cox regression analysis indicated a significant interaction between platelet transfusion and subphenotype, with more platelet transfusion associated with a decreased risk of in-hospital mortality in subphenotype 3 [hazard ratio (HR): 0.66, 95% confidence interval (CI): 0.46–0.94]. In addition, there was a significant interaction between fluid intake and subphenotype, with a higher fluid intake being associated with a decreased risk of in-hospital mortality for subphenotype 3 (HR: 0.94, 95% CI: 0.89–0.99 per 1l increase in fluid intake) and an increased risk of in-hospital mortality for high fluid intake in subphenotypes 1 (HR: 1.10, 95% CI: 1.03–1.18 per 1l increase in fluid intake) and 2 (HR: 1.19, 95% CI: 1.08–1.32 per 1l increase in fluid intake).

Conclusion: Four subphenotypes of TP in critically ill patients with different clinical characteristics and outcomes and differential responses to therapeutic interventions were identified using routine clinical data. These findings can help improve the identification of different subphenotypes in patients with TP for better individualized treatment of patients in the ICU.

KEYWORDS

thrombocytopenia, subphenotypes, fluid resuscitation, artificial intelligence, latent profile analysis, critically ill

1. Introduction

Thrombocytopenia (TP) is generally defined as having a platelet count of $<100 \times 109$/L. This condition is common among critically ill patients in both medical and surgical intensive care units (ICUs), with a global prevalence of 21–77% (1, 2). The causes of TP in ICU patients vary, including sepsis, trauma, surgery, and medication (2, 3). Most patients develop TP within 4 days of admission to the ICU. A long duration of TP is associated with a poor prognosis (4, 5). Numerous studies have demonstrated that TP is an independent risk factor for mortality in ICU patients, being associated with severe bleeding events and increased transfusion requirements as well as with the duration of ICU stay and an increased incidence of acute kidney injury (AKI) (6, 7). Unfortunately, the efficacy of current interventions and treatment methods for TP in ICU patients is limited (8, 9).

Previous studies based on standardized treatment regimens for patients with TP have failed to yield satisfactory treatment outcomes. For example, a meta-analysis of the therapeutic efficacy of recombinant human thrombopoietin in patients with TP with sepsis by Zhang et al. revealed no significant difference in 28-day mortality (10). A recent review reported that the use of platelet transfusion, glucocorticoids, and intravenous immune globulin for the treatment of immune TP requires further study (11). The possible reasons for the unsatisfactory treatment outcomes in patients with TP include the significant heterogeneity of TP, which is associated with the presence of multiple pathogenic factors, such as inflammation, endothelial dysfunction, coagulopathy, hemodilution, and altered platelet production, in critically ill patients (12). Subphenotyping, a precision medicine-based treatment option, is currently a very common approach for addressing disease heterogeneity and has been applied to common critical illnesses, such as sepsis, AKI, and acute respiratory distress syndrome (ARDS) (13–15).

Most studies have focused on determining prognosis by staging, and only few studies have focused on different responses to treatment after staging. For example, Zhang et al. retrospectively analyzed 14,993 patients with severe sepsis and identified four subphenotypes of sepsis using latent profile analysis, each of which responded differently to fluid resuscitation (16). Bhatraju et al. used latent class analysis to classify a critically ill AKI population and applied it to AKI patients in the Vasopressin and Septic Shock Trial. The result of the initial analysis was negative, but subphenotyping revealed that vasopressin therapy had survival benefits in patients with subphenotype 1 (17). However, only few studies have reported on the subphenotypes of severe TP, and even fewer studies have reported on its response to different therapeutic interventions (12).

Therefore, this study aimed to identify different subphenotypes in TP patients admitted to the ICU of our hospital over the last 10 years with different clinical outcomes and different responses to therapeutic interventions, using latent profile analysis based on routine clinical data, with the aim of improving prognosis prediction and treatment of critically ill patients and providing guidance for clinicians to achieve individualized management of patients.

2. Materials and methods

2.1. Study design

This study followed the Strengthening the Reporting of Observational Studies in Epidemiology guidelines (Supplementary Table S1). In this retrospective study, 1,666 patients with TP who were first admitted to the ICU of Dongyang People's Hospital between January 1, 2010, and October 31, 2020, were included. The inclusion criteria were first admission to the ICU and ICU stay of $\geq$48 h. The exclusion criteria were age $<$ 18 years, hematological malignancy, liver cirrhosis, or previous splenectomy.

2.2. Data collection and grouping

2.2.1. Data collection

Data were collected using the medical record information mining software provided by Shanghai Le9 Healthcare Technology Co., Ltd. (Shanghai, China). The following information was collected: (1) age, sex, Acute Physiology and Chronic Health Evaluation (APACHE)-II score, complications, vasopressor use, renal replacement therapy, fluid intake and urine output for 24 h after ICU admission; and biochemical indexes and first vital signs at ICU admission.

The therapeutic interventions include glucocorticoid use, immunoglobulin use, platelet transfusion during ICU stay, and fluid intake for 24 h after ICU admission.

The primary outcome was hospital mortality. The secondary outcomes included duration of mechanical ventilation, length of ICU stay, length of hospital stay, and hospitalization cost.

2.2.2. Diagnostic criteria

We defined TP as a platelet count of $<100 \times 10^9$/L in the first 48 h after ICU admission (2, 3).

2.3. Data processing

Variables with >20% missing values were deleted. If the incidence of missing values was <2%, the mean value of the variable was substituted for the missing values. The missing values of variables with

Abbreviations: AKI, Acute kidney injury; APACHE, Acute physiology and chronic health evaluation; ARDS, Acute respiratory distress syndrome; BIC, Bayesian information criterion; CI, Confidence interval; ICU, Intensive care unit; LPA, Latent profile analysis; LRT, Likelihood ratio test; HR, Hazard ratio; TP, Thrombocytopenia.

loss rates of >2 and < 15% were replaced using multiple imputations. Outliers were handled as missing values.

2.4. Latent profile analysis

Latent profile analysis (LPA), an unsupervised machine learning algorithm, is a modeling approach for classifying latent variables that focuses on identifying potential subgroups within a population, based on a specific set of variables, using an expectation–maximization algorithm to estimate the parameters of the latent class model (18). The variables included in LPA modeling are clinical and are incorporated from domain expertise and from the relevant literature (16, 19, 20). Pearson's correlation analysis was used to determine the correlations among characteristic variables, and variables with correlation coefficients >0.7 were removed. Finally, the following 15 common clinical variables were selected: platelet count at initial admission to ICU, age, creatinine level, glucose concentration, systolic blood pressure, respiratory rate, oxygen saturation, heart rate, white blood cell count, hematocrit level, lactate level, pH, partial pressure of oxygen, partial pressure of carbon dioxide, and bicarbonate level. The number of categories was determined using the Bayesian information criterion (BIC), entropy, and bootstrap likelihood ratio tests. Lower BIC values indicated a better model fit. Entropy ranged from 0 to 1, with higher values indicating higher accuracy of categorization. The Vuong–Lo–Mendell–Rubin likelihood ratio test (LRT) was used to assess the number of mixture components in a given finite mixture model parameterization, and value of ps were reported to compare n-class and $(n-1)$-class models (21). A value of p of <0.05 indicated statistical significance in the LRT. In addition, the proportion of patients in each potential class with a number of patients of >5% of any other potential class should be assigned to a class with a minimum probability greater than 0.8, otherwise members of this class were considered unstable (22). The number of potential classes was determined in conjunction with clinical interpretation.

2.5. Statistical analyses

Descriptive statistics were analyzed conventionally using the CBCgrps package in R[1] (23). Normally distributed measurement data are expressed as mean and standard deviation ($x \pm s$), and non-normally distributed data are expressed as median [interquartile range (IQR): P25, P75]. Comparisons across groups on baseline characteristics were performed using analysis of variance for continuous variables and the chi-square tests for categorical variables. All statistical analyses were performed using R (software version 4.1.3; https://www.r-project.org/). A value of p <0.05 was considered statistically significant.

The Kaplan–Meier method was used to analyze the relationships of the four subphenotypes with in-hospital 30-day mortality. Multivariate Cox regression models were used to investigate the independent association between therapeutic interventions and mortality. Variables with $p < 0.1$ in the univariate regression analysis and the important clinical variables were selected for the Cox model to test for interactions between different categories and therapeutic interventions. The model was adjusted for the following covariates: age, sex, APACHE II score, vasopressor used, surgery, sepsis and white blood cell count. Platelet transfusion and fluid intake separately interacted with each category. The hazard ratio (HR) and associated 95% confidence interval (CI) for the effect of platelet transfusion and each 1 l increase in fluid intake on mortality outcomes are reported.

2.6. Ethics approval

This study was approved by the Ethics Committee of Dongyang People's Hospital (DRY-2023-YX-016) and followed all related local guidelines and regulations, including the human genetics-related regulations. The need for obtaining informed consent was waived by the Ethical Committee of Dongyang People's Hospital, due to the retrospective nature of this study, and the study involved no human tissue collection and storage process. The data were analyzed anonymously by removing personal information of the patients.

3. Results

3.1. Study population

The flow diagram of this study is shown in Figure 1. After excluding 8,702 patients, 1,666 participants with a mean age of 61.5 ± 16.6 years were finally included. Of these, 61.6% were male. The overall mortality rate was 23.4%.

3.2. Selection of optimal categories

The Akaike information criterion and sample size-adjusted BIC value decreased from the 2-class model to the 10-class model, but the decrease began to slow from the 4-class model to the 5-class model. The 4-class model had the largest entropy and minimum probability of <0.8, starting at the 5-class model, suggesting that the minimum probability assigned to this class was <0.8, and the 5–10-class models were considered unstable (Figure 2). Therefore, the optimal selection was a 4-class model.

3.3. Clinical characteristics and outcomes of subphenotypes

The characteristics of the four subphenotypes are shown in Figure 3 and Table 1. Subphenotype 1 was the most abundant one of the four categories, with a total of 1,097 patients, accounting for 66% of all patients. The values of all variables were approximate of the means. Thus, subphenotype 1 was considered as the baseline category. Subphenotype 2 was characterized by low oxygen saturation [94, IQR: 93–96%], low partial pressure of oxygen (97.4 ± 41.9 mmHg), and the highest partial pressure of carbon dioxide (36.1 ± 8.2 mmHg) and was considered as the respiratory failure category. Subphenotype 3 was characterized by the highest serum creatinine level (272, IQR: 216–272 mmol/L) and low bicarbonate levels (17.6 ± 3.7 mmol/L) and was considered as the renal insufficiency category. Subphenotype 4 was characterized by the highest lactate level (7.90, IQR, 6.40–10.05 mmol/L),

1 https://www.r-project.org/

low systolic blood pressure (116.4 ± 23.8 mmHg), and low bicarbonate level (17.0 ± 2.7 mmol/L) and was considered as the shock category.

Table 2 shows a comparison of clinical outcomes. Subphenotype 1 had the lowest mortality rate (17.4%), the lowest duration of mechanical ventilation, the shortest duration of ICU stay and hospital stay, and the lowest hospitalization cost. Subphenotype 3 had the highest mortality rate (47.4%), the highest APACHE II score (25.0 ± 8.1), and the highest proportion of renal replacement therapy (47.4%). Subphenotype 4 had a mortality rate of 31.3%, the longest duration of hospital stay (23 days, IQR: 12–34 days), and the highest hospital cost (CNY 108×10^3, IQR: CNY 52×10^3–149×10^3). Subphenotypes 2 and 4 had similar mortality rates (Figure 4).

3.4. Therapeutic interventions

There were significant differences in the proportion of platelet transfusion among the four subphenotypes ($p < 0.001$). Subphenotype 1 had the lowest platelet transfusion rate (11.9%), and others had a platelet transfusion rate of >20%. After adjusting for age, sex, APACHE II score, vasopressor used, white blood cell count, surgery, and sepsis, multivariate Cox regression models indicated that there was a significant interaction between platelet transfusion and each category, with higher platelet transfusion associated with a decreased risk of in-hospital mortality in subphenotype 3 (HR: 0.66, 95% CI: 0.46–0.94; Table 3). The total fluid intake at 24 h after admission to the ICU was 4.0 (IQR: 3.3–5.2) L, and the total urine output was 2.3 (IQR: 1.6–3.1) L in all patients. Subphenotype 1 had the lowest fluid intake (3.91, IQR: 3.28–4.84) L and highest urine output (2.4, IQR: 1.7–3.1) L. Subphenotype 4 had the highest fluid intake (4.66, IQR: 3.72–6.51) L, and subphenotype 3 had the lowest urine output (1.3, IQR: 0.4–2.3) L (Table 2). However, a significant interaction was noted between fluid intake and each category, with higher fluid intake associated with a decreased risk of in-hospital mortality in subphenotype 3 (HR: 0.94, 95% CI: 0.89–0.99 per 1 l increase in fluid intake) but associated with an increased risk of in-hospital mortality in subphenotypes 1 (HR: 1.10, 95% CI: 1.03–1.18 per 1 l increase in fluid intake) and 2 (HR: 1.19, 95% CI: 1.08–1.32 per 1 l increase in fluid intake; Table 4). Figure 5 shows platelet transfusion and risk of hospital mortality, stratified by four subphenotypes, whereas Figure 6 shows fluid intake and risk of hospital mortality, stratified by four subphenotypes.

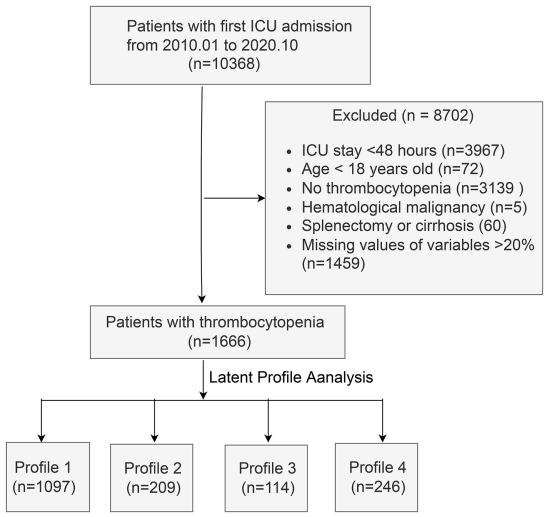

FIGURE 1
Flow chart of the study. ICU, intensive care unit.

3.5. Sensitivity analysis

We deleted 203 patients with missing data, retained outliers for sensitivity analysis, and obtained similar results in LPA analysis (Supplementary Figure S1). The maximum value of entropy was in four categories; therefore, the best classification was four categories, and the features of the four categories were also similar.

4. Discussion

In this study, we identified four clinical subphenotypes of TP, with different physiological characteristics and in-hospital mortality, using only routine clinical data. We also found an interaction between subphenotypes and platelet transfusion and fluid intake, suggesting the involvement of these subphenotypes in precision medicine-based approaches to the treatment of TP.

Best Number of Classes for FMM

Classes	AIC	SABIC	Entropy	prob_min	prob_max	n_min	n_max	BLRT_p
2	69115	69218	0.807	0.889	0.968	0.284	0.716	0.01
3	67290	67429	0.885	0.883	0.981	0.13	0.648	0.01
4	66660	66835	0.933	0.892	0.978	0.068	0.658	0.01
5	66379	66589	0.815	0.793	0.983	0.076	0.436	0.01
6	66091	66338	0.791	0.753	0.979	0.065	0.339	0.01
7	65988	66271	0.769	0.719	0.986	0.058	0.291	0.01
8	65929	66247	0.752	0.707	0.981	0.059	0.245	0.01
9	65719	66073	0.76	0.646	0.97	0.025	0.261	0.01
10	65655	66045	0.764	0.644	0.983	0.047	0.219	0.01

FIGURE 2
Best number of classes for latent profile analysis. The value of *p* was reported for the bootstrap likelihood ratio test comparing the current model (k class) to the model with k-1 class. AIC, Akaike information criterion; SABIC, sample size-adjusted Bayesian information criteria.

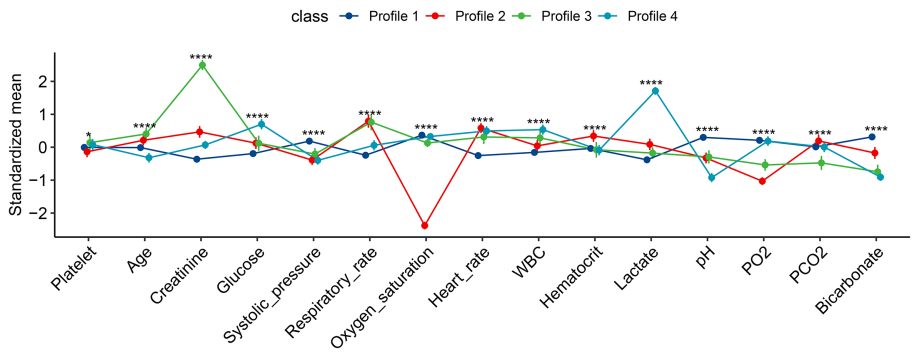

FIGURE 3
Characteristics of the four subphenotypes identified by latent profile analysis. All numeric values were scaled for better visualization on the vertical axis. Profile 1 is the largest class over all study days with all variables in average value (the baseline class). Profile 2 is characterized by low oxygen saturation and partial pressure of oxygen, the highest partial pressure of carbon dioxide (the respiratory failure class). Profile 3 is characterized by the highest serum creatinine and low bicarbonate levels (renal dysfunction class). Profile 4 is characterized by the highest lactate level, and low systolic pressure and bicarbonate level (the shock class). FMM, finite mixture modeling; WBC, white blood cell; PO_2, partial pressure of oxygen; PCO_2, partial pressure of carbon dioxide. *p<0.05, ****p<0.001.

TABLE 1 Continuous variables included in the mixture modeling.

Characteristic	Profile 1 (n=1,097)	Profile 2 (n=209)	Profile 3 (n=114)	Profile 4 (n=246)	p
Age (years)	61.3 ± 16.6	65.0 ± 15.8	68.2 ± 14.2	56.2 ± 16.5	<0.001
Platelet (×10⁹/L)	87.1 ± 28.7	83.0 ± 33.1	91.7 ± 38.2	89.8 ± 36.8	0.041
White blood cell (×10⁹/L)	11.2 ± 5.0	12.3 ± 6.7	13.7 ± 7.7	15.2 ± 6.1	<0.001
Hematocrit	0.3 ± 0.1	0.3 ± 0.1	0.3 ± 0.1	0.3 ± 0.1	<0.001
pH	7.4 ± 0.1	7.4 ± 0.1	7.4 ± 0.1	7.3 ± 0.1	<0.001
PO2 (mmHg)	171.0 ± 54.9	97.4 ± 41.9	126.6 ± 49.0	169.6 ± 57.8	<0.001
PCO2 (mmHg)	34.8 ± 6.4	36.1 ± 8.2	31.5 ± 7.0	34.8 ± 7.4	<0.001
Bicarbonate (mmol/L)	21.4 ± 2.9	19.7 ± 4.1	17.6 ± 3.7	17.0 ± 2.7	<0.001
Lactate (mmol/L)	2.20 (1.50, 3.20)	2.70 (1.50, 5.00)	2.60 (1.50, 3.68)	7.90 (6.40, 10.05)	<0.001
Creatinine (mmol/L)	70 (55, 91)	106 (70, 170)	272 (216, 272)	90 (71, 121)	<0.001
Glucose (mmol/L)	8.3 (6.8, 10.1)	8.9 (7.2, 11.5)	9.2 (7.2, 11.5)	11.4 (8.9, 13.9)	<0.001
Systolic pressure (mmHg)	133.1 ± 28.8	116.7 ± 26.1	121.9 ± 24.7	116.4 ± 23.8	<0.001
Heart rate (/min)	88.0 ± 19.0	105.4 ± 19.9	99.8 ± 21.2	103.6 ± 20.2	<0.001
Respiratory rate (/min)	14 (12, 16)	18 (14, 26)	18 (14, 25)	14 (12, 19)	<0.001
Oxygen saturation (%)	100 (100, 100)	94 (93, 96)	100 (99, 100)	100 (99, 100)	<0.001

Continuous variables are described by means and quarterbacks. Categories variables are analyzed by χ^2 test and continuous variables are analyzed by Wilcoxon rank sum test. PO2, partial pressure of oxygen; PCO2, partial pressure of carbon dioxide.

Platelet transfusion is a common treatment for patients with PT, but it can be ineffective for various reasons, such as infection, medication, disseminated intravascular coagulation, etc. (24–26). In some cases, platelet counts transiently increase after transfusion, and several studies have demonstrated that platelet transfusion does not improve patient outcomes (27). Our study demonstrated that platelet transfusion can improve in-hospital mortality rates in patients with subphenotype 3 of TP, indicating that identifying subphenotypes is a potential method for addressing platelet transfusion in critically ill patients.

Intravenous fluids are the cornerstone of patient care in the ICU; both inadequate fluid intake and fluid overload increased mortality. Overall, in-hospital mortality increased with higher fluid intake in our study, which is consistent with the finding of previous studies (28, 29). However, in subphenotype 3 cases, increased fluid intake was associated with improved outcomes. This may be associated with the clinical characteristics of subphenotype 3, including renal dysfunction, metabolic acidosis. Most clinicians are now aware that AKI patients require fluid restriction; however, excessive fluid restriction may lead to insufficient effective blood volume (30, 31). Therefore, a more precise volume assessment is necessary for this patient subpopulation. Subphenotype 4 exhibited the highest lactate level, a high fluid intake, and a high urine output but lower mortality than that exhibited by subphenotype 3, which may be related to less fluid overload. Previous studies have demonstrated that fluid overload is positively correlated with mortality in critically ill patients (32, 33). Therefore, we believe that precise fluid management based on subphenotypic classification is a promising future direction.

TABLE 2 Categorical variables and outcome variables not included in the mixture modeling.

Characteristic	Profile 1 (*n*=1,097)	Profile 2 (*n*=209)	Profile 3 (*n*=114)	Profile 4 (*n*=246)	*p*
Male [*n*(%)]	646.0 (58.9%)	143.0 (68.4%)	90.0 (78.9%)	147.0 (59.8%)	<0.001
Smoking [*n*(%)]	377.0 (34.4%)	89.0 (42.6%)	49.0 (43.0%)	90.0 (36.6%)	0.056
Alcohol drinking [*n*(%)]	404.0 (36.8%)	83.0 (39.7%)	48.0 (42.1%)	91.0 (37.0%)	0.637
Comorbidities [*n*(%)]					
Hypertension	312.0 (28.4%)	78.0 (37.3%)	60.0 (52.6%)	58.0 (23.6%)	<0.001
Diabetes	86.0 (7.8%)	28.0 (13.4%)	24.0 (21.1%)	27.0 (11.0%)	<0.001
Congestive heart failure	32.0 (2.9%)	15.0 (7.2%)	13.0 (11.4%)	16.0 (6.5%)	<0.001
Chronic obstructive pulmonary disease	75.0 (6.8%)	35.0 (16.7%)	17.0 (14.9%)	9.0 (3.7%)	<0.001
Input_24h (L/h)	3.91 (3.28, 4.84)	4.02 (3.27, 5.14)	4.27 (2.98, 5.97)	4.66 (3.72, 6.51)	<0.001
Uo_24h (L/h)	2.4 (1.7, 3.1)	2.3 (1.3, 3.2)	1.3 (0.4, 2.3)	2.3 (1.6, 3.1)	<0.001
APACHE-II score	18.2 ± 7.0	21.9 ± 8.3	25.0 ± 8.1	20.8 ± 7.4	<0.001
Vasopressor used [n(%)]	651.0 (59.3%)	167.0 (79.9%)	98.0 (86.0%)	202.0 (82.1%)	<0.001
Glucocorticoid used [n(%)]	460 (41.9)	104 (49.8)	38 (33.3)	123 (50)	0.004
Immunoglobulin used [*n*(%)]	9 (0.8)	13 (6.2)	4 (3.5)	3 (1.2)	<0.001
Platelet infusion [*n*(%)]	130 (11.9)	43 (20.6)	29 (25.4)	59 (24)	<0.001
Renal replacement therapy [*n*(%)]	20.0 (1.8%)	31.0 (14.8%)	54.0 (47.4%)	28.0 (11.4%)	<0.001
Biochemical indexes on ICU admission					
Red blood cell ($\times 10^9$/L)	3.4 ± 0.6	3.6 ± 0.8	3.3 ± 0.8	3.3 ± 0.8	<0.001
Potassium (mmol/L)	4.1 ± 0.5	4.1 ± 0.6	4.5 ± 0.7	4.0 ± 0.6	<0.001
Sodium(mmol/L)	142.0 ± 4.1	142.5 ± 4.6	141.8 ± 5.4	144.5 ± 4.3	<0.001
Calcium (mmol/L)	2.0 ± 0.2	1.9 ± 0.2	1.9 ± 0.2	1.9 ± 0.2	0.013
Urea (mmol/L)	7.3 (5.5, 9.3)	10.3 (7.1, 15.2)	19.6 (14.6, 20.7)	7.9 (6.0, 10.3)	<0.001
Prothrombin time (s)	15.6 (14.5, 16.8)	15.9 (14.4, 18.5)	17.0 (14.9, 19.4)	16.7 (15.0, 20.0)	<0.001
International normalized ratio)	1.25 (1.13, 1.38)	1.28 (1.13, 1.55)	1.38 (1.18, 1.65)	1.35 (1.19, 1.68)	<0.001
Activated partial thromboplastin time (s)	40 (36, 46)	46 (40, 55)	47 (41, 58)	43 (37, 59)	<0.001
D.dimer (μg/L)	5.3 (2.1, 13.6)	7.0 (2.6, 16.0)	6.6 (2.7, 16.0)	8.1 (2.6, 16.0)	0.001
Outcome					
Hospital_mortality [*n*(%)]	191.0 (17.4%)	68.0 (32.5%)	54.0 (47.4%)	77.0 (31.3%)	<0.001
Ventilation duration (days)	2 (1, 7)	4 (1, 9)	4 (0, 10)	3 (1, 10)	0.017
ICU length of stay (days)	5 (3, 11)	7 (4, 13)	7 (3, 14)	7 (4, 13)	<0.001
Length of hospital stay (days)	21 (13, 31)	17 (10, 30)	16 (8, 26)	23 (12, 34)	<0.001
Cost ($\times 10^3$ yuan)	67 (40, 101)	57 (32, 99)	57 (32, 97)	108 (52, 149)	<0.001

Continuous variables are described by means and quarterbacks. Categories variables are analyzed by χ^2 test and continuous variables are analyzed by Wilcoxon rank sum test. APACHE, acute physiology and chronic health evaluation; Input_24h, fluid input for 24 h on ICU admission; Uo_24h, urine volume for 24 h on ICU admission; ICU, intensive care unit; Hosp. LOS, length of hospital stay.

Previous classifications of TP were based only on the severity of platelet count decrease, and some critically ill patients often presented with transient TP that was not well reflective of patient prognosis or therapeutic efficacy. Wu et al. reviewed three subphenotypes based on possible mechanisms of sepsis-associated TP: increased platelet consumption, decreased platelet production, and increased platelet destruction (34). In a similar study, Bedet et al. used hierarchical clustering of 60 patients with septic shock and identified five subphenotypes of patients with septic TP, which facilitated further understanding of the mechanisms of TP (12). However, their study included 27 endogenous mediators associated with sepsis, and the clinical applicability of this classification system may be limited.

In the present study, the classification of clinical subphenotypes of TP was based on LPA, which can be used to assess continuous indicators commonly measured in clinics. In contrast to cluster analysis, LPA considers measurement errors and uses objective criteria to determine the optimal categories, making it more robust and reliable, with a minimum class membership probability of >0.8 indicating good model stability (35). Similar techniques have been

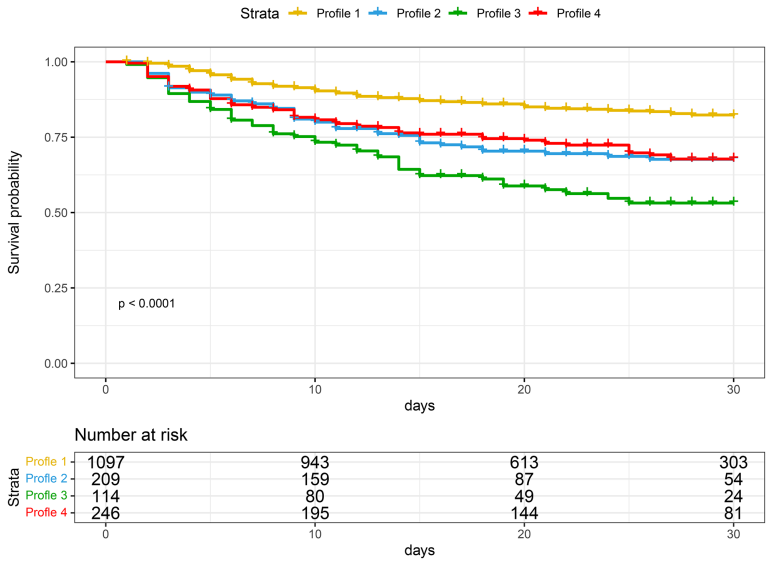

FIGURE 4
Kaplan–Meier curves for 30-day survival, stratified by four subphenotypes.

TABLE 3 Cox's proportional hazard models for platelet transfusion and hospital mortality in different profiles.

Characteristic	HR	95% CI	*p*
Age	0.84	0.68, 1.03	0.10
Sex	1.14	0.92, 1.42	0.2
APACHE-II score	1.10	1.08, 1.11	<0.001
Vasopressor used	3.43	2.42, 4.84	<0.001
White blood cell	0.97	0.95, 0.99	<0.001
Surgery	0.47	0.38, 0.59	<0.001
Sepsis	0.71	0.56, 0.88	0.002
Class			
Profile 1	—	—	
Profile 2	1.01	0.71, 1.44	0.9
Profile 3	1.48	1.03, 2.13	0.036
Profile 4	1.38	1.00, 1.91	0.053
Interaction between profile and platelet transfusion			
Profile 1	1.17	0.81,1.67	0.4
Profile 2	0.88	0.48, 1.64	0.7
Profile 3	0.66	0.46, 0.94	0.023
Profile 4	0.69	0.39, 1.23	0.2

HR, hazard ratio; CI, confidence interval; APACHE, acute physiology and chronic health evaluation; Input_24h, fluid input for 24 h on ICU admission; ICU, intensive care unit.

TABLE 4 Cox's proportional hazard models for fluid input and hospital mortality in different profiles.

Characteristic	HR	95% CI	*p*
Age	0.88	0.72, 1.09	0.2
Sex	1.12	0.90, 1.40	0.3
APACHE-II score	1.09	1.08, 1.11	<0.001
Vasopressor used	3.38	2.39, 4.79	<0.001
White blood cell	0.97	0.95, 0.99	<0.001
Surgery	0.47	0.38, 0.59	<0.001
Sepsis	0.72	0.57, 0.90	0.005
Class			
Profile 1	—	—	
Profile 2	0.81	0.44, 1.48	0.5
Profile 3	2.30	1.21, 4.36	0.011
Profile 4	1.76	0.94, 3.28	0.077
Interaction between profile and Input_24h			
Profile 1	1.10	1.03, 1.18	0.005
Profile 2	1.19	1.08, 1.32	0.001
Profile 3	0.94	0.89, 0.99	0.029
Profile 4	1.02	0.93, 1.11	0.7

HR, hazard ratio; CI, confidence interval; APACHE, acute physiology and chronic health evaluation; Input_24h, fluid input for 24 h on ICU admission; ICU, intensive care unit.

successfully applied to analyze therapeutic heterogeneity among subgroups of ARDS patients (36, 37).

This study had some limitations. First, the nature of the study was retrospective, and no causal inferences could be drawn. Moreover, the variables investigated were selected with reference to previous studies. Information on some underlying variables (such as height and weight) and inflammation-related variables (such as C-reactive protein and procalcitonin levels) was not available. Thus, further validation of our results in prospective studies is required. Second, the study was conducted at a single center and lacked external validation, which may limit the generalizability and reproducibility of the findings. Future research may explore external validation to ensure the robustness and reliability of the subphenotypes identified. Third, while LPA is a useful technique for identifying subgroups within a population, it is still a relatively new and evolving methodology.

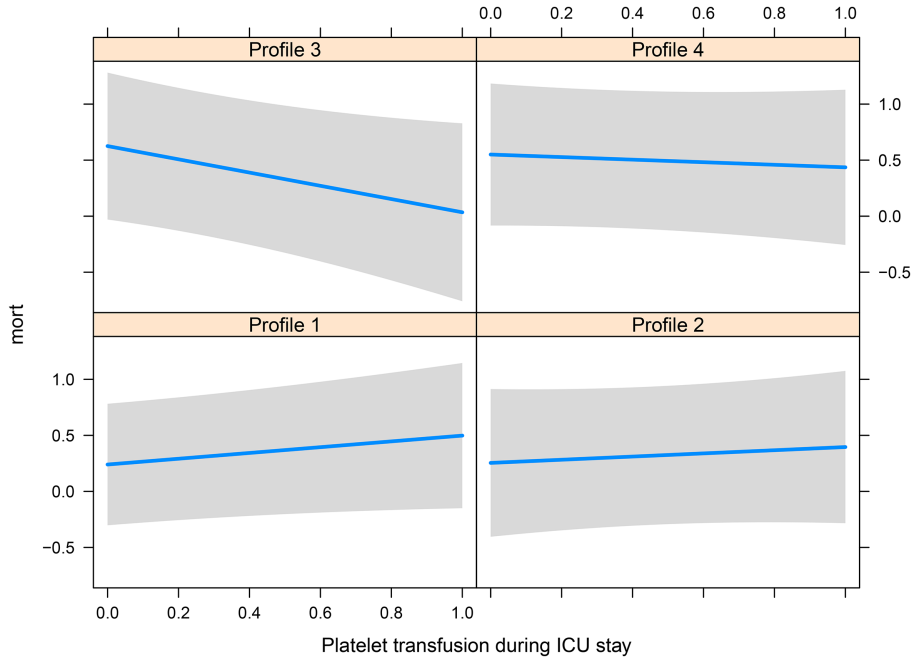

FIGURE 5
Platelet transfusion and risk of hospital mortality, stratified by four subphenotypes.

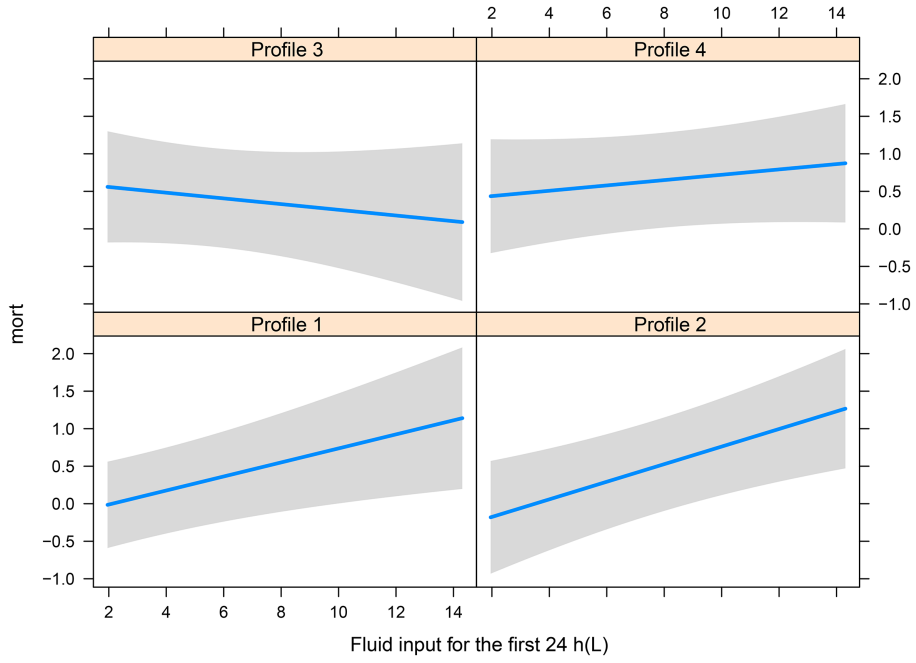

FIGURE 6
Fluid intake and risk of hospital mortality, stratified by four subphenotypes.

Further validation and refinement of this technique may be required to ensure its accuracy and reproducibility. Finally, we were unable to exclude patients with specific types of TP, such as TP due to pharmacological factors and immune-related TP. Fortunately, the overall proportion of such cases was small and did not affect the final results.

5. Conclusion

We identified four subphenotypes of patients with TP in the ICU, with different prognoses and different responses to therapeutic interventions, using common biochemical indicators and vital signs. These findings can improve our understanding of the heterogeneity of

patients with TP and can be used as a basis for future studies. In addition, these findings may facilitate the identification of different subphenotypes of TP for better individualized treatment of patients in the ICU.

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 DRY-2023-YX-016. Written informed consent for participation was not required for this study in accordance with the national legislation and the institutional requirements.

Author contributions

WZ and XJ carried out the design. XJ analyzed the data and drafted the manuscript. YP revised the manuscript. XC supervised the study. All authors contributed to the article and approved the submitted version.

Funding

This study was supported by the Conba Hospital Management project of the Zhejiang Hospital association (grant number 2021ZHA-KEB335) and the Dongyang Science and Technology Bureau (grant number 21-337).

Acknowledgments

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

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.2023.1166896/full#supplementary-material

References

1. 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

2. Ostadi Z, Shadvar K, Sanaie S, Mahmoodpoor A, Saghaleini SH. Thrombocytopenia in the intensive care unit. *Pak J Med Sci*. (2019) 35:282–7. doi: 10.12669/pjms.35.1.19

3. Menard CE, Kumar A, Houston DS, Turgeon AF, Rimmer E, Houston BL, et al. Evolution and impact of thrombocytopenia in septic shock: a retrospective cohort study. *Crit Care Med*. (2019) 47:558–65. doi: 10.1097/CCM.0000000000003644

4. Thiele T, Selleng K, Selleng S, Greinacher A, Bakchoul T. Thrombocytopenia in the intensive care unit-diagnostic approach and management. *Semin Hematol*. (2013) 50:239–50. doi: 10.1053/j.seminhematol.2013.06.008

5. Jiang X, Zhang W, Ma X, Cheng X. Risk of hospital mortality in critically ill patients with transient and persistent thrombocytopenia: a retrospective study. *Shock*. (2022) 58:471–5. doi: 10.1097/SHK.0000000000002005

6. 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

7. Zarychanski R, Houston DS. Assessing thrombocytopenia in the intensive care unit: the past, present, and future. *Hematol Am Soc Hematol Educ Program*. (2017) 2017:660–6. doi: 10.1182/asheducation-2017.1.660

8. Hamada SR, Garrigue D, Nougue H, Meyer A, Boutonnet M, Meaudre E, et al. Impact of platelet transfusion on outcomes in trauma patients. *Crit Care*. (2022) 26:49. doi: 10.1186/s13054-022-03928-y

9. Knöbl P. Thrombocytopenia in the intensive care unit: diagnosis, differential diagnosis, and treatment. *Med Klin Intensivmed Notfmed*. (2016) 111:425–33. doi: 10.1007/s00063-016-0174-8

10. Zhang J, Lu Z, Xiao W, Hua T, Zheng Y, Yang M. Efficacy and safety of recombinant human thrombopoietin on sepsis patients with thrombocytopenia: a systematic review and meta-analysis. *Front Pharmacol*. (2020) 11:940. doi: 10.3389/fphar.2020.00940

11. Cooper N, Ghanima W. Immune thrombocytopenia. *N Engl J Med*. (2019) 381:945–55. doi: 10.1056/NEJMcp1810479

12. Bedet A, Razazi K, Boissier F, Surenaud M, Hue S, Giraudier S, et al. Mechanisms of thrombocytopenia during septic shock: a multiplex cluster analysis of endogenous sepsis mediators. *Shock*. (2018) 49:641–8. doi: 10.1097/SHK.0000000000001015

13. Vaara ST, Forni LG, Joannidis M. Subphenotypes of acute kidney injury in adults. *Curr Opin Crit Care*. (2022) 28:599–604. doi: 10.1097/MCC.0000000000000970

14. Soussi S, Sharma D, Jüni P, Lebovic G, Brochard L, Marshall JC, et al. Identifying clinical subtypes in sepsis-survivors with different one-year outcomes: a secondary latent class analysis of the FROG-ICU cohort. *Crit Care*. (2022) 26:114. doi: 10.1186/s13054-022-03972-8

15. Maddali MV, Churpek M, Pham T, Rezoagli E, Zhuo H, Zhao W, et al. Validation and utility of ARDS subphenotypes identified by machine-learning models using clinical data: an observational, multicohort, retrospective analysis. *Lancet Respir Med*. (2022) 10:367–77. doi: 10.1016/S2213-2600(21)00461-6

16. 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

17. Bhatraju PK, Zelnick LR, Herting J, Katz R, Mikacenic C, Kosamo S, et al. Identification of acute kidney injury subphenotypes with differing molecular signatures and responses to vasopressin therapy. *Am J Respir Crit Care Med*. (2019) 199:863–72. doi: 10.1164/rccm.201807-1346OC

18. Nylund KL, Asparouhov T, Muthén BO. Deciding on the number of classes in latent class analysis and growth mixture modeling: a Monte Carlo simulation study. *Struct Equ Model Multidiscip J*. (2007) 14:535–69. doi: 10.1080/10705510701575396

19. 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

20. Zhang Z, Yao M, Ho KM, Hong Y. Subphenotypes of cardiac arrest patients admitted to intensive care unit: a latent profile analysis of a large critical care database. *Sci Rep*. (2019) 9:13644. doi: 10.1038/s41598-019-50178-0

21. Kim SY. Determining the number of latent classes in single- and multi-phase growth mixture models. *Struct Equ Model*. (2014) 21:263–79. doi: 10.1080/10705511.2014.882690

22. Nasserinejad K, van Rosmalen J, de Kort W, Lesaffre E. Comparison of criteria for choosing the number of classes in bayesian finite mixture models. *PLoS One*. (2017) 12:e0168838. doi: 10.1371/journal.pone.0168838

23. 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

24. Cohn CS. Platelet transfusion refractoriness: how do I diagnose and manage? *Hematology Am Soc Hematol Educ Program*. (2020) 2020:527–32. doi: 10.1182/hematology.2020000137

25. Slichter SJ, Davis K, Enright H, Braine H, Gernsheimer T, Kao KJ, et al. Factors affecting posttransfusion platelet increments, platelet refractoriness, and platelet transfusion intervals in thrombocytopenic patients. *Blood*. (2005) 105:4106–14. doi: 10.1182/blood-2003-08-2724

26. Aster RH, Curtis BR, McFarland JG, Bougie DW. Drug-induced immune thrombocytopenia: pathogenesis, diagnosis, and management. *J Thromb Haemost*. (2009) 7:911–8. doi: 10.1111/j.1538-7836.2009.03360.x

27. Hod E, Schwartz J. Platelet transfusion refractoriness. *Br J Haematol*. (2008) 142:348–60. doi: 10.1111/j.1365-2141.2008.07189.x

28. Shen Y, Huang X, Zhang W. Association between fluid intake and mortality in critically ill patients with negative fluid balance: a retrospective cohort study. *Crit Care*. (2017) 21:104. doi: 10.1186/s13054-017-1692-3

29. Marik PE, Linde-Zwirble WT, Bittner EA, Sahatjian J, Hansell D. Fluid administration in severe sepsis and septic shock, patterns and outcomes: an analysis of a large national database. *Intensive Care Med*. (2017) 43:625–32. doi: 10.1007/s00134-016-4675-y

30. Ostermann M, Liu K, Kashani K. Fluid management in acute kidney injury. *Chest*. (2019) 156:594–603. doi: 10.1016/j.chest.2019.04.004

31. Perner A, Prowle J, Joannidis M, Young P, Hjortrup PB, Pettilä V. Fluid management in acute kidney injury. *Intensive Care Med*. (2017) 43:807–15. doi: 10.1007/s00134-017-4817-x

32. Messmer AS, Zingg C, Müller M, Gerber JL, Schefold JC, Pfortmueller CA. Fluid overload and mortality in adult critical care patients-a systematic review and meta-analysis of observational studies. *Crit Care Med*. (2020) 48:1862–70. doi: 10.1097/CCM.0000000000004617

33. Silversides JA, Fitzgerald E, Manickavasagam US, Lapinsky SE, Nisenbaum R, Hemmings N, et al. Deresuscitation of patients with iatrogenic fluid overload is associated with reduced mortality in critical illness. *Crit Care Med*. (2018) 46:1600–7. doi: 10.1097/CCM.0000000000003276

34. Wu X, Li Y, Tong H. Research advances in the subtype of sepsis-associated thrombocytopenia. *Clin Appl Thromb Hemost*. (2020) 26:1076029620959467. doi: 10.1177/1076029620959467

35. Zhang Z, Abarda A, Contractor AA, Wang J, Dayton CM. Exploring heterogeneity in clinical trials with latent class analysis. *Ann Transl Med*. (2018) 6:119. doi: 10.21037/atm.2018.01.24

36. Famous KR, Delucchi K, Ware LB, Kangelaris KN, Liu KD, Thompson BT, et al. Acute respiratory distress syndrome subphenotypes respond differently to randomized fluid management strategy. *Am J Respir Crit Care Med*. (2017) 195:331–8. doi: 10.1164/rccm.201603-0645OC

37. Sinha P, Delucchi KL, Thompson BT, McAuley DF, Matthay MA, Calfee CS, et al. Latent class analysis of ARDS subphenotypes: a secondary analysis of the statins for acutely injured lungs from sepsis (SAILS) study. *Intensive Care Med*. (2018) 44:1859–69. doi: 10.1007/s00134-018-5378-3

frontiers | Frontiers in Medicine

TYPE Original Research
PUBLISHED 18 May 2023
DOI 10.3389/fmed.2023.1165129

OPEN ACCESS

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

REVIEWED BY
Bin Yi,
Army Medical University, China
Jun Lyu,
First Affiliated Hospital of Jinan University, China

*CORRESPONDENCE
Li Xie
✉ redcat8851@163.com

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

RECEIVED 13 February 2023
ACCEPTED 02 May 2023
PUBLISHED 18 May 2023

CITATION
Yang J, Peng H, Luo Y, Zhu T and Xie L (2023) Explainable ensemble machine learning model for prediction of 28-day mortality risk in patients with sepsis-associated acute kidney injury.
Front. Med. 10:1165129.
doi: 10.3389/fmed.2023.1165129

Explainable ensemble machine learning model for prediction of 28-day mortality risk in patients with sepsis-associated acute kidney injury

Jijun Yang[1†], Hongbing Peng[2†], Youhong Luo[1], Tao Zhu[1] and Li Xie[3*]

[1]Department of Critical Care Medicine, Loudi Central Hospital, Loudi, China, [2]Department of Pulmonary and Critical Care Medicine, Loudi Central Hospital, Loudi, China, [3]Patient Service Center, Loudi Central Hospital, Loudi, China

Background: Sepsis-associated acute kidney injury (S-AKI) is a major contributor to mortality in intensive care units (ICU). Early prediction of mortality risk is crucial to enhance prognosis and optimize clinical decisions. This study aims to develop a 28-day mortality risk prediction model for S-AKI utilizing an explainable ensemble machine learning (ML) algorithm.

Methods: This study utilized data from the Medical Information Mart for Intensive Care IV (MIMIC-IV 2.0) database to gather information on patients with S-AKI. Univariate regression, correlation analysis and Boruta were combined for feature selection. To construct the four ML models, hyperparameters were tuned via random search and five-fold cross-validation. To evaluate the performance of all models, ROC, K-S, and LIFT curves were used. The discrimination of ML models and traditional scoring systems was compared using area under the receiver operating characteristic curve (AUC). Additionally, the SHapley Additive exPlanation (SHAP) was utilized to interpret the ML model and identify essential variables. To investigate the relationship between the top nine continuous variables and the risk of 28-day mortality. COX regression-restricted cubic splines were utilized while controlling for age and comorbidities.

Results: The study analyzed data from 9,158 patients with S-AKI, dividing them into a 28-day mortality group of 1,940 and a survival group of 7,578. The results showed that XGBoost was the best performing model of the four ML models with AUC of 0.873. All models outperformed APS-III 0.713 and SAPS-II 0.681. The K-S and LIFT curves indicated XGBoost as the most effective predictor for 28-day mortality risk. The model's performance was evaluated using ROCpr curves, calibration curves, accuracy, precision, and F1 scores. SHAP force plots were utilized to interpret and visualize the personalized predictive power of the 28-day mortality risk model. Additionally, COX regression restricted cubic splines revealed an interesting non-linear relationship between the top nine variables and 28-day mortality.

Conclusion: The use of ensemble ML models has shown to be more effective than the LR model and conventional scoring systems in predicting 28-day

mortality risk in S-AKI patients. By visualizing the XGBoost model with the best predictive performance, clinicians are able to identify high-risk patients early on and improve prognosis.

KEYWORDS

sepsis-associated acute kidney injury, ensemble machine learning, prediction model, XGBoost, MIMIC-IV database

Introduction

Sepsis continues to be a major cause of life-threatening conditions in critically ill patients. The excessive pro- or anti-inflammatory response can lead to cellular and organ dysfunction, ultimately resulting in death (1, 2). The most significant sepsis-associated organ disorder is acute kidney injury (AKI), which has a high prevalence (2, 3). AKI is an independent risk factor for high mortality (3, 4), and contributes to 58.6% of the excess attributable mortality (5). Sepsis-associated acute kidney injury (S-AKI) can be caused by microvascular dysfunction, inflammation, and metabolic reorganization. These play a crucial role in the development of S-AKI (3). However, the high heterogeneity in S-AKI is associated with multiple pathogenic mechanisms (3, 4), and there are currently no effective preventive or therapeutic measures available. The treatment for S-AKI is reactive and non-specific, which can result in a high mortality rate due to the difficulty in predicting AKI at the time of patient presentation. As such, salvage therapy is often the primary treatment option (3). However, providing early warning to patients at high mortality risk can help clinicians stratify patient management and improve the prognosis of patients with S-AKI.

Acute kidney injury, is a frequently encountered clinical syndrome that often accompanies critical illness. Its developmental process is complex and multifaceted. It is not sufficient to rely on a single variable to predict the mortality rate associated with AKI. Instead, combining multiple factors would be a more accurate way to forecast the prognosis of AKI (3). In the field of intensive care, conventional scoring systems that integrate clinical symptoms and laboratory data have been extensively utilized to forecast the prognosis of severely ill patients. Notably, the Sequential Organ Failure Assessment (SOFA), Acute Physiology Score III (APS-III), and Simplified Acute Physiology Score II (SAPS-II) scores have demonstrated robust predictive capabilities (6, 7). The prediction of 90-day mortality caused by severe infection-related AKI in China was carried out using COX regression analysis. The study identified several independent predictor variables including age, emergency ICU admission, post-surgical cases, admission diagnosis, AKI etiology, disease severity score, mechanical ventilation, use of boosters and blood outcomes such as albumin, potassium, and pH (8). In a study analyzing 30-day mortality in elderly patients with sepsis, a multivariate logistic regression-based analysis was conducted and resulted in a more accurate prediction with an AUC of 0.831 (9). Additionally, a multivariate prediction model for ICU and in-hospital death in AKI patients undergoing continuous renal replacement therapy found to be more accurate than SOFA, APACHE-II, and SAP-II scores (10). Recent trends suggest that the implementation of big data technologies in healthcare, specifically machine learning, has led to an improvement in the quality of care and optimization of healthcare processes and management strategies (11, 12). Studies have shown that machine learning prediction models have been successful in early warning of AKI occurrence and mortality risk (13, 14), with the XGBoost model achieving a high performance in predicting S-AKI (AUC 0.821) (15). Zhou et al. utilized data from the MIMIC III database to create a machine learning model for predicting AKI within 48 h of sepsis-related ARDS cases. Their model outperformed the discriminatory ability of SOFA (16). This highlights the potential of machine learning algorithms in accurately predicting the development of S-AKI.

Recent studies have shown that machine learning algorithms have achieved better performance in predicting S-AKI prognosis. For instance, the XGBoost model was constructed in a recent study to outperform the SOFA score and SAP-II in predicting mortality at different periods based on dynamic data of S-AKI cases updated every 12 h in the MIMICIV public database (14). However, there is a lack of research comparing multiple ensemble machine learning algorithms for early predict on of the high risk of 28-day mortality in S-AKI. Ensemble ML algorithms differ from traditional prediction models like logistic regression in that they do not involve rigorous screening of variables or adjustment for data imbalance during the construction process. This can lead to overfitting and classification boundary shifting in the resulting models. Previous studies on ML models have not extensively explored the linear or non-linear relationships between significant individual variables of the prediction model and the resulting outcomes.

This project aims to train and test multiple ensemble ML models using S-AKI data from the MIMIC-IV library. The goal is to select the best model that can provide early warning of the 28-day mortality risk in S-AKI cases. The interpretation and visualization of the prediction models are done using SHAP

Abbreviations: DM-without-cc, diabetes mellitus without complications; DM-with-cc, diabetes mellitus with complications; AMI, acute myocardial infarction; CHF, congestive heart failure; LMR, lymphocyte to monocyte ratio; CeVD, cerebrovascular disease; NLR, neutrophil to lymphocyte ratio; SBP, systolic blood pressure; DBP, diastolic blood pressure; MBI, body mass index; ROX_HR, the ratio of ROX index over HR (beats/min), multiplied by a factor of 100; PF_ratio, PaO2/FiO2 ratio; M_solid_tumor, metastatic solid tumor; BUN, blood urea nitrogen; S-AKI, sepsis-associated acute kidney injury; ICU, intensive care units; ML, machine learning; MIMIC-IV, the Medical Information Mart for Intensive Care IV; SMOTE, The Synthetic Minority Oversampling Technique; RF, random forest; GBM, Gradient Boosting Machine; XGBoost, Extreme Gradient Boosting; LR, logistic regression; AUC, the area under the receiver operating characteristic curve; SHAP, SHapley Additive exPlanation.

values. Specifically, SHAP force plots are analyzed to identify important mortality-related variables for individual cases. We utilized COX regression-restricted cubic spline plots to analyze the correlation between crucial, independent variables and 28-day mortality. Our ultimate goal is to develop a prediction model that can aid in treatment decisions for patients with S-AKI who are at a high risk of 28-day mortality, ultimately improving their chances of survival.

Materials and methods

Participants

The subject case dataset was obtained from the Medical Information Mart Intensive Care IV (MIMIC IV 2.0) database, which provides extensive information on more over 250,000 patients who were admitted to Beth Israel Deaconess Medical Center in Boston, Massachusetts, USA, from 2008 to 2021. The MIMIC IV public database was approved by the Institutional Review Board (IRB) and has undergone a thorough deidentification process. The database is freely available to researchers worldwide after receiving joint approval from the ethics review boards of MIT and Harvard Medical School. Informed consent was waived as the study was retrospective. To request access to the database, one of the investigators (HP) obtained a certificate (certification number 50527660) by passing the Human Research Participant Protection Examination.

Patients

The study included adult patients aged ≥18 years or older who met the Third International Consensus Definitions for Sepsis and Septic Shock (Sepsis-3) criteria (2), which requires the presence of known or suspected infection along with organ dysfunction wand a Sequential Organ Failure Assessment (SOFA) score of 2 or higher. Additionally, the study also included patients with AKI that was diagnosed and staged according to the 2012 Kidney Disease: Improving Global Outcomes (KDIGO) guidelines. The study excluded patients with renal disease, such as glomerulonephritis, diabetic nephropathy, hypertensive nephropathy, hereditary nephritis, and chronic renal failure caused by various other diseases. Additionally, only the first hospitalization was considered and patients with ICU stays (LOS) of less than 24 h were also excluded.

The study collected a comprehensive set of data on each patient, including their demographics (3 items), vital signs, blood gas analysis, blood cell count, blood biochemistry, hemodialysis phase reduction, and co-morbidities. Additionally, the study recorded information on AKI staging, use of an invasive ventilator, and urine output 24 h after ICU admission, resulting in a total of 60 variables. The study measured disease severity score (SOFA, ASPIII, SAPAII) within 24 h of ICU admission, length of stay, ICU time, 90-day mortality subgroup, in-hospital mortality subgroup, and follow-up time from hospitalization to death. Participants were divided into mortality and groups based on whether death occurred within 28 days.

Outcomes

The primary outcome after ICU admission was death within 28 days. Secondary outcomes included hospital mortality, length of stayin both the hospital and ICU, and COX regression-restricted cubic spline analysis.

Statistical methods

In our study, we excluded any variables with missing data greater than 20% of the case data. For the remaining missing values, we used the random forest method to interpolate. We recorded physiological data of patients every hour and used the mean value. For laboratory data, we selected the maximum or minimum value based on the basis that had the greatest impact on outcome in the clinic.

In the baseline data table, continuous variables are presented as median (IQR), and categorical variables as n (%). Appropriate statistical tests such as the Mann–Whitney U test, Student's t-test, chi-square test, or Fisher's exact test were used to compare baseline characteristic variables.

In the variable screening process, we first eliminated variables with $P > 0.05$ using univariate logistic regression analysis as they were deemed less likely to be relevant for 28-day mortality. We then removed variables with correlations greater than 0.75 through eliminated by correlation analysis. Finally, we used the "Boruta" package with the random forest algorithm to screen for essential characteristic variables to be included in the final model.

To calculate the lambda value of each variable in the right-skewed distribution, we used the Box-Cox method. We then performed a series of transformations, including square root, inverse, log, and inverse transformations, to obtain the transformed data-set.

To address data imbalance, we utilized the Synthetic Minority Oversampling Technique (SMOTE) algorithm during the ensemble machine learning model fitting process. The data-set was divided into training and testing sets at a 7:3 ratio. Ensemble learning algorithms, known for their superior performance in machine learning, were employed for the model fitting process. We utilized four models to construct the prediction model: Logistic regression (LR) as the baseline model, and Random Forest (RF), Gradient Boosting Machine (GBM), and Extreme Gradient Boosting Tree (XGBoost) representing the Bagging and Boosting algorithms. The hyperparameters were tuned using the random search method, and the ensemble machine learning model was fitted using the 5-fold cross-validation method. These models were automatically constructed using the 'creditmodel' data package. The performance of the four ensemble learning models was evaluated using ROC, K-S, and LIFT curves. AUC values were utilized to compare the differentiation ability of the prediction models with two traditional scoring systems, ASP III and SAPS II, ultimately selecting the best prediction model Additional evaluation of the prediction models' performance was conducted using ROCpr curves, calibration curves, accuracy, precision, and F1-score. SHapley's Additional exPlanation (SHAP) is a model-agnostic technique based on cooperative game theory. It is used to explain the predictions filtered through the best ensemble machine learning model. The model construction process was shown in **Figure 1A**.

The study will also analyze hospital mortality, hospital length of stay, and ICU length of stay as secondary outcomes. In particular, the COX regression analysis will focus on the relationship between important continuous variables and the 28-day risk of death. To analyze the relationship between important continuous data variables and 28-day mortality, we will use COX regression restricted cubic splines with 3 knots. This will be done after adjusting for age and comorbidities, based on the ranking of the most important variables in the prediction model. Both linear and non-linear relationships will be examined.

The analyses were conducted using R version 4.2.1. Our findings are fully reproducible, and the data is available online through the MIMIC-IV(2.0) database.

Result

Baseline characteristics

In this study, we extracted data from 69,639 first ICU admissions in MIMIC IV (2.0), and identified 9,158 patients who were diagnosed with sepsis and AKI, had no previous renal disease, and were aged ≥18 years based on nadir criteria. The patients were then divided into two groups: a mortality group (1,940 cases) and a survival group (7218 cases) based on whether they died within 28 days (Figure 1B). The 28-day mortality rate for S-AKI from the MIMIC IV (2.0) dataset was found to be 21.2%. The mortality and survival groups showed significant differences in most baseline variables, as indicated by Tables 1, 2. Patients who died had higher SOFA, APSIII, and SAPS II scores compared to those who survived, as shown in Table 1.

Data cleaning and features selection

In our study, we excluded variables with missing values greater than 20%. For the remaining variables, we used the random forest method to perform multiple interpolations on the missing values (Supplementary Figure 1). The interpolation density plot demonstrated that the five interpolated datasets closely matched the distribution of the original data set (Supplementary Figure 2). The complete data set, consisting of 60 independent variables, was obtained after selecting the best-interpolated data set.

Univariate regression analysis was performed for all variables, and those with a *p*-value greater than 0.05 were removed. The following variables were excluded: ROX, Platelets, Basophils, Lymphocytes, PLR, Peripheral vascular disease, Chronic pulmonary disease, Rheumatic disease, Peptic ulcer disease, AIDS, Dialysis, and Dialysis type. The study conducted a univariate regression analysis on all variables and presented the results using forest plots (Supplementary Figure 3). In the correlation study of continuous variables, those with a correlation greater than 0.75 were eliminated, leaving only the variable with the most significant impact on 28-day death. As a result, Base excess was eliminated (Supplementary Figure 4). The Boruta algorithm, which is based on random forest, was used to sort the importance of variables for further variable screening. This resulted in the identification of 38 variables that were deemed appropriate for model fitting (as shown in Figure 2). Prior to fitting the machine learning model, data distribution analysis was performed on all continuous variables, and box plots were obtained (as demonstrated in Supplementary Figure 5A). To address right-skewed distribution, we utilized the Box-Cox method to calculate the lambda value for each variable. For variables with a lambda value close to -0.5, such as BMI, Pao2, Lactate, Creatinine, BUN, and Anion gap, we performed a square root inverse conversion. We performed log conversion for the lambda values of Glucose, LMR, Pao2/Fio2 ratio, Urine output, WBC, Neutrophils, NLR, ROX-HR, and Monocytes, which were close to 0. For Bicarbonate, which had a lambda value close to 0.5, we performed square root conversion. The lambda value of Respiration rate, Paco2, and Potassium were close to -1, so we performed log conversion for these variables as well. The effect after the transfer was shown by a box plot (as demonstrated in Supplementary Figure 5B). The transformed data were integrated with other untransformed data to create a new dataset for building and internally validating the model.

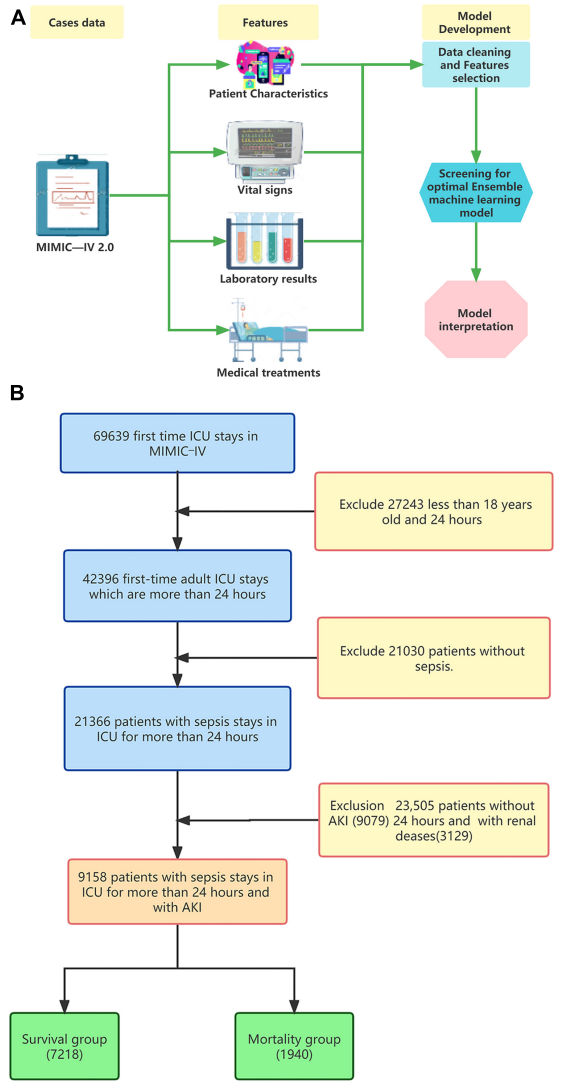

FIGURE 1
(A) Model development process and **(B)** flowchart of the study.

TABLE 1 Demographic and clinical characteristics of 28-day survival and mortality group.

	[ALL]	Survival group	Mortality group	*p*-Value
	N = 9,158	*N* = 7,218	*N* = 1,940	
Characteristics (median [IQR] and n (%))				
Age (years)	67.0 [57.0–78.0]	67.0 [56.0–77.0]	71.0 [59.0–82.0]	<0.001
Gender				<0.001
Female	3,916 (42.8%)	2,992 (41.5%)	924 (47.6%)	
Male	5,242 (57.2%)	4,226 (58.5%)	1,016 (52.4%)	
BMI[a]	29.2 [25.0–34.3]	29.7 [25.4–34.7]	27.6 [23.4–32.7]	<0.001
Urine output (ml)	1,280 [815–1,875]	1,390 [931–1,975]	862 [439–1,415]	<0.001
Severity score (median [IQR] and mean (SD))[c]				
APSIII	52.0 [37.0–75.0]	47.0 [34.0–66.0]	79.0 [59.0–101]	0.000
SOFA	3.00 [2.00–5.00]	3.00 [2.00–4.00]	4.00 [2.00–5.00]	<0.001
SAPSII	39.0 [31.0–50.0]	37.0 [29.0–46.0]	50.0 [40.0–61.0]	<0.001
Vital signs (median [IQR] and mean (SD))				
Respiratory rate (cpm[d])	28.0 [24.0–32.0]	27.0 [24.0–31.4]	30.0 [25.0–34.0]	<0.001
Heart rate (cpm[d])	105 [92.0–120]	103 [91.0–118]	113 [97.0–128]	<0.001
Systolic blood pressure (mmHg)	86.0 [78.0–94.0]	87.0 [79.0–95.0]	82.0 [73.0–92.0]	<0.001
Diastolic blood pressure (mmHg)	44.0 [38.0–50.0]	44.0 [39.0–50.0]	43.0 [36.0–49.0]	<0.001
Mean arterial pressure (mmHg)	57.0 [50.0–63.0]	57.0 [51.0–63.0]	54.0 [47.0–61.0]	<0.001
Temperature (°C)	37.4 [37.0–38.0]	37.4 [37.1–38.0]	37.3 [36.9–38.0]	<0.001
SpO2 (%)	93.0 [90.0–95.0]	93.0 [91.0–95.0]	92.0 [87.0–94.0]	<0.001
ROX index[b]	7.11 [4.89–9.89]	7.18 [5.04–9.83]	6.81 [4.26–10.1]	<0.001
ROX–HR index[b]	6.82 [4.48–9.95]	6.98 [4.71–10.0]	6.12 [3.66–9.55]	<0.001
Breathing assistance (median [IQR]), n (%)				
Ventilation				0.001
No vetilation	3,832 (41.8%)	3,087 (42.8%)	745 (38.4%)	
Ventilation	5,326 (58.2%)	4,131 (57.2%)	1,195 (61.6%)	
Dialysis				0.434
No	8,753 (95.6%)	6,892 (95.5%)	1,861 (95.9%)	
Yes	405 (4.42%)	326 (4.52%)	79 (4.07%)	
Dialysis type				0.465
No	8,753 (95.6%)	6,892 (95.5%)	1,861 (95.9%)	
CRRT[e]	248 (2.71%)	196 (2.72%)	52 (2.68%)	
IHD[e]	157 (1.71%)	130 (1.80%)	27 (1.39%)	
In-hospital mortality				0.000
Survival	7,467 (81.5%)	7,103 (98.4%)	364 (18.8%)	
Mortality	1,691 (18.5%)	115 (1.59%)	1,576 (81.2%)	
Recorded time of death (days)	24.0 [7.00–178]	202 [73.0–681]	7.00 [2.00–14.0]	0.000
Hospitalization time (days)	9.17 [5.55–16.5]	9.84 [6.05–17.8]	6.88 [3.07–12.6]	<0.001
ICU time (days)	3.43 [1.92–7.11]	3.24 [1.86–6.92]	4.16 [2.20–7.94]	<0.001
AKI stage[f]				<0.001
1	2873 (31.4%)	2382 (33.0%)	491 (25.3%)	
2	4765 (52.0%)	3789 (52.5%)	976 (50.3%)	
3	1520 (16.6%)	1047 (14.5%)	473 (24.4%)	

(Continued)

TABLE 1 (Continued)

	[ALL]	Survival group	Mortality group	*p*-Value
	N = 9,158	*N* = 7,218	*N* = 1,940	
Comorbidities				
Myocardial infarct				0.054
No	7,636 (83.4%)	6,047 (83.8%)	1,589 (81.9%)	
Yes	1,522 (16.6%)	1,171 (16.2%)	351 (18.1%)	
Congestive heart failure				0.011
No	6,792 (74.2%)	5,397 (74.8%)	1,395 (71.9%)	
Yes	2,366 (25.8%)	1,821 (25.2%)	545 (28.1%)	
Peripheral vascular disease				0.128
No	8,103 (88.5%)	6,367 (88.2%)	1,736 (89.5%)	
Yes	1,055 (11.5%)	851 (11.8%)	204 (10.5%)	
Cerebrovascular disease				<0.001
No	7,958 (86.9%)	6,370 (88.3%)	1,588 (81.9%)	
Yes	1,200 (13.1%)	848 (11.7%)	352 (18.1%)	
Dementia				<0.001
No	8,808 (96.2%)	6,994 (96.9%)	1,814 (93.5%)	
Yes	350 (3.82%)	224 (3.10%)	126 (6.49%)	
Chronic pulmonary disease				0.372
No	6,806 (74.3%)	5,380 (74.5%)	1,426 (73.5%)	
Yes	2,352 (25.7%)	1,838 (25.5%)	514 (26.5%)	
Rheumatic disease				1.000
No	8,841 (96.5%)	6,968 (96.5%)	1,873 (96.5%)	
Yes	317 (3.46%)	250 (3.46%)	67 (3.45%)	
Peptic ulcer disease				0.075
No	8,890 (97.1%)	7,019 (97.2%)	1,871 (96.4%)	
Yes	268 (2.93%)	199 (2.76%)	69 (3.56%)	
Mild liver disease				<0.001
No	7,610 (83.1%)	6,201 (85.9%)	1,409 (72.6%)	
Yes	1,548 (16.9%)	1,017 (14.1%)	531 (27.4%)	
Diabetes mellitus without complications				0.015
No	6,934 (75.7%)	5,424 (75.1%)	1,510 (77.8%)	
Yes	2,224 (24.3%)	1,794 (24.9%)	430 (22.2%)	
Diabetes mellitus with complications				0.023
No	8,737 (95.4%)	6,867 (95.1%)	1,870 (96.4%)	
Yes	421 (4.60%)	351 (4.86%)	70 (3.61%)	
Paraplegia				<0.001
No	8,755 (95.6%)	6,935 (96.1%)	1,820 (93.8%)	
Yes	403 (4.40%)	283 (3.92%)	120 (6.19%)	
Malignant cancer				<0.001
No	7,917 (86.4%)	6,385 (88.5%)	1,532 (79.0%)	
Yes	1,241 (13.6%)	833 (11.5%)	408 (21.0%)	
Severe liver disease				<0.001
No	8,314 (90.8%)	6,684 (92.6%)	1,630 (84.0%)	
Yes	844 (9.22%)	534 (7.40%)	310 (16.0%)	

(Continued)

TABLE 1 (Continued)

	[ALL]	Survival group	Mortality group	*p*-Value
	N = 9,158	*N* = 7,218	*N* = 1,940	
Metastatic solid tumor				<0.001
No	8,586 (93.8%)	6,933 (96.1%)	1,653 (85.2%)	
Yes	572 (6.25%)	285 (3.95%)	287 (14.8%)	
AIDS				0.487
No	9,117 (99.6%)	7,188 (99.6%)	1,929 (99.4%)	
Yes	41 (0.45%)	30 (0.42%)	11 (0.57%)	

Continuous variable data are presented as median (SD or interquartile ranges, IQR). Classified variable data are presented as n (%). Unless otherwise stated, the Mann–Whitney U test is used for the continuous variable, the χ^2 test, or the Fisher's exact test for the categorical variable.
[a]BMI, body mass index.
[b]ROX, ratio of SpO2/FIO2 to respiratory rate; ROX-HR, the ratio of ROX index over HR (beats/min), multiplied by a factor of 100.
[c]APSIII, Acute Physiology Score III; SAPSII, Simplified Acute Physiology Score II; SOFA, Sequential Organ Failure Assessment.
[d]cpm, counts per minute.
[e]CRRT: continuous renal replacement therapy; IHD: Intermittent Hemodialysis.
[f]AKI, acute kidney injury.

Development of 28-day mortality risk prediction model

Out of the total number of patients, 1,940 individuals passed away within 28 days, resulting in a mortality rate of 21.2% in the dataset. The balanced dataset was created using the SMOTE algorithm and then divided into a training and testing set with a ratio of 7:3. The AUC values for the four prediction models in the testing set were as follows: XGBoost model had an AUC value of 0.873 (with a range of 0.860-0.886), GBM model had an AUC value of 0.865 (with a range of 0.851-0.878), RF model had an AUC value of 0.849 (with a range of 0.834-0.863), and LR model had an AUC value of 0.850 (with a range of 0.836-0.864). The study found that all four machine learning models performed similarly and were more accurate than the traditional scoring systems ASPIII (0.713 95% CI 0.694-0.733) and SAPS II (0.681 95% CI 0.661-0.701). The ROC curve analysis demonstrated that the ensemble machine learning algorithm was significantly better than outperforms the traditional scoring system in predicting the 28-day mortality risk (as shown in Figure 3A). The K-S curves depicted in Figures 3B–E indicate that XGBoost exhibits a slightly superior differentiation ability compared to the other prediction models. Additionally, the LIFT curve (Figure 4) demonstrates that XGBoost outperforms the other models in the 40-50% position of the testing set. This could be attributed to XGBoost's algorithm, which has demonstrated exceptional learning performance in tabular data, and its robustness to noise, which is attributed to its regularization technique. The ensemble machine learning algorithm, XGBoost, was selected to build the 28-day mortality risk prediction model for S-AKI.

XGBoost model optimization and visualization

The XGBoost model was optimized and evaluated using the "xgboost" package. The area under the precision-recall curve (AUCpr) was found to be 0.873, which was similar to the area under the ROC curve (Figure 5A). This suggests that the model has comparable predictive ability for both death and survival. The model's accuracy, precision, recall, and F1-score were 0.773, 0.724, 0.896, and 0.801, respectively. The results indicate that the XGBoost model performed well in predicting mortality and survival groups. Additionally, the Recall metric outperformed Accuracy, which minimizes the possibility of under diagnosing mortality cases. The calibration curve analysis demonstrated that the model was accurately calibrated for predicting 28-day mortality risk, with no significant overestimation or underestimation (Figure 5B).

To determine the contribution of each variable to the XGBoost model, SHAP values were utilized. The importance of each feature was calculated using the Shapley value, which compared the model's prediction with and without the feature using the "shapviz" package (Figure 5C). The logarithm of urine output during the first 24 h of ICU admission was found to be the most important variable in predicting the 28-day mortality risk in patients with S-AKI. Among the important variables, pulse oxygen, temperature, age, and pH et al. are included. Cerebrovascular disease is one of the most significant comorbidities that affect the risk of death within 28 days. In Figures 5D–F, SHAP explanatory force plots were used to analyze three cases in the test group (#266, #1066, and #2066), Each variable's Shapley value is represented by an arrow that indicates an increase (red positive values) or decrease (yellow negative values) in the prediction. The force plots also show the main variables and their corresponding values. The variables that have a significant influence on the prediction vary from case to case.

Secondary outcomes

Our analysis of essential patient information revealed that the in-hospital mortality rate of S-AKI was 18.2%. Of these patients, 81.2% died within 28 days, with the primary time of death occurring within this timeframe. Additionally, 364 cases (18.8%) resulted in death within 28 days after discharge from the hospital. The death group had a shorter hospitalization duration compared to the survival group, by three days (6.88 [3.07–12.6] vs. 9.84 [6.05–17.8]). However, the death group had a slightly longer duration of ICU stay compared to (4.16 [2.20–7.94] vs. 3.24 [1.86–6.92]). It was observed

TABLE 2 Laboratory results of all patients within 24 h after admission to ICU.

	[ALL]	Survival group	Mortality group	*p*-Value
	N = 9158	*N* = 7218	*N* = 1940	
Arterial blood gas analysis (median [IQR] and mean (SD))				
pH	7.31 [7.24–7.36]	7.31 [7.26–7.36]	7.28 [7.17–7.36]	<0.001
PaO_2 (mmHg)	75.0 [46.0–103]	80.0 [52.0–108]	55.0 [39.0–84.0]	<0.001
$PaCO_2$ (mmHg)	47.0 [41.0–54.0]	47.0 [42.0–54.0]	47.0 [40.0–57.0]	0.384
PaO_2/FiO_2 ratio	168 [100–248]	175 [108–254]	134 [78.0–226]	<0.001
Base excess (mmol/L)	−3.00 [−7.00 to 0.00]	−3.00 [−6.00 to 0.00]	−5.00 [−10.00 to 0.00]	<0.001
Lactate (mmol/L)	2.40 [1.70–3.80]	2.40 [1.60–3.50]	3.10 [1.80–5.90]	<0.001
Anion gap (mmol/L)	16.0 [13.0–19.0]	15.0 [13.0–18.0]	18.0 [15.0–23.0]	<0.001
Bicarbonate (mmol/L)	24.0 [22.0–26.0]	24.0 [22.0–26.0]	23.0 [20.0–26.0]	<0.001
Complete blood cell count (median [IQR])				
White cell count (× 109/L)	14.8 [10.8–19.8]	14.6 [10.8–19.3]	15.8 [10.9–21.8]	<0.001
Neutrophil count (× 10^9/L)	8.87 [5.12–13.6]	8.57 [5.01–13.0]	10.3 [5.75–15.9]	<0.001
Eosinophils count (× 10^9/L)	0.06 [0.01–0.15]	0.07 [0.01–0.16]	0.03 [0.00–0.11]	<0.001
Lymphocyte count (× 10^9/L)	1.02 [0.58–1.65]	1.07 [0.62–1.71]	0.87 [0.47–1.42]	<0.001
Monocytes count (× 10^9/L)	0.53 [0.31–0.85]	0.52 [0.31–0.83]	0.59 [0.34–0.97]	<0.001
Platelets count (× 10^9/L)	155 [105–218]	155 [109–216]	154 [86.0–229]	0.018
NLR ratio[a]	7.88 [3.99–15.9]	7.33 [3.80–14.7]	10.5 [5.03–21.1]	<0.001
PLR ratio[a]	145 [77.8–281]	140 [76.5–274]	167 [81.2–318]	<0.001
LMR ratio[a]	1.90 [0.96–3.53]	2.00 [1.00–3.75]	1.44 [0.72–2.75]	<0.001
Hemoglobin (g/L)	9.80 [8.30–11.3]	9.80 [8.40–11.3]	9.60 [7.97–11.4]	<0.001
Blood chemistry results (median [IQR] and mean (SD))				
Blood glucose (mg/dl)	101 [86.0–124]	100 [86.0–121]	107 [85.0–134]	<0.001
Albumin (mg/dl)	3.20 [2.60–3.80]	3.30 [2.70–3.90]	3.00 [2.40–3.60]	<0.001
Blood urea nitrogen (mmol/L)	21.0 [16.0–32.0]	20.0 [15.0–29.0]	30.0 [20.0–46.0]	<0.001
Creatinine (mg/dl)	1.10 [0.80–1.60]	1.00 [0.80–1.40]	1.40 [1.00–2.20]	<0.001
Blood chemistry results (median [IQR])				
Calcium (mmol/L)	7.90 [7.40–8.40]	8.00 [7.40–8.40]	7.80 [7.10–8.40]	<0.001
Chloride (mmol/L)	103 [99.0–106]	104 [100–107]	101 [97.0–105]	<0.001
Sodium (mmol/L)	137 [134–140]	137 [135–139]	137 [133–140]	<0.001
Potassium (mmol/L)	4.50 [4.10–5.00]	4.50 [4.10–4.90]	4.60 [4.10–5.30]	<0.001

Continuous variable data are presented as median (SD or interquartile ranges, IQR). Classified variable data are presented as n (%). Unless otherwise stated, the Mann–Whitney U test is used for the continuous variable, the χ^2 test, or the Fisher's exact test for the categorical variable.
[a]NLR, neutrophil to lymphocyte ratio; PLR, platelet to lymphocyte ratio; LMR, lymphocyte to monocyte ratio.

that the severity of S-AKI condition was directly proportional to the length of ICU stay and increased the risk of early death.

The study found that the independent prediction performance of the top nine continuous variables in the XGBoost model for 28-day death risk was unclear. To detect non-linear or linear relationships between these variables and 28-day mortality, restricted cubic splines of COX regression were used. The model was adjusted for age (67 years) and comorbidities such as cerebrovascular disease, mild liver injury, and metastatic solid tumors. The results are presented in Figure 6. The study found that SpO2 and pH had a nearly linear relationship with a higher risk of death associated lower values. Additionally, variables such as 24-h urine volume (approximately 1500 ml), temperature (approximately 37.3°C), age (approximately 67 years), glucose (approximately 100 mg/dl), and sodium (approximately 136 mmol/L) showed a U-shaped change, with the risk of death being higher at the highest or lowest values relative to the bottom of the curve. The initial levels of BUN (around 37 mg/dl) and WBC (around 20 × 10^9/L) showed a steep increase, but later on, they remained relatively stable. Moreover, there was no significant rise in the mortality risk with the increase in these values.

Discussion

Acute kidney injury is a significant contributor to high mortality rates in sepsis patients. Early recognition and management are crucial in preventing the need for salvage

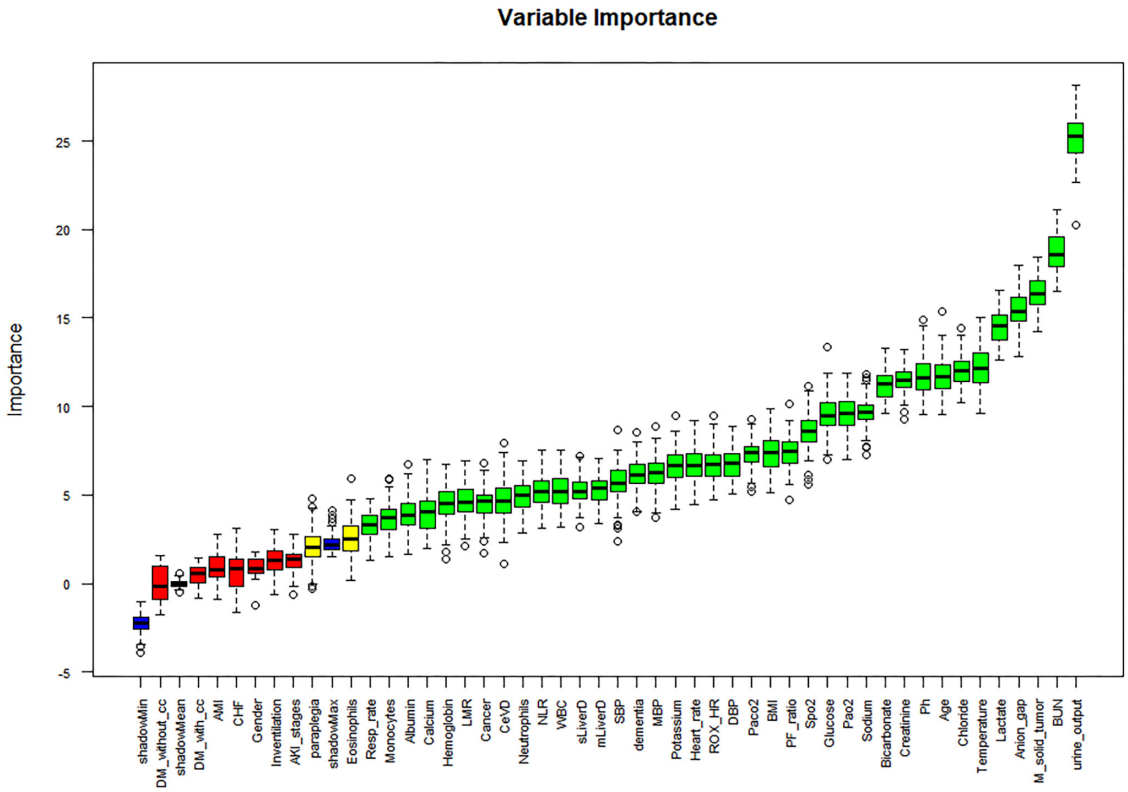

FIGURE 2
Boruta-based feature selection results. DM-without-cc, diabetes mellitus without complications; DM-with- cc, diabetes mellitus with complications; AMI, acute myocardial infarction; CHF, congestive heart failure; LMR, lymphocyte to monocyte ratio; CeVD, cerebrovascular disease; NLR, neutrophil to lymphocyte ratio; SBP, systolic blood pressure; DBP, diastolic blood pressure; MBI, body mass index; ROX_HR, the ratio of ROX index over HR (beats/min), multiplied by a factor of 100; PF_ratio, PaO2/FiO2 ratio; M_solid_tumor, Metastatic solid tumor; BUN, blood urea nitrogen.

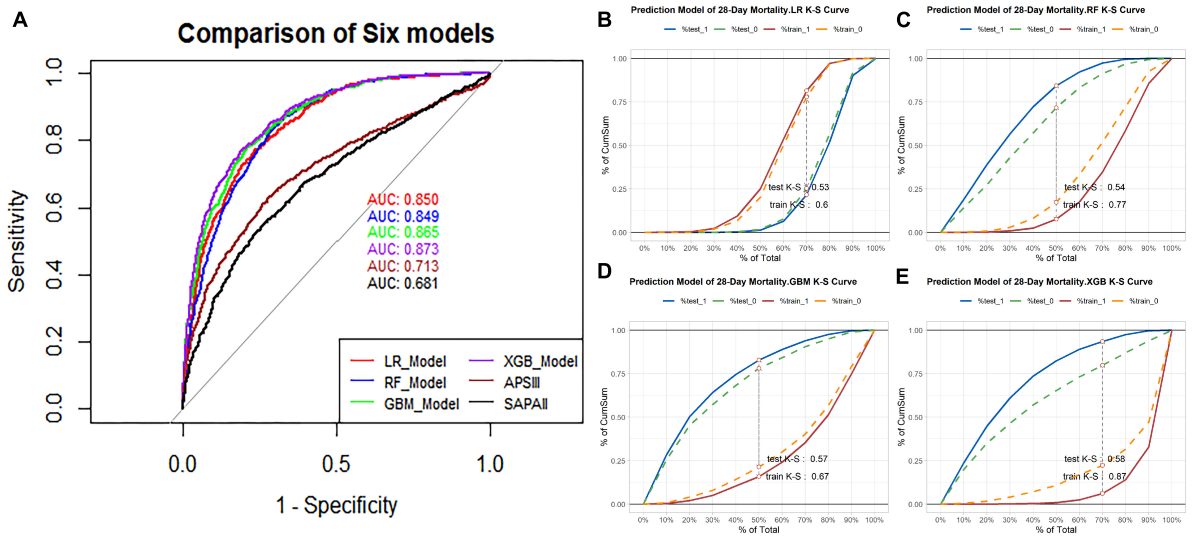

FIGURE 3
(A) Using receiver operating characteristic (ROC) curve and area under the receiver operating characteristic curve (AUC) to compare the discriminant ability of four models and traditional scoring. **(B)** K-S curve of 28-day mortality risk prediction model based on Logistic regression, test K-S 0.53 and train K-S 0.6. **(C)** K-S curve of 28-day mortality risk prediction model based on the Random Forest, test K-S 0.54 and train K-S 0.77. **(D)** K-S curve of 28-day mortality risk prediction model based on the GBM, test K-S 0.57 and train K-S 0.67. **(E)** K-S curve of 28-day mortality risk prediction model based on the XGBoost, test K-S 0.58 and train K-S 0.87.

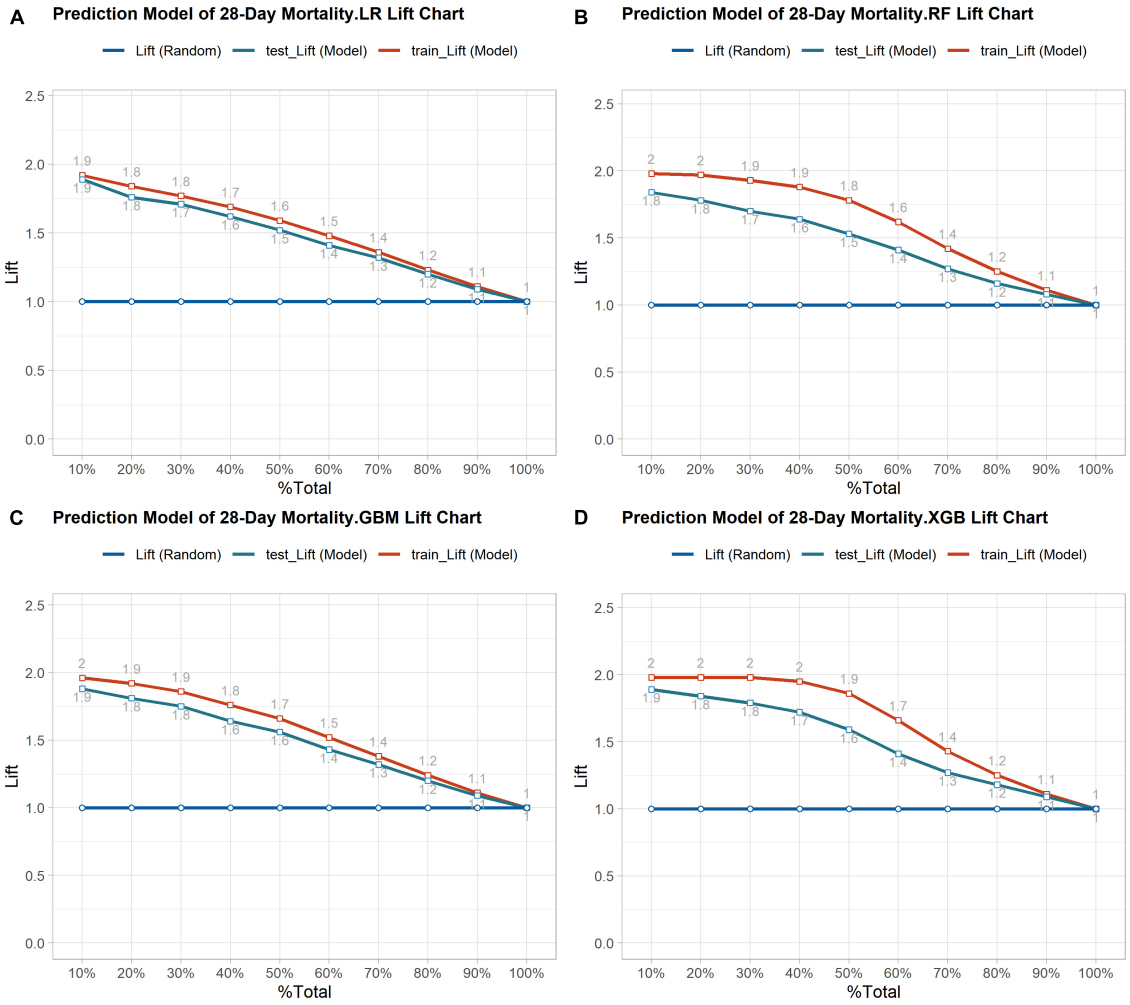

FIGURE 4
(A) Lift curve of 28-day mortality risk prediction model based on Logistic regression. **(B)** Lift curve of 28- day mortality risk prediction model based on the Random Forest. **(C)** Lift curve of 28-day mortality risk prediction model based on the GBM. **(D)** Lift curve of 28-day mortality risk prediction model based on the XGBoost.

treatment and reducing mortality. However, traditional scoring systems do not adequately meet clinical needs. This study proposes using machine learning to predict the 28-day risk of death from S-AKI in ICU inpatients, providing personalized predictions to guide clinical stratification and grading management. The risk of death in these patients has been a challenging aspect to predict in the past.

Clinical symptoms and laboratory tests are frequently employed in traditional scoring systems to predict critical patient outcomes. Two representative methods are the ASP III and SAPS II scores, both of which exhibit strong performance in predicting in-hospital patient mortality (17–19). Previous research has indicated that traditional scores were slightly less reliable in predicting hospitalization due to Acute Kidney Injury (AKI) or mortality within 60 days (6, 7). And it has not been used to predict death within 28 days. In recent years, there has been a growing interest in utilizing machine learning (ML) algorithms for diagnostic and prognostic disease studies. These ML models have shown to surpass traditional scoring methods in terms of predictive accuracy (15, 16). In our study, we also observed that machine learning models outperformed conventional scoring systems in all 28-day mortality prediction for S-AKI patients. The XGBoost algorithm-based 28-day mortality risk prediction model for S-AKI achieved better prediction performance with an AUPR value of 0.873 and good calibration performance. Our XGBoost model demonstrated a slightly better predictive performance compared to another study that utilized the same database (MIMIC-IV), study endpoint and ML algorithm. The area under the curve (AUC) was 0.850, while the other study achieved an AUC of 0.818 (14). Our model's superior performance of our model may be attributed to the inclusion of co-morbidities in our predictor variables. It is known that cases with co-morbidities have a higher mortality rate in patients with sepsis. Compared with the traditional scoring system, The use of machine learning prediction models can potentially enhance clinicians' decision-making and improve disease prognosis.

The most critical step in training machine learning models is data engineering, particularly data preprocessing. This process plays a vital role in preventing the risk of overfitting and classification boundary shifts, ultimately leading to improved

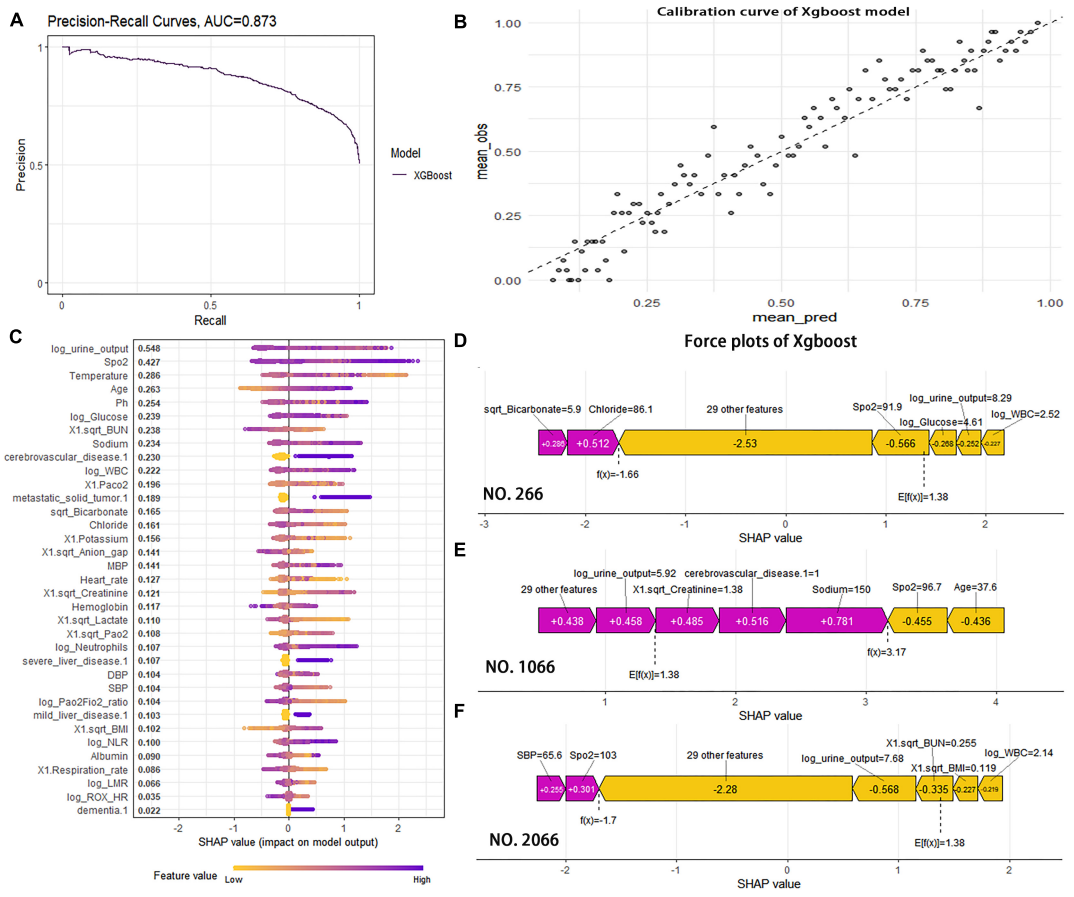

FIGURE 5
(A) The precision-recall (PR) curve was used further to evaluate the classification ability of the XGBoost model; AUCPR (0.873) indicated the model performed well in predicting case classification. **(B)** The calibration curve showed high coherence between the predicted and actual probability of XGBoost model. **(C)** The features are ranked according to the sum of the SHAP values for all patients, and the SHAP values are used to show the distribution of the effect of each feature on the XGBoost model outputs. **(D)** The sharp value force plot of case 266 was used to individually predict the characteristic variables. **(E)** The sharp value force plot of case 1066 was used to individually predict the characteristic variables. **(F)** The sharp value force plot of case 2066 was used to individually predict the characteristic variables.

predictive performance of the models. Despite the importance of data preprocessing, it is often overlooked, and most machine learning models still require thorough investigation in this area (14–16). The S-AKI model we created to predict 28-day mortality risk underwent thorough data processing. We utilized a combination of univariate regression, correlation analysis, and variable screening with Boruta of the random forest algorithm. The Boruta algorithm is a powerful and robust variable screening method that is sensitive to detecting causal variables while minimizing the number of false positives, making it suitable for both high-dimensional and low-dimensional datasets (20). When working with unbalanced categorical datasets, machine learning algorithms may not be reliable and their predictions may be biased, leading to misleading accuracy. To address this issue, we apply the SMOTE algorithm to discard the practice of randomly oversampling replicate samples, which can prevent the problem of random oversampling prone to overfitting. Studies have shown that this approach can improve classifier performance (21, 22). The synthetic data algorithm addresses the issue of data imbalance by avoiding information loss in both undersampling and oversampling methods.

Structured data dominates medical databases, and XGBoost has emerged as a top-performing integrated machine learning algorithm for prediction and classification based on this data (15, 16, 23). Hou, et al. (24) utilized MIMIC III (V1.4) sepsis patient data to develop an algorithm based on XGBoost for predicting 30-day mortality in septic patients. Their algorithm outperformed the logistic regression model and SAPS-II score prediction model with an AUC of 0.857 compared to 0.819 and 0.797, respectively. Additionally, the XGBoost algorithm demonstrated superior accuracy for sepsis diagnosis compared to the SOFA score with an AUC of 0.89 versus 0.596 (25). Liu, J and colleagues (26) utilized eICU data to develop a mortality prediction model for ICU AKI patients. Their study found that the XGBoost model outperformed LR, SVM, and RF machine learning algorithms. Previous research has demonstrated the efficacy of XGBoost as an ensemble machine learning algorithm in disease diagnosis and prognosis studies, particularly structured data. In this study, the performance of RF based on Bagging ensemble machine learning algorithm and XGBoost and GBM based on Boosting method were compared to traditional logistic regression in predicting 28-day mortality in S-AKI. The results indicated that

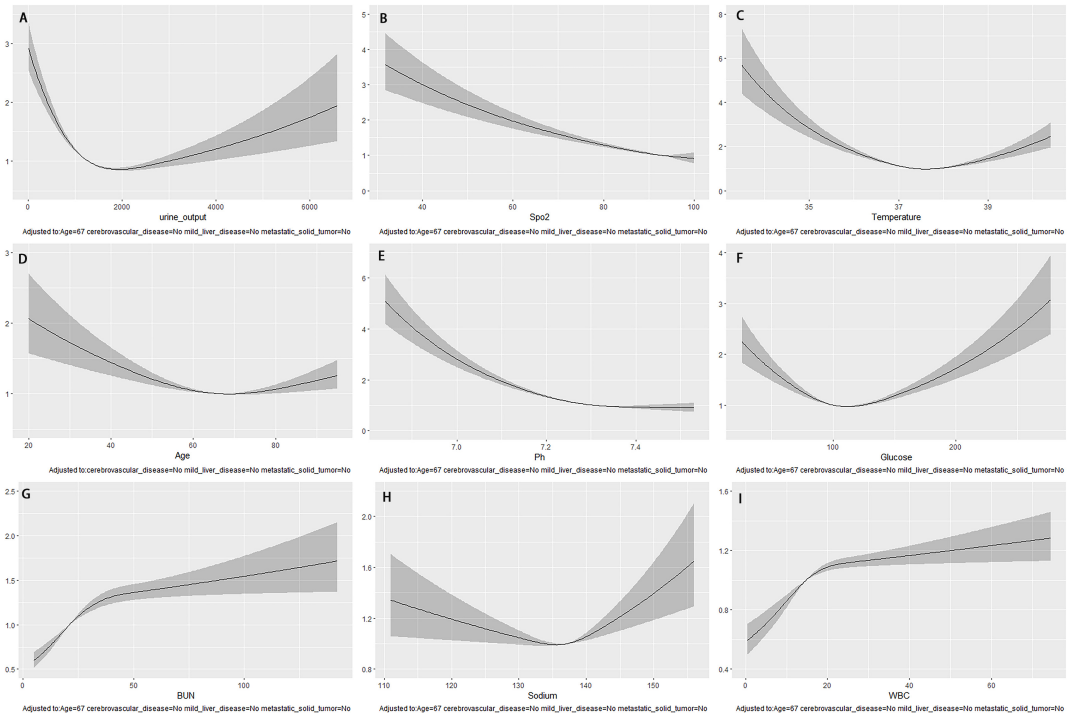

FIGURE 6
After adjusting for age and underlying disease, the COX regression-restricted cubic spline examined the nonlinear relationship between nine continuous variables and 28-day mortality risk. **(A)** Urine output within the first 24 h; **(B)** Spo2; **(C)** temperature; **(D)** age; **(E)** Ph; **(F)** glucose; **(G)** BUN; **(H)** sodium; **(I)** WBC.

the ensemble learning algorithms outperformed logistic regression. Among the ensemble algorithms, XGBoost demonstrated the best performance, as evidenced by the ROC, K-S, and LIFT curves.

The prediction model for 28-day mortality risk characteristics was ranked using SHAP values, with the logarithm of the 1st 24-h urine volume being identified as the most important variable. According to a study conducted on 183 intensive care units in Australia and New Zealand (27), a urine output threshold of less than 0.5 ml/kg/h within 24 h of ICU admission was found to be predictive of mortality in intensive care unit patients. Furthermore, the study trained an XGBoost machine learning model to predict in-hospital mortality, and discovered that low urine output was strongly associated with mortality in patients with sepsis. In patients with S-AKI receiving continuous renal replacement therapy (CRRT), urine output within the first 24 h of CRRT initiation was found to be a significant predictor of death (HR 2.6 95% CI 1.6–4.3 $p < 0.001$)among the various clinical variables related to mortality (28). Our study revealed that the logarithmic value of urine volume within the first 24 h is closely linked with the highest weight in the 28-day mortality risk model. Additionally, utilizing COX regression-restricted cubic splines and adjusting for age and underlying disease, we discovered a non-linear relationship between 24-h urine volume and 28-day mortality risk. The inflection point was observed at a 24-h urine volume of approximately 1,800 ml. Below this threshold, the risk of death decreased as urine volume increased, while above it, the risk of death increased with increasing urine volume.

Previous research has established that SpO2 is a risk factor for sepsis-related death (29). Similarly, our study discovered that SpO2 was linked to a higher likelihood of 28-day mortality in S-AKI cases. Using COX regression-restricted cubic splines study, we observed a near-linear negative correlation between SpO2, pH, and the risk of 28-day mortality. The relationship between temperature, age, glucose, BUN, sodium, and WBC and 28-day mortality risk was found to be non-linear. Specifically, body temperature, age, blood glucose, and sodium ions showed U-shaped changes, while BUN and WBC exhibited a post-phase plateau.

The variables that determine death risk differ between cases due to their non-linear relationship. In our study, SHAP force plots provide a direct graphical illustration for ensemble learning visualization interpretation. The color yellow represents a negative association with 28-day mortality risk, while red represents a positive association. The ability of machine learning predictions to show individualization is further illustrated by the fact that the variables that play a significant role in three different cases are not perfectly correlated. In some cases, the same variable may have opposite effects, such as the logarithmic value of 24-h urine volume, which is negatively correlated in #266 and #2066 and positively correlated in #1066. This may be due to a U-shaped relationship between urine volume and the risk of 28-day death.

Limitations

While this study provides valuable insights, it is important to acknowledge its limitations. It is a single-center retrospective data modeling study that relies solely on the MIMIC-IV (2.0) database and lacks external validation. Future studies will incorporate a

multicenter dataset and prospective study data to optimize and externally validate the model. Second, it is important to consider that there may be other factors that can affect the 28-day mortality risk in patients with S-AKI that were not measured or extracted, such as imaging data and treatment strategy. To improve the accuracy of predictive models, it may be beneficial to incorporate different types of data and use multimodal algorithms. Third, the data engineering process involves several steps, including data interpolation, feature selection, variable transformation, and data imbalance processing. However, these steps can sometimes lead to model overfitting and misrepresentation of important features. In our next study, we will focus on ensuring the completeness of the data set. Additionally, different types of variables are sequentially incorporated into the construction of the model to observe the effects of different variables on the prediction performance of the model. Finally, we utilize two integration algorithms, bagging and Boosting, and may introduce stacking integration algorithms in the future.

Conclusion

In this study, we have showcased the effectiveness of ensemble machine learning algorithms in predicting the risk of mortality within 28 days of patients with S-AKI. The SHAP approach has been used to enhance the interpretability of these models, thereby enabling clinicians to gain a better understanding of the underlying reasons behind the results. This knowledge will aid clinicians in making informed clinical decisions with regard to the stratification and management of S-AKI patients.

Data availability statement

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

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

JY, HP, and LX conceived the study and designed the trial. All authors involved in data collection, data management, data quality control, and data statistical analysis, participated in the revision of the manuscript, and reviewed and approved the final manuscript.

Funding

JY was supported by Loudi Science and Technology Innovation Project [Grant Loukefa (2022) No. 32]. HP was supported by Hunan Clinical Medical Technology Innovation Guidance Project (Grant 2021SK51607).

Acknowledgments

We thank the ICU nursing team for their contribution.

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.2023.1165129/full#supplementary-material

References

1. Evans L, Rhodes A, Alhazzani W, Antonelli M, Coopersmith C, French C, et al. Surviving sepsis campaign: international guidelines for management of sepsis and septic shock 2021. *Intensive Care Med.* (2021) 47:1181–247. doi: 10.1007/s00134-021-06506-y

2. Singer M, Deutschman C, Seymour C, 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

3. Peerapornratana S, Manrique-Caballero C, Gómez H, Kellum J. Acute kidney injury from sepsis: current concepts, epidemiology, pathophysiology, prevention and treatment. *Kidney Int.* (2019) 96:1083–99. doi: 10.1016/j.kint.2019.05.026

4. Manrique-Caballero C, Del R, Gomez H. Sepsis-associated acute kidney injury. *Crit Care Clin.* (2021) 37:279–301. doi: 10.1016/j.ccc.2020.11.010

5. Wang Z, Weng J, Yang J, Zhou X, Xu Z, Hou R, et al. Acute kidney injury-attributable mortality in critically ill patients with sepsis. *Peerj.* (2022) 10:e13184. doi: 10.7717/peerj.13184

6. Demirjian S, Chertow G, Zhang J, O'Connor T, Vitale J, Paganini E, 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

7. Gong Y, Ding F, Zhang F, Gu Y. Investigate predictive capacity of in-hospital mortality of four severity score systems on critically ill patients with acute kidney injury. *J Investig Med.* (2019) 67:1103–9. doi: 10.1136/jim-2019-001003

8. Shum H, Kong H, Chan K, Yan W, Chan T. Septic acute kidney injury in critically ill patients - a single-center study on its incidence, clinical characteristics, and outcome predictors. *Ren Fail.* (2016) 38:706–16. doi: 10.3109/0886022X.2016.1157749

9. Xin Q, Xie T, Chen R, Wang H, Zhang X, Wang S, et al. Construction and validation of an early warning model for predicting the acute kidney injury in elderly patients with sepsis. *Aging Clin Exp Res.* (2022) 34:2993–3004. doi: 10.1007/s40520-022-02236-3

10. Järvisalo M, Kartiosuo N, Hellman T, Uusalo P. Predicting mortality in critically ill patients requiring renal replacement therapy for acute kidney injury in a retrospective single-center study of two cohorts. *Sci Rep.* (2022) 12:10177. doi: 10.1038/s41598-022-14497-z

11. Wu W, Li Y, Feng A, Li L, Huang T, Xu A, 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

12. 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

13. Chang H, Chiang J, Wang C, Chiu P, Abdel-Kader K, Chen H, et al. Predicting mortality using machine learning algorithms in patients who require renal replacement therapy in the critical care unit. *J Clin Med.* (2022) 11:5289. doi: 10.3390/jcm11185289

14. Luo X, Yan P, Duan S, Kang Y, Deng Y, Liu Q, et al. Development and validation of machine learning models for real-time mortality prediction in critically ill patients with sepsis-associated acute kidney injury. *Front Med.* (2022) 9:853102. doi: 10.3389/fmed.2022.853102

15. Yue S, Li S, Huang X, Liu J, Hou X, Zhao Y, et al. Machine learning for the prediction of acute kidney injury in patients with sepsis. *J Transl Med.* (2022) 20:215. doi: 10.1186/s12967-022-03364-0

16. Zhou Y, Feng J, Mei S, Zhong H, Tang R, Xing S, et al. Machine learning models for predicting acute kidney injury in patients with sepsis associated acute respiratory distress syndrome. *Shock.* (2023) 59:352–9. doi: 10.1097/SHK.0000000000002065

17. Nassar A, Malbouisson L, Moreno R. Evaluation of simplified acute physiology score 3 performance: a systematic review of external validation studies. *Crit Care.* (2014) 18:R117. doi: 10.1186/cc13911

18. Ohno-Machado L, Resnic F, Matheny M. Prognosis in critical care. *Annu Rev Biomed Eng.* (2006) 8:567–99. doi: 10.1146/annurev.bioeng.8.061505.095842

19. Tong-Minh K, Welten I, Endeman H, Hagenaars T, Ramakers C, Gommers D, et al. Predicting mortality in adult patients with sepsis in the emergency department by using combinations of biomarkers and clinical scoring systems: a systematic review. *BMC Emerg Med.* (2021) 21:70. doi: 10.1186/s12873-021-00461-z

20. Degenhardt F, Seifert S, Szymczak S. Evaluation of variable selection methods for random forests and omics data sets. *Brief Bioinform.* (2019) 20:492–503. doi: 10.1093/bib/bbx124

21. Varotto G, Susi G, Tassi L, Gozzo F, Franceschetti S, Panzica F. Comparison of resampling techniques for imbalanced datasets in machine learning: application to epileptogenic zone localization from interictal intracranial eeg recordings in patients with focal epilepsy. *Front Neuroinform.* (2021) 15:715421. doi: 10.3389/fninf.2021.715421

22. Shi X, Qu T, Van Pottelbergh G, van den Akker M, De Moor B. A resampling method to improve the prognostic model of end-stage kidney disease: a better strategy for imbalanced data. *Front Med.* (2022) 9:730748. doi: 10.3389/fmed.2022.730748

23. Liu S, Fu B, Wang W, Liu M, Sun X. Dynamic sepsis prediction for intensive care unit patients using xgboost-based model with novel time-dependent features. *IEEE J Biomed Health Inform.* (2022) 26:4258–69. doi: 10.1109/JBHI.2022.3171673

24. 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

25. Yuan K, Tsai L, Lee K, Cheng Y, Hsu S, Lo Y, et al. The development an artificial intelligence algorithm for early sepsis diagnosis in the intensive care unit. *Int J Med Inform.* (2020) 141:104176. doi: 10.1016/j.ijmedinf.2020.104176

26. Liu J, Wu J, Liu S, Li M, Hu K, Li K. Predicting mortality of patients with acute kidney injury in the icu using xgboost model. *PLoS One.* (2021) 16:e246306. doi: 10.1371/journal.pone.0246306

27. Heffernan A, Judge S, Petrie S, Godahewa R, Bergmeir C, Pilcher D, et al. Association between urine output and mortality in critically ill patients: a machine learning approach. *Crit Care Med.* (2022) 50:e263–71. doi: 10.1097/CCM.0000000000005310

28. Pérez-Fernández X, Sabater-Riera J, Sileanu F, Vázquez-Reverón J, Ballús-Noguera J, Cárdenas-Campos P, et al. Clinical variables associated with poor outcome from sepsis-associated acute kidney injury and the relationship with timing of initiation of renal replacement therapy. *J Crit Care.* (2017) 40:154–60. doi: 10.1016/j.jcrc.2017.03.022

29. Bakhtawar S, Sheikh S, Qureshi R, Hoodbhoy Z, Payne B, Azam I, et al. Risk factors for postpartum sepsis: a nested case-control study. *BMC Pregnancy Childbirth.* (2020) 20:297. doi: 10.1186/s12884-020-02991-z

frontiers | Frontiers in Medicine

TYPE Original Research
PUBLISHED 25 July 2023
DOI 10.3389/fmed.2023.1221711

OPEN ACCESS

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

REVIEWED BY
Chang Hu,
Zhongnan Hospital of Wuhan University, China
Guo-wei Tu,
Fudan University, China

*CORRESPONDENCE
Yuan Gao
✉ rj_gaoyuan@163.com
Zhengyu He
✉ zhengyuheshsmu@163.com

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

RECEIVED 12 May 2023
ACCEPTED 06 July 2023
PUBLISHED 25 July 2023

CITATION
Zhou Y, Feng J, Mei S, Tang R, Xing S, Qin S, Zhang Z, Xu Q, Gao Y and He Z (2023) A deep learning model for predicting COVID-19 ARDS in critically ill patients. *Front. Med.* 10:1221711. doi: 10.3389/fmed.2023.1221711

A deep learning model for predicting COVID-19 ARDS in critically ill patients

Yang Zhou†, Jinhua Feng†, Shuya Mei, Ri Tang, Shunpeng Xing, Shaojie Qin, Zhiyun Zhang, Qiaoyi Xu, Yuan Gao* and Zhengyu He*

Department of Critical Care Medicine, Renji Hospital, School of Medicine, Shanghai Jiaotong University, Shanghai, China

Background: The coronavirus disease 2019 (COVID-19) is an acute infectious pneumonia caused by a severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) infection previously unknown to humans. However, predictive studies of acute respiratory distress syndrome (ARDS) in patients with COVID-19 are limited. In this study, we attempted to establish predictive models to predict ARDS caused by COVID-19 via a thorough analysis of patients' clinical data and CT images.

Method: The data of included patients were retrospectively collected from the intensive care unit in our hospital from April 2022 to June 2022. The primary outcome was the development of ARDS after ICU admission. We first established two individual predictive models based on extreme gradient boosting (XGBoost) and convolutional neural network (CNN), respectively; then, an integrated model was developed by combining the two individual models. The performance of all the predictive models was evaluated using the area under receiver operating characteristic curve (AUC), confusion matrix, and calibration plot.

Results: A total of 103 critically ill COVID-19 patients were included in this research, of which 23 patients (22.3%) developed ARDS after admission; five predictive variables were selected and further used to establish the machine learning models, and the XGBoost model yielded the most accurate predictions with the highest AUC (0.94, 95% CI: 0.91–0.96). The AUC of the CT-based convolutional neural network predictive model and the integrated model was 0.96 (95% CI: 0.93-0.98) and 0.97 (95% CI: 0.95–0.99), respectively.

Conclusion: An integrated deep learning model could be used to predict COVID-19 ARDS in critically ill patients.

KEYWORDS

COVID-19, ARDS, deep learning, artificial intelligence, computated tomography

Introduction

The coronavirus disease 2019 (COVID-19) is an acute infectious pneumonia caused by a severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) infection (1). Evidence has shown that 33% of COVID-19 patients are at high risk of progressing into severe cases, which are accompanied by increasing mortality and morbidity (2, 3). Moreover, severe SARS-CoV-2 infection may directly lead to acute respiratory distress syndrome (ARDS), and the manifestations could be viewed as a combination of pneumonia and ARDS (4).

Although significant advances have been made in understanding and managing ARDS, the morbidity and mortality of patients diagnosed with ARDS still remain high (5).

Unfortunately, the benefits of different therapies for established ARDS are limited (6–8). Since then, the paradigm for the management of ARDS has been shifted from treatment to prevention. Identification of patients at high risk of ARDS is important for clinicians to implement effective, preventive therapies to reduce the burden of ARDS. It is reported that the median time from the onset of COVID-19 symptoms to intubation is 8.5 days when COVID-19 ARDS occurs (9). There have been several studies focusing on the early prediction of ARDS, which described well-known risk factors associated with ARDS (10–12). However, COVID-19 ARDS is a serious complication of COVID-19, which has different clinical features from pre-COVID-19 ARDS (13). Hence, a clinical tool tailored for predicting COVID-19 ARDS is urgently needed.

In recent years, artificial intelligence (AI) has emerged as a promising tool in the medical field. The remarkable advantage of artificial intelligence in handling massive data could help with disease diagnostics and prognostics, radiographic recognition, and personalized treatment, etc. (14). During the COVID-19 pandemic, first-hand CT and clinical datasets helped clinicians make decisions and better understand the viral infection. For example, elevated levels of inflammatory cytokines and a reduction of T-cell subsets are closely related to COVID-19 pneumonia (15). The radiology features of COVID-19 pneumonia include a peripheral distribution of opacification, frosted glass opacities, and vascular thickening and enlargement (16). In spite of the distinct features observed in COVID-19 patients, the clinician may find it hard to figure out the underlying correlations between the clinical features and the features of CT slices, hindering the comprehensive understanding of the disease. Here, we aimed to provide a method pooling all the patients' features including CT and clinical features for improving the precision of the prediction of COVID-19 ARDS.

Methods

This is a retrospective study approved by the institutional Ethics Committees at Shanghai Renji Hospital, and informed patient consent was waived.

Study patients

All patients admitted to the intensive care unit in Shanghai Renji Hospital between April 2022 and June 2022 were screened for eligibility. Inclusion criteria were as follows: (1) patients who were 18 years old and above; and (2) patients who met the diagnosis of COVID-19 ARDS. Exclusion criteria were as follows: (1) patients who were diagnosed with ARDS within the first day of admission; (2) missing clinical data were more than 20%; and (3) without any CT scan results.

Diagnosis of COVID-19 ARDS

SARS-CoV-2 infection can be identified by the detection of viral RNA in nasopharyngeal secretions via PCR test. The diagnosis of COVID-19 was confirmed by the patients' clinical history, epidemiological contact, and a positive SARS-CoV-2 test.

The diagnosis of ARDS followed the Berlin definition: (1) requirement of mechanical ventilation and positive end-expiratory pressure or continuous positive airway pressure $\geq$ 5 cmH_2O; (2) a certain degree of hypoxemia: severe ($PaO_2/FiO_2 \leq 100$ mmHg), moderate (PaO_2/FiO_2 between 100 mmHg and 200 mmHg), or mild (PaO_2/FiO_2 between 200 mmHg and 300 mmHg); and (3) without evidence of pleural effusion, lung collapse, lung nodules, or cardiogenic pulmonary edema from the chest radiography (16). A patient who satisfied the criteria of COVID-19 and ARDS was diagnosed with COVID-19 ARDS.

Data collection

We collected the first sets of chest CT images and clinical data after the patients' admission to the intensive care unit. The clinical data included demographic information, comorbidity conditions, respiratory support methods, onset symptoms, vital signs at admission, aeration variables, routine blood tests, inflammation tests, biochemical tests, blood coagulation tests, lymphocyte subset tests, and cytokine profile tests. Original CT images both in JPG and DICOM format of the included patients were collected. In this study, we randomly divided the patients into training and validation cohorts in a ratio of 7:3.

Statistical analysis

The categorical variables were presented as counts and corresponding proportions and were further compared using the chi-square test or Fisher's exact test. The continuous variables were reported as the median and the interquartile range; the Mann–Whitney U-test was applied to compare the differences between the groups. The multivariate logistic regression was performed to figure out the independent risk factors associated with COVID-19 ARDS. A nomogram plot was further established based on the result of the multivariate logistic regression. A two-tailed *P*-value of <0.05 was considered significant. The data analysis in this study was completed via Python version 3.8 and R version 4.0.5.

The COVID-19 ARDS prediction based on clinical features

Four different machine learning algorithms were implemented to establish the predictive models for COVID-19 ARDS, including logistic regression (LR), support vector machine (SVM), random forest (RF), and extreme gradient boosting (XGBoost). The training cohort was divided into five partitions, of which four-fifths were used to train the models, and the remaining part was used to validate the models. The hyperparameters of all the models were fine-tuned for the highest area under the receiver operating to avoid the problem of overfitting. We followed two specific rules when searching for the best hyperparameters, which were as follows: (1) the training loss was the lowest after the test of all combinations

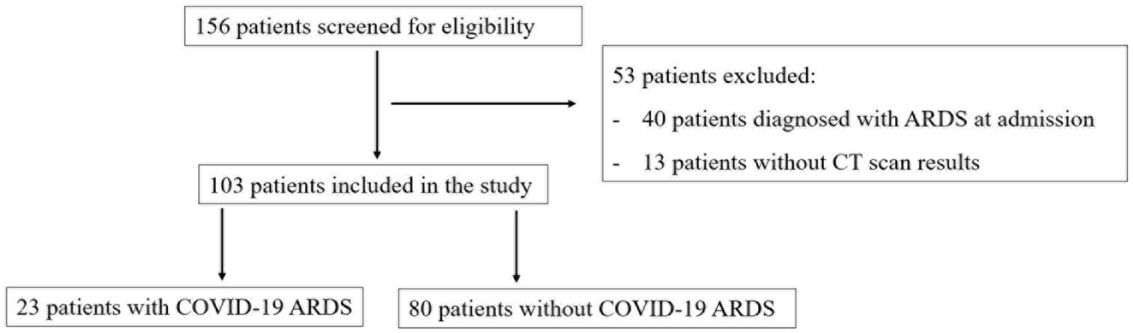

FIGURE 1
The flowchart of patients' selection.

of hyperparameters; (2) the log loss in the validation cohort was less than –log 0.5 and higher than the training cohort. Grid search with 5-fold cross-validation was applied to search for the most appropriate hyperparameters in the training cohort. Finally, the predictive performance of established models was compared in the validation cohort.

The labeling of individual CT slices

We first manually labeled 897 slices of 30 patients to train the classification model for individual CT slices. The CT slices were classified into two types: (1) normal CT, in which the image features in lungs were consistent with healthy lungs; (2) abnormal CT, in which image features were associated with COVID-19 pneumonia. Two senior ICU clinicians (ZHand YG) independently labeled individual CT slices. Any disagreements were resolved through discussion. The deep learning framework was based on the architecture of VGG-16, which consisted of 13 convolutional layers and 3 fully connected layers. We further internally validated the classification model and used it to label the remaining 2, 300 CT slices. Finally, every CT slice was classified into a normal CT image or an abnormal CT image.

The COVID-19 ARDS prediction based on CT images

After the auto-labeling of individual CT slices, we assumed that an abnormal CT slice classified by the model was a positive case. Then, the possibility of being an abnormal CT slice for every CT image was calculated. The 10 most probable abnormal CT slices of a single patient were viewed as the representative CT images and were input into the second VGG-16 network. This convolutional neural network (CNN) allows for the shift from the prediction of COVID-19 ARDS based on individual CT slices to the prediction based on a single patient. The VGG-16 network consists of 1 input layer, 13 convolutional layers, 3 fully connected layers, and 1 output layer. The convolutional layers were used to handle feature extraction and presentation. The pooling layers were used for filtering abundant information under the max-pooling strategy. In the last three output layers, the possibility of being a positive case was calculated for each CT slice. For the individual CT-based prediction, the possibility ranged from 0 to 1, representing a CT slice classified into a normal CT image or an abnormal CT slice. For the single patient-based prediction, the possibility ranged from 0 to 1, representing a patient being predicted to develop COVID-19 ARDS or not.

The integration of predictions models

The integration of two prediction models based on CT images and clinical data was achieved by the penalized logistic regression algorithm. The L2 regularization of the penalized logistic regression algorithm was used. To be specific, the machine learning model based on clinical features and the CNN model based on CT images individually generated two scores for the prediction of COVID-19 ARDS, which were taken as input features for the penalized logistic regression algorithm. At last, the penalized logistic regression algorithm calculated a prediction score for the COVID-19 ARDS outcome.

The evaluation of model performance

We randomly divided the patients into the training cohort and the validation cohort in a ratio of 7:3. The overall predictive performance of the integrated model was measured in the test cohort. The receiver operating characteristics (ROC) curve and the confusion matrices of all established predictive models were depicted to compare the performance of the predictive models. A ROC curve is a graphic plot used to illustrate a binary classifier's diagnostic ability as the discrimination threshold varies. It is created by plotting the true-positive rate against the false-positive rate at different discrimination thresholds. The calibration plots were also depicted to assess the predictive performance of all the models.

10.3389/fmed.2023.1221711

TABLE 1 Baseline characteristics of included patients.

Characteristics	Total ($n = 103$)	Non-ARDS cohort ($n = 80$)	ARDS cohort ($n = 23$)	*P*-value
Demographic variables				
Age (years)	75 (64, 87)	74 (63, 87)	84 (75, 89)	0.014
Gender, *n* (%)				0.917
Male	48 (46.6%)	38 (47.5%)	10 (43.5%)	
Female	55 (53.4%)	42 (52.5%)	13 (56.5%)	
BMI (kg/m2)	26.9 (22.7, 33.5)	26.5 (23.0, 33.9)	27.2 (24.1, 34.3)	0.346
Marital status, *n* (%)				0.269
Single	8 (7.8%)	6 (7.5%)	2 (8.7%)	
Married	93 (90.3%)	72 (90.0%)	21 (91.3%)	
Other	2 (1.9%)	2 (2.5%)	0 (0%)	
Comorbidities				
Congestive heart failure, *n* (%)	30 (29.1%)	23 (28.8%)	7 (30.4%)	1.000
Hypertension, *n* (%)	54 (52.4%)	41 (51.2%)	13 (56.5%)	0.834
Diabetes, *n* (%)	25 (24.3%)	18 (22.5%)	7 (30.4%)	0.613
Chronic kidney disease, *n* (%)	21 (20.4%)	15 (18.8%)	6 (26.1%)	0.634
Arrhythmia, *n* (%)	13 (12.6%)	11 (13.8%)	2 (8.7%)	0.774
Respiratory support				0.002
Spontaneous breathing, *n* (%)	38 (36.9%)	38 (47.5%)	0 (0%)	
Nasal cannula, *n* (%)	24 (23.3%)	17 (21.2%)	7 (30.4%)	
Mask ventilation, *n* (%)	7 (6.8%)	4 (5%)	3 (13%)	
High flow, *n* (%)	31 (30.1%)	19 (23.8%)	12 (52.2%)	
Non-invasive ventilator, *n* (%)	1 (0.97%)	1 (1.2%)	0 (0%)	
Intubation, *n* (%)	2 (1.94%)	1 (1.2%)	1 (4.3%)	
Onset symptoms				
Fever, *n* (%)	42 (40.8%)	37 (46.2%)	5 (21.7%)	0.062
Cough, *n* (%)	56 (54.4%)	43 (53.8%)	13 (56.5%)	1.000
Sore throat, *n* (%)	9 (8.7%)	9 (11.2%)	0 (0%)	0.206
Nausea, *n* (%)	2 (1.94%)	1 (1.2%)	1 (4.3%)	0.927
Headache, *n* (%)	7 (6.8%)	4 (5%)	3 (13%)	0.378
Chest distress, *n* (%)	2 (1.94%)	2 (2.5%)	0 (0%)	1.000
Vital signs at admission				
T (°C)	36.80 (36.50 to 37.30)	36.80 (36.50 to 37.50)	36.70 (36.55 to 37.05)	0.430
SBP (mmHg)	129.00 (111.00, 145.00)	126.09 ± 24.77	134.65 ± 26.13	0.152
DBP (mmHg)	69.00 (64.50 to 80.00)	72.11 ± 15.27	73.22 ± 13.43	0.754
HR (/min)	96.00 (81.00 to 113.50)	96.00 (80.00 to 111.00)	105.00 (85.00 to 125.00)	0.139
RR (/min)	20.00 (18.00 to 25.00)	20.00 (18.00 to 25.00)	21.00 (19.50 to 24.00)	0.293
Aeration variables				
PaO2 (mmHg)	89 (66, 118)	94.5 (76, 137.5)	94.5 (76, 137.5)	<0.001
PaCO2 (mmHg)	37.5 (32.5, 44.5)	37.2 (31.9, 44.5)	37.2 (31.9, 44.5)	0.358
SpO2 (%)	97 (95, 99)	98 (95, 99)	98 (95, 99)	0.022
PaO2/FiO2	192.0 (159.0, 252.0)	201.0 (171.0, 265.9)	201.0 (171.0, 265.9)	<0.001

(Continued)

10.3389/fmed.2023.1221711

TABLE 1 (Continued)

Characteristics	Total ($n = 103$)	Non-ARDS cohort ($n = 80$)	ARDS cohort ($n = 23$)	*P*-value
Routine blood test			38 (34.1, 45.3)	
White blood cell (K/UL)	8.59 (6.30, 13.21)	8.16 (6.03, 13.14)	96 (91, 98)	0.152
Neutrophil (K/UL)	7.30 (4.96, 12.25)	6.36 (4.41 to 11.38)	159.5 (140.0, 171.5)	0.024
Monocyte (K/UL)	0.47 (0.32, 0.74)	0.48 (0.32 to 0.74)		0.800
Reb blood cell (K/UL)	3.46 ± 0.90	3.47 ± 0.90	10.72 (7.59, 14.26)	0.988
Platelet (K/UL)	163.00 (108.50, 256.00)	158.50 (107.50, 256.00)	9.63 (7.22 to 14.98)	0.553
Hemoglobin (g/dL)	10.41 ± 2.92	10.34 ± 2.96	10.72 (7.59, 14.26)	0.835
Glucose (mg/l)	7.10 (5.70 to 10.15)	6.65 (5.55 to 9.70)	9.63 (7.22 to 14.98)	0.023
Inflammation				
C-reactive protein (mg/L)	46.49 (22.49, 77.53)	38.50 (19.29 to 75.03)	0.39 (0.30 to 0.77)	0.021
Procalcitonin (ng/mL)	0.32 (0.08, 0.81)	0.26 (0.07 to 0.77)	3.46 ± 1.02	0.212
Serum Amyloid A (mg/L)	155.01 (49.84, 350.00)	100.26 (27.21 to 350.00)	186.00	0.004
Biochemical test				
ALT (U/L)	22.00 (11.00, 43.50)	20.00 (10.50 to 45.50)	22.00 (14.50 to 37.50)	0.590
AST (U/L)	31.00 (21.50, 48.50)	29.50 (21.50 to 49.00)	37.00 (21.00 to 47.00)	0.791
LDH (U/L)	276.00 (220.00, 88.00)	263.50 (210.00 to 388.00)	291.00 (236.50 to 388.50)	0.289
Bilirubin (mg/dl)	13.70 (9.45, 19.60)	13.75 (9.65 to 19.90)	12.20 (9.05 to 17.95)	0.571
Urea (mmol/L)	8.78 (5.37, 15.85)	7.06 (4.96 to 13.77)	15.85 (9.00 to 24.14)	<0.001
Creatine (mg/l)	79.00 (49.50, 146.50)	71.50 (48.00 to 126.00)	92.00 (60.00 to 193.00)	0.183
eGFR (ml/min)	76.00 (36.50, 95.50)	77.50 (38.00 to 102.00)	67.00 (19.50 to 86.00)	0.128
PH	7.40 (7.35 to 7.45)	7.42 (7.37 to 7.45)	7.35 (7.30 to 7.42)	0.003
Sodium (mmol/L)	139.00 (135.00, 144.00)	139.00 (134.00, 142.00)	141.00 (137.00, 151.00)	0.060
Potassium (mmol/L)	3.60 (3.10, 4.00)	3.50 (3.10, 4.00)	3.90 (3.50, 4.15)	0.112
Chlorine (mmol/L)	105.00 (99.00, 112.00)	102.50 (98.00, 110.50)	110.00 (104.00, 119.50)	0.003
Calcium (mmol/L)	1.09 (1.06, 1.14)	1.09 (1.04, 1.13)	1.12 (1.08, 1.17)	0.030
Albumin (g/dl)	2.8 (2.2, 3.6)	2.8 (2.3, 3.6)	2.7 (2.0, 3.5)	0.418
TG (mmol/L)	1.42 (0.88, 1.81)	1.48 (0.90, 1.80)	1.37 (0.68, 1.86)	0.568
TC (mmol/L)	3.36 (2.82, 4.55)	3.47 (2.82, 4.22)	3.28 (2.70, 5.35)	0.994
HDL (mmol/L)	0.86 (0.62, 1.06)	0.87 (0.59, 1.07)	0.84 (0.64, 1.01)	0.862
LDL (mmol/L)	2.30 (1.64 to 2.58)	2.30 (1.64, 2.57)	2.27 (1.90, 3.27)	0.724
Non-HDL (mmol/L)	2.47 (1.95, 3.05)	2.47 (1.95, 3.00)	2.56 (2.13, 3.73)	0.360
BNP (pg/ml)	190.00 (88.00, 492.50)	187.00 (86.00, 562.00)	195.00 (126.00, 313.50)	0.871
TNI (ng/ml)	0.04 (0.01, 0.07)	0.03 (0.01, 0.07)	0.04 (0.02, 0.07)	0.397
Mb (μg/L)	92.40 (43.50, 247.25)	81.50 (39.15, 233.25)	149.20 (87.50, 350.45)	0.021
CKMB (ng/ml)	2.50 (1.50, 5.10)	2.40 (1.40, 3.50)	4.60 (2.30, 8.10)	0.008
Blood coagulation test				
TT (seconds)	15.20 (14.60 to 16.40)	15.15 (14.60 to 16.45)	15.30 (14.60 to 16.10)	0.994
APTT (seconds)	31.60 (27.30 to 36.15)	31.25 (27.50 to 36.30)	32.50 (27.05 to 35.90)	0.698
PT (seconds)	12.90 (11.90 to 14.80)	12.75 (11.70 to 14.75)	13.20 (12.35 to 15.45)	0.139
INR	1.10 (1.02 to 1.27)	1.10 (1.00 to 1.27)	1.13 (1.05 to 1.33)	0.212
FG (g/L)	3.98 (3.01 to 4.61)	3.76 (2.79 to 4.58)	4.54 (3.89 to 4.72)	0.005

(Continued)

TABLE 1 (Continued)

Characteristics	Total ($n = 103$)	Non-ARDS cohort ($n = 80$)	ARDS cohort ($n = 23$)	*P*-value
DD (mg/L)	1.56 (0.68 to 3.19)	1.36 (0.59 to 2.48)	2.42 (1.44 to 3.58)	0.017
FDP (mg/L)	11.90 (5.90 to 20.95)	11.15 (5.40 to 20.10)	18.10 (10.55 to 24.55)	0.040
Lymphocyte subsets				
Lymphocyte (10e9/L)	0.73 (0.50 to 1.01)	0.79 (0.54 to 1.26)	0.63 (0.40 to 0.76)	0.017
T lymphocyte (10e6/L)	424.30 (268.90, 679.40)	526.65 (291.05 to 858.10)	309.90 (220.00, 429.75)	<0.001
B lymphocyte (10e6/L)	82.00 (41.65, 156.60)	82.05 (38.50 to 152.35)	79.50 (44.35 to 172.65)	0.994
Th lymphocyte (10e6/L)	280.90 (154.40, 434.35)	294.20 (174.15 to 482.20)	199.40 (86.80 to 361.25)	0.025
Ts lymphocyte (10e6/L)	149.90 (88.30 to 244.50)	158.30 (97.35 to 244.50)	129.10 (48.35 to 236.15)	0.167
Natural killer cell (10e6/L)	115.60 (63.40 to 184.60)	126.00 (69.30 to 201.35)	87.60 (52.90 to 143.95)	0.033
CD4/CD8 ratio	1.60 (1.14 to 2.56)	1.58 (1.12 to 2.42)	1.60 (1.14 to 2.92)	0.669
Cytokine profiles				
IL1 (pg/ml)	1.22 (0.83 to 1.69)	1.22 (0.76 to 1.57)	1.37 (0.94 to 2.55)	0.224
IL2 (pg/ml)	1.03 (0.61 to 1.69)	1.03 (0.66 to 1.47)	1.06 (0.58 to 1.94)	0.571
IL4 (pg/ml)	1.45 (1.08 to 2.17)	1.35 (1.07 to 2.09)	1.67 (1.27 to 2.54)	0.132
IL5 (pg/ml)	0.79 (0.38 to 1.14)	0.76 (0.37 to 1.14)	0.97 (0.63 to 1.20)	0.226
IL6 (pg/ml)	46.91 (20.91 to 113.00)	37.58 (17.13 to 81.49)	118.00 (50.28 to 279.58)	<0.001
IL8 (pg/ml)	16.07 (6.26 to 53.18)	13.13 (5.93 to 51.21)	48.32 (14.11 to 91.98)	0.034
IL10 (pg/ml)	4.12 (2.28 to 6.26)	3.58 (2.28 to 6.14)	5.16 (2.49 to 10.01)	0.328
IL17A (pg/ml)	3.28 (1.31 to 4.58)	3.02 (1.27 to 4.42)	3.44 (1.35 to 5.58)	0.542
TNF (pg/ml)	1.98 (1.26 to 2.79)	1.90 (1.26 to 2.66)	2.48 (1.06 to 3.42)	0.169
IFN-α (pg/ml)	1.04 (0.66 to 2.06)	0.98 (0.65 to 1.69)	1.36 (0.95 to 2.60)	0.083
IFN-γ (pg/ml)	1.53 (1.11 to 1.89)	1.53 (1.11 to 1.94)	1.53 (1.14 to 1.79)	0.921

Results

Baseline clinical features of included patients

In total, 103 patients were enrolled in the study after the screening for eligibility, of whom 23 patients (22.3%) developed COVID-19 ARDS. The flowchart of the patients' selection is provided in Figure 1. The baseline clinical features of the included patients are presented in Table 1. There were no missing data in our study.

A summary of collected CT images

Original chest CT images containing fields of the lung parenchyma were obtained from 103 patients. The total number of included CT images was 3,187, of which 690 CT slices were from COVID-19 ARDS patients and 2,497 CT slices were from non-COVID-19 ARDS patients. We manually classified 897 CT slices from 30 patients into normal CT images or abnormal CT images.

TABLE 2 Multivariate logistic regression analysis of risk factors of COVID-19 ARDS based on selected variables in the training cohort.

Variable	Coefficient	OR (95% CI)	*P*-value
Age	0.089	1.093 (1.015, 1.177)	0.018
P/F ratio	−0.024	0.977 (0.963, 0.991)	0.001
CRP	0.017	1.017 (1.001, 1.033)	0.036
T lymphocyte	−0.004	0.996 (0.993, 0.999)	0.021
IL-6	0.008	1.008 (1.002, 1.017)	0.045

OR, odds ratio; CI, confidence interval.

The multivariate logistic regression analysis of clinical features

After the multivariate logistic regression analysis, five risk factors were figured out to be independently associated with COVID-19 ARDS. We concluded that age (OR, 1.093; 95% CI, 1.015–1.177), PaO2/FiO2 ratio (OR, 0.977; 95% CI, 0.963–0.991), C-reactive protein (OR, 1.017; 95% CI, 1.001–1.033), the count of total T lymphocytes (OR, 0.996; 95% CI, 0.993–0.999), and IL-6 (OR, 1.008; 95% CI, 1.002–1.017) were independent risk factors of COVID-19 ARDS. The detailed results of the multivariate logistic regression analysis are shown in Table 2. A nomogram

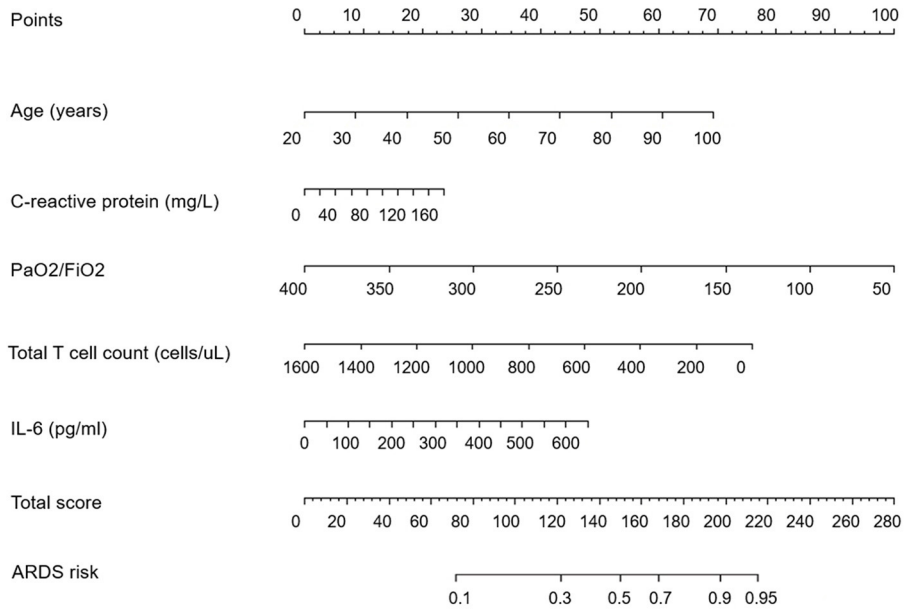

FIGURE 2
The nomogram plot for the prediction of COVID-19 ARDS.

plot was illustrated based on the result of the multivariate logistic regression model (Figure 2). We could calculate the risk score and the corresponding possibility of COVID-19 ARDS using the nomogram.

The predictive performance of models based on clinical features

We developed four machine learning models to predict COVID-19 ARDS, including logistic regression, support vector machine, random forest, and extreme gradient boosting. The ROC curves of all the machine learning models are shown in Figure 3A. The area under the ROC curve of the XGBoost model was 0.94, which outperformed the logistic regression model (AUC = 0.82), the support vector machine model (AUC = 0.77), and the random forest model (AUC = 0.92). We also performed the Delong test to compare the AUCs of the XGBoost model against the other three models (XGBoost model vs. logistic regression model, $P<0.001$; XGBoost model vs. support vector machine model, $P < 0.001$; and XGBoost vs. random forest model, $P = 0.002$). The calibration curves are provided in Figure 3B. The XGBoost model was finally chosen to be the best machine learning model to predict COVID-19 ARDS in our study.

The predictive performance of the CNN model based on CT images

In total, 897 manually labeled CT slices were used to train the classification CNN model based on individual CT images. Figure 4A shows the ROC curve of the classification CNN model (AUC = 0.99). The confusion matrix of the classification CNN model is shown in Figure 4B. The normal CT slices and the abnormal CT slices were correctly distinguished by the classification CNN model.

The predictive performance of the integrated deep learning model

The integrated deep learning model consisted of the XGBoost model based on the clinical features and the CNN model based on the selected CT slices from the individual patients. The ROC curves of the two individual models and the integrated deep learning model are shown in Figure 5A. The area under the ROC curve values of the XGBoost model, the CNN model, and the integrated model were 0.94 (95% CI: 0.91–0.96), 0.96 (95% CI: 0.93–0.98), and 0.97 (95% CI: 0.95–0.99), respectively.

The calibration curve plot indicated a good agreement between the predicted probabilities of COVID-19 ARDS calculated by the predictive models and the actual outcome (Figure 5B). The confusion matrices were plotted using clinical features, CT images, and integrated data to predict COVID-19 ARDS (Figure 6). We found that the integrated deep learning model could yield more accurate predictions than the individual model based on clinical features or CT images. More details about the predictive performance of the models are provided in Table 3.

Discussion

The outbreak of COVID-19 led to a global pandemic, and the main causes of the deaths were pulmonary complications such as acute respiratory distress syndrome. A comprehensive analysis of

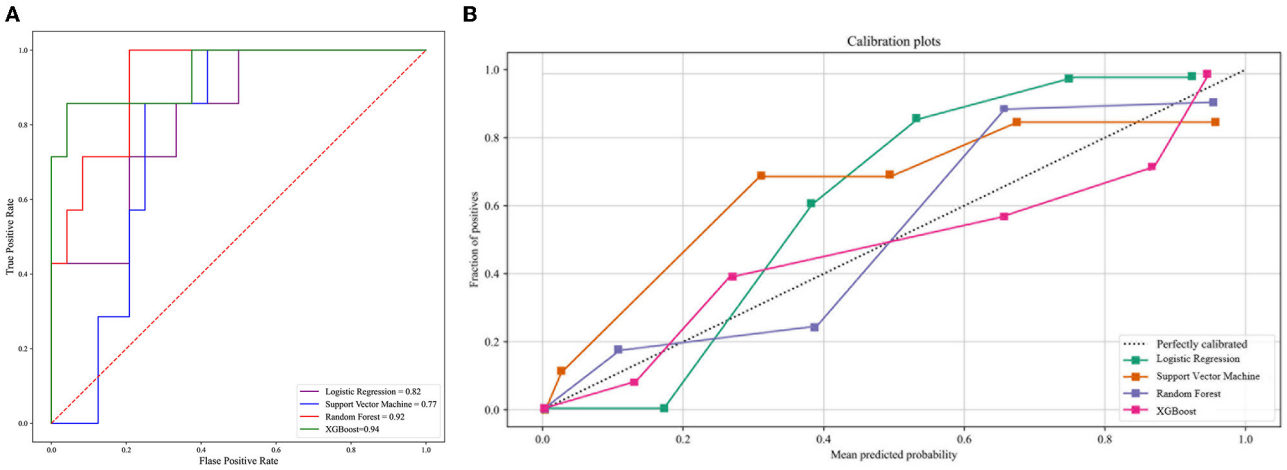

FIGURE 3
(A) The ROC curve of the four machine learning models. **(B)** The calibration curve of the four machine learning models.

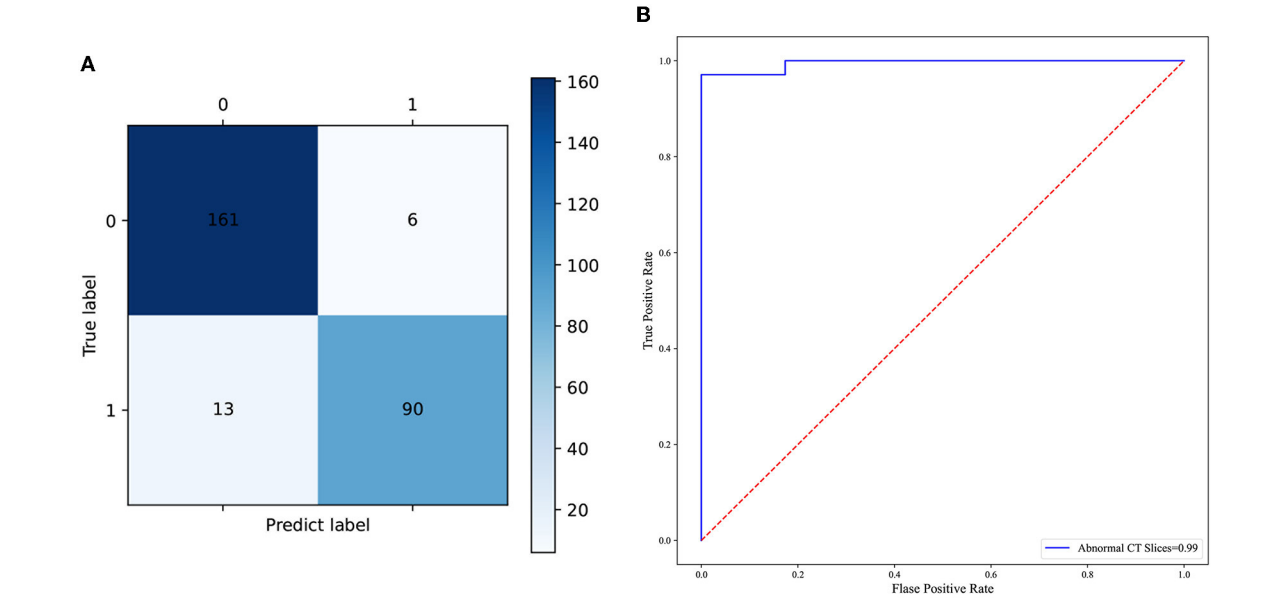

FIGURE 4
(A) The confusion matrix of the predictive performance of the individual CT slices classification model. **(B)** The ROC curve of the classification model of the individual CT slices.

the clinical symptoms, laboratory test results, and CT images is crucial to help understand the scope of COVID-19. We believe that an ensemble predictive model based on the integrated data from the patients could provide more information about the risk factors of complications such as ARDS brought on by COVID-19. Moreover, detailed and accurate risk evaluation of COVID-19 ARDS is important for clinicians to provide more personalized treatment to patients. Some published studies have applied advanced artificial intelligence methods to predict the prognosis of COVID-19 (17–20). They demonstrated the value of machine learning algorithms for predicting the outcomes of COVID-19, but no radiology information was included in the studies (21, 22). Lee et al. developed a deep learning model comprising the chest radiology score and clinical information to predict severe illness in COVID-19 patients (23). However, chest radiology is not suitable for the confirmation of diagnosis or evaluation of COVID-19 outcomes (24). Wang et al. reported an automatic quantitative model based on CT images to predict ARDS in COVID-19 patients (25). In this study, the infection fields of the lung were segmented for the quantitative analysis of the volume and density. We thought the quantitative analysis of CT images could not make the most of the CT information and thus may yield less accurate predictions.

In this retrospective study, we developed three models for the prediction of COVID-19 ARDS. Two individual models were established based on the clinical features data and the CT images, respectively; the third deep learning model was integrated by

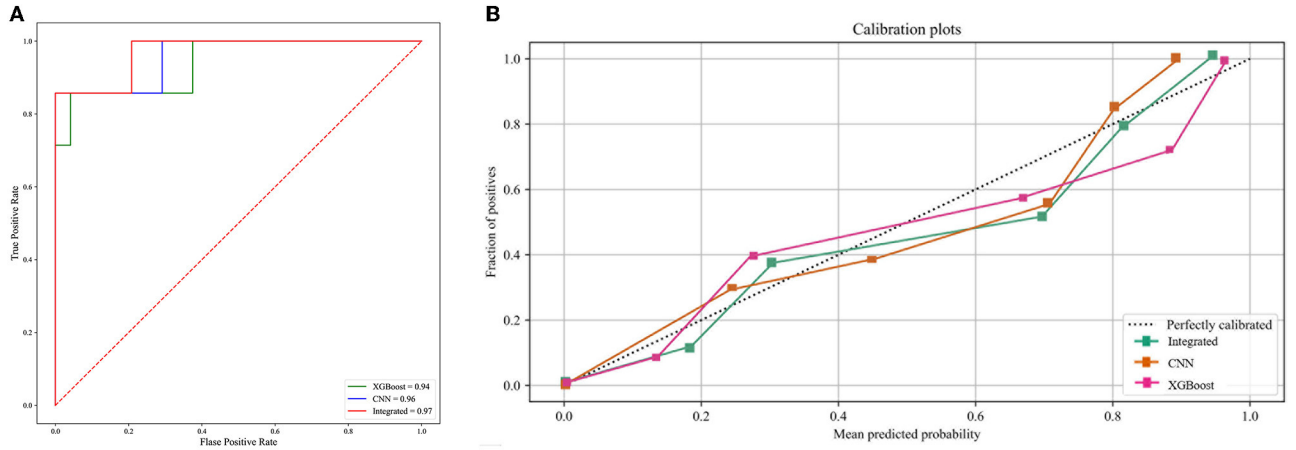

FIGURE 5
(A) The ROC curve of the three predictive models. **(B)** The calibration curve of the three predictive models.

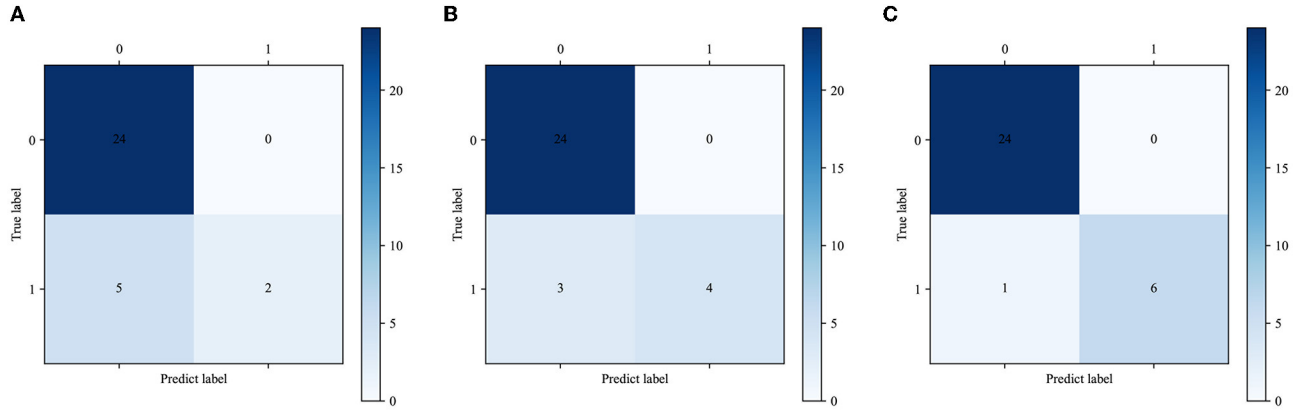

FIGURE 6
(A) The confusion matrix of the XGBoost model. **(B)** The confusion matrix of the CNN model. **(C)** The confusion matrix of the integrated model.

TABLE 3 **Predictive performance of established models in validation cohort.**

Models	Accuracy	Precision	Sensitivity	Specificity	AUC (95% CI)
XGB	0.84	0.29	1	0.83	0.94 (0.91, 0.96)
CNN	0.90	0.57	1	0.89	0.96 (0.93, 0.98)
XGB + CNN	0.97	0.86	1	0.96	0.97 (0.95, 0.99)

AUC, area under curve; XGB, extreme gradients boosting; CNN, convolutional neural network; CI, confidence interval.

the two individual models. We found that the integrated deep learning model could offer better discriminatory performance for predicting COVID-19 ARDS than the two individual models. To strengthen the understanding of COVID-19 ARDS, we performed the multivariate logistic regression to find out the independent risk factors associated with COVID-19 ARDS and depicted the nomogram plot for it. We found that age, the concentration of c-reactive protein, PaO2/FiO2 ratio, the count of total T lymphocytes, and the level of IL-6 were related to COVID-19 ARDS. The inevitable deterioration in immunity response in senior citizens may be the reason for advanced age being a risk factor for COVID-19 ARDS (26). COVID-19 is manifested as a multisystemic disease, and the hyperinflammatory response is extremely associated with its outcome (27). COVID-19 ARDS also causes typical lung pathological changes, which are accompanied by acute and chronic inflammation (28, 29). High concentrations of CRP and IL-6 may indicate a pro-inflammatory state, which has been reported as a risk factor for a severe outcome (26, 27). It is reported that critically ill COVID-19 patients exhibited a status of immune cell hyporesponsiveness when compared to healthy people (28). Several studies have highlighted the values of T-lymphocyte subset absolute counts in predicting morbidity

in COVID-19 patients (29–31). The XGBoost model was selected as the best model to handle the clinical features data because of the best predictive performance tested in the validation cohort. XGBoost stands for "Extreme Gradient Boosting" and was first proposed by Friedman (32). The XGBoost model is one of the ensembling learning algorithms, which makes precise predictions based on a series of weak classifiers, and it has been applied in many studies to deal with massive medical data.

The CT scan procedure can provide more information about the severity of lung damage and acute respiratory failure with a much faster turnaround time (2, 33). The distinctive characteristics of CT slices from COVID-19 ARDS patients could be captured by the convolutional neural network. In our study, the predictive performance of the VGG-16 model was better than that of the model based on the clinical features data. VGG architecture was first proposed by the Visual Geometry Group from Oxford and ranges from 11 to 19 layers (34). The VGG models are widely used as image classifiers or the fundamental basis of newly developed models, which also use images as input data. The VGG-16 network was first used to classify the individual CT slices into normal and abnormal images. Furthermore, the individual patient-based prediction of COVID-19 ARDS was also fulfilled by the VGG-16 network. The XGBoost model and the VGG-16 network model are complementary to each other. The predictive performance of the integrated model was superior to the individual ones. The integrated deep learning model we proposed was demonstrated to be reliable in predicting COVID-19 ARDS with high accuracy in our study. The tremendous progress made in the field of artificial intelligence facilitated the analysis of massive medical data. Our deep learning model may be one example of an automatic analysis tool that can be used for various medical data or alarming systems of adverse events in critically ill patients. Once the integrated deep learning model is fused into the information system of the hospitals, it could rapidly and correctly identify patients at high risk of COVID-19 ARDS without redundant operations.

There are some limitations in our study. First, this is a single-center retrospective study with a relatively small sample size. Second, the validation of the predictive model was only performed in the internal cohort. It is unclear whether similar predictive performance can be observed in other medical centers when our models are applied.

Conclusion

In our study, we tried to establish different models to predict COVID-19 ARDS. We found that the models based on the clinical features or the CT images could provide accurate predictions of COVID-19 ARDS. Moreover, the integrated model combining the two individual models exhibited the best predictive performance with the highest accuracy and ROC value.

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 Ethics Committees at Shanghai Renji Hospital. Written informed consent for participation was not required for this study in accordance with the national legislation and the institutional requirements. Written informed consent was not obtained from the individual(s) for the publication of any potentially identifiable images or data included in this article.

Author contributions

YG and ZH contributed to the conception and design of the study. YZ and SQ organized the database. YZ and JF performed the statistical analysis. SM, SX, and QX wrote the first draft of the manuscript. YZ, JF, ZH, ZZ, and RT wrote sections of the manuscript. All authors contributed to the manuscript revision, read, and approved the submitted version.

Funding

This study was supported by the Shanghai Science and Technology Commission (22YF1423300) and the Renji Hospital Clinical Research Innovation and Cultivation Fund (RJPY-DZX-008).

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. Guan WJ, Ni ZY, Hu Y, Liang WH, Ou CQ, He JX, et al. Clinical characteristics of Coronavirus disease 2019 in China. *New Eng J Med.* (2020) 382:1708–20. doi: 10.1056/NEJMoa2002032

2. 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

3. Huang C, Wang Y, Li X, Ren L, Zhao J, Hu Y, et al. Clinical features of patients infected with 2019 novel coronavirus in Wuhan, China. *Lancet.* (2020) 395:497–506. doi: 10.1016/S0140-6736(20)30183-5

4. Gibson PG, Qin L, Puah SH. COVID-19 acute respiratory distress syndrome (ARDS): clinical features and differences from typical pre-COVID-19 ARDS. *Med J Aust.* (2020) 213:54–6. doi: 10.5694/mja2.50674

5. Williams GW, Berg NK, Reskallah A, Yuan X, Eltzschig HK. Acute respiratory distress syndrome. *Anesthesiology.* (2021) 134:270–82. doi: 10.1097/ALN.0000000000003571

6. Adhikari NK, Burns KE, Friedrich JO, Granton JT, Cook DJ, Meade MO. Effect of nitric oxide on oxygenation and mortality in acute lung injury: systematic review and meta-analysis. *BMJ.* (2007) 334:779. doi: 10.1136/bmj.39139.716794.55

7. Fuller BM, Mohr NM, Skrupky L, Fowler S, Kollef MH, Carpenter CR. The use of inhaled prostaglandins in patients with ARDS: a systematic review and meta-analysis. *Chest.* (2015) 147:1510–22. doi: 10.1378/chest.14-3161

8. Adhikari NK, Dellinger RP, Lundin S, Payen D, Vallet B, Gerlach H, et al. Inhaled nitric oxide does not reduce mortality in patients with acute respiratory distress syndrome regardless of severity: systematic review and meta-analysis. *Crit Care Med.* (2014) 42:404–12. doi: 10.1097/CCM.0b013e3182a27909

9. Ioannidis J. Global perspective of COVID-19 epidemiology for a full-cycle pandemic. *Eur J Clin Invest.* (2020). doi: 10.1111/eci.13423

10. Liu W, Tao Z-W, Wang L, Yuan M-L, Liu K, Zhou L, et al. Analysis of factors associated with disease outcomes in hospitalized patients with 2019 novel coronavirus disease. *Chin Med J.* (2020) 133:1032–8. doi: 10.1097/CM9.0000000000000775

11. Bai Y, Xia J, Huang X, Chen S, Zhan Q. Using machine learning for the early prediction of sepsis-associated ARDS in the ICU and identification of clinical phenotypes with differential responses to treatment. *Front Physiol.* (2022) 13:1050849. doi: 10.3389/fphys.2022.1050849

12. Huang L, Song M, Liu Y, Zhang W, Pei Z, Liu N, et al. Acute respiratory distress syndrome prediction score: derivation and validation. *Am J Crit Care.* (2021) 30:64–71. doi: 10.4037/ajcc2021753

13. Koski E, Murphy J. AI in Healthcare. *Stud Health Technol Inform.* (2021) 284:295–9. doi: 10.3233/SHTI210726

14. Chan JF, Yuan S, Kok KH, et al. A familial cluster of pneumonia associated with the 2019 novel coronavirus indicating person-to-person transmission: a study of a family cluster. *Lancet.* (2020) 395:514–23. doi: 10.1016/S0140-6736(20)30154-9

15. Bai HX, Hsieh B, Xiong Z, Halsey K, Choi JW, Tran TML, et al. Performance of radiologists in differentiating COVID-19 from non-COVID-19 viral pneumonia at chest CT. *Radiology.* (2020) 296:E46–54. doi: 10.1148/radiol.2020200823

16. ARDS Definition Task Force, Ranieri VM, Rubenfeld GD, Thompson BT, Ferguson ND, Caldwell E, et al. Acute respiratory distress syndrome: the Berlin definition. *JAMA.* (2012) 307:2526–33. doi: 10.1001/jama.2012.5669

17. Xu W, Sun NN, Gao HN, Chen ZY, Yang Y, Ju B, et al. Risk factors analysis of COVID-19 patients with ARDS and prediction based on machine learning. *Sci Rep.* (2021) 11:2933. doi: 10.1038/s41598-021-82492-x

18. Sang S, Sun R, Coquet J, Carmichael H, Seto T, Hernandez-Boussard T. Learning from past respiratory infections to predict COVID-19 outcomes: retrospective study. *J Med Internet Res.* (2021) 23:e23026. doi: 10.2196/23026

19. Lee HW, Yang HJ, Kim H, Kim U-H, Kim DH, Yoon SH, et al. Deep learning with chest radiographs for making prognoses in patients with COVID-19: retrospective cohort study. *J Med Internet Res.* (2023) 25:e42717. doi: 10.2196/42717

20. Wang Y, Chen Y, Wei Y, Li M, Zhang Y, Zhang N, et al. Quantitative analysis of chest CT imaging findings with the risk of ARDS in COVID-19 patients: a preliminary study. *Ann Transl Med.* (2020) 8:594. doi: 10.21037/atm-20-3554

21. Chan LL, Tan EK. Evidence of added value of chest CT in Coronavirus disease (COVID-19) pneumonia with initial negative RT-PCR results. *Am J Roentgenol.* (2020) 215:W41. doi: 10.2214/AJR.20.23815

22. Chen N, Zhou M, Dong X, Qu J, Gong F, Han Y, et al. Epidemiological and clinical characteristics of 99 cases of 2019 novel corona virus pneumonia Wuhan, China: a descriptive study. *Lancet.* (2020) 395:507–13. doi: 10.1016/S0140-6736(20)30211-7

23. Silva MJA, Ribeiro LR, Gouveia MIM, Marcelino BDR, Santos CSD, Lima KVB, et al. Hyperinflammatory response in COVID-19: a systematic review. *Viruses.* (2023) 15:553. doi: 10.3390/v15020553

24. Xu Z, Shi L, Wang Y, Zhang J, Huang L, Zhang C, et al. Pathological findings of COVID-19 associated with acute respiratory distress syndrome. *Lancet Respir Med.* (2020) 8:420–2. doi: 10.1016/S2213-2600(20)30076-X

25. Tian S, Xiong Y, Liu H, Niu L, Guo J, Liao M, et al. Pathological study of the 2019 novel coronavirus disease (COVID-19) through postmortem core biopsies. *Mod Pathol.* (2020) 33:1007–14. doi: 10.1038/s41379-020-0536-x

26. Li X, Xu S, Yu M, Wang K, Tao Y, Zhou Y, et al. Risk factors for severity and mortality in adult COVID-19 inpatients in Wuhan. *J Allergy Clin Immunol.* (2020) 146:110–8. doi: 10.1016/j.jaci.2020.04.006

27. Forsblom E, Helanne H, Kortela E, Silén S, Meretoja A, Järvinen A. Inflammation parameters predict fatal outcome in male COVID-19 patients in a low case-fatality area - a population-based registry study. *Infect Dis.* (2022) 54:558–71. doi: 10.1080/23744235.2022.2055786

28. Morrell ED, Bhatraju PK, Sathe NA, et al. Chemokines, soluble PD-L1, and immune cell hyporesponsiveness are distinct features of SARS-CoV-2 critical illness. *Am J Physiol Lung Cell Mol Physiol.* (2022) 323:L14–26. doi: 10.1152/ajplung.00049.2022

29. Zhang J, Wang Z, Wang X, Hu Z, Yang C, Lei P. Risk factors for mortality of COVID-19 patient based on clinical course: a single center retrospective case-control study. *Front Immunol.* (2021) 12:581469. doi: 10.3389/fimmu.2021.581469

30. Xiong L, Zang X, Feng G, Zhao F. Wan, S, Zeng W, et al. Clinical characteristics and peripheral immunocyte subsets alteration of 85 COVID-19 Deaths. *Aging.* (2021) 13:6289–97. doi: 10.18632/aging.202819

31. Kazancioglu S, Yilmaz FM, Bastug A, Sakalli A, Ozbay BO, Buyuktarakci C, et al. Lymphocyte subset alteration and monocyte CD4 expression reduction in patients with severe COVID-19. *Viral Immunol.* (2021) 34:342–51. doi: 10.1089/vim.2020.0166

32. Friedman JH. Greedy function approximation: a gradient boosting machine. *Ann Stat.* (2001) 29:1189–232. doi: 10.1214/aos/1013203451

33. Shi H, Han X, Jiang N, Cao Y, Alwalid O, Gu J, et al. Radiological findings from 81 patients with COVID-19 pneumonia in Wuhan, China: a descriptive study. *Lancet Infect Dis.* (2020) 20:425–34. doi: 10.1016/S1473-3099(20)30086-4

34. Simonyan K, Zisserman A. Very deep convolutional networks for large-scale image recognition. *Computer Science.* (2014). doi: 10.48550/arXiv.1409.1556

frontiers | Frontiers in Artificial Intelligence

TYPE Original Research
PUBLISHED 28 July 2023
DOI 10.3389/frai.2023.1179226

OPEN ACCESS

EDITED BY
Farah Kidwai-Khan,
Yale University, United States

REVIEWED BY
Joao Sousa,
University of Lisbon, Portugal
Rixin Wang,
Yale University School of Medicine,
United States

*CORRESPONDENCE
Addisu Jember Zeleke
✉ addisu.zeleke2@unibo.it

RECEIVED 06 March 2023
ACCEPTED 10 July 2023
PUBLISHED 28 July 2023

CITATION
Zeleke AJ, Palumbo P, Tubertini P, Miglio R and Chiari L (2023) Machine learning-based prediction of hospital prolonged length of stay admission at emergency department: a Gradient Boosting algorithm analysis.
Front. Artif. Intell. 6:1179226.
doi: 10.3389/frai.2023.1179226

Machine learning-based prediction of hospital prolonged length of stay admission at emergency department: a Gradient Boosting algorithm analysis

Addisu Jember Zeleke[1]*, Pierpaolo Palumbo[1], Paolo Tubertini[2], Rossella Miglio[3] and Lorenzo Chiari[1,4]

[1]Department of Electrical, Electronic, and Information Engineering Guglielmo Marconi, University of Bologna, Bologna, Italy, [2]Enterprise Information Systems for Integrated Care and Research Data Management, Istituto di Ricovero e Cura a Carattere Scientifico (IRCCS) Azienda Ospedaliero—Universitaria di Bologna, Bologna, Italy, [3]Department of Statistical Sciences, University of Bologna, Bologna, Italy, [4]Health Sciences and Technologies Interdepartmental Center for Industrial Research (CIRI SDV), University of Bologna, Bologna, Italy

Objective: This study aims to develop and compare different models to predict the Length of Stay (LoS) and the Prolonged Length of Stay (PLoS) of inpatients admitted through the emergency department (ED) in general patient settings. This aim is not only to promote any specific model but rather to suggest a decision-supporting tool (i.e., a prediction framework).

Methods: We analyzed a dataset of patients admitted through the ED to the "Sant"Orsola Malpighi University Hospital of Bologna, Italy, between January 1 and October 26, 2022. PLoS was defined as any hospitalization with LoS longer than 6 days. We deployed six classification algorithms for predicting PLoS: Random Forest (RF), Support Vector Machines (SVM), Gradient Boosting (GB), AdaBoost, K-Nearest Neighbors (KNN), and logistic regression (LoR). We evaluated the performance of these models with the Brier score, the area under the ROC curve (AUC), accuracy, sensitivity (recall), specificity, precision, and F1-score. We further developed eight regression models for LoS prediction: Linear Regression (LR), including the penalized linear models Least Absolute Shrinkage and Selection Operator (LASSO), Ridge and Elastic-net regression, Support vector regression, RF regression, KNN, and eXtreme Gradient Boosting (XGBoost) regression. The model performances were measured by their mean square error, mean absolute error, and mean relative error. The dataset was randomly split into a training set (70%) and a validation set (30%).

Results: A total of 12,858 eligible patients were included in our study, of whom 60.88% had a PloS. The GB classifier best predicted PloS (accuracy 75%, AUC 75.4%, Brier score 0.181), followed by LoR classifier (accuracy 75%, AUC 75.2%, Brier score 0.182). These models also showed to be adequately calibrated. Ridge and XGBoost regressions best predicted LoS, with the smallest total prediction error. The overall prediction error is between 6 and 7 days, meaning there is a 6–7 day mean difference between actual and predicted LoS.

Conclusion: Our results demonstrate the potential of machine learning-based methods to predict LoS and provide valuable insights into the risks behind prolonged hospitalizations. In addition to physicians' clinical expertise, the results of these models can be utilized as input to make informed decisions, such as

predicting hospitalizations and enhancing the overall performance of a public healthcare system.

KEYWORDS

emergency department, prolonged length of stay, machine learning, prediction, classification, regression

1. Introduction

1.1. Importance of addressing hospitalization LoS after an emergency department visit

The Length of Stay (LoS) measures the time a patient spends in a hospital, from admission to discharge. It is a key indicator of the quality of hospital services, including the speed and efficiency of patient treatment, the prevention of hospital-acquired infections, the ability to anticipate prolonged stays due to pre-existing medical conditions, resource utilization, and the cost of inpatient care. LoS can also be used to evaluate the success of surgical procedures and patient outcomes. With an in-depth understanding of LoS and potential adverse events, hospitals can make informed decisions and improve patients' overall quality of care. Accurate LoS prediction enables the efficient use of medical resources, better clinical decision-making, and provision of useful prognostic information. In hospital management, LoS is critical in determining hospital costs and patient satisfaction. Furthermore, it is associated with disease severity and mortality (Paterson et al., 2006). During an ED visit, some predictors of hospital LoS were known before admission to the hospital. Prior studies have shown that patients in EDs have a longer LoS (Krochmal and Riley, 1994; Liew et al., 2003). It has been demonstrated that extended hospital stays negatively affect clinical outcomes: according to Sud et al. (2017), long LoS is associated with increased mortality and readmission rates; the results of Bo et al. (2016) indicated that PLoS is associated with cognitive impairment, functional limitations, and higher burdens of comorbidity; the results of Emori and Gaynes (1993) also indicated that PLoS increased the risk of hospital-acquired infections. Patients are prioritized based on their level of medical need in a triage plan to enhance healthcare and reduce mortality. Models that predict patient-related outcome measures and LoS are useful tools for maximizing healthcare utilization (Gellman, 1974). As a result, policymakers and clinicians could determine how to allocate resources among different approaches by comparing treatments across disciplines.

1.2. Methodological review/predictive modeling of PLoS

Machine learning (ML) provides innovative methods in data predictions that are widely used. Numerous studies have examined how different predictive models can predict LoS more accurately (Lu et al., 2015). A prediction model based on factors affecting LoS has been developed in previous studies using multiple supervised learning techniques. For categorical outcomes, including logistic regression (LoR), random forest (RF), Gradient Boosting (GB), K-nearest neighbors (KNN), support vector machine (SVM), decision tree (DT), and artificial neural networks (ANN; Hachesu et al., 2013; LaFaro et al., 2015) were used to predict LoS. In a study by Chuang et al. (2018), LoR, SVM, RF, multivariate adaptive regression splines (MARS), classification and regression tree (CART), etc. were used to study the prediction of PLoS in patients undergoing general surgery. The RF classifier showed the highest performance. In another interesting study, for a continuous outcome, Caetano et al. (2015) used and compared regressors, including multiple regression (MR), RF regression, decision tree (DT), neural network (NN), and support vector machine (SVM) regression. The RF regression showed the highest performance. The performance of ensemble learning models (like RF, GB, AdaBoost) is usually better than that of single learning models (Han et al., 2019). An alternative, data-driven approach to predictive analytics in emergency care is available through preprocessing, data mining, and machine learning techniques applied to big data stored in electronic health records (EHRs; Yu et al., 2018). In other clinical data from inpatients with lower limb fractures, Colella et al. (2021) employed similar ML techniques to predict PLoS, by dividing the outcome variable into two classes. Kirchebner et al. (2020) conducted an exploratory study on hospitalized schizophrenic patients to predict PLoS. This study selected the most significant features using a forward selection procedure. Then various machine learning classification algorithms were used for binary outcomes: with and without prolonged LoS. Overall in the literature, SVM, GB, LoR, NN, and RF are the most common and widely used supervised ML classifier algorithms used to estimate LoS (Jiang et al., 2010; Morton et al., 2014). Table 1 provides a brief overview of ML models, prediction outcomes, and the target groups for which LoS was predicted.

1.3. Related works

Previous research has investigated various methods of predicting LoS with varying scopes and settings. LoS can be predicted for all patients admitted to the hospital based on non-medical factors such as type of admission, gender, race, insurance status, place of residence, and the cost of hospitalization, as well as medical characteristics like risk/severity measures, primary condition groups, emergency degree, and prior admissions. It is also possible to predict LoS for specific diseases or surgical procedures. The most frequently reported factors that affect the ED LoS are patient age, gender, triage category, mode of arrival, the requirement for an interpreter, admission, diagnostic

TABLE 1 Brief review of ML models and patients groups for predicting hospital patients' LoS.

References	Outcome: prediction type	ML models	Target group	Results
Mekhaldi et al. (2020)	Regression	RF Regressor, Gradient Boosting Regressor	General patients	GB performed better than RF; performance were checked by MSE, the R-squared and the Adjusted R-squared.
Daghistani et al. (2019)	Classification	RF, ANN, SVM, BN	Cardiac patients	RF model outperformed all other models: sensitivity (0.80), accuracy (0.80), and AUROC (0.94).
Tsai et al. (2016)	Regression	LR, ANN	Cardiac patients	LR model performed slightly better than ANN models, with a MAE value of 3.76 and 3.87
Symum and Zayas-Castro (2020)	Classification	DT C5.0, linear SVM, KNN, RF, and multi-layered artificial neural net	Chronic disease (congestive heart failure, acute myocardial infarction, COPD, pneumonia, type 2 diabetes).	For all patient groups, LSVM (Lagrangian SVM) with wrapper feature selection performed well.
Tanuja et al. (2011)	Classification	Naive Bayes; KNN; DT classifiers; Multi-layer backpropagation	General patients	MLP and NB models had the best classification accuracy of around 85%, while KNN performed poorly with only 63.6% accuracy
Combes et al. (2014)	Regression and classification	Two based models: *Classifier*: RF, LMT (Logistical model tree), MP, DT (C4.5-J48), NBTree, REPTree, and SVM. *Regression*: LR, SV regression, MLP, IRM (Isotonic regression model), M5P, PRLM (Pace regression linear models)	Pediatric	Using 10-fold cross-validation, obtained the best performances in using logistic regression, and in continuous outcome SVM Regression showed a lower prediction error.
Etu et al. (2022)	Classification	LoR, GB, DT, and RF	COVID-19 Patients	The GB model outperformed the baseline classifier (LoR) and tree-based classifiers (DT and RF) with an accuracy of 85% and F1-score of 0.88 for predicting ED LoS
Alsinglawi et al. (2020a)	Regression	RF Regressor; GB Regressor; Stacking Regressor; DNN	Cardiovascular patients in the ICU	GB regressor outweighed the other proposed models, and showed a higher R-squared.
Kirchebner et al. (2020)	Classification	BT; KNN; SVM	Schizophrenic patients	Two factors have been identified as particularly influential for a prolonged forensic LoS, namely (attempted) homicide and the extent of the victim's injuries.
Thongpeth et al. (2021)	Regression	LR with three penalized linear (ridge, lasso, elastic net), and 4 ML model types: SVR, NN, RF, and XGBoost	Chronic disease	The RF model had the best predictive performance with the smallest prediction errors, while linear ridge regression had the poorest prediction performance with the largest prediction errors.

LoR, logistic regression; LR, linear regression; RF, random forest; NB, Naive Bayes; ANN, artificial neural network; SVM, support vector machine; MLP, Multi-layer backpropagation; DT, decision tree; GB, Gradient Boosting; XGBoost, eXtreme Gradient Boosting; KNN, K-nearest neighbors; BN, Bayesian network; COPD, chronic obstructive pulmonary disease; MSE, mean square error; ICU, intensive care unit.

complexity necessitating extra testing, and the availability of resources, including staff and beds (Asaro et al., 2007; Biber et al., 2013; Rahman et al., 2020). Patient characteristics influencing LoS, such as demographics and comorbidities, are often available at triage and admission (Tsai et al., 2016). Several studies in the literature have examined the LoS trends in general patients (Tanuja et al., 2011; Mekhaldi et al., 2020), or in particular patient populations, focusing, for instance, on a certain age group (Ackroyd-Stolarz et al., 2011; Launay et al., 2018; Marfil-Garza et al., 2018; Sir et al., 2019) or specific health conditions (e.g., cardiology; García-González et al., 2014; Tsai et al., 2016; Chuang et al., 2018; Daghistani et al., 2019), peritoneal dialysis (Wu et al.,

2020), schizophrenia (Kirchebner et al., 2020), knee arthroplasty (Song et al., 2020), COVID-19 (Vekaria et al., 2021; Etu et al., 2022; Zeleke et al., 2022), abdominal pain (Dadeh and Phunyanantakorn, 2020), mental health (Wolff et al., 2015), cardiovascular diseases (Almashrafi et al., 2016; Alsinglawi et al., 2020a), or in specific discipline areas or specialties such as spine surgery (Basil and Wang, 2019) and cancer surgeries (Laky et al., 2010; Gohil et al., 2014; Jo et al., 2021). However, most of these studies have had limited sample sizes and have not considered a wide range of clinical factors. In-hospital adverse events are known to increase the risk of prolonged Length of Stay (LoS) in older patients (Ackroyd-Stolarz et al., 2011).

A study of Length of Stay (LoS) in the emergency department of a tertiary care center (van der Veen et al., 2018) found a significant association between multiple chief complaints, including headaches and chest pain, laboratory/radiology testing, and consultation with prolonged hospitalization in the ED. Another population-based study conducted in Osaka, Japan (Katayama et al., 2021) showed that factors such as old age, traffic accidents, lack of a permanent address, need for nursing care, and being solitary were associated with prolonged hospitalization for patients transported by ambulance. Another retrospective study of prolonged LoS in a tertiary healthcare center in Mexico (Marfil-Garza et al., 2018) showed that demographic and disease-specific differences, such as younger age, male gender, lower physician-to-patient ratio, emergency and weekend admissions, surgery, number of comorbidities, and lower socioeconomic status, were associated with a prolonged LoS. Diseases with the greatest risk for prolonged LoS included complex conditions like bone marrow transplant, systemic mycoses, parasitosis, and complex abdominal diseases like intestinal fistulas.

1.4. Aims

This study used various supervised machine learning algorithms to predict the length of stay for patients admitted through the emergency department in general patient settings. The outcome was analyzed as both a dichotomous (PLoS) and continuous (LoS) variable. Data was gathered from routine triage and ED admission processes and recorded in the hospital's electronic medical records. The best-performing model was selected to make predictions and gain meaningful insights for future patients.

2. Materials and methods

2.1. Study design and population

We screened for eligibility for all admissions to the hospital through the ED of the public University Hospital of Bologna Sant'Orsola-Malpighi (AOSP), Bologna, Italy, between January 1, 2022, and October 26, 2022. AOSP is a 1,500-bed tertiary care teaching hospital in Central-Northern Italy with 70,000 emergency department visits per year, this is one of the largest hospitals in the country (Fridman et al., 2022). All the necessary steps of the clinical pathway: ED triage, medical examination, hospital admission, and hospital discharge, are shown in Figure 1. We included all patients who visited the ED, were admitted to the hospital, and stayed until they got formal permission to discharge. Any patients who left the ED, were transferred to another hospital, refused the hospitalization, died, went away after the medical examination, left without being seen, or left without notice (detail as shown in Figure 2) were excluded from the analysis.

2.1.1. Outcome variable

The primary outcome of this study was hospital length of stay (LoS) and prolonged length of stay (PloS). LoS is calculated as the number of days between admission and discharge. We defined PLoS threshold as any LoS that is longer than the reported average LoS (i.e., 6 days; Zoller et al., 2014; Song et al., 2020; Wu et al., 2020). The LoS was reclassified as binary (i.e., either "without PLoS$<$ 6 'days' or with PLoS" $\geq$6 "days") for classification analysis, and LoS as a continuous outcome for regression analysis.

2.1.2. Independent variables

Any information collected at triage and available from ED admissions was considered as a predictor of LoS or PLoS. These include demographic factors (such as gender and age), mode of arrival/source of admission, risk categories as determined by triage at the entrance, and current problems or chief complaints. A detailed description of each independent feature, measure category, and outcome is presented in Supplementary Table 1.

2.2. Model development

2.2.1. Predictive models fitting and evaluation: binary outcome

The diagram in Figure 3 shows the data analysis framework we followed for developing and evaluating our predictive model. The main objective is to predict the categorical class labels of new data points or instances based on past observations. Based on the literature, six common classification algorithms were selected for comparison: GradientBoosting (GB), random forests (RF), support vector machine (SVM), K-Nearest Neighbors (KNN), AdaBoost, and logistic regression (LoR). The model with the highest prediction performance was used to identify predictive factors contributing to the outcome. We randomly divided the data into training (70%) and testing or validation (30%) sets. The analyses were performed in *Scikit-learn* in Python (Jupyter notebook version; Pedregosa et al., 2011).

Hastie et al. (2009) provide detailed explanations, but here we provide a brief overview of ML techniques, and hyperparameters tuning settings.

2.2.1.1. Random Forests (RF)

In statistical applications, Random Forests (RF) are a commonly used type of supervised machine learning that can be utilized for both classification and regression tasks (Breiman, 2001; Genuer et al., 2010). RF predicts outcome labels for a group of samples by building several decision trees using a random set of covariates. The weak classifier can be transformed into a

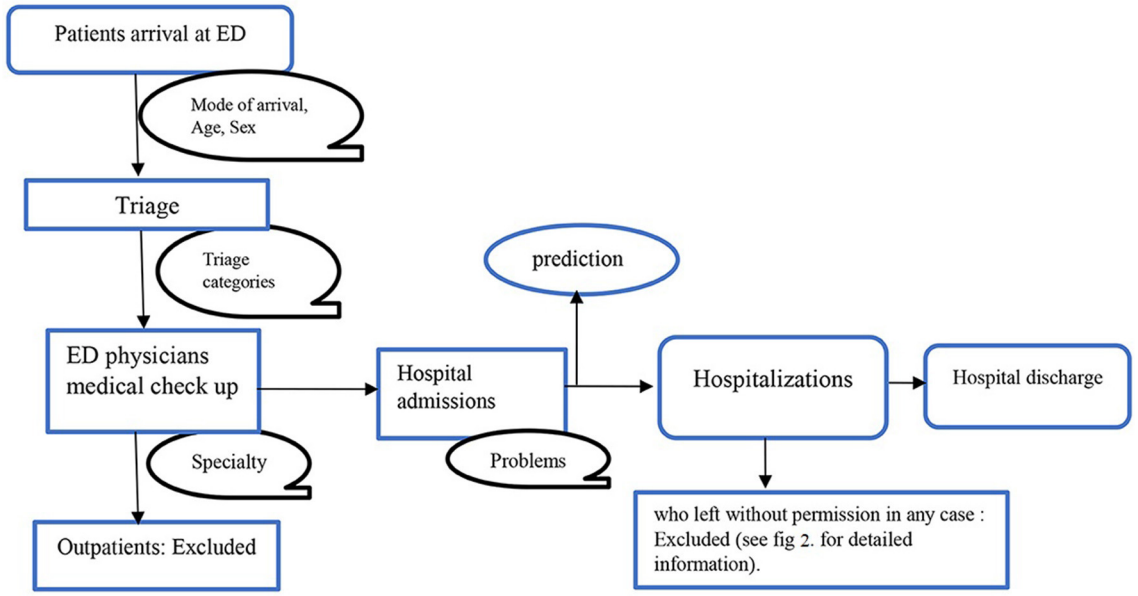

FIGURE 1
Clinical pathway.

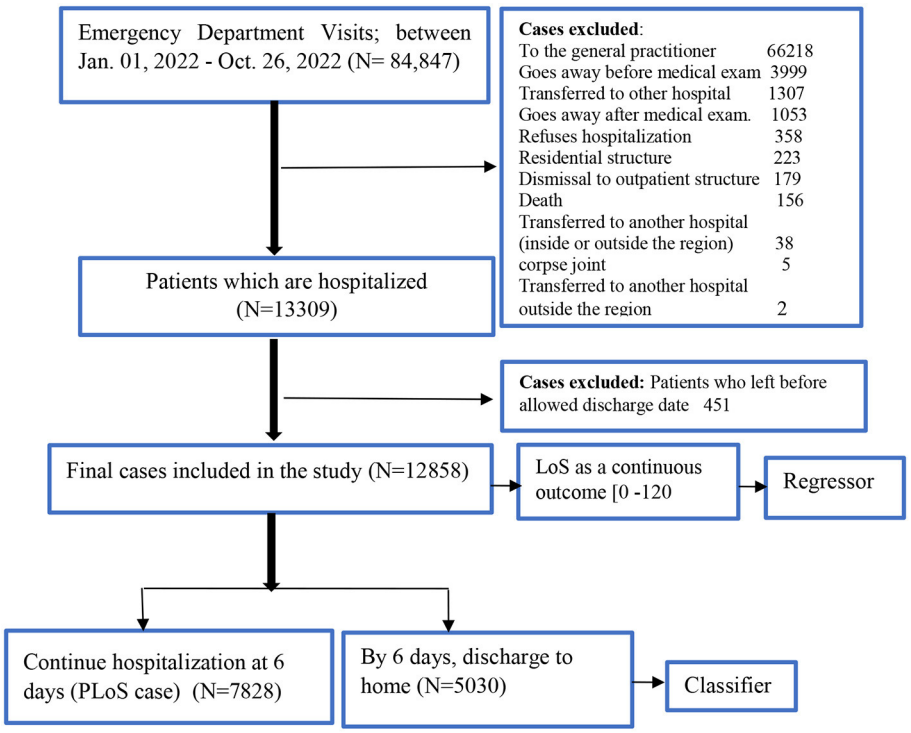

FIGURE 2
Flowchart of patients selection.

strong one by taking the majority of votes for classification and averaging in regression. To enhance the classification accuracy, multiple decision trees are combined in RF to form an ensemble classification algorithm. Each tree is grown using a bootstrapped sample from the original data (Qi, 2012). An ensemble ML method combines a series of underperforming classifiers to produce an improved classifier. The mechanism for this combination differs between ensemble algorithms. In this study, the RF model was created using the *sklearn.RandomForestClassifier* package in Python (Pedregosa et al., 2011).

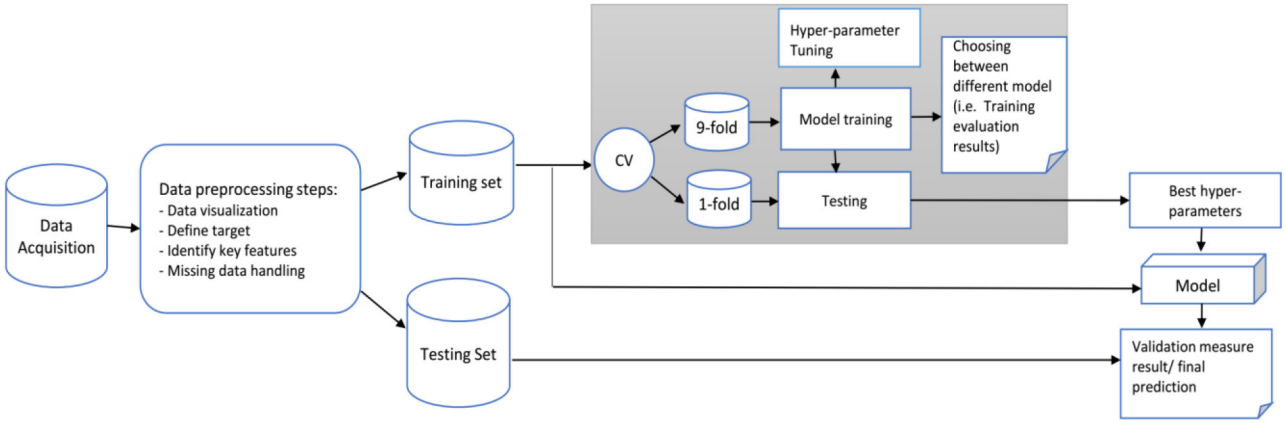

FIGURE 3
Proposed framework for our prediction model.

2.2.1.2. Gradient Boosting (GB)

Gradient Boosting is an ensemble learning model that employs decision trees as its base classifier, without bootstrap sampling (Luo et al., 2020). GB aims to create a robust predictive model by combining weak learning models, considering the bias of all previous decision trees in the model. Furthermore, unlike randomization in other methods, GB focuses on fixing the target outcomes in order to minimize errors. In this study, the GB model was constructed using the *sklearn.GradientBoostingClassifier* package in Python (Pedregosa et al., 2011).

2.2.1.3. Support vector machines (SVMs)

In SVMs, the data is separated using a large gap or hyperplane to deal with linearly non-separable problems. It works by finding an optimal separating hyperplane in the feature space for classification. The Python *sklearn.SVC* package was used to build the SVM model for this study (Pedregosa et al., 2011).

2.2.1.4. AdaBoost classifier

Similar to GB, AdaBoost classifier is also a boosting algorithm, converting a set of weak learners into a single strong learner. However, they differ on how they create weak learners during the iterative process. In GB, as mentioned, it is to minimize the cumulative predicted errors. Still, in AdaBoost it focuses on training the prior miscalculated observations and alters the data distribution to improve sample weight values. The Python *sklearn.AdaBoostingClassifier* package was used to build the AdaBoost model for this study (Pedregosa et al., 2011).

2.2.1.5. K-Nearest Neighbors (KNN)

KNN is an instance-based algorithm, which labels the test record based on its distance from similar data during training (i.e., which analyzes the similarities between the new data and the existing data and adds the new data into the category that is highly similar to the available categories). The only step in building the model is storing the training dataset. Then, the algorithm finds the closest data points in the training dataset, or its "nearest neighbors" to predict a new data point (Keller et al., 1985). Python *sklearn.KNeighborsClassifier* package was used to build the AdaBoost model for this study (Pedregosa et al., 2011).

TABLE 2 Hyperparameter tuning summary.

Model classifiers	Hyperparameter tuning description
RF	# of _estimators = 200; longest path between root node and leaf node, max_ depth = 15; class_ weight = "balanced;" Number of maximum features for each tree, max_ features = sqrt; min_ samples_ split = 2; min_ samples_ leaf = 1; random_ state = 42
GB	# of estimators = 200, max_depth = 4, and loss = ls
KNN	Number of neighbors = 10; algorithms = "auto;" leaf_ size = 1; p = 1; weights = "uniform"
AdaBoost	Similar to RF, define the Decision tree (Dt) classifier first in the same setting and then boost the Dt fit by AdaBoostClassifier.
SVM	Kernel = linear; degree of similarity, gamma = 0.01; regularization, C = 10
LoR	No critical hyperparameters need to be tuned.

2.2.1.6. Logistic regression (LoR)

The LoR model is widely used in binary classification problems. The parameter of interest is estimated using maximum likelihood estimation. Similarly, Python *sklearn.LogisticRegression* package was used for this classifier.

Every machine learning (ML) technique requires the optimization of hyperparameters to enhance its performance. To develop a well-performing generalized model, it is crucial to carefully select the hyperparameters. Different algorithms will have distinct sets of hyperparameters.

The hyperparameter tuning summary for each type of classifier and their descriptions used for this analysis are shown in Table 2.

In building a prediction model, evaluating its performance and accuracy is important. Various metrics were used to assess the model's accuracy, including the Brier score, AUC, accuracy,

sensitivity, specificity, precision, and F1-measure (Steyerberg et al., 2010). Calibration curve plots were also employed to visualize the calibration power of each model and ensure that the model fitted the data optimally. By carefully evaluating the predictive power of the model, we can ensure that the results produced by the model are reliable and can be trusted for decision-making purposes in the healthcare system.

Brier score is an overall performance measure, a measure of the accuracy of a predicted probability score (i.e., mean squared error of probability estimate). A low Brier score suggests an excellent overall performance (Steyerberg et al., 2010).

$$BS = \frac{\sum_{i=1} \left(\hat{p}\left(y_i\right) - y_i\right)^2}{n}$$

An evaluation metric like accuracy calculates the proportion of correct predictions (both positive and negative) out of all the predictions made by the model. Achieving the highest accuracy level is important. Sensitivity or recall reflects the number of positive predictions that were accurately identified, while specificity measures the same for negative predictions. A higher recall indicates that more true values were correctly predicted. The F1-score balances precision and recall by taking the harmonic mean of both values. The overall predictive accuracy of the model was evaluated by determining the area under the receiver operating characteristic curve (AUC). Calibration is crucial in developing and validating clinical prediction models, which refers to the match between predicted and observed risks (Steyerberg, 2019). In the case of binary outcomes, calibration measures the agreement between estimated and observed probabilities of occurrence. Calibration curves were used to assess calibration. A perfect model's calibration curve would be diagonal, meaning that the predicted probabilities align with the observed probabilities.

2.2.1.7. Variables importance

The most effective prediction model was utilized to determine the importance of variables. Identifying key factors in machine learning predictions is crucial. The metric used to evaluate this is the mean decrease in impurity, which calculates the average change in the impurity of nodes across all trees in the ensemble, taking into account the proportion of samples that reach each node. A higher value generally means that the feature is more significant. With high-dimensional datasets, it is crucial to properly select and rank covariates for both prediction and interpretation purposes.

2.2.2. Predictive models fitting and evaluation: continuous outcome

In order to minimize information loss in a classification task, we also explored it as a continuous outcome and employed regression models. Our study employed eight different learning algorithms, including linear regression (LR) and its penalized versions (Lasso, Ridge, and Elastic Net regression), as well as Support Vector Regression, Random Forest Regression, K-Nearest Neighbors (KNN), and XGBoost Regression.

2.2.2.1. Linear regression (LR)

This method involves fitting a linear equation to the data to establish a relationship between the independent variables and the dependent or outcome variable. The equation can then be used to make predictions based on the input data. The linear regression model is typically expressed in the following form:

$$y_i = \beta_0 + \sum_{j=1}^{n} \beta_j x_{ij}$$

where y_i is the continuous outcome value of subject i, β_0 is intercept, β_j is the coefficient of feature j, and x_{ij} is feature j of subject i.

It is possible to estimate the regression parameter of a linear regression model using the least square method by minimizing the error term in the unknown β_j.

$$\hat{\beta} = argmin_{\beta} \left\{ \frac{1}{n} \sum_{i=1}^{n} \left(y_i - \hat{y}_i\right)^2 \right\}$$

2.2.2.2. Ridge regression

It works by finding the coefficients that minimize the sum of error squares by applying a penalty to those coefficients (Tibshirani, 1996).

$$\hat{\beta} = argmin_{\beta} \left\{ \frac{1}{n} \sum_{i=1}^{n} \left(y_i - \hat{y}_i\right)^2 + \lambda \sum_{j=1}^{p} {\beta_j}^2 \right\}$$

λ is the regularization parameter that we are going to optimize.

2.2.2.3. Lasso regression

The same task but uses the sum of absolute values of the weights for the penalty (Tibshirani, 1996).

$$\hat{\beta} = argmin_{\beta} \left\{ \frac{1}{n} \sum_{i=1}^{n} \left(y_i - \hat{y}_i\right)^2 + \lambda \sum_{j=1}^{p} \left|\beta_j\right| \right\}$$

2.2.2.4. Elastic-Net

A combination of lasso and ridge regression that reduces bias, better than lasso or ridge regressions (Friedman et al., 2009).

$$\hat{\beta} = argmin_{\beta} \left\{ \frac{1}{n} \sum_{i=1}^{n} \left(y_i - \hat{y}_i\right)^2 + \lambda_1 \sum_{j=1}^{p} {\beta_j}^2 + \lambda_2 \sum_{j=1}^{p} \left|\beta_j\right| \right\}$$

In contrast to prediction models, regression models focus on estimating the relationship between a set of independent variables and a continuous outcome variable. Instead of categorizing the outcome into specific classes, the regression models aim to predict the continuous value of the outcome based on the given set of independent variables. The performance measure used in regression models is typically the mean squared error, or the root mean squared error, which represents the average deviation between the predicted and actual values of the outcome variable. Regression models aim to minimize these errors, thereby providing a more accurate prediction of the continuous outcome.

Using a loss function helps us evaluate the performance of a prediction model by quantifying the difference between the predicted and the actual values. Mean square error (MSE), mean absolute error (MAE), and mean relative error (MRE) were calculated to measure the prediction performance of each

10.3389/frai.2023.1179226

TABLE 3 Presenting characteristics of patients who visited the ED of ASOP, Bologna, Italy, 2022 ($n = 12,858$).

Factor	Total, *n* (%) ($n = 12,858$)	PLoS (i.e., ≥6 days)		Proportion difference (%) (with PLoS—without PLoS)
		With PLoS ($n = 7,828$) 60.88%	Without PLoS ($n = 5,030$) 39.12%	
Age, median	72	-	-	-
Age categories, *n* (%)				
(0–17)	1,170 (9.1)	329 (4.2)	841 (16.7)	−12.5
(18–29)	679 (5.3)	158 (2.1)	521 (10.4)	−8.3
(30–49)	1,772 (13.8)	616 (7.9)	1,176 (23.4)	−15.5
(50–69)	2,364 (18.4)	1,554 (19.9)	810 (16.1)	3.8
70 or older	6,873 (53.5)	5,171 (66.1)	1,702 (33.9)	32.2
Gender, *n* (%)				
Male	6,101 (47.4)	3,928 (50.2)	2,173 (43.2)	7.0
Female	6,757 (52.6)	3,900 (49.8)	2,857 (56.8)	−7.0
Mode of arrival, *n* (%)				
Ambulance–118	6,645 (51.7)	4,624 (59.1)	2,021 (40.2)	18.9
Own vehicle/walk-in	4,769 (37.2)	2,204 (28.2)	2,565 (51.0)	−22.8
Others[a]	1,444 (11.2)	1,000 (12.8)	444 (8.8)	4.0
Triage category				
Red	807 (6.3)	539 (6.9)	268 (5.3)	1.6
Orange	4,360 (33.9)	2,367 (30.2)	1,993 (39.6)	−9.4
Light blue	4,253 (33.1)	3,065 (39.2)	1,188 (23.6)	15.6
Green	3,224 (25.1)	1,784 (22.8)	1,440 (28.6)	−5.8
White	214 (1.7)	73 (0.9)	141 (2.8)	−1.9
Specialty, *n* (%)				
General medicine	3,757 (29.2)	2,995 (38.3)	762 (15.1)	23.2
Geriatric	1,624 (12.6)	1,252 (16.0)	372 (7.4)	8.6
Astanteria/casualty department	1,450 (10.7)	809 (10.3)	641 (12.7)	−2.4
Obstetrics and gynecology	1,159 (9.0)	114 (1.5)	1,045 (20.8)	−19.3
Pediatrics	609 (4.7)	193 (2.5)	416 (8.3)	−5.8
General surgery	571 (4.4)	276 (3.5)	295 (5.9)	−2.4
Infectious and tropical diseases	533 (4.1)	372 (4.8)	161 (3.2)	1.6
Orthopedics and traumatology	481 (3.7)	378 (4.8)	103 (2.0)	2.8
Urology	405 (3.2)	99 (1.3)	306 (6.1)	−4.8
Coronary unit	377 (2.9)	283 (3.6)	94 (1.9)	1.7
Pediatric surgery	376 (2.9)	77 (1.0)	299 (5.9)	−4.9
Gastroenterology	308 (2.4)	237 (3.0)	71 (1.4)	1.6
Cardiology	150 (1.2)	96 (1.2)	54 (1.1)	0.1
Intensive care	141 (1.1)	113 (1.4)	28 (0.6)	0.8
Pneumology	135 (1.1)	111 (1.4)	24 (0.5)	0.9
Nephrology	105 (0.8)	91 (1.2)	14 (0.3)	0.9
Oncology	93 (0.7)	67 (0.9)	26 (0.5)	0.4
Vascular surgery	89 (0.7)	65 (0.8)	24 (0.5)	0.3
Missing values	76 (0.6)	30 (0.4)	46 (0.9)	−0.5

(Continued)

TABLE 3 (Continued)

Factor	Total, *n* (%) (*n* = 12,858)	PLoS (i.e., ≥6 days)		Proportion difference (%) (with PLoS—without PLoS)
		With PLoS (*n* = 7,828) 60.88%	Without PLoS (*n* = 5,030) 39.12%	
Others[b]	419 (3.3)	170 (2.2)	249 (5.0)	−2.8
Problems, *n* (%)				
Dyspnea	1,954 (15.2)	1,446 (18.5)	508 (10.1)	8.4
Abdominal pain	1,268 (9.9)	739 (9.4)	529 (10.5)	−1.1
Fever/hyperpyrexia/hyperthermia	1,090 (8.5)	761 (9.7)	329 (6.5)	3.2
Problems in pregnancy > 20th week	944 (7.3)	70 (0.9)	847 (16.8)	−15.9
Non-specific minor disorders	579 (4.5)	395 (5.0)	184 (3.7)	1.4
Chest pain of suspected cardiovascular cause	524 (4.1)	329 (4.2)	195 (3.9)	0.3
Sincope/pre-sincope	344 (2.7)	220 (2.8)	114 (2.3)	0.5
Generalized asthenia	325 (2.5)	257 (3.3)	68 (1.4)	1.9
Politrauma—contusive	301 (2.3)	198 (2.5)	103 (2.0)	0.5
Pain at the side	278 (2.2)	100 (1.3)	178 (3.5)	−2.3
Nausea and/or vomiting repeated	269 (2.1)	150 (1.9)	119 (2.4)	−0.5
Heart palm/irregular wrist	251 (2.0)	156 (2.0)	95 (1.9)	0.1
Altered level of consciousness	234 (1.8)	165 (2.1)	69 (1.4)	0.7
State of confusion	213 (1.7)	162 (2.1)	51 (1.0)	1.1
Hematochezia/rectorrage/melena	194 (1.5)	136 (1.7)	58 (1.2)	0.6
Lower limbs injury	187 (1.5)	157 (2.0)	30 (0.6)	1.4
Cough/congestion	181 (1.4)	105 (1.3)	76 (1.5)	−0.2
Lower limbs pain	160 (1.2)	137 (1.8)	23 (0.5)	1.3
Chest pain not suspected due to cardiovascular cause	158 (1.2)	92 (1.2)	66 (1.3)	−0.1
Pallor/anemia	137 (1.1)	108 (1.4)	29 (0.6)	0.8
Request for urgent specialist advice	135 (1.0)	94 (1.2)	41 (0.8)	0.4
Macro-hematuria	130 (1.0)	70 (0.9)	60 (1.2)	−0.3
Diarrhea	121 (0.9)	85 (1.1)	36 (0.7)	0.4
Request for prescription or performance	120 (0.9)	75 (1.0)	45 (0.9)	0.1
Swollen/edematous leg	119 (0.9)	104 (1.3)	15 (0.3)	1.0
Weakness of extremities/symptoms associated with cerebrovascular disease	118 (0.9)	89 (1.1)	29 (0.6)	0.6
Symptoms of infection of the urinary tract	115 (0.9)	78 (1.1)	37 (0.7)	0.3
Diagnostics for biochemical images/examinations	108 (0.8)	77 (1.1)	31 (0.6)	0.4
Head trauma	99 (0.8)	54 (0.7)	45 (0.9)	−0.2
Other[c]	2,212 (17.2)	1,219 (15.6)	993 (19.7)	−4.2

[a]Taxi, helicopter 118, army ambulances, fire brigade, police, etc.
[b]Damages, Ent (ear, nose, and throat) problem, nephrology (enabled for transplantation), neonatology, pediatric oncology, semi-intensive therapy, maxillo facial surgery, hematology, thoracic surgery ophthalmology, heart surgery, neonatal intensive care, pediatric heart surgery, and dermatology.
[c]More than 135 cases.

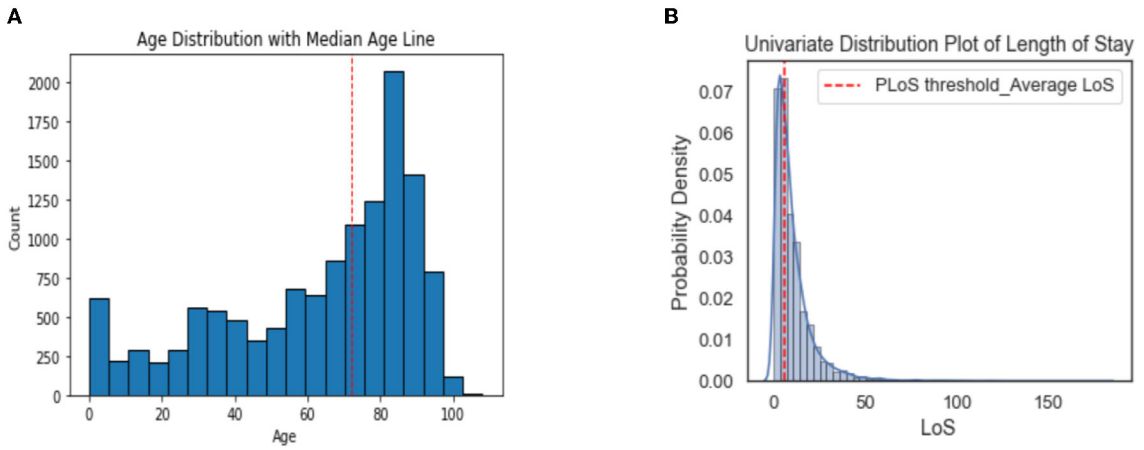

FIGURE 4
Histograms showing the distribution of Age **(A)** and LoS **(B)** in all patients.

model. MSE is the most widely used loss function for continuous outcomes. Still, we also considered MAE and MRE to get a more comprehensive understanding of the performance. The lower the value of the loss function, the better the model's prediction performance.

$MSE = \frac{\sum_{i=1}^{n}(\hat{y}_i - y_i)^2}{n}$; $MAE = \frac{\sum_{i=1}^{n}|\hat{y}_i - y_i|}{n}$; and $MRE = \frac{\sum_{i=1}^{n}\left(\frac{|\hat{y}_i - y_i|}{y_i}\right)}{n}$ where $\hat{y}_i$ and y_i are the predicted LoS and actual LoS for the ith test data.

3. Results

3.1. Patient selection

Figure 1 illustrates a flowchart of patients' eligibility for analysis in the emergency department of triaging system. A total of 84,847 patient visits were recorded at the ED between January 1 and October 26, 2022. After filtering for exclusion criteria, 12,858 patients were available for analysis.

3.2. Descriptive statistics

3.2.1. Patients characteristics summary

Of the 12,858 eligible patients included in the study, 60.88% had a prolonged length of stay (LoS). The median age of the patients was 72 years, and the elderly age groups (50–69 and 70+) had longer LoS than the other age groups. The male patients comprised 52.6% (6,757/12,858) of the total population. 51.7% of the patients arrived at the hospital via ambulance and had a longer stay compared to those who arrived by car or on foot. In the triage categories, patients with red codes, which indicate an higher severity at the ED admission, had a longer LoS, while green and white codes showed shorter stays. Light blue codes were also associated with prolonged LoS.

The most common problems among the patients were dyspnea (15.2%), abdominal pain (9.9%), and fever/hyperpyrexia/hyperthermia (8.5%). The majority of patients were seen by specialists in general medicine (29.2%), geriatrics (12.6%), astanteria or casualty department (10.7%), obstetrics and gynecology (9.0%), and pediatrics (4.7%). A detailed breakdown of patient characteristics can be found in Table 3. The count plots for each patient for each specialty and problems are included in the Appendix, in Supplementary Figures 1, 2, respectively.The distribution of length of stay (LoS) for the patients is depicted in a histogram in Figure 4. The distribution of LoS values was found to be right-skewed, with a majority of patients having an LoS ranging from 1 to 20 days. To further explore the impact of different factors on LoS, a visualization of the dichotomous outcome result for each factor is presented in Figure 5, while Figure 6 shows the continuous outcome for each factor. By examining these visualizations, we can gain insights into which factors may significantly impact LoS and further investigate the relationships between these factors and patient outcomes. Overall, these figures provide a clear and concise way to understand the distribution of LoS values and their relationship with different factors.

Figure 7 displays the average LoS for each problem and specialty. The highest average LoS was observed in Intensive Care, Vascular Surgery, Nephrology, General Medicine, Gastroenterology, Infectious Diseases, Orthopedics and Traumatology, Pneumology, Geriatrics, Cardiology, Oncology, and the Coronary Unit, respectively. The average LoS was also higher for patients experiencing issues such as swollen/edematous legs, lower limb pain, generalized weakness, requests for urgent specialist advice, altered levels of consciousness, diagnostic tests for biochemical exams or images, non-specific minor disorders, dyspnea, lower limb injuries, requests for prescription refills, and pallor/anemia.

3.3. Prediction and model performance results: binary outcome

The AUCs for all machine learning methods ranged from 0.643 for AdaBoost to 0.754 for GB (see Figure 8). GB was the best-performing classifier, followed by LoR (AUC = 0.752) and

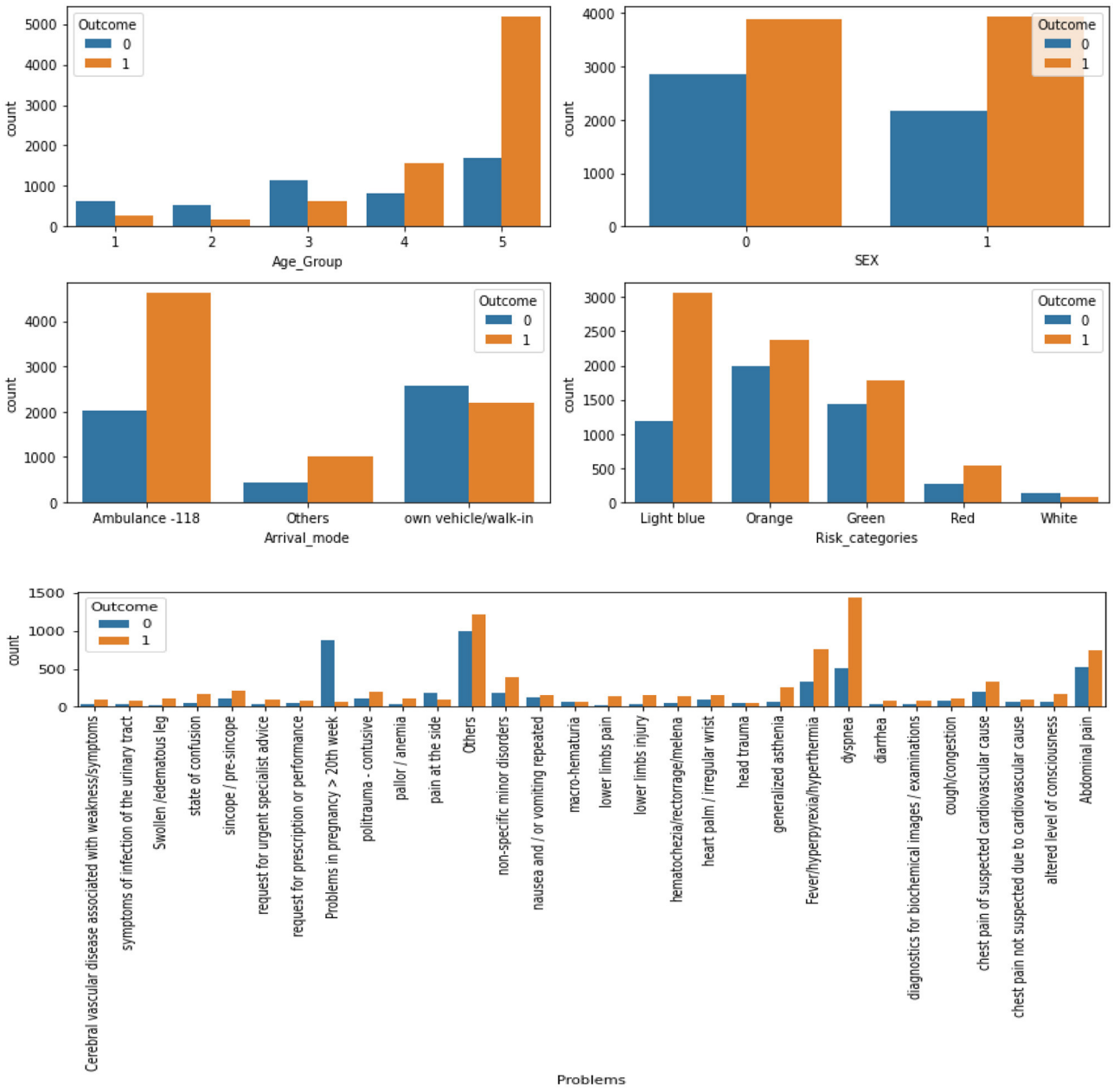

FIGURE 5
The results for each factor's dichotomous outcome (0, LoS < 6, without PLoS; 1, LoS ≥ 6, with PLoS).

SVM (AUC = 0.726). The F1-scores ranged from 0.65 (AdaBoost) to 0.73 in GB, and 0.74 in LoR (see Table 4), indicating a high capability of these models to predict the prolonged length of stay.

Of the six models, the Gradient Boosting (GB) classifier demonstrated the best prediction performance in terms of accuracy (75.4%), Area Under the Curve (AUC; 0.754), and Brier score (0.181). The Logistic Regression (LoR) model had the second-best performance, with an accuracy of 75%, AUC of 0.752, and a Brier score of 0.182. Based on these results, GB and LoR were chosen as the final models due to their better performance. However, the Ada Boost model showed poor performance with the highest Brier score, lowest accuracy, and lowest AUC values. Despite attempting hyperparameter optimization, the model's accuracy did not significantly improve.

The calibration plots for each model are displayed in Figure 9. The graph shows that GB and LoR have an almost ideal calibration or optimal fit. The Random Forest (RF) and K-Nearest Neighbor (KNN) models are well-calibrated but tend to overestimate the probabilities of a prolonged length of stay (PLoS) for most patients. Conversely, the Ada Boost and Support Vector Machine (SVM) models are poorly calibrated, with Ada Boost

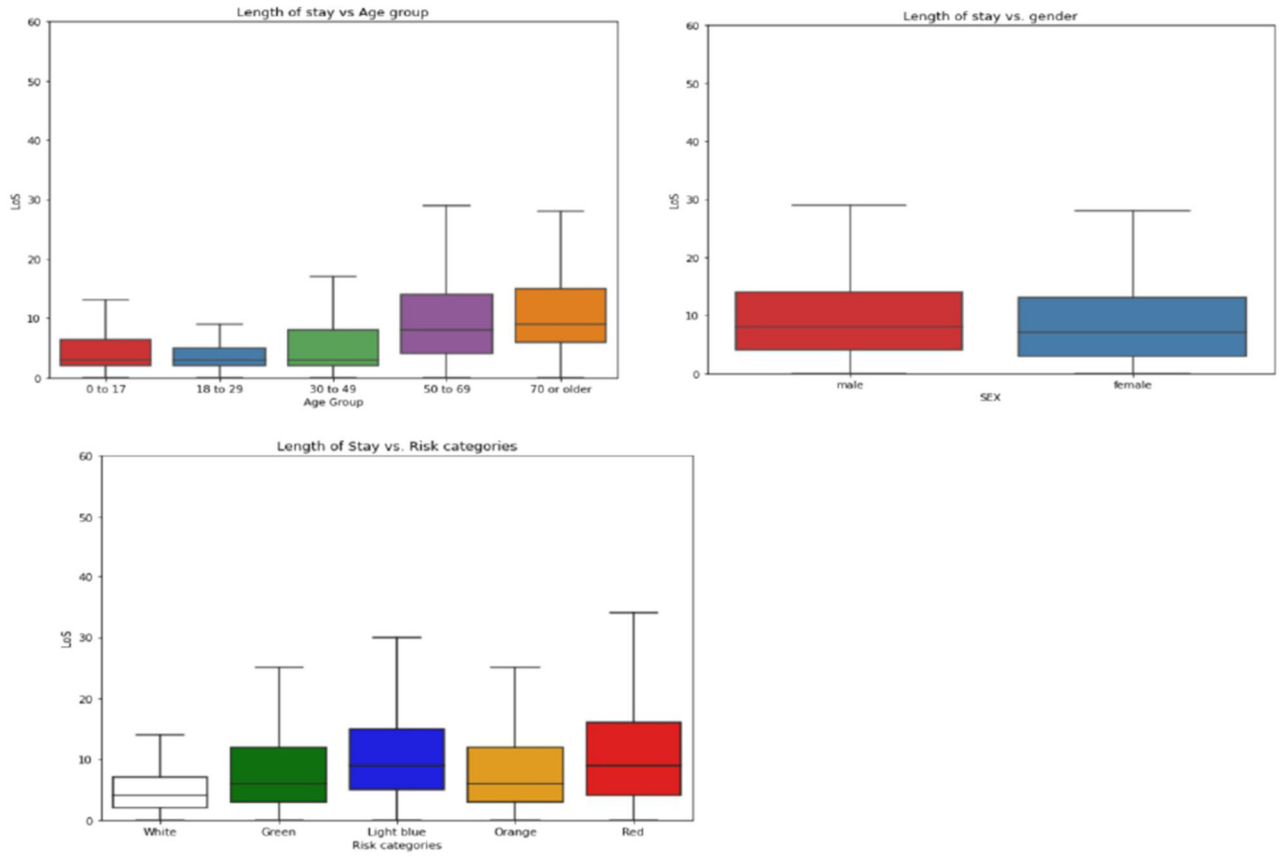

FIGURE 6
Boxplots of length of stay (LoS) on demographic factors, separated into two panels. **(Upper panel)** Shows LoS boxplots for age groups **(left)** and sex **(right)**. **(Lower panel)** Shows the box plot for risk categories of triage evaluations.

underestimating the probability of a PLoS for patients identified as low risk and overestimating it for patients in the two highest risk deciles.

The model with the highest prediction accuracy, Gradient Boosting (GB), was used to determine the relative importance of features. Figure 10 displays the results of the variable importance ranking generated by the GB model. In order of importance, the most important features were: Age Group 5 (Individuals over 70 years old), Problems in pregnancy after 20 weeks, Sex, and Age Group 4 (Individuals between 50 and 69 years old).

3.4. Prediction and model performance results: continuous outcome

The models used for predicting Length of Stay (LoS) were compared in Table 5, including various linear, penalized linear, and other machine learning models using different loss functions or total error measures. Ridge Regression and XGBoost Regression are identified as the best models based on their lower loss function values. The loss function or the total error performance measure is also visualized in Figure 11, where RMSE is on the left and MAE is on the right.

4. Discussion

In this study, we aimed to compare and evaluate predictive models using supervised machine learning algorithms for predicting prolonged length of stay in patients admitted through the emergency department (ED) in general patients settings. It is intended to promote a specific model and suggest or propose a decision-support tool as part of a predictive framework. It is well-established that reducing the length of inpatient hospital stays is one of the ways to improve the quality of life and sustainability of healthcare systems (Baek et al., 2018). Therefore, our study aims to assist physicians and doctors in making informed decisions that enable personalized interventions and guide their decision-making process to predict hospitalizations and enhance healthcare quality.

In most PLoS prediction models, predicting the outcome relies on either classification or regression. Our study utilized two separate modeling methods to predict the outcome, employing both a dichotomous value (PLoS), and a continuous value (LoS)—that is to minimize information loss in a classification task. Adopting precise and accurate modeling techniques improves

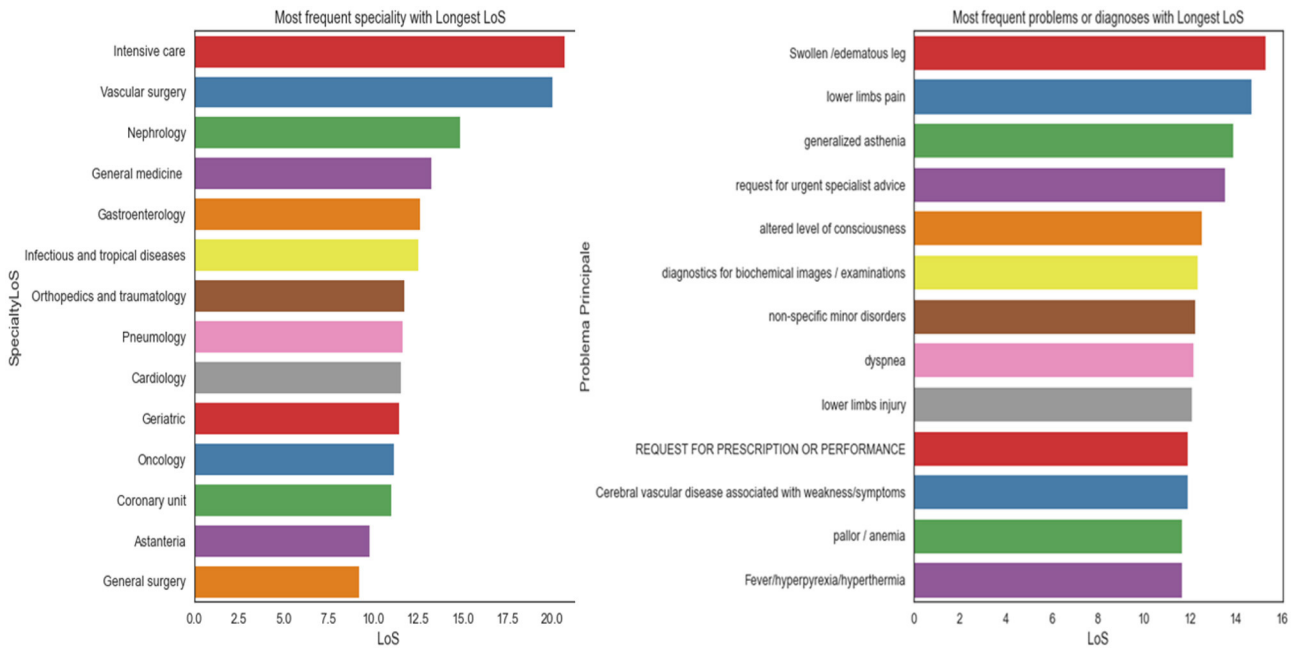

FIGURE 7
The average LoS for each problem **(right)** and for each specialty **(left)**.

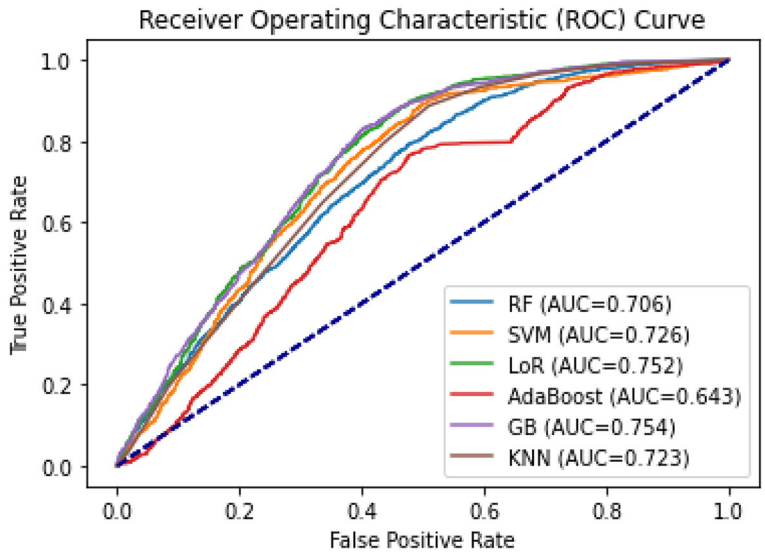

FIGURE 8
ROC curves and AUC of the six classification models for PLoS prediction.

the results and interpretations. In recent years, the prediction of patient LoS for various diseases and scenarios has been extensively explored using a variety of statistical and machine learning methods such as Logistic Regression (LoR), Random Forest (RF), Support Vector Machines (SVM), K-Nearest Neighbors (KNN), decision tree-based methods, among others (Barsasella et al., 2022).

Of the six classifiers evaluated in this study (LoR, RF, SVM, GB, AdaBoost, and KNN), five of them, excluding AdaBoost, had AUCs > 0.7, suggesting them as effective tools to predict the outcome (Florkowski, 2008). The predictive performance of the classifier models was evaluated using popular statistical indicators such as accuracy, AUC, and Brier score. GB performed the best among the six classifier models, followed by LoR. AdaBoost showed poor performance as it underestimated the probability of PLoS in patients identified as low risk and overestimated it in two patient deciles classified as high risk. Similar results were observed in other studies (Alsinglawi et al., 2020b), which used ML models to predict LoS for adult ICU cardiovascular

TABLE 4 The prediction performance of the six classification models for PLoS prediction.

Classifier algorithms	Brier score	AUC	Precision	Recall	F1-score	Accuracy
LoR	0.182	0.752	0.75	0.75	0.74	0.75
RF	0.226	0.706	0.67	0.68	0.68	0.68
GB	0.181	0.754	0.75	0.75	0.73	0.75
SVM	0.192	0.726	0.74	0.74	0.72	0.74
AdaBoost	0.255	0.643	0.65	0.65	0.65	0.65
KNN	0.198	0.723	0.70	0.71	0.70	0.71

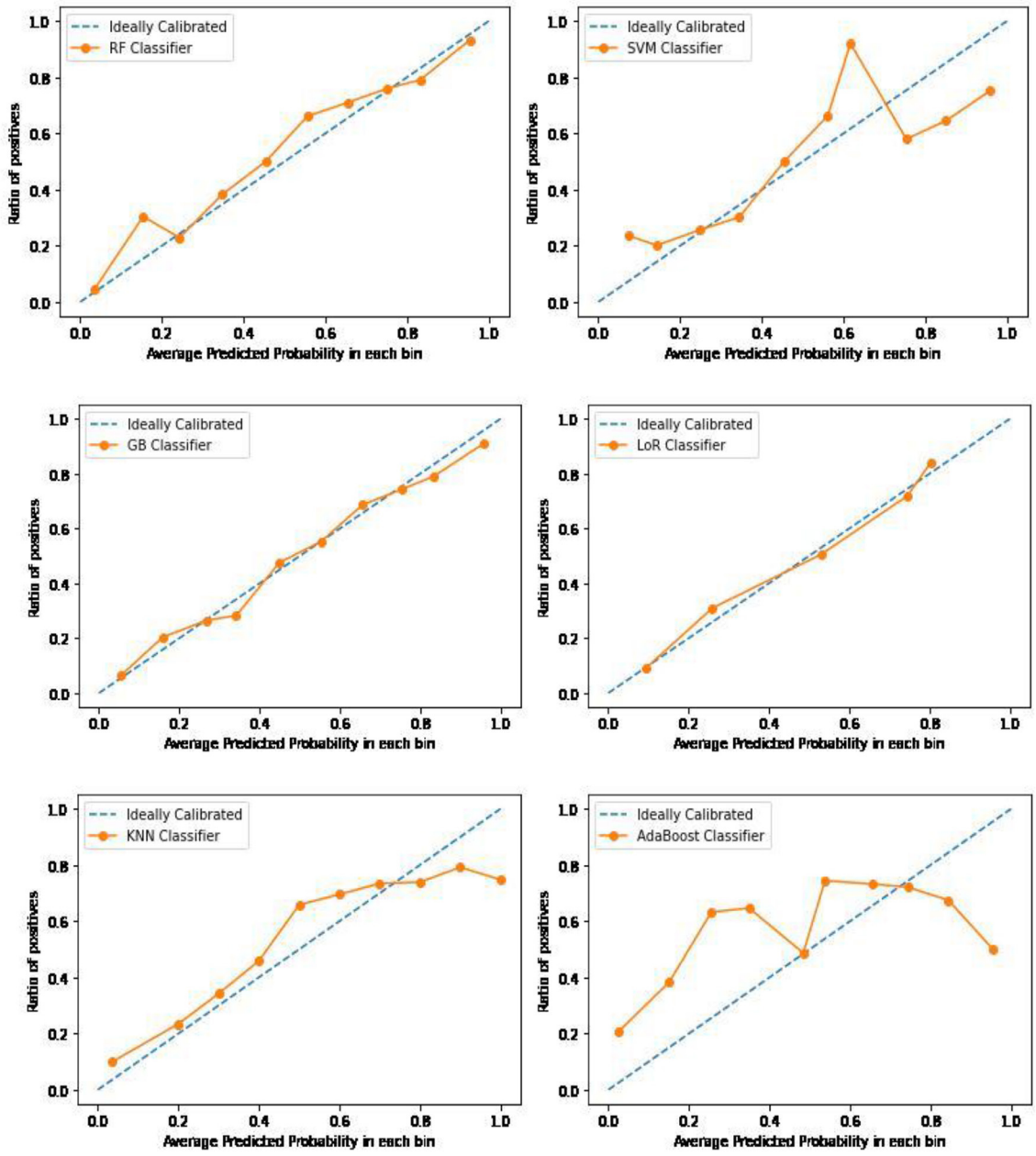

FIGURE 9
Calibration curve plots of the six classification models for PLoS prediction.

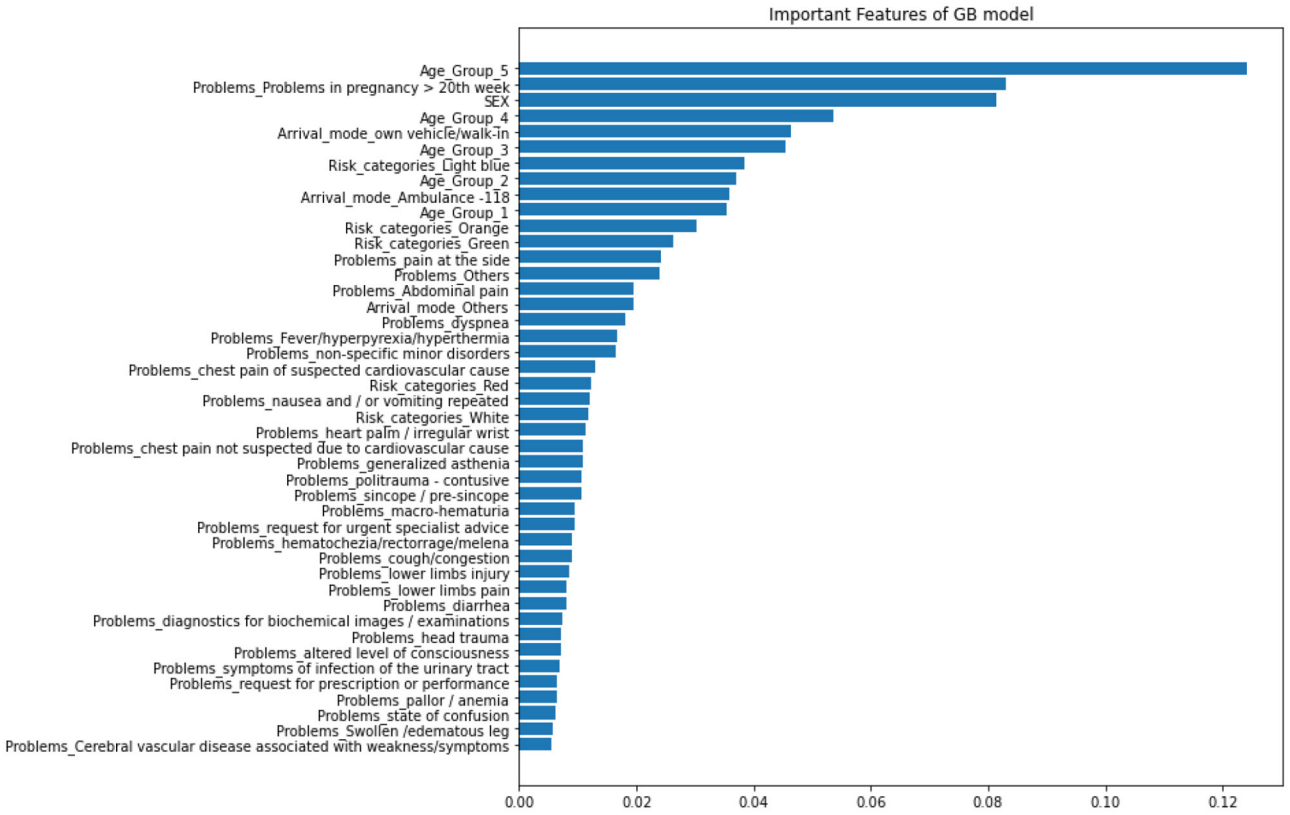

FIGURE 10
Gradient Boosting variable importance.

TABLE 5 Comparisons of classifier methods with continuous target variables for statistical and ML models applied to our datasets.

Model	Loss function			
	MSE	RMSE	MAE	MRE
Linear regression	107.045	10.346	6.671	1.283
Penalized linear models	-	-	-	-
Lasso regression	109.034	10.441	6.730	1.319
Ridge regression	**107.044**	**10.346**	**6.670**	**1.283**
Elastic net regression	109.034	10.442	6.741	1.322
Other ML learning models	-	-	-	-
Support vector regression	119.103	10.913	6.188	0.854
XGBoost regression	**107.209**	**10.354**	**6.589**	**1.213**
Random forest regressor	132.899	11.528	7.393	1.332
K-nearest neighbors regression	129.045	11.359	7.331	1.315

MSE, mean square error; RMSE, root mean square error; MAE, mean absolute error; MRE, mean relative error.
The bolded values indicate the lowest values of prediction error (e.g. Ridge and XGBoost regressions) for continuous outcomes, LoS.

hospitalizations, with the best results obtained using the GB algorithm.

Several studies, including (Kong et al., 2020; Jo et al., 2021; Wu et al., 2021; Xiong et al., 2022), have shown that the GB classifier outperforms other algorithms in predicting PLoS, with reported accuracy, AUC, and Brier score ranging from 75.3 to 82.9%, 0.74 to 0.873, and 0.122 to 0.156, respectively. Our study's findings are consistent with these results. In contrast to some other studies, Random Forest (RF), a widely used ensemble model, has been shown to perform well in certain contexts. For instance, in Xue et al. (2022), RF achieved high accuracy, AUC, and Brier scores of 0.822, 85.8%, and 0.137, respectively, suggesting its efficacy for predicting length of stay in hospital patients. These findings highlight the importance of carefully selecting the appropriate machine learning

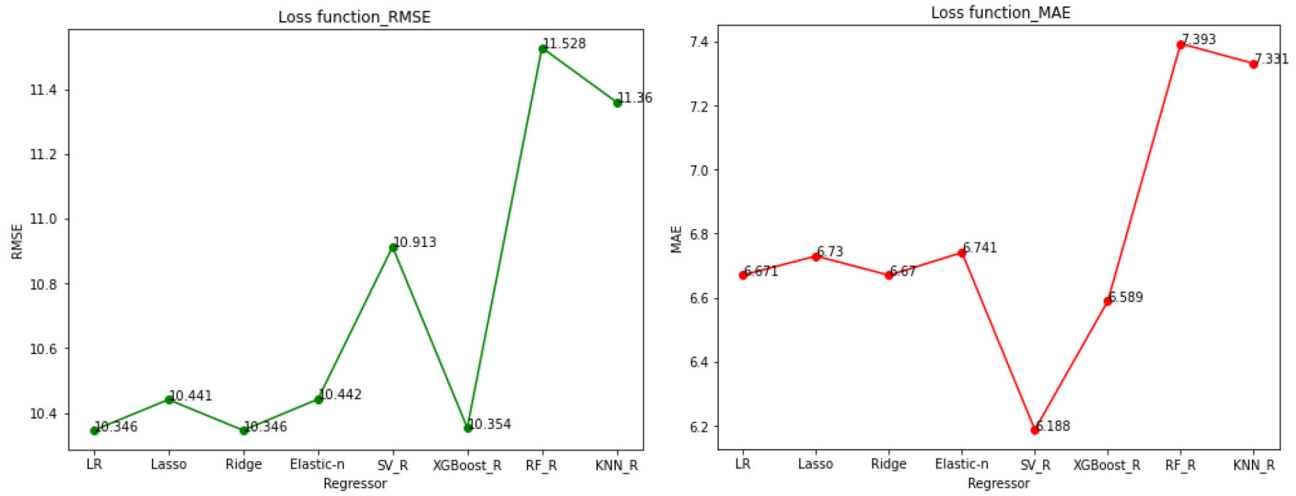

FIGURE 11
The loss function/total error visualization.

algorithm based on the specific data and problem being addressed. While RF may be a strong choice for certain applications, it may not necessarily be the best option in all cases. Therefore, it is crucial to systematically compare the performance of different algorithms and identify the optimal model for a given dataset. Such efforts can ultimately lead to more accurate and reliable predictions for clinical decision-making. Moreover, RF has demonstrated superior performance in predicting the outcome in various healthcare contexts. For example, RF has been shown to perform well in predicting LoS in newborns (Thompson et al., 2018), patients undergoing general surgery (Chuang et al., 2015), and individuals with COPD (chronic obstructive pulmonary disease; Luo et al., 2017). However, the results may vary depending on the specific patient population, clinical variables included in the model, and machine learning algorithm used. Moreover, we analyzed the importance of the features used in our best models, i.e., GB. In order of importance, the most important features were: Age Group 5 (Individuals over 70 years old), Problems in pregnancy after 20 weeks, Sex, and Age Group 4 (Individuals between 50 and 69 years old).

In addition, our study also aimed to predict continuous outcomes using eight ML regression models, as described in the methodology. After evaluating the models' performance, we found that Ridge and XGBoost regressions outperformed the others, resulting in lower prediction errors. Our findings align with previous studies, such as Chen and Klasky (2022), which reported similar results with lower prediction errors or loss functions. For instance, they reported the lowest mean absolute error between prediction and actual duration to be around 4 days, while our study showed a similar result of around 6 days. In addition, the XGBoost regression model also showed better results in Gabriel et al. (2023) for spine surgery LoS prediction. In another study on regression outcomes (Caetano et al., 2014), which examined the general patient population, six regression techniques were compared, including average prediction, decision trees, multiple regression, ANN ensembles, RF, and SVM. The RF regression model was found to yield the most accurate results with the lowest loss. Overall, our study adds to the existing body of literature highlighting the effectiveness of machine learning regression models in predicting continuous outcomes in healthcare. In particular, our results demonstrate the potential of Ridge and XGBoost regressions in improving the accuracy of LoS prediction.

To summarize, selecting the most appropriate ML algorithm that matches the specific data and problem at hand and comparing the performance of different algorithms are crucial steps in identifying the optimal model for a given dataset to ensure accurate and reliable clinical decisions. The best-performing models can then be selected as the final models. As a result, GB followed by LoR is our best-performing classification model, while Ridge Regression and XGBoost Regression were the regression model choices. These final models can now be utilized to make informed decisions or derive meaningful insights for future patients. It is important to note that the choice of the optimal model may depend on various factors, such as the type of data, the problem being addressed, and the specific goals of the analysis. Therefore, it is recommended to evaluate and compare the performance of different models when developing predictive models for various clinical applications.

One of the strengths of our study was that we used all data from ED-admitted patients, so heterogeneous patients were included in the analysis. Moreover, we evaluated several ML techniques for predicting both a categorical and a continuous outcome. However, our study has some limitations that should be recognized. One limitation of the study is that vital signs for triage evaluation information and laboratory test results were not available, which is probably one of the most important indicators (Calzavacca et al., 2010); and data was only collected from one hospital so we were not able to validate the prediction model externally. Moreover, the results of this study may be biased toward other normative periods since the data were collected during the COVID-19 pandemic. Furthermore, interpreting ML results can be difficult due to the black-box nature of some models, which can make it challenging to understand the factors that contribute to the final prediction. However, linear models such as LASSO, Ridge,

Elastic-Net Regression, and Logistic Regression provide regression coefficients, making them transparent and easily interpretable (Kotsiantis et al., 2006; Deo, 2015). Other techniques like feature selection and model-agnostic interpretability methods can also improve transparency.

In future work, we will focus on a specific specialty or disease that is prevalent in the hospital. In addition, efforts will be made to incorporate missing features such as vital signs in triage evaluation and laboratory test results. The aim is to enhance the dataset by adding more information regarding features and patients to produce better results and tackle more advanced prediction tasks such as Length of Stay (LoS) after surgeries and utilization of critical hospital resources.

5. Conclusions

As a result of our research, we have found that ML models are effective in predicting outcomes. Our findings showed that the GB classifier performed best, followed by LoR. These models can be utilized as a decision-support tool to inform healthcare decisions and predict new patient hospitalizations. Additionally, for continuous outcomes, Ridge regression and XGBoost regression displayed the best prediction performance with the lowest total prediction error. Healthcare providers can utilize our models to predict the hospitalization of new patients or to drive quality improvement initiatives. It is worth mentioning that this study is the first of its kind conducted in this hospital and can serve as a reference for future similar studies and provide valuable insights for informed decision-making.

Data availability statement

The data presented in this article are not publicly available due to ethical restrictions. Requests to access the data should be directed to the corresponding author.

Ethics statement

The studies involving human participants were reviewed and ethically approved by the Bioethics Committee of the University of Bologna, Italy (approval number: 0058022, February 24, 2023). Written informed consent from the participants' legal guardian/next of kin was not required to participate in this study in accordance with the national legislation and the institutional requirements.

Author contributions

AZ: methodology, formal analysis, data curation, and writing—original draft preparation. PP: methodology, data curation, and writing—review and editing. PT: data curation and writing—review and editing. RM: methodology, supervision, and writing—review and editing. LC: conceptualization, methodology, supervision, and writing—review and editing. All authors have read and agreed to the published version of the manuscript.

Funding

This research was partly supported by Policlinico Sant'Orsola-Malpighi through funding of the Ph.D. scholarship of AZ.

Acknowledgments

We wish to express our gratitude to all the individuals who are engaged in data organization activities at Policlinico Sant'Orsola-Malpighi, including nurses, physicians, and other administrative staff members.

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/frai.2023.1179226/full#supplementary-material

References

Ackroyd-Stolarz, S., Read Guernsey, J., Mackinnon, N. J., and Kovacs, G. (2011). The association between a prolonged stay in the emergency department and adverse events in older patients admitted to hospital: a retrospective cohort study. *Br. Med. J. Qual. Saf.* 20, 564–569. doi: 10.1136/bmjqs.2009.034926

Almashrafi, A., Elmontsri, M., and Aylin, P. (2016). Systematic review of factors influencing length of stay in ICU after adult cardiac surgery. *BMC Health Serv. Res.* 16, 318. doi: 10.1186/s12913-016-1591-3

Alsinglawi, B., Alnajjar, F., Mubin, O., Novoa, M., Alorjani, M., Karajeh, O., et al. (2020a). "Predicting length of stay for cardiovascular hospitalizations in the intensive care unit: machine learning approach," in *2020 42nd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)*. (IEEE), 5442–5445.

Alsinglawi, B., Alnajjar, F., Mubin, O., Novoa, M., Alorjani, M., Karajeh, O., et al. (2020b). Predicting length of stay for cardiovascular hospitalizations in the intensive care unit: Machine learning approach. *Annu. Int. Conf. IEEE Eng. Med. Biol. Soc.* 2020, 5442–5445. doi: 10.1109/EMBC44109.2020.9175889

Asaro, P. V., Lewis, L. M., and Boxerman, S. B. (2007). The impact of input and output factors on emergency department throughput. *Acad. Emerg. Med.* 14, 235–242. doi: 10.1197/j.aem.2006.10.104

Baek, H., Cho, M., Kim, S., Hwang, H., Song, M., Yoo, S., et al. (2018). Analysis of length of hospital stay using electronic health records: a statistical and data mining approach. *PLoS ONE* 13, e0195901. doi: 10.1371/journal.pone.0195901

Barsasella, D., Bah, K., Mishra, P., Uddin, M., Dhar, E., Suryani, D. L., et al. (2022). A machine learning model to predict length of stay and mortality among diabetes and hypertension inpatients. *Medicina* 58, 111568. doi: 10.3390/medicina581 11568

Basil, G. W., and Wang, M. Y. (2019). Trends in outpatient minimally invasive spine surgery. *J. Spine Surg.* 5, S108–S114. doi: 10.21037/jss.2019.04.17

Biber, R., Bail, H. J., Sieber, C., Weis, P., Christ, M., Singler, K., et al. (2013). Correlation between age, emergency department length of stay and hospital admission rate in emergency department patients aged ≥70 years. *Gerontology* 59, 17–22. doi: 10.1159/000342202

Bo, M., Fonte, G., Pivaro, F., Bonetto, M., Comi, C., Giorgis, V., et al. (2016). Prevalence of and factors associated with prolonged length of stay in older hospitalized medical patients. *Geriatr. Gerontol. Int.* 16, 314–321. doi: 10.1111/ggi.12471

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

Caetano N., Laureano R., and Cortez P. (2014). "A data-driven approach to predict hospital length of stay - a portuguese case study," in *Proceedings of the 16th International Conference on Enterprise Information Systems - Volume 1: ICEIS* (SciTePress), 407–414. doi: 10.5220/0004892204070414

Caetano, N., Cortez, P., and Laureano, R. M. S. (2015). Using data mining for prediction of hospital length of stay: an application of the CRISP-DM. *Methodology* 9, 149–166. doi: 10.1007/978-3-319-22348-3_9

Calzavacca, P., Licari, E., Tee, A., Egi, M., Downey, A., Quach, J., et al. (2010). The impact of rapid response system on delayed emergency team activation patient characteristics and outcomes—a follow-up study. *Resuscitation* 81, 31–35. doi: 10.1016/j.resuscitation.2009.09.026

Chen, L., and Klasky, H. (2022). *Six Machine-Learning Methods for Predicting Hospital-Stay Duration for Patients With Sepsis: A Comparative Study*. Mobile, AL: IEEE.

Chuang, M. T., Hu, Y., and Lo, C. L. (2018). Predicting the prolonged length of stay of general surgery patients: a supervised learning approach. *Int. Trans. Operat. Res.* 25, 75–90. doi: 10.1111/itor.12298

Chuang, M. T., Hu, Y. H., Tsai, C. F., Lo, C. L., and Lin, W. C. (2015). "The identification of prolonged length of stay for surgery patients," in *2015 IEEE International Conference on Systems, Man, and Cybernetics.* (Hong Kong: IEEE), 3000–3003. doi: 10.1109/SMC.2015.522

Colella, Y., Scala, A., de Lauri, C., Bruno, F., Cesarelli, G., Ferrucci, G., et al. (2021). "Studying variables affecting the length of stay in patients with lower limb fractures by means of Machine Learning," in *2021 5th International Conference on Medical and Health Informatics.* (New York, NY: ACM), 39–43.

Combes, C., Kadri, F., and Chaabane, S. (2014). *Predicting Hospital Length of Stay Using Regression Models: Application to Emergency Department*. 10ème Conférence Francophone de Modélisation, Optimisation et Simulation–MOSIM'14, Nov 2014, Nancy, France (hal-01081557) HAL (Hyper Articles en Ligne).

Dadeh, A., and Phunyanantakorn, P. (2020). Factors affecting length of stay in the emergency department in patients who presented with abdominal pain. *Emerg. Med. Int.* 2020, 1–7. doi: 10.1155/2020/5406516

Daghistani, T. A., Elshawi, R., Sakr, S., Ahmed, A. M., Al-Thwayee, A., Al-Mallah, M. H., et al. (2019). Predictors of in-hospital length of stay among cardiac patients: a machine learning approach. *Int. J. Cardiol.* 288, 140–147. doi: 10.1016/j.ijcard.2019.01.046

Deo, R. C. (2015). Machine learning in medicine. *Circulation* 132, 1920–1930. doi: 10.1161/CIRCULATIONAHA.115.001593

Emori, T. G., and Gaynes, R. P. (1993). An overview of nosocomial infections, including the role of the microbiology laboratory. *Clin. Microbiol. Rev.* 6, 428–442. doi: 10.1128/CMR.6.4.428

Etu, E. E., Monplaisir, L., Arslanturk, S., Masoud, S., Aguwa, C., Markevych, I., et al. (2022). Prediction of length of stay in the emergency department for COVID-19 patients: a machine learning approach. *IEEE Access* 10, 42243–42251. doi: 10.1109/ACCESS.2022.3168045

Florkowski, C. M. (2008). Sensitivity, specificity, receiver-operating characteristic (ROC) curves and likelihood ratios: communicating the performance of diagnostic tests. *Clin. Biochem. Rev.* 29, S83–S87. Available online at: https://pubmed.ncbi.nlm.nih.gov/18852864/

Fridman, S. E., di Giampietro, P., Sensoli, A., Beleffi, M., Bucce, C., Salvatore, V., et al. (2022). Prediction of conventional oxygen therapy failure in COVID-19 patients with acute respiratory failure by assessing serum lactate concentration, PaO_2/FiO_2 ratio, and body temperature. *Cureus* 2022, 21987. doi: 10.7759/cureus.21987

Friedman, J., Latash, M. L., and Zatsiorsky, V. M. (2009). Prehension synergies: a study of digit force adjustments to the continuously varied load force exerted on a partially constrained hand-held object. *Exp. Brain Res.* 197, 1–13. doi: 10.1007/s00221-009-1818-1

Gabriel, R. A., Harjai, B., Simpson, S., Du, A. L., Tully, J. L., George, O., et al. (2023). An ensemble learning approach to improving prediction of case duration for spine surgery: algorithm development and validation. *JMIR Perioper. Med.* 6, e39650. doi: 10.2196/39650

García-González, P., Fácila Rubio, L., Montagud, V., Chacón-Hernández, N., Fabregat-Andrés, Ó., Morell, S., et al. (2014). Predictors of prolonged hospitalization in cardiology. *Revista Española de Cardiología* 67, 62–63. doi: 10.1016/j.recesp.2013.05.024

Gellman, D. D. (1974). Cost-benefit in health care: we need to know much more. *Can. Med. Assoc. J.* 111, 988–989.

Genuer, R., Poggi, J. M., and Tuleau-Malot, C. (2010). Variable selection using random forests. *Pattern Recognit. Lett.* 31, 2225–2236. doi: 10.1016/j.patrec.2010.03.014

Gohil, R., Rishi, M., and Tan, B. H. L. (2014). Pre-operative serum albumin and neutrophil-lymphocyte ratio are associated with prolonged hospital stay following colorectal cancer surgery. *Br. J. Med. Med. Res.* 4, 481–487. doi: 10.9734/BJMMR/2014/5444

Hachesu, P. R., Ahmadi, M., Alizadeh, S., and Sadoughi, F. (2013). Use of data mining techniques to determine and predict length of stay of cardiac patients. *Healthc. Inform. Res.* 19, 121. doi: 10.4258/hir.2013.19. 2.121

Han, X., Zheng, X., Wang, Y., Sun, X., Xiao, Y., Tang, Y., et al. (2019). Random forest can accurately predict the development of end-stage renal disease in immunoglobulin a nephropathy patients. *Ann. Transl. Med.* 7, 234–234. doi: 10.21037/atm.2018. 12.11

Hastie, T., Tibshirani, R., and Friedman, J. (2009). *The Elements of Statistical Learning: Data Mining, Inference, and Prediction.* Springer. doi: 10.1007/978-0-387-84858-7

Jiang, X., Qu, X., and Davis, L. B. (2010). "Using data mining to analyze patient discharge data for an urban hospital," in *Proceedings of the 2010 International Conference on Data Mining (DMIN).* (Las Vegas, NV), 139–144.

Jo, Y. Y., Han, J., Park, H. W., Jung, H., Lee, J. D., Jung, J., et al. (2021). Prediction of prolonged length of hospital stay after cancer surgery using machine learning on electronic health records: retrospective cross-sectional study. *JMIR Med. Inform.* 9, e23147. doi: 10.2196/23147

Katayama, Y., Kitamura, T., Tanaka, J., Nakao, S., Nitta, M., Fujimi, S., et al. (2021). Factors associated with prolonged hospitalization among patients transported by emergency medical services: a population-based study in Osaka, Japan. *Medicine* 100, e27862. doi: 10.1097/MD.0000000000027862

Keller, J. M., Gray, M. R., and Givens, J. A. (1985). A fuzzy K-nearest neighbor algorithm. *IEEE Trans. Syst. Man. Cybern.* 15, 580–585. doi: 10.1109/TSMC.1985.6313426

Kirchebner, J., Günther, M. P., Sonnweber, M., King, A., and Lau, S. (2020). Factors and predictors of length of stay in offenders diagnosed with schizophrenia—A machine-learning-based approach. *BMC Psychiatry* 20, 201. doi: 10.1186/s12888-020-02612-1

Kong, G., Lin, K., and Hu, Y. (2020). Using machine learning methods to predict in-hospital mortality of sepsis patients in the ICU. *BMC Med. Inform. Decis. Mak.* 20, 251. doi: 10.1186/s12911-020-01271-2

Kotsiantis, S. B., Zaharakis, I. D., and Pintelas, P. E. (2006). Machine learning: a review of classification and combining techniques. *Artif. Intell. Rev.* 26, 159–190. doi: 10.1007/s10462-007-9052-3

Krochmal, P., and Riley, T. A. (1994). Increased health care costs associated with ED overcrowding. *Am. J. Emerg. Med.* 12, 265–266. doi: 10.1016/0735-6757(94)9 0135-X

LaFaro, R. J., Pothula, S., Kubal, K. P., Inchiosa, M. E., Pothula, V. M., Yuan, S. C., et al. (2015). Neural network prediction of ICU length of stay following cardiac surgery based on pre-incision variables. *PLoS ONE* 10, e0145395. doi: 10.1371/journal.pone.0145395

Laky, B., Janda, M., Kondalsamy-Chennakesavan, S., Cleghorn, G., and Obermair, A. (2010). Pretreatment malnutrition and quality of life - association with prolonged length of hospital stay among patients with gynecological cancer: a cohort study. *BMC Cancer* 10, 232. doi: 10.1186/1471-2407-10-232

Launay, C. P., Kabeshova, A., Lanoé, A., Chabot, J., Levinoff, E. J., Beauchet, O., et al. (2018). Age effect on the prediction of risk of prolonged length hospital stay in older patients visiting the emergency department: results from a large prospective geriatric cohort study. *BMC Geriatr.* 18, 127. doi: 10.1186/s12877-018-0820-5

Liew, D., Liew, D., and Kennedy, M. P. (2003). Emergency department length of stay independently predicts excess inpatient length of stay. *Med. J. Aust.* 179, 524–526. doi: 10.5694/j.1326-5377.2003.tb05676.x

Lu, M., Sajobi, T., Lucyk, K., Lorenzetti, D., and Quan, H. (2015). Systematic review of risk adjustment models of hospital length of stay (LOS). *Med. Care*. 53, 355–365. doi: 10.1097/MLR.0000000000000317

Luo, L., Lian, S., Feng, C., Huang, D., and Zhang, W. (2017). "Data mining-based detection of rapid growth in length of stay on COPD patients," in *2017 IEEE 2nd International Conference on Big Data Analysis (ICBDA)* (Beijing: IEEE), 254–258. doi: 10.1109/ICBDA.2017.8078819

Luo, R., Tan, X., Wang, R., Qin, T., Chen, E., Liu, T. Y., et al. (2020). Accuracy prediction with non-neural model for neural architecture search. *arXiv 2020*. doi: 10.48550/arXiv.2007.04785

Marfil-Garza, B. A., Belaunzarán-Zamudio, P. F., Gulias-Herrero, A., Zuñiga, A. C., Caro-Vega, Y., Kershenobich-Stalnikowitz, D., et al. (2018). Risk factors associated with prolonged hospital length-of-stay: 18-year retrospective study of hospitalizations in a tertiary healthcare center in Mexico. *PLoS ONE* 13, e0207203. doi: 10.1371/journal.pone.0207203

Mekhaldi, R. N., Caulier, P., Chaabane, S., Chraibi, A., and Piechowiak, S. (2020). "Using machine learning models to predict the length of stay in a hospital setting," in *Trends and Innovations in Information Systems and Technologies. WorldCIST 2020. Advances in Intelligent Systems and Computing*, Vol. 1159, eds A. Rocha, H. Adeli, L. Reis, S. Costanzo, I. Orovic, and F. Moreira (Cham: Springer). doi: 10.1007/978-3-030-45688-7_21

Morton, A., Marzban, E., Giannoulis, G., Patel, A., Aparasu, R., Kakadiaris, I. A. A., et al. (2014). "Comparison of supervised machine learning techniques for predicting short-term in-hospital length of stay among diabetic patients," in *2014 13th International Conference on Machine Learning and Applications* (Detroit, MI: IEEE), 428–431. doi: 10.1109/ICMLA.2014.76

Paterson, R., MacLeod, D., Thetford, D., Beattie, A., Graham, C., Lam, S., et al. (2006). Prediction of in-hospital mortality and length of stay using an early warning scoring system: clinical audit. *Clin. Med*. 6, 281–284. doi: 10.7861/clinmedicine.6-3-281

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., et al. (2011). Scikit-learn: machine learning in python. *J. Mach. Learn. Res*. 12, 2825–2830.doi: 10.5555/1953048.2078195

Qi, Y. (2012). "Random forest for bioinformatics," in *Ensemble Machine Learning,* eds C. Zhang and Y. Ma (New York, NY: Springer), 307–323. doi: 10.1007/978-1-4419-9326-7_11

Rahman, M. A., Honan, B., Glanville, T., Hough, P., and Walker, K. (2020). Using data mining to predict emergency department length of stay greater than 4 hours: derivation and single-site validation of a decision tree algorithm. *Emerg. Med. Australas*. 32, 416–421. doi: 10.1111/1742-6723.13421

Sir, Ö., Hesselink, G., van den Bogaert, M., Akkermans, R. P., and Schoon, Y. (2019). Risk factors for prolonged length of stay of older patients in an academic emergency department: a retrospective cohort study. *Emerg. Med. Int*. 2019, 4937827. doi: 10.1155/2019/4937827

Song, X., Xia, C., Li, Q., Yao, C., Yao, Y., Chen, D., et al. (2020). Perioperative predictors of prolonged length of hospital stay following total knee arthroplasty: a retrospective study from a single center in China. *BMC Musculoskelet. Disord*. 21, 62. doi: 10.1186/s12891-020-3042-x

Steyerberg, E. W. (2019). *Clinical Prediction Models. 2nd Edn*. New York, NY: Springer-Verlag.

Steyerberg, E. W., Vickers, A. J., Cook, N. R., Gerds, T., Gonen, M., Obuchowski, N., et al. (2010). Assessing the performance of prediction models: a framework for traditional and novel measures. *Epidemiology* 21, 128–138. doi: 10.1097/EDE.0b013e3181c30fb2

Sud, M., Yu, B., Wijeysundera, H. C., Austin, P. C., Ko, D. T., Braga, J., et al. (2017). Associations between short or long length of stay and 30-day readmission and mortality in hospitalized patients with heart failure. *JACC Heart Fail*. 5, 578–588. doi: 10.1016/j.jchf.2017.03.012

Symum, H., and Zayas-Castro, J. L. (2020). Prediction of chronic disease-related inpatient prolonged length of stay using machine learning algorithms. *Healthc. Inform. Res*. 26, 20–33. doi: 10.4258/hir.2020.26.1.20

Tanuja, S., Acharya, D. U., and Shailesh, K. R. (2011). Comparison of different data mining techniques to predict hospital length of stay. *J. Pharm. Biomed. Sci*. 7.

Thompson, B., Elish, K. O., and Steele, R. (2018). "Machine learning-based prediction of prolonged length of stay in newborns," in *2018 17th IEEE International Conference on Machine Learning and Applications (ICMLA)* (Orlando, FL: IEEE), 1454–1459. doi: 10.1109/ICMLA.2018.00236

Thongpeth, W., Lim, A., Wongpairin, A., Thongpeth, T., and Chaimontree, S. (2021). Comparison of linear, penalized linear and machine learning models predicting hospital visit costs from chronic disease in Thailand. *Inform. Med. Unlocked* 26, 100769. doi: 10.1016/j.imu.2021.100769

Tibshirani, R. (1996). Regression shrinkage and selection via the Lasso. *J. Royal Stat. Soc*. 58, 267–288. doi: 10.1111/j.2517-6161.1996.tb02080.x

Tsai, P. F. J., Chen, P. C., Chen, Y. Y., Song, H. Y., Lin, H. M., Lin, F. M., et al. (2016). Length of hospital stay prediction at the admission stage for cardiology patients using artificial neural network. *J. Healthc. Eng*. 2016, 7035463. doi: 10.1155/2016/7035463

van der Veen, D., Remeijer, C., Fogteloo, A. J., Heringhaus, C., and de Groot, B. (2018). Independent determinants of prolonged emergency department length of stay in a tertiary care centre: a prospective cohort study. *Scand. J. Trauma Resusc. Emerg. Med*. 26, 81. doi: 10.1186/s13049-018-0547-5

Vekaria, B., Overton, C., Wiśniowski, A., Ahmad, S., Aparicio-Castro, A., Curran-Sebastian, J., et al. (2021). Hospital length of stay for COVID-19 patients: data-driven methods for forward planning. *BMC Infect. Dis*. 21, 700. doi: 10.1186/s12879-021-06371-6

Wolff, J., McCrone, P., Patel, A., Kaier, K., and Normann, C. (2015). Predictors of length of stay in psychiatry: analyses of electronic medical records. *BMC Psychiatry* 15, 238. doi: 10.1186/s12888-015-0623-6

Wu, J., Kong, G., Lin, Y., Chu, H., Yang, C., Shi, Y., et al. (2020). Development of a scoring tool for predicting prolonged length of hospital stay in peritoneal dialysis patients through data mining. *Ann. Transl. Med*. 8, 1437. doi: 10.21037/atm-20-1006

Wu, J., Lin, Y., Li, P., Hu, Y., Zhang, L., Kong, G., et al. (2021). Predicting prolonged length of ICU stay through machine learning. *Diagnostics* 11, 2242. doi: 10.3390/diagnostics11122242

Xiong, F., Cao, X., Shi, X., Long, Z., Liu, Y., and Lei, M. (2022). A machine learning–based model to predict early death among bone metastatic breast cancer patients: a large cohort of 16,189 patients. *Front. Cell Dev. Biol*. 10, 1059597. doi: 10.3389/fcell.2022.1059597

Xue, X., Liu, Z., Xue, T., Chen, W., and Chen, X. (2022). Machine learning for the prediction of acute kidney injury in patients after cardiac surgery. *Front. Surg*. 9, 946610. doi: 10.3389/fsurg.2022.946610

Yu, K. H., Beam, A. L., and Kohane, I. S. (2018). Artificial intelligence in healthcare. *Nat. Biomed. Eng*. 2, 719–731. doi: 10.1038/s41551-018-0305-z

Zeleke, A. J., Moscato, S., Miglio, R., and Chiari, L. (2022). Length of stay analysis of COVID-19 hospitalizations using a count regression model and quantile regression: a study in Bologna, Italy. *Int. J. Environ. Res. Public Health* 19, 2224. doi: 10.3390/ijerph19042224

Zoller, B., Spanaus, K., Gerster, R., Fasshauer, M., Stehberger, P. A., Klinzing, S., et al. (2014). ICG-liver test versus new biomarkers as prognostic markers for prolonged length of stay in critically ill patients—a prospective study of accuracy for prediction of length of stay in the ICU. *Ann. Intensive Care* 4, 19. doi: 10.1186/s13613-014-0019-7

frontiers | Frontiers in Medicine

TYPE Original Research
PUBLISHED 01 August 2023
DOI 10.3389/fmed.2023.1190125

OPEN ACCESS

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

REVIEWED BY
Roman Pfeifer,
University Hospital Zürich, Switzerland
Lubna Samad,
Indus Hospital, Pakistan

*CORRESPONDENCE
Shi-Ke Hou
houshike@tju.edu.cn
Hao-Jun Fan
haojunfan86@163.com
Wei Han
sugh_hanwei@szu.edu.cn

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

RECEIVED 20 March 2023
ACCEPTED 10 July 2023
PUBLISHED 01 August 2023

CITATION
Han W, Yuan J-Y, Li R, Yang L, Fang J-Q, Fan H-J and Hou S-K (2023) Clinical application of a body area network-based smart bracelet for pre-hospital trauma care.
Front. Med. 10:1190125.
doi: 10.3389/fmed.2023.1190125

Clinical application of a body area network-based smart bracelet for pre-hospital trauma care

Wei Han[1,2*†], Jin-Yang Yuan[2†], Rui Li[2], Le Yang[2], Jia-Qin Fang[3], Hao-Jun Fan[1*] and Shi-Ke Hou[1*]

[1]Institute of Disaster and Emergency Medicine, Tianjin University, Tianjin, China, [2]Emergency Department of Shenzhen University General Hospital, Shenzhen, Guangdong, China, [3]School of Microelectronics, South China University of Technology, Guangzhou, Guangdong, China

Objective: This study aims to explore the efficiency and effectiveness of a body area network-based smart bracelet for trauma care prior to hospitalization.

Methods: To test the efficacy of the bracelet, an observational cohort study was conducted on the clinical data of 140 trauma patients pre-admission to the hospital. This study was divided into an experimental group receiving smart bracelets and a control group receiving conventional treatment. Both groups were randomized using a random number table. The primary variables of this study were as follows: time to first administration of life-saving intervention, time to first administration of blood transfusion, time to first administration of hemostatic drugs, and mortality rates within 24 h and 28 days post-admission to the hospital. The secondary outcomes included the amount of time before trauma team activation and the overall length of patient stay in the emergency room.

Results: The measurement results for both the emergency smart bracelet as well as traditional equipment showed high levels of consistency and accuracy. In terms of pre-hospital emergency life-saving intervention, there was no significant statistical difference in the mortality rates between both groups within 224 h post-admission to the hospital or after 28-days of treatment in the emergency department. Furthermore, the treatment efficiency for the group of patients wearing smart bracelets was significantly better than that of the control group with regard to both the primary and secondary outcomes of this study. These results indicate that this smart bracelet has the potential to improve the efficiency and effectiveness of trauma care and treatment.

Conclusion: A body area network-based smart bracelet combined with remote 5G technology can assist the administration of emergency care to trauma patients prior to hospital admission, shorten the timeframe in which life-saving interventions are initiated, and allow for a quick trauma team response as well as increased efficiency upon administration of emergency care.

KEYWORDS

wearable electronic devices, body area network, pre-hospital emergency care, trauma, clinical application

1. Introduction

A common term called the "golden hour," which is based on the "trauma death curve" theory refers to an approximately 60-min window following a severe injury in which effective treatment is needed to reduce morbidity and mortality rates (1). In complex or difficult to reach areas, traditional emergency response systems may struggle to arrive at the scene of an accident in a timely manner, leading to missed opportunities for prompt care to be administered. In such situations, a device with more portability and effectiveness is needed to provide life support on site during transport to the hospital (2). This is especially crucial for large-scale emergencies in which a large number of patients require treatment within a short period of time or when there is potential for the limited availability of emergency response because the number of patients is higher than usual. As a result, additional methods of professional management and communication were needed during these events (3). Therefore, the development of a more efficient trauma care system pre-hospitalization was of great importance.

Wireless body area network (WBAN) is an emerging technology that allows for local area network communication while consuming low quantities of energy. Remote life-sign monitoring systems developed based on WBAN technology have been shown to significantly increase the data transmission rate compared to traditional healthcare systems (4). In traditional healthcare systems, information for most patients is collected and transmitted *via* wired methods, which lack flexibility and limit the users' normal range of activities. WBAN technology can automatically collect and record physiological signals from the patient in different environments, such as home, office, or a hospital, without affecting normal activities. Various physiological parameters can be transmitted to hospitals or servers, promoting a more efficient and timely treatment. Furthermore, sensor nodes can be used to monitor the sudden onset of conditions in the patient and promptly notify hospitals and family members to provide timely treatment.

Vital signs such as blood pressure, heart rate, body temperature, and blood oxygen saturation are external readouts of various physiological activities in the human body and are basic indicators for judging whether the body is healthy. When abnormalities occur, vital signs show different degrees of change, corresponding to dynamic changes caused by disease occurrence, development, and resolution. Therefore, real-time monitoring and recording of human vital signs provide an important scientific basis for clinical diagnosis and timely treatment of patients and to ensure correct guidance is given to the nursing staff caring for patients. At present, conventional monitoring methods often require patients to stay still for a few seconds. Furthermore, medical staff is required to bring monitoring equipment to the patient's bedside in order to measure and record specific data for each patient, which is quite inefficient. WBAN technology enables intelligent monitoring through distributed sensor nodes, collecting vital sign information from the human body in real-time and transmitting them online to hospital servers (5–9). This technology is particularly useful when needed in operating rooms, intensive care units (ICU), and other hospital wards (6, 10). However, to date, there have only been a few studies on applying this technology for use in emergency medical services (4).

Wearable devices can integrate various biosensors to monitor and record physiological information such as blood pressure, pulse, blood oxygen saturation, respiratory rate, body temperature, electrocardiogram data, or electromyogram data through attachment to the body. These devices have excellent mobility (7, 8) and use body area network technologies in addition to other new technologies, such as remote 5G interaction, to provide a remote, real-time monitoring solution for pre-hospitalized patients, thus informing both diagnosis and treatment in emergency care (8, 9). This study aimed to explore the impact of a multi-parameter integrated life-monitoring smart bracelet based on BAN technology for efficient and effective emergency treatment of patients prior to hospitalization. Our findings provided evidence for the development of wearable monitoring devices and remote emergency medical technology based on WBAN, as well as for improving the quality of trauma treatment for patients before hospitalization.

2. Materials and methods

2.1. General information

Clinical data from 140 pre-hospitalized trauma patients who were admitted to Shenzhen University General Hospital between June 10, 2022, and January 31, 2023, were analyzed in this observational cohort study. Inclusion criteria were: (1) trauma patients who were transported by the Shenzhen University General Hospital 120 Center and received treatment in the emergency department; (2) aged 18 to 80 years; (3) those who provided informed consent. Exclusion criteria were the following: (1) patients with mental disorders or unwilling to cooperate; (2) pregnant women; (3) patients who were confirmed dead after their initial assessments. The inclusion and exclusion process for this study is further detailed in Figure 1.

The sample size was calculated by GPower 3.1.9.7 software. The statistical method of t-test for two independent samples was applied; effect size (d) was set to 0.5, power of the test (1-β) was 0.8, and significance level (α) was set to 0.05. Each group required at least 64 participants. Therefore, 70 participants were included in each of the two groups (control and test groups) in this study; the experimental group consisted of 70 pre-hospitalized trauma patients who were treated using smart bracelets containing body area network technology, while the control group consisted of 70 pre-hospitalized trauma patients who were treated using traditional methods. Before data collection, grouping was completed by a random method. Specifically, samples were numbered 1–140 in advance, and each sample was randomly assigned a random three-digit number using the random number table. Then, the samples are sorted based on their three-digit number. According to the sorting results, the top 70 samples are divided into a control group, while the rest of them are divided into a test group. The experimenter decides whether to use the test equipment according to the group of patients who are presented sequentially.

This study was approved by the ethics committee of Shenzhen University General Hospital (Ethics Approval No. SUGHKYLL2022061001). It was conducted in strict compliance with relevant regulations and ethical guidelines. Informed consent was obtained from all patients or their family members. Obtaining written informed consent at the pre-hospital scene can be challenging and may hinder emergency rescue work. Therefore, we only obtained verbal informed consent from patients or family members at the

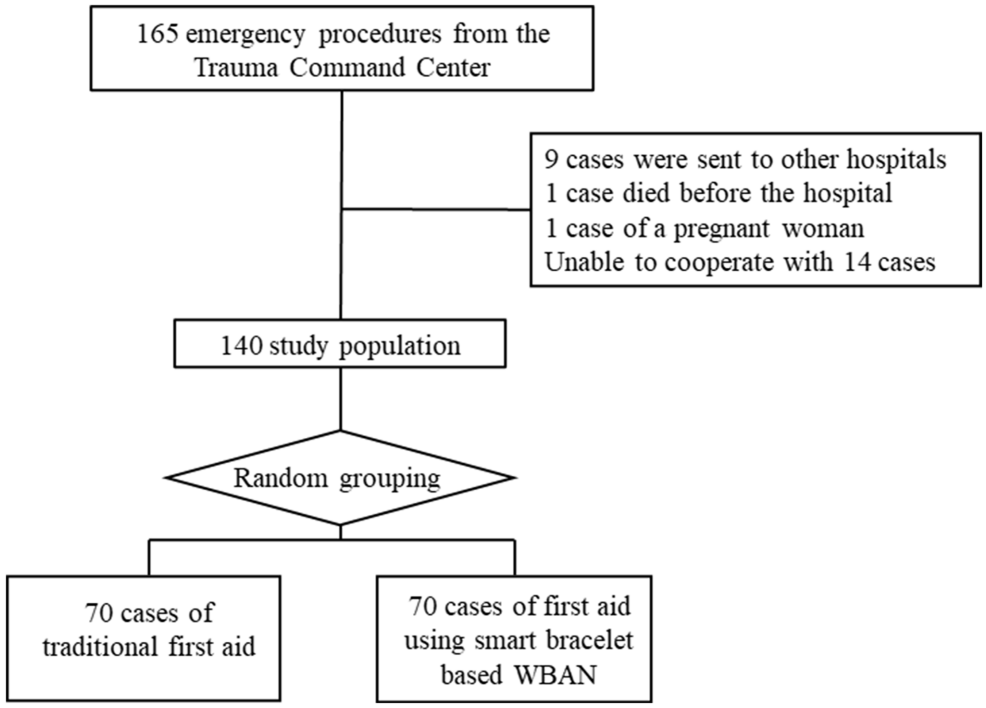

FIGURE 1
The study flow chart.

FIGURE 2
Emergency smart bracelet.

scene, with written informed consent signed at the hospital. In cases where patients could not provide verbal consent and had no family members present, such as those who are unconscious, the patients still wore bracelets, and written informed consent was given by the family members at the hospital.

2.2. Experimental equipment

The experimental device used for this study was a multi-parameter integrated life-monitoring smart bracelet based on BAN technology, which was independently developed by our team, as shown in Figure 2. This smart bracelet can simultaneously monitor blood pressure, heart rate, blood oxygen saturation, body temperature, and respiratory rate, and perform single-lead electrocardiography.

The sensor component used in the bracelet was based on a Nordic52832 control chip, which includes an oxygen chip (TI high-performance analog front end AFE4404 + 2*Osram2703 PD + Osram three-in-one LED), temperature sensor (CT1711 array), electrocardiogram chip (Ti chip 129X), photoelectric chip (Ti AFE4404 + double Osram2703), heart rate chip (Yiguang PD70), and a gravity sensor (Rome KXTJ3-1057). The installed communication module uses a low-power 4.2BLE Bluetooth module, which requires the central node device (Figure 3) to be compatible with Android 4.4 or higher, IOS 8.0 or higher, as well as support Bluetooth 4.0. The hardware performance parameters were as follows: (1) the bracelet contains a memory of 512 KB (Flash 64 M); (2) the screen display was

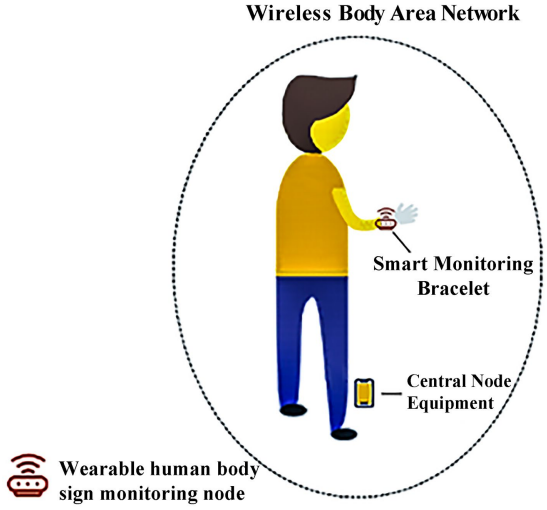

FIGURE 3
Wireless body area network.

approximately 1.3" IPS 240×240; (3) the battery capacity was 240mAh, which allowed for 15-day standby periods or 5–7 days of full-time monitoring; (4) the bracelet supported physical buttons; (5) a built-in motor for vibration reminders; (6) it uses magnetic charging interface; (7) the waterproof rating for the body of the bracelet easily met IP67 standards. The bracelet can collect patient vital signs (blood pressure, blood oxygen saturation, heart rate, respiratory rate, temperature) in real-time. After wearing and completing the first-time measurement, we obtained blood pressure, blood oxygen saturation, and respiratory rate measurements at a frequency of 20 Hz, and obtained heart rate and temperature at 60 Hz.

The equipment used in this study included traditional life sign monitoring equipment that is commonly employed in the hospital prior to admission, which included: electronic blood pressure cuffs (Yuwell YE680A), pulse oximeters (Edan H100B), infrared thermometers (Fudakang KM-WD01), 12-lead electrocardiograph machines (Edan SE1201), as well as a vehicle-mounted defibrillator monitor (Mindray BeneHeart D6). The respiratory rate of patients was measured prior to hospital admission through visual estimation or stethoscope.

2.3. Emergency rescue methods

The control group underwent standard emergency rescue. Upon receiving a trauma emergency rescue task from the center, staff from the emergency department performed pre-admission vital sign monitoring *via* traditional emergency equipment upon arrival at the injury scene. The patients' medical histories were obtained, their vital signs were measured, and a physical examination was performed to assess initial patient conditions. On-site treatment was provided as needed, and it included: the opening of patient's airways, establishing venous access, oxygen supplementation, as well as other interventions such as tracheal intubation, cricothyroidotomy, needle decompression, and fluid replacement. After the staff completed on-site treatment, patients were transported by ambulance to the nearest trauma center. In the ambulance, patients' cardiac statuses were monitored using a vehicle-mounted electrocardiogram measuring heart rate, blood pressure, pulse, oxygen saturation, and respiratory rate. The trauma team was activated upon arrival at the hospital, and a treatment plan was prepared based on the patients' condition *via* phone or direct network communication.

For the experimental group, a smart wristband based on BAN technology combined with traditional equipment was applied for vital sign monitoring. Furthermore, remote communication was conducted through 5G internet technology before and after hospitalization. The study researchers did not interfere in any routine emergency rescue procedures. After obtaining consent from the patients or their family members upon arrival at the scene, the wristband was put on to monitor the patients' blood pressure, heart rate, blood oxygen saturation, respiratory rate, and body temperature. The wristband data was connected to the BAN of the central node device and synchronized in real-time to the emergency physicians' terminal in the hospital *via* 5G signaling. The active emergency physician in the hospital guided patient treatment using an online screen video according to the patients' condition. Furthermore, the emergency department doctor activated the trauma team while preparing a patient rescue plan based on their conditions.

2.4. Variable definitions

The amount of time needed to administer the patient's first rescue intervention, the amount of time needed to start a blood transfusion, the amount of time until the first use of hemostatic drugs, as well as 24-h and 28-day mortality rates were the primary variables. The secondary variables included the time necessary to activate the trauma team as well as the length of stay in the emergency department.

The evaluation indicators for treatment efficiency included: (1) rescue intervention measures, such as endotracheal intubation, cricothyrotomy, needle decompression, fluid replacement, use of hemostatic drugs (tranexamic acid), and blood transfusion; (2) the amount of time needed to begin patients' first rescue intervention after their initial encounter with medical personnel prior to hospitalization; (3) amount of time necessary to begin a blood transfusion after emergency department admission; (4) amount of time between emergency department admission and the patients' first use of hemostatic agents; (5) amount of time between the initial encounter with medical personnel prior to hospitalization to activation of the in-hospital trauma team.

The evaluation indicators of treatment effectiveness included: (1) mortality within 24-h of hospitalization, defined as the proportion of patients who died for any reason within 24-h after admission to the emergency department in each group; (2) 28-day mortality, defined as the proportion of patients who died for any reason within 28-days after injury in each group; (3) total time spent in the emergency department from admission to discharge.

2.5. Data collection

Data collection was performed by the research team prior to and after hospitalization. Pre-hospitalization data was collected in the ambulance and included vital signs measured by the smart bracelet and conventional equipment (blood pressure, heart rate, oxygen

saturation, respiratory rate, and temperature), the site and type of injury, the injury severity score (ISS), time of arrival at the scene, time life-saving interventions were initiated, the time of trauma team contact at the hospital, the time of ambulance entry, as well as the time of emergency department arrival. In-hospital data was collected by a thorough review of patient records and nursing documents and included blood transfusion times, the use of hemostatic drugs, as well as the time patients left the operating room. The research team did not participate in clinical decision-making or treatment during these processes.

2.6. Statistical analyses

Statistical analyses were performed using SPSS Statistics 27.0 software (International Business Machines Corporation, United States). Continuous variables were expressed as mean ± standard deviation if normally distributed or as median values (interquartile range) if not normally distributed and were compared using student's t-test or Mann–Whitney U test as appropriate. Categorical variables were expressed as frequencies or percentages and compared using a chi-square test or Fisher's exact test. Kendall's tau-b test was used to assess the consistency of the first measurement results from each type of equipment. $p < 0.05$ represented statistical significance.

3. Results

3.1. Characteristics of the study population

There were no statistically significant differences ($p > 0.05$) in the general characteristics between the test and control groups. The main mechanism of trauma in both groups was car accident injury and falling injury, without a statistical difference between the two groups ($p > 0.05$). The most common trauma sites in both the control group (28.57%) and test group (35.71%) were the limbs. The control group consisted of 8 patients with head and neck trauma (11.43%), 13 patients with thoracic trauma (18.57%), 18 patients with abdominal trauma (25.71%), and 6 patients with trauma in multiple areas (8.57%). The test group consisted of 5 patients with head and neck trauma (7.14%), 12 patients with thoracic trauma (17.14%), 15 patients with abdominal trauma (21.43%), and 8 patients with trauma in multiple areas (11.43%). There was no significant statistical difference ($p > 0.05$) in the main trauma sites between the two groups. There were 30 patients (42.86%) in the control group and 28 patients (40.00%) in the experimental group who had severe trauma (ISS > 16 points); there was no significant statistical difference ($p > 0.05$) in the proportion of patients with severe trauma between the two groups (Table 1).

TABLE 1 Comparison of general characteristics between groups.

Project	Control group	Test group	*p*
N	70	70	
Age [$\bar{x}$ ± s, years]	43.31 ± 13.87	44.17 ± 14.48	0.721
Gender			0.290
Male [*n* (%)]	48 (68.57)	42 (60.00)	
Female [*n* (%)]	22 (31.43)	28 (40.00)	
Mechanism of trauma			0.716
Falling injury [*n* (%)]	21 (30.00)	23 (32.86)	
Car accident injury [*n* (%)]	33 (47.14)	28 (40.00)	
Violent injury [*n* (%)]	5 (7.14)	8 (11.43)	
Sharp object injury [*n* (%)]	2 (2.86)	4 (5.71)	
Other [*n* (%)]	9 (12.86)	7 (10.00)	
Major site of trauma			0.888
Head and neck [*n* (%)]	8 (11.43)	5 (7.14)	
Face [*n* (%)]	4 (5.71)	3 (4.29)	
Thorax [*n* (%)]	13 (18.57)	12 (17.14)	
Abdomen [*n* (%)]	18 (25.71)	15 (21.43)	
Limbs [*n* (%)]	20 (28.57)	25 (35.71)	
Surface [*n* (%)]	1 (1.43)	2 (2.86)	
Multiple areas [*n* (%)]	6 (8.57)	8 (11.43)	
ISS pre-hospitalization score	17.36 ± 13.44	18.49 ± 12.86	0.612
Trauma severity			0.731
ISS score ≤ 16[*n* (%)]	40 (57.14)	42 (60.00)	
ISS score >16[*n* (%)]	30 (42.86)	28 (40.00)	

3.2. Consistency and accuracy of the smart bracelet

Patients' blood pressure, heart rate, blood oxygen saturation, respiratory rate, and temperature were measured *via* a smart bracelet and compared with the same metrics obtained *via* traditional devices. A paired rank-sum test was performed; the result is shown in Table 2. No significant differences were found between groups ($p > 0.05$), which indicates a high consistency between the smart bracelet and traditional methods. Yet, the results of the first measurements for blood pressure ($K = 0.862$), heart rate ($K = 0.899$), blood oxygen saturation ($K = 0.605$), respiratory rate ($K = 0.751$), and temperature ($K = 0.635$) prior to hospitalization measured *via* smart bracelet were more accurate, and these results were considered statistically significant ($p < 0.001$).

3.3. Comparison of rescue efficiency

The rescue interventions and treatment efficiencies of both patient groups were compared. The results showed that the time to administration of first-aid first life-saving intervention ($t = 2.040$, $p = 0.049$) and blood transfusions ($t = 2.310$, $p = 0.048$), as well as the use of hemostatic drugs ($t = 4.416$, $p < 0.001$) were significantly shorter

TABLE 2 Analysis of consistency and accuracy in initial measurement results between the smart bracelets and traditional devices in the pre-hospital setting.

Project	Traditional device	Smart bracelet	Z/Kendall coefficient	*p*
Differences in initial measurements between both device types				
Systolic blood pressure [Media(IQR),mmHg]	126 (111.5–146.3)	127.5 (111.0–148.3)	−1.704	0.088
Heart rate [Median(IQR),/min]	91 (77.8–106.0)	92 (78.5–104.8)	−0.003	0.997
Blood oxygen saturation [Median(IQR),%]	97 (95.8–98.0)	97 (97.0–98.0)	−1.653	0.098
Respiratory rate [Median(IQR),%]	15 (13.0–19.0)	15 (12.0–19.3)	−1.238	0.216
Temperature [Median(IQR),°C]	36.6 (36.2–36.9)	36.6 (36.5–36.7)	−0.281	0.779
Consistency of initial measurements between both devices in the pre-hospital setting				
Systolic blood pressure [Media(IQR),mmHg]	126 (111.5–146.3)	127.5 (111.0–148.3)	0.862	<0.001
Heart rate [Median(IQR), /min]	91 (77.8–106.0)	92 (78.5–104.8)	0.899	<0.001
Blood oxygen saturation [Median(IQR),%]	97 (95.8–98.0)	97 (97.0–98.0)	0.605	<0.001
Respiratory rate [Median(IQR),%]	15 (13.0–19.0)	15 (12.0–19.3)	0.751	<0.001
Temperature [Median(IQR),°C]	36.6 (36.2–36.9)	36.6 (36.5–36.7)	0.635	<0.001

for patients with smart bracelets compared to the control group (Table 3), thus suggesting that smart bracelets may improve pre-hospital life-saving interventions ($p < 0.05$). However, when the efficiency of pre-hospital life-saving interventions was discussed separately, including tracheal intubation, fluid replenishment, and needle decompression, there was no significant difference between the two groups (all $p > 0.05$). The efficiency of in-hospital life-saving interventions, including blood transfusion ($p < 0.05$) and the use of hemostatic drugs ($p < 0.05$), for patients in the experimental group was better than that of the control group. Furthermore, the time to trauma team engagement for patients with smart bracelets was 3.

3.4. Comparison of treatment effects

When comparing the treatment effects in both groups of patients, the duration of stay in the emergency room (ER) was significantly shorter for patients wearing the smart bracelet compared to the control group ($t = 2.075$, $p = 0.043$). Furthermore, there were no significant differences in mortality rates between both groups within 24-h post-admission to the ER or on day-28 of patient care ($p > 0.05$) (Table 4).

4. Discussion

This study validated the consistency and accuracy of a multi-parameter integrated life monitoring smart bracelet based on WBAN technology for use prior to hospitalization and studied the impact of combined WBAN and remote 5G technology on treatment efficiency and outcomes for these trauma patients.

Compared with traditional equipment, small and integrated monitoring devices benefit medical personnel performing treatments on trauma patients while increasing overall patient compliance (2, 11). Wearable devices have been widely used in healthcare for personalized diagnosis and treatment systems, and their effectiveness has been demonstrated in rehabilitation medicine, intraoperative monitoring, sports medicine, and other fields of research (8, 12, 13). However, the application of a BAN to emergency medical care has not yet been reported. Moreover, the literature on the accuracy and clinical benefits of wearable devices is still limited (14).

The results from this study provide additional information on the accuracy of wearable devices for use in the field of emergency medical care. In this study, we found no significant statistical difference ($p > 0.05$) between blood pressure, heart rate, blood oxygen saturation, respiratory rate, and temperature measurements in the experimental group (with smart bracelet) and control patients (with traditional devices) prior to hospitalization. Yet, the consistency of smart bracelet measurements for blood pressure ($K = 0.862$), heart rate ($K = 0.899$), blood oxygen saturation ($K = 0.605$), respiratory rate ($K = 0.751$), and temperature ($K = 0.635$) was superior compared with the measurements obtained *via* traditional devices (all $p < 0.001$). Although our results suggest that the smart bracelet demonstrates a high degree of accuracy with regard to the measurement of vital signs, measurement errors cannot be ruled out. Yet, to the best of our knowledge, no study has validated the accuracy of wearable devices for use in trauma patients prior to hospitalization.

In China, there is a shortage of the equipment used for emergency care before hospitalization. Therefore, using integrated and portable devices may significantly improve the efficiency of emergency care for these patients. Liu et al. showed that using a portable wireless life monitoring device during trauma care before hospitalization could improve the accuracy of predicting life-saving interventions for patients (15). Furthermore, wearable devices achieve real-time data transmission through wireless and human-computer interaction technology, thus allowing medical staff to

TABLE 3 Comparison of rescue efficiencies between groups.

Project	Control group	Test group	χ^2/t	*p*
Administration of at least one life-saving intervention prior to hospitalization [*n* (%)]	20 (28.57)	18 (25.17)	0.144	0.704
Amount of time before use of first life-saving intervention prior to hospitalization [$\overline{x}$ ±s, Min]	6.65 ± 3.12	4.83 ± 2.24	2.040	0.049
Tracheal intubation prior to hospitalization [n (%)]	6 (8.57)	5 (7.14)	0.099	0.753
Amount of time before intubation prior to hospitalization [$\overline{x}$ ±s, Min]	4.67 ± 3.51	3.68 ± 2.58	0.520	0.616
Fluid replenishment prior to hospitalization [*n* (%)]	16 (22.86)	13 (18.57)	0.391	0.532
Time to initiation of fluid replenishment prior to hospitalization [$\overline{x}$ ±s, Min]	6.97 ± 2.84	5.27 ± 2.03	1.802	0.083
Needle decompression prior to hospitalization [*n* (%)]	2 (2.86)	1 (1.43)		1
Emergency blood transfusion [*n* (%)]	8 (11.43)	6 (8.57)	0.317	0.573
Starting time of blood transfusion [$\overline{x}$ ±s, Min]	163.25 ± 83.44	91.67 ± 23.27	2.310	0.048
Emergency use of hemostatic drugs [*n* (%)]	33 (47.14)	29 (41.43)	0.365	0.546
Duration of emergency hemostatic drug use [$\overline{x}$ ±s, Min]	36.91 ± 7.70	25.62 ± 11.72	4.416	<0.001
The situation of trauma team activation prior to arriving at the hospital				
Start a Trauma Team [*n* (%)]	30 (42.86)	28 (40.00)	0.118	0.731
Time to trauma team activation [$\overline{x}$ ±s, Min]	8.22 ± 3.76	5.80 ± 3.04	2.709	0.009

TABLE 4 Comparison of treatment effects for both groups of patients.

Project	Control group	Test group	χ^2/t	*p*
24-h mortality rate [*n* (%)]	2 (2.86)	0 (0.00)		0.496
28-day mortality rate [*n* (%)]	3 (4.29)	1 (1.43)		0.620
Number of patients needing resuscitation [*n* (%)]	25	20	0.819	0.366
Patient length of stay in the emergency room [$\overline{x}$ ±s, Min]	199.60 ± 71.67	159.36 ± 65.29	2.075	0.043

remotely and instantaneously understand a patient's physical condition. Furthermore, high levels of integration and the compactness of wearable devices make them more environmentally friendly (16). The smart bracelet used in our study not only monitors vital signs in real-time during emergencies but can also be used for remote medical assistance through the use of body area networks and remote 5G technology. Our results suggested that the use of a BAN-based smart bracelet in emergency care prior to hospitalization can implement life-saving interventions in a more timely manner compared to conventional emergency care techniques, including first life-saving intervention ($t = 2.040$, $p = 0.049$), blood transfusion ($t = 2.310$, $p = 0.048$) and the use of hemostatic drugs ($t = 4.416$, $p < 0.001$). When multiple life-saving interventions (i.e., tracheal intubation, fluid resuscitation, needle decompression) from our study were separately analyzed, no significant difference was found between the groups. On-site tracheal intubation is a challenging procedure, with questionable short-term benefits. First responders often lack experience in this technique, leading to delayed or repeated intubation, which increases the risk of death (17). Therefore, using efficient and portable devices to shorten on-site assessment time may lead to quicker intubation, fluids and needle decompression administration. However, the small number of patients in our study introduced significant variability in the results, making it impossible to draw a clear conclusion.

Overall, the experimental group received life-saving interventions faster than the control group. Furthermore, we also found that patients in the experimental group received assistance from the trauma teams in a shorter period thanks to the 5G remote medical assistance ($t = 2.709$, $p = 0.009$). Previous studies have shown that timely and effective life-saving interventions can reduce mortality rates among trauma patients and that remote communication with emergency surgeons significantly improves the effect of life-saving interventions as well as reduces overall mortality rates in trauma patients (18, 19). Collaborative treatment between on-site and intra-hospital care can improve the diagnosis efficiency and treatment of severely injured patients (20). It is currently undisputed that minimizing the time from a severe injury to treatment is important; however, our results showed no significant difference in 24-h and 28-day mortality rates between groups. The overall number of patients who died in our study was small, and the results we obtained contained significant variation. Therefore, we could draw no clear conclusion from this data. The smart bracelets group had shorter stays in the emergency department than the control group ($t = 2.075$, $p = 0.043$). This is most likely due to the smart bracelet technology that reduced patient admission time and

increased the number of resources available to patients in the emergency department (20, 21). For patients receiving emergency care before hospitalization, BAN can be used to perform simultaneous multi-user monitoring, which is more effective for monitoring the health statuses of patients on-site and coordinating large-scale casualty treatment when necessary (10, 20, 21).

The present study has a few limitations: (1) this is an observational study, and the results are inevitably subject to confounding factors. However, we effectively controlled these factors by using random grouping for the experiment. The general conditions of both patient groups (i.e., age, gender, trauma type, pre-hospitalization ISS score, and trauma severity) were compared, showing no statistical differences. (2) Although we compared the baseline data of the two groups of patients and found no significant statistical difference in the results (Table 1), not all samples were subjected to life-saving interventions (Table 3), and there may be some bias in the baseline data of those who subjected to life-saving interventions between two groups. To some extent, group randomization reduces the possibility of such bias, and further study should have more specific trauma samples or larger samples for stratified analysis. (3) There are differences in clinical experience among different clinical decision-makers, and the difference in enthusiasm for implementing life-saving interventions may have a certain degree of interference with the results, which were not evaluated. (4) In order to identify the advantages of using a BAN-based smart bracelet, future studies should include a separate experimental group that will use this system so as to reduce bias. However, there is currently insufficient evidence to determine whether the results of wearable devices used in pre-hospital settings are reliable. In a major accident, medical staff may be more inclined to focus on traditional equipment during the pre-hospital treatment period for each patient, which we did not evaluate during this study.

5. Conclusion

A first aid smart bracelet based on body area network technology can improve the treatment efficiency and effectiveness of trauma care in patients pre-hospitalization. Emergency smart bracelets can shorten the start time of a patient's first life-saving intervention, such as a blood transfusion, administering hemostatic drugs, and notification of the trauma team, and reduce the time spent in the emergency room. However, the results of this study did not suggest that smart bracelets made a significant difference concerning patient survival. Therefore, we provided an effective technical mean for emergency doctors to improve both efficiency and efficacy of emergency treatment; however, further research and verification are needed.

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 Review Committee of the General Hospital of Shenzhen University. The patients/participants provided their written informed consent to participate in this study.

Author contributions

WH, J-YY, RL, LY, and J-QF completed the data collection and organization. WH, J-YY, H-JF, and S-KH analyzed the data and completed the first draft of the manuscript. All authors contributed to the article and approved the submitted version.

Funding

This research was supported by Shenzhen Medical and Health Three Project Support (Project no. SZSM201911007); Shenzhen University Stability Support Program, (Project no. 20200824145152001) "5G+ Medical and Health Application Pilot Project" of Ministry of Industry and Information and National Health Commission (Guangdong Province, Direction 1:5G + Emergency Treatment, 11).

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. Hsieh SL, Hsiao CH, Chiang WC, Shin SD, Jamaluddin SF, Son DN, et al. Association between the time to definitive care and trauma patient outcomes: every minute in the golden hour matters. *Eur J Trauma Emerg Surg*. (2022) 48:2709–16. doi: 10.1007/s00068-021-01816-8

2. Lin S., Zhang N., Zhao Y. "Development and application of portable general life support system". *Chinese J Emergency* (2021) 16: 1422–1426. doi: 10.3969/j.issn.1673-6966.2021.12.022

3. Ahmad S. Mass casualty incident management. *Mo Med*. (2018) 115:451–5.

4. Jamil F, Ahmad S, Iqbal N, Kim D-H. Towards a remote monitoring of patient vital signs based on IoT-based Blockchain integrity management platforms in smart hospitals. *Sensors*. (2020) 20:2195. doi: 10.3390/s20082195

5. Han W, Wang J, Hou S, Bai T, Jeon G, Rodrigues JJPC. An PPG signal and body channel based encryption method for WBANs. *Futur Gener Comput Syst*. (2023) 141:704–12. doi: 10.1016/j.future.2022.11.020

6. Han W, Wang J, Zhou Y, Jiang Y. Research status and Progress of first aid intelligent wearable system based on body area network. *Chinese disaster Res Med*. (2021) 9:1113–7. doi: 10.13919/j.issn.2095-6274.2021.07.008 (Chinese).

7. Mamdiwar SD, R A, Shakruwala Z, Chadha U, Srinivasan K, Chang CY. Recent advances on IoT-assisted wearable sensor Systems for Healthcare Monitoring. *Biosensors*. (2021) 11:372. doi: 10.3390/bios11100372

8. Lu L, Zhang J, Xie Y, Gao F, Xu S, Wu X, et al. Wearable health devices in health care: narrative systematic review. *JMIR Mhealth Uhealth*. (2020) 8:e18907. doi: 10.2196/18907

9. Koceska N, Komadina R, Simjanoska M, Koteska B, Strahovnik A, Jost A, et al. Mobile wireless monitoring system for prehospital emergency care. *Eur J Trauma Emerg Surg*. (2020) 46:1301–8. doi: 10.1007/s00068-019-01130-4

10. Tung HC, Tsang KF, Lam KL, Tung HY, Li BY, Yeung LF, et al. A mobility enabled inpatient monitoring system using a ZigBee medical sensor network. *Sensors (Basel)*. (2014) 14:2397–416. doi: 10.3390/s140202397

11. Cancela J, Pastorino M, Tzallas AT, Tsipouras MG, Rigas G, Arredondo MT, et al. Wearability assessment of a wearable system for Parkinson's disease remote monitoring based on a body area network of sensors. *Sensors (Basel)*. (2014) 14:17235–55. doi: 10.3390/s140917235

12. Li RT, Kling SR, Salata MJ, Cupp SA, Sheehan J, Voos JE. Wearable performance devices in sports medicine. *Sports Health*. (2016) 8:74–8. doi: 10.1177/1941738115616917

13. Semiz B, Carek AM, Johnson JC, Ahmad S, Heller JA, Vicente FG, et al. Non-invasive wearable patch utilizing Seismocardiography for Peri-operative use in surgical patients. *IEEE J Biomed Health Inform*. (2021) 25:1572–82. doi: 10.1109/JBHI.2020.3032938

14. Bent B, Goldstein BA, Kibbe WA, Dunn JP. Investigating sources of inaccuracy in wearable optical heart rate sensors. *NPJ Digit Med*. (2020) 3:18. doi: 10.1038/s41746-020-0226-6

15. Liu NT, Holcomb JB, Wade CE, Darrah MI, Salinas J. Evaluation of standard versus nonstandard vital signs monitors in the prehospital and emergency departments: results and lessons learned from a trauma patient care protocol. *J Trauma Acute Care Surg*. (2014) 77:S121–6. doi: 10.1097/TA.0000000000000192

16. Nan X, Wang X, Kang T, Zhang J, Dong L, Dong J, et al. Review of flexible wearable sensor devices for biomedical application. *Micromachines (Basel)*. (2022) 13:1395. doi: 10.3390/mi13091395

17. Arnason B, Hertzberg D, Kornhall D, Gunther M, Gellerfors M. Pre-hospital emergency anaesthesia in trauma patients treated by anaesthesiologist and nurse anaesthetist staffed critical care teams. *Acta Anaesthesiol Scand*. (2021) 65:1329–36. doi: 10.1111/aas.13946

18. Mitra B, Bade-Boon J, Fitzgerald MC, Beck B, Cameron PA. Timely completion of multiple life-saving interventions for traumatic haemorrhagic shock: a retrospective cohort study. *Burns T*. (2019) 7:22. doi: 10.1186/s41038-019-0160-5

19. Endo A, Kojima M, Uchiyama S, Shiraishi A, Otomo Y. Physician-led prehospital management is associated with reduced mortality in severe blunt trauma patients: a retrospective analysis of the Japanese nationwide trauma registry. *Scand J Trauma Resusc Emerg Med*. (2021) 29:9. doi: 10.1186/s13049-020-00828-4

20. Zhang H, Zhao J, Zhang Y, Liu D, Hu B, Wang H, et al. The application effect of the new internet of things linkage model of 5G cloud + medical treatment in the treatment of patients with severe trauma. *Chinese J Trauma*. (2022) 38:359–64. doi: 10.3760/cma.j.cn501098-20211102-00565 (Chinese).

21. Nino V, Claudio D, Schiel C, Bellows B. Coupling wearable devices and decision theory in the United States emergency department triage process: a narrative review. *Int J Environ Res Public Health*. (2020) 17:17249561. doi: 10.3390/ijerph17249561

frontiers | Frontiers in Medicine

TYPE Original Research
PUBLISHED 03 August 2023
DOI 10.3389/fmed.2023.1151996

OPEN ACCESS

EDITED BY
Qinghe Meng,
Upstate Medical University, United States

REVIEWED BY
Yi Zi Ting Zhu,
First Affiliated Hospital of Chongqing Medical University, China
Hamidreza Bolhasani,
Islamic Azad University, Iran

*CORRESPONDENCE
Yu-wen Chen
✉ chenyuwen@cigit.ac.cn
Bin Yi
✉ yibin1974@163.com

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

RECEIVED 01 February 2023
ACCEPTED 10 July 2023
PUBLISHED 03 August 2023

CITATION
Hu X-y, Li Y-j, Shu X, Song A-l, Liang H, Sun Y-z, Wu X-f, Li Y-s, Tan L-f, Yang Z-y, Yang C-y, Xu L-q, Chen Y-w and Yi B (2023) A new, feasible, and convenient method based on semantic segmentation and deep learning for hemoglobin monitoring.
Front. Med. 10:1151996.
doi: 10.3389/fmed.2023.1151996

A new, feasible, and convenient method based on semantic segmentation and deep learning for hemoglobin monitoring

Xiao-yan Hu[1†], Yu-jie Li[1†], Xin Shu[1], Ai-lin Song[1], Hao Liang[1], Yi-zhu Sun[1], Xian-feng Wu[1], Yong-shuai Li[1], Li-fang Tan[1], Zhi-yong Yang[1], Chun-yong Yang[1], Lin-quan Xu[2], Yu-wen Chen[2*] and Bin Yi[1*]

[1]Department of Anesthesiology, Southwest Hospital, Third Military Medical University (Army Medical University), Chongqing, China, [2]Chongqing Institute of Green and Intelligent Technology, Chinese Academy of Science, Chongqing, China

Objective: Non-invasive methods for hemoglobin (Hb) monitoring can provide additional and relatively precise information between invasive measurements of Hb to help doctors' decision-making. We aimed to develop a new method for Hb monitoring based on mask R-CNN and MobileNetV3 with eye images as input.

Methods: Surgical patients from our center were enrolled. After image acquisition and pre-processing, the eye images, the manually selected palpebral conjunctiva, and features extracted, respectively, from the two kinds of images were used as inputs. A combination of feature engineering and regression, solely MobileNetV3, and a combination of mask R-CNN and MobileNetV3 were applied for model development. The model's performance was evaluated using metrics such as R^2, explained variance score (EVS), and mean absolute error (MAE).

Results: A total of 1,065 original images were analyzed. The model's performance based on the combination of mask R-CNN and MobileNetV3 using the eye images achieved an R^2, EVS, and MAE of 0.503 (95% CI, 0.499–0.507), 0.518 (95% CI, 0.515–0.522) and 1.6 g/dL (95% CI, 1.6–1.6 g/dL), which was similar to that based on MobileNetV3 using the manually selected palpebral conjunctiva images (R^2: 0.509, EVS:0.516, MAE:1.6 g/dL).

Conclusion: We developed a new and automatic method for Hb monitoring to help medical staffs' decision-making with high efficiency, especially in cases of disaster rescue, casualty transport, and so on.

KEYWORDS

continuous hemoglobin monitoring, deep learning, semantic segmentation, mask R-CNN, MobileNetV3

1. Introduction

Continuous monitoring of hemoglobin (Hb) helps doctors make better decisions regarding blood transfusions. The most frequently used methods for Hb monitoring are automatic blood analysis and arterial blood gas (ABG) analysis, which require professional operators and devices. Therefore, they are not ideal for continuous Hb monitoring, especially during disaster rescue scenes, field rescue, emergent public health events (e.g., COVID-19), casualty transport, and battlefield rescue. Pulse co-oximetry hemoglobin (SpHb) was

developed by Masimo Corporation, which is continuous and non-invasive and is used for providing additional and relatively precise information between measurements of Hb by invasive blood samples. However, its accuracy depends on the blood flow and temperature of the tested fingers (1). Additionally, the SpHb cannot be used with other monitors, thus restricting its clinical application.

Recently, non-invasive methods for continuous Hb monitoring based on computer vision technology have shown great potential (Supplementary Table 1). The basis of these methods is that the palpebral conjunctiva and the nailbed pallor could be used to diagnose anemia (2). Most of the studies focused on using the image of the palpebral conjunctiva to detect anemia. The typical characteristics of research in this area were as follows: first, images were obtained using special devices (fundus cope or macro-lens) (3–6) or consumer-grade smartphones or cameras (7, 8), among which models based on images obtained by fundus cope achieved the best performance (with an R^2 value of 0.52, and area under the receiver operating characteristic curve (AUROC) of 0.93) (6); second, instead of estimating the exact concentration of Hb, detecting anemia patients was more common (5, 9–11), which may be associated with the small sample size of images (Supplementary Table 1); third, most of the model inputs were features extracted from the manually selected palpebral conjunctiva (7, 12); however, recently semantic segmentation algorithms were also applied to realize automatic estimation (3, 4, 13).

Above all, new methods for continuous Hb monitoring with the three advantages are badly needed: no requirement for a special device or position during image acquisition, automation presented by using eye images as model input; the ability to estimate the exact concentration of Hb; and the ability to detect anemia with different thresholds. Therefore, we aimed to develop a new method that combines semantic segmentation and deep learning algorithms to estimate the exact concentration of Hb for surgical patients with the eye images obtained using smartphones, to compare the model's performance with models based on feature engineering and solely deep learning methods using the eye and manually selected palpebral conjunctiva images, respectively, and to find out whether it would be promising for clinical and special situations.

2. Materials and methods

The study protocol was approved by the institutional ethics committee of the First Affiliated Hospital of the Third Military Medical University (also called Army Medical University, KY2021060) on February 20, 2021, and written informed consent was obtained from each patient. The clinical trial was registered on the Chinese Clinical Trial Registry (No. ChiCTR2100044138) on March 11, 2021. The principal researcher was Prof. Bin Yi. Patient enrollment and image acquisition were completed at the First Affiliated Hospital of the Third Military Medical University in Chongqing, China, between March 18, 2021, and April 26, 2021.

2.1. Patient enrollment and image acquisition

The inclusion criteria were as follows: volunteering to participate in the research; ABG analysis needed according to routine clinical practice; Hb variance larger than 1.5 g/dL perioperatively. The exclusion criteria were as follows: suffering eye diseases, eye irradiation, or receiving facial radiation therapy, suffering carbon monoxide poisoning, nitrite poisoning, jaundice, or other systemic diseases that would change the color of the palpebral conjunctiva.

There were eight researchers who participated in the research: one for patient enrollment, two for image acquisition, two for data collection and collation, one for palpebral conjunctiva identification, and two for quality control. One day before the operation, all patients who met the criteria and were willing to participate in the study signed written informed consent. On the surgical day, when the enrolled patients were undergoing ABG analysis, two researchers came to the operation room or the post-anesthetic care unit (PACU) to take pictures of the right and left faces with the standard exposing way of the palpebral conjunctiva in the routine light of the operation room and PACU. The time between ABG analysis and image acquisition was within 10 min. All the images were obtained when patients were in a supine position and by the rear camera of the same smartphone (20.00 megapixels and f/1.8 aperture) with the same parameters. At the same time, the other two researchers collected patients' information. After the whole day of image acquisition, the two researchers, for data collection and collation, picked out images obtained from the patients whose Hb variation was larger than 1.5 g/dL. The unselected images were all deleted permanently. The selected half-face images were cut as eye images following the criteria shown in Supplementary Figure 1. During the whole process, the two researchers for quality control checked the enrollment, images, basic information, and so on.

2.2. Image pre-processing

As shown in Figure 1, after image acquisition, manual palpebral conjunctiva recognition, image pre-processing, and up-sampling were conducted. To keep the same standard of palpebral conjunctiva identification, one researcher worked on the manual segmentation of palpebral conjunctiva via Photoshop (Photoshop cs 6.0, Adobe Systems, California, USA) and Colabeler (version 2.0.4, Hangzhou Kuaiyi Technology Co. Ltd., Hangzhou, China). Subsequently, the eye and the palpebral conjunctiva images by Photoshop and Colabeler were normalized to a fixed size (500 × 500) to avoid possible loss of useful information as previously described (7). Due to that, different shapes and sizes of bright spots on the images were unavoidable, and denoising was also conducted. In the current study, K-means clustering was applied to identify the bright spot area in the Gray-level image converted from a corresponding RGB color image, and then the values of all pixels in the bright spot area were replaced by the mean value of all pixels in the non-bright spot area as previously described (7).

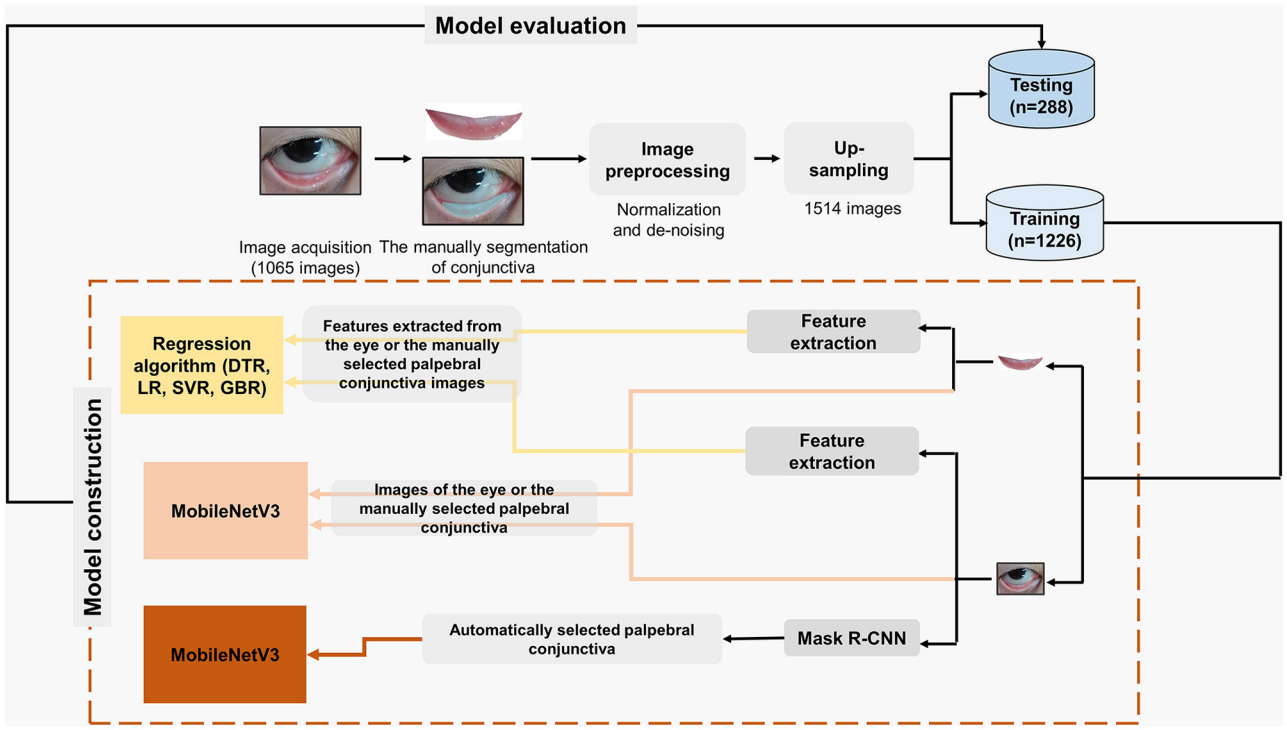

FIGURE 1
Flow diagram for estimating the exact concentration of Hb based on different algorithms with different inputs. DTR, Decision Tree regression; LR, Linear Regression; SVR, Support Vector regression; GBR, Gradient Boosting regression; CNN, Convolutional Neural Network.

2.3. Feature extraction for the eye and the palpebral conjunctiva images

As shown in Figure 1, features extracted from the eye and the palpebral conjunctiva images were inputted for regression. The methods for feature extraction from the palpebral conjunctiva were relatively mature, so we applied the same algorithm for feature extraction as the study conducted by Miaou et al. (7). However, in the current study, we utilized normalized eye images and the palpebral conjunctiva as inputs for feature extraction, rather than relying on a manually selected fixed rectangular area. We extracted 18 features, including Hue Ratio, Pixel Values in the Middle, Entropy H to describe the distribution of the blood vessels and Binarization of the High Hue Ratio.

2.3.1. Automatically segmentation of the palpebral conjunctiva by mask R-CNN

Herein, automatic recognition of the palpebral conjunctiva from eye images was achieved using mask R-CNN (14). Mask R-CNN is an instance segmentation framework extended by Faster RCNN (15), which could simultaneously perform pixel-level object segmentation and target recognition. It operates in two stages: the first stage scans the image and generates suggestions, and the second stage classifies the suggestions, generates bounding boxes, creates masks for accurate delineation of the recognized objects. Except for the original Faster RCNN network structure, the mask R-CNN also included the feature pyramid network (16) and the region of interest alignment algorithm (ROI Align) (14). Detailed information is given in the Supplementary Methods and Supplementary Figure 2. For semantic segmentation performance. We report the average precision (AP) and average recall (AR) over mask Intersection-over-Union (mIoU) thresholds (50%, 75%, 50%, and 95%). The segmentation work was conducted using ubuntu16.04TSL, Pytorch 1.3, and CUDA 11.0 platforms.

2.4. Establishment of models for the exact concentration of Hb based on different algorithms

As shown in Figure 1, all the extracted features were inputted to develop models. Models were fitted with decision tree regression, linear regression, support vector regression, and gradient boosting regression, respectively. MobileNetV3 (17) was applied to models directly using the eye and the palpebral conjunctiva images. In the current study, the classification structure of the mobilenetV3 tail was changed to a regression structure for the exact concentration of Hb. The mean square error loss function was used for training. These experiments used the open-source PyTorch learning framework and Python programming to realize the algorithm network. The hardware environment is a Dawning workstation from Chongqing Institute of Green and Intelligent Technology, Chinese Academy of Sciences, equipped with dual NVIDIA 2080Ti graphics cards (11 GB) and a 64-bit Ubuntu16.04 operating system

(detailed information is shown in Supplementary Methods and Supplementary Figure 2).

2.5. Establishment of models based on the combination of mask R-CNN and MobileNetV3

We attempted to estimate the exact concentration of Hb based on the mask R-CNN and MobileNetV3 in two steps: semantic segmentation and regression (Supplementary Figure 2). First, semantic segmentation was performed to automatically recognize the palpebral conjunctiva from the eye images. Then, the recognized palpebral conjunctiva images were entered into the MobileNetV3 network to estimate the exact concentration of Hb. This two-step method could automatically estimate the exact concentration of Hb with eye images.

2.6. Model evaluation

For estimating the exact concentration of Hb, we evaluated the model's performance with the mean absolute error (MAE), R^2 and Explained variance score (EVS). The MAE is used to describe the average difference between the estimated value and the actual value. The EVS describes the similarity between the dispersion degree of the difference between all predicted values and samples. EVS was calculated by the following formula: $EVS\left(y,\hat{y}\right) = 1 - \frac{Var\{y- \hat{y}\}}{Var\{y\}}$, where y is the Hb measured by ABG analysis, $\hat{y}$ is the estimated Hb, and Var is the square of the standard deviation. R^2 is also called the coefficient of determination. The closer the value to 1, the stronger the ability to interpret the output and the better the model fitting. Furthermore, we paid more attention to whether the new method could provide a relatively precise trend of Hb and recognize anemia with different thresholds. In addition to evaluating the model's performance using regression parameters, we investigated the correlation between the estimated and actual Hb and the ability to recognize anemia patients (Hb <10.0 g/dL, 11.0 g/dL, and 12.0 g/dL) according to the estimated Hb. Moreover, we also evaluated the accuracy when the accurate estimation was determined by the set range of absolute value of the difference (e.g., within 1.5 g/dL, 2.0 g/dL) between the estimated and the actual Hb. All the detailed information on image pre-processing and the main code for this study has been provided on GitHub (https://github.com/keyan2017/hemoglobin-prediction).

2.7. Statistical analysis

All the statistical analysis was conducted on the R platform (R Studio, version 1.4.1717, USA). For quantitative variables, the mean, standard deviation (SD), and range are presented. For the primary effectiveness variables, 95% confidence intervals (Cis) are presented. The correlation between estimated and actual Hb was tested via Pearson analysis, wherein the $r_{pearson}$, P, 95% CI were provided [ggstatsplot (18), version: 0.9.0]. Meanwhile, density distribution and scatter plots were completed with R packages [ggplot2 (19), version: 3.3.5]. All statistical tests were two-sided, and $P < 0.05$ indicated statistical significance.

3. Results

In the current study, 1,073 pieces of eye images from 284 patients with an average age of 51.5 years old for elective surgery (M/F: 117/167) were obtained (Supplementary Table 2). Finally, 1,065 images were analyzed; three images were excluded for inadequate exposure, and five were excluded due to overexposure. After image pre-processing and up-sampling, 1,226 images were in the training dataset, and 288 were in the test dataset (Figure 1). The mean and the distribution of Hb in the training dataset were similar to those in the test dataset (Figure 2A).

Using features extracted from the manually selected palpebral conjunctiva as input to detect anemia was the most common in this area. Models directly using the manually selected palpebral conjunctiva images as input based on MobileNetV3 yielded R^2, EVS, and MAE of 0.509 (95% CI, 0.505–0.512), 0.516 (95% CI, 0.513–0.519) and 1.6 g/dL (95% CI, 1.6–1.6), which was much better than those using features as input (Table 1). However, when the inputs were eye images, the model's performance was poorer, even based on MobileNetV3.

To further improve the model's performance with the eye images as input, mask R-CNN was applied for automatic segmentation of the palpebral conjunctiva from the eye images. As shown in the representative images in Figure 2B, despite the concentration of Hb (anemia or not) and the shape of the palpebral conjunctiva (wide or slender), the IoU of manually and automatically selected conjunctiva was relatively satisfied. Meanwhile, regardless of the thresholds of the mIoU, the AP and AR were relatively accepted (Table 2), which was well-matched with the existing research (4). The model based on the combination of mask R-CNN and MobileNetV3 achieved a good consequence with an R^2 of 0.503 (95% CI, 0.499–0.507), EVS of 0.518 (95% CI, 0.515–0.522), MAE of 1.6 g/dL (95% CI, 1.6–1.6), which was similar to the model's performance using manually selected palpebral conjunctiva and was better than that directly using eye images (Table 1). The correlation between the estimated and the actual Hb was 0.77 (95% CI, 0.72–0.82); moreover, for different thresholds, the correlation between the estimated between and actual Hb remaining satisfied (Figures 2C–F). Meanwhile, when we determined the range of absolute value of the difference between the estimated and actual Hb within 2.0 g/dL as the standard of accurate estimation, the accuracy was 72.2% (Supplementary Table 3). Moreover, according to the estimated Hb, we re-evaluated the model's performance for recognizing anemia patients with different thresholds (Hb <10.0 g/dL, 11.0 g/dL, and 12.0 g/dL). When the threshold was 10.0 g/dL, the accuracy, specificity, and AUROC were 85.4%, 97.2%, and 0.752 (95% CI, 0.698–0.801) (Supplementary Table 4).

4. Discussion

Herein, we developed a new method that could not only automatically estimate the exact concentration of Hb but also

10.3389/fmed.2023.1151996

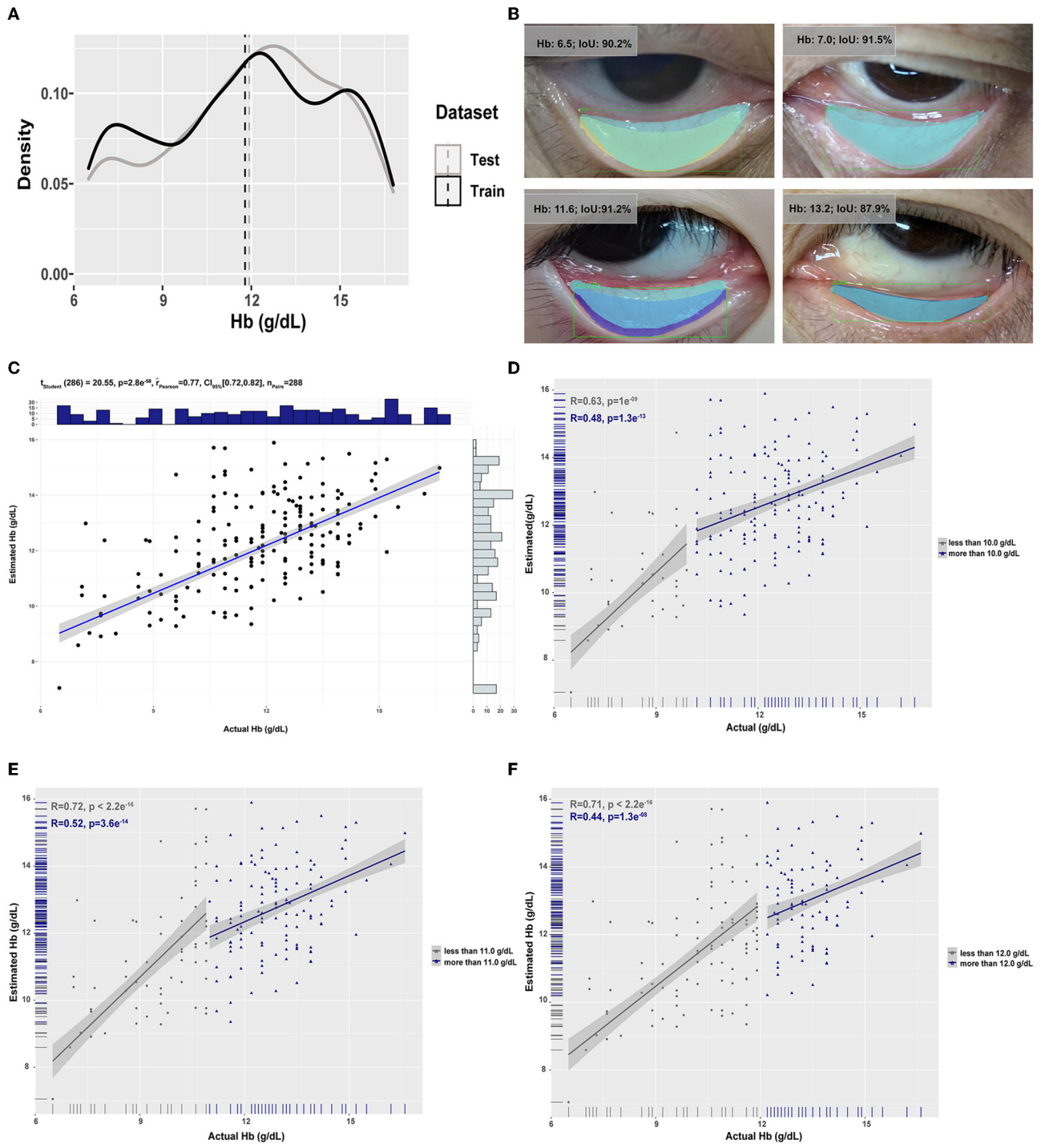

FIGURE 2
The data distribution in the training and test datasets and the performance of models are based on the combination of mask R-CNN and MobileNetV3. **(A)** The distribution of concentration of Hb in the training and test datasets. The vertical dashed lines were the mean concentration of Hb in the two datasets. **(B)** Representative overlay images of manually selected conjunctiva (light blue) and automatically recognized conjunctiva (other colors) in cases of different concentrations of Hb. The correlation between estimated and actual Hb was analyzed by Pearson analysis with different thresholds [**(C)**: no threshold; **(D)**: the threshold was 10g/dL; **(E)**: the threshold was 11 g/dL; **(F)**: the threshold was 12 g/dL].

achieve a similar performance using manually selected palpebral conjunctiva as input.

As previously described, a quick and non-invasive method for Hb monitoring that can provide additional and relatively precise information between measurements of Hb using invasive blood samples is badly needed, especially for situations such as disaster rescue scenes, field rescue, emergent public health events (e.g., COVID-19), casualty transport, and battlefield rescue. Though

TABLE 1 Performance of models based on different methods with different inputs on test dataset.

Algorithm	R^2 (95% CI)	EVS (95% CI)	MAE (95% CI), g/dL
Models with the input of manually selected conjunctiva images			
Decision tree regression	0.262 (0.242, 0.283)	0.267 (0.247, 0.287)	2.1 (2.0, 2.1)
Linear regression	0.300 (0.288, 0.312)	0.304 (0.292, 0.315)	2.0 (2.0, 2.0)
Support vector regression	0.267 (0.248, 0.286)	0.270 (0.252, 0.289)	2.0 (2.0, 2.0)
Gradient boosting regression	0.296 (0.283, 0.308)	0.298 (0.287, 0.31)	2.0 (2.0, 2.0)
MobileNetV3	0.509 (0.505, 0.512)	0.516 (0.513, 0.519)	1.6 (1.6, 1.6)
Models with the input of eye images			
Decision tree regression	−0.013 (−0.032, 0.006)	−0.001 (−0.021, 0.018)	2.4 (2.4, 2.4)
Linear regression	−0.077 (−0.113, −0.041)	−0.064 (−0.101, −0.027)	2.5 (2.4, 2.5)
Support vector regression	−0.052 (−0.081, −0.024)	−0.039 (−0.068, −0.01)	2.4 (2.4, 2.5)
Gradient boosting regression	−0.053 (−0.083, −0.024)	−0.037 (−0.067, −0.007)	2.4 (2.4, 2.5)
MobileNetV3	0.306 (0.296, 0.317)	0.338 (0.329, 0.348)	2.0 (2.0, 2.0)
Models with the input of conjunctiva by automatically selection from eye images			
Mask R-CNN combined with MobileNetV3	0.503 (0.499, 0.507)	0.518 (0.515, 0.522)	1.6 (1.6, 1.6)

Data were presented with 95% CIs. MAE, mean absolute error; EVS, Explained Variance Score; CI, confidence interval.

TABLE 2 The average precision and recall under different thresholds of mIoU when automatically segmentation of conjunctiva.

Threshold of mIoU	AP	AR	maxDets
0.5	0.989	0.996	100
0.5–0.95	0.672	0.720	100
0.75	0.845	0.891	100

mIoU, mask Intersection over Union; AP, average precision; AR, average recall; maxDets, maxdectetions.

SpHb is a non-invasive, continuous device for Hb monitoring, its application, and promotion were restricted due to its inability to be used on other platforms except Massimo's.

Numerous teams have been working on developing non-invasive methods to detect anemia or estimate the exact concentration of Hb based on computer vision technology in the last few years (Supplementary Table 1). Initially, researchers attempted to find features associated with anemia or Hb based on the manually selected palpebral conjunctiva images. The erythema index [EI = log (S_{red}) – log (S_{green})], where S is the brightness of the palpebral conjunctiva in the relevant color channel) was found to be significantly associated with measured Hb (the r^2 could be up to 0.397), based on which the sensitivity and specificity for anemia (Hb < 11.0 g/dL) were 57.0 and 83.0% (20). Meanwhile, Miaou et al. (7) determined three important features, including entropy, binarization of the high Hue ratio, and PVM of G components, for detecting anemia with the palpebral conjunctiva images. Models based on these features achieved higher sensitivity and κ values than previous studies. Afterward, ANN (7, 21), Elman neural network (22), and CNN (23) were applied to detect anemia or estimate the exact concentration of Hb and achieved high accuracy. However, most of these studies were not "real" deep learning because the inputs were features extracted by feature engineering. It may be associated with the sample size being too small to fulfill the number of images needed for deep learning. However, their studies still showed that deep learning may help elevate the model's performance. Herein, we used the same method as Professor Miaou's for feature extraction and applied selected features to estimate the exact concentration of Hb using traditional regression algorithms and observed poorer performance than those directly using images as input based on MobileNetV3. It suggested that images were more informative and effective than extracted features when estimating the exact concentration of Hb. Meanwhile, deep learning algorithms may be more helpful when the inputs are images rather than features.

Despite the difference in input (features *vs.* images) and estimations (classification *vs.* regression) between previous research on models based on deep learning and ours, we compared our results with previous studies in Table 3. Though models based on MobileNetV3 with the manually selected palpebral conjunctiva achieved the best performance in the current study, the performance was much poorer when the eye images were used as input. It was suggested that the palpebral conjunctiva images as input were the most important to estimate the exact concentration of Hb or detect anemia in patients. Thus, we applied mask R-CNN to automatically segment the palpebral conjunctiva to help elevate the performance of models with eye images as input. Afterward, we got satisfactory results from segmentation, and the two-step model achieved a similar performance to that using the manually selected palpebral conjunctiva as input. Dimauro et al. (13) made great efforts to develop non-invasive and continuous Hb monitoring based on computer vision technology. In 2019, they attempted to obtain the relevant sections of the palpebral conjunctiva automatically by contour detection and

TABLE 3 Performance comparison between our method with previous works.

References	Inputs	Sample Size	Classification/ regression/segmentation	Main results
Kasiviswanathan et al. (4)	The eye images	135	Segmentation	The accuracy of the automatic segmentation was 85.7%.
Dimauro et al. (13)	The eye images	65	Segmentation	The correlation between feature "a" and Hb was 0.74.
Jain et al. (21)	The conjunctiva images	99	Classification	The accuracy, sensitivity and specificity for prediction anemia was 97.00%, 99.21% and 95.42%.
Saldivar-Espinoza et al. (23)	The conjunctiva images	300	Classification	The Sensitivity, accuracy, and specificity were 77.6%, 43.0%, and 36.0%.
Muthalagu (22)	The conjunctiva images	127	Classification	The sensitivity and specificity for detecting anemia were 77.3% and 96.1%.
Collings et al. (20)	The conjunctiva images	94	Classification	The sensitivity and specificity were 57.0% and 83% in the internal validation datasets.
Our method	The eye images	1065	Segmentation	The accuracy of the automatic segmentation was 82.6%.
			Regression	The correlation, MAE between the estimated and the actual Hb was 0.77, 1.6 g/dL. The accuracy, specificity, and AUROC were 85.4%, 97.2%, and 0.752 (Hb threshold was 10.0 g/dL)

Hb, hemoglobin; MAE, mean absolute error; AUROC, area under the receiver operating characteristic curve.

feature extraction, of which the correlation between automatically extracted features and the exact concentration of Hb could be up to 0.74 (13). Recently, they attempted to apply the Biased Normalized Cuts Approach (3) and CNN (4) to automatically segment the palpebral conjunctiva from the eye images obtained using the special device and consumer-grade cameras, respectively. For images obtained using a special device, feature extraction and regression were conducted after automatic segmentation of palpebral conjunctiva, with similar results to those of manually selected palpebral conjunctiva images (3). As for the images obtained using consumer-grade cameras, the IoU score between the ground truth and the segmented mask was 85.7% (4), which is similar to ours (82.6%) (Table 3). Dimauro et al. (13) study suggested that automatically estimating the exact concentration of Hb with eye images from customer-grade cameras or smartphones is the new trend in the area of non-invasive and continuous Hb monitoring. Our results also showed that a combination of semantic segmentation and deep learning methods might be a new strategy for this area.

Our method was more convenient and simpler than previous ones since manually selecting the conjunctiva is no longer needed before inputting the images. There were some other advantages to our study. First of all, the sample size of the original images was larger compared with previous research (Supplementary Table 1), which would reduce overfitting and increase robustness. Second, smartphones obtained images when patients were lying on their backs awake or anesthetized, which would be more convenient for promotion and application in various situations. Third, herein, we estimated the exact concentration of Hb, which was seldom conducted in previous studies. Estimating the exact concentration of Hb could not only indicate the trend change of Hb but also easily detect anemia according to various thresholds without repeated image labeling (Supplementary Table 3). In summary, the combination of mask RCNN and MobileNetV3 to automatically estimate the exact concentration of Hb is quite promising in the area of non-invasive and continuous Hb monitoring in a variety of situations.

There are some limitations to the current study. First, though we tried to enroll more images for analysis and model development, the amounts of images from anemia and non-anemia were still imbalanced. The model's performance might be better if more images were enrolled, especially those from patients with anemia. Second, the images were obtained from one center, so external validation was not conducted. Multicenter research should be conducted to further increase the model's performance and robustness. Third, there is a significant difference in the mean Hb concentration between the Hb level from ABG and the standard venous analyzers, so images labeled with Hb measured by the standard venous analyzers should be enrolled to correct bias.

5. Conclusion

In summary, we developed a method to estimate the exact concentration of Hb based on a combination model of mask R-CNN and MobileNetV3, which achieved an R^2 of 0.503 (95% CI, 0.499–0.507) and an MAE of 1.6 g/dL (95% CI, 1.6–1.6). It can help medical staff's decision-making with high efficiency, especially in disaster rescue scenes, field rescue, emergent public health events, casualty transport, and battlefield rescue. Furthermore, our method was more convenient and simpler than previous ones since manually selecting the conjunctiva is no longer needed before inputting the images.

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 the Institutional Ethics Committee of the First Affiliated Hospital of Third Military Medical University (KY2021060). The patients/participants provided their written informed consent to participate in this study.

Author contributions

BY and Y-wC supervised the study designation, analysis, and manuscript edits. X-yH performed data acquisition, image processing, and manuscript drafting. Y-jL performed study designation, statistical analysis, and manuscript drafting. XS and Y-zS acquired data and screened samples. A-lS made implementation, figure creation, and manuscript edits. HL has made data acquisition, sample screening, and manuscript edits. Y-sL has made contributions to enroll participants and sign informed consent. X-fW has made contributions to image processing and obtained funding. L-fT performed the acquisition and interpretation of the data. C-yY made contributions to the discussion of study designation and data acquisition. Z-yY performed data analysis and drafted the manuscript. L-qX contributed to the technical, implementation, figure creation, and manuscript edits. All authors contributed to the article and approved the submitted version.

Funding

The study is supported by BY's National Key R&D Program of China (No. 2018YFC0116702), National Natural Science Foundation of China (No. 82070630), Medical Science and Technology Innovation Special Project (2023DZXZZ006), and Chongqing Science and Health Joint Medical Research Project (2020FYYX076). Y-wC's Youth innovation promotion association of the Chinese Academy of Sciences (2020377). Y-jL's Special support for Chongqing postdoctoral research project in 2020 and National Natural Science Foundation of China (No. 82100658). X-fW's Undergraduate research and training program of Third Military Medical University (No. 2021XBK19).

Acknowledgments

We thanked all participants for their contributions to this research.

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.2023.1151996/full#supplementary-material

References

1. Miller RD. *Patient Blood Management: Transfusion Therapy. Miller's Anesthesia, 8/E*. Singapore: Elesiver (2016), p. 2968.

2. Sheth TN CN, Bowes M, Detsky AS. The relation of conjunctival pallor to the presence of anemia. *J Gen Intern Med.* (1997) 12:102–6. doi: 10.1007/s11606-006-5004-x

3. Dimauro G, Simone L. Novel biased normalized cuts approach for the automatic segmentation of the conjunctiva. *Electronics.* (2020) 9:997. doi: 10.3390/electronics9060997

4. Kasiviswanathan S, Bai Vijayan T, Simone L, Dimauro G. Semantic segmentation of conjunctiva region for non-invasive anemia detection applications. *Electronics.* (2020) 9:1309. doi: 10.3390/electronics9081309

5. Noor NB, Anwar MS, Dey M. Comparative study between decision tree, svm and knn to predict anaemic condition. *BECITHCON.* (2019) 5:24–8. doi: 10.1109/BECITHCON48839.2019.9063188

6. Mitani A, Huang A, Venugopalan S, Corrado GS, Peng L, Webster DR, et al. Detection of anaemia from retinal fundus images via deep learning. *Nat Biomed Eng.* (2020) 4:18–27. doi: 10.1038/s41551-019-0487-z

7. Chen YM, Miaou SG, Bian H. Examining palpebral conjunctiva for anemia assessment with image processing methods. *Comput Methods Programs Biomed.* (2016) 137:125–35. doi: 10.1016/j.cmpb.2016.08.025

8. Gerson Delgado-Rivera AR-G, Alva-Mantari A, Saldivar-Espinoza B, Mirko Z, Franklin BP, Mario SB. *Method for the Automatic Segmentation of the Palpebral Conjunctiva using Image Processing. IEEE International Conference on Automation/XXIII Congress of the Chilean Association of Automatic Control (ICA-ACCA).* Piscataway, NJ: IEEE (2018).

9. Bevilacqua V, Dimauro G, Marino F, Brunetti A, Cassano F, Di Maio A, et al. A novel approach to evaluate blood parameters using computer vision techniques. In: *IEEE International Symposium on Medical Measurements and Applications, MeMeA 2016 - Proceedings.* (2016). p. 7533760. doi: 10.1109/MeMeA.2016.7533760

10. Dimauro G, Caivano D, Girardi F. A new method and a non-invasive device to estimate anemia based on digital images of the conjunctiva. *IEEE Access.* (2018) 6:46968–75. doi: 10.1109/ACCESS.2018.2867110

11. Dimauro G, Guarini A, Caivano D, Girardi F, Pasciolla C, Iacobazzi A. Detecting clinical signs of anaemia from digital images of the palpebral conjunctiva. *IEEE Access.* (2019) 7:113488–98. doi: 10.1109/ACCESS.2019.2932274

12. Suner S, Crawford G, McMurdy J, Jay G. Non-invasive determination of hemoglobin by digital photography of palpebral conjunctiva. *J Emerg Med.* (2007) 33:105–11. doi: 10.1016/j.jemermed.2007.02.011

13. Dimauro G, Baldari L, Caivano D, Colucci G, Girardi F. Automatic segmentation of relevant sections of the conjunctiva for non-invasive anemia detection. In: *International Conference on Smart and Sustainable Technologies (SpliTech).* Split: IEEE (2018). p. 1–5. Available online at: https://ieeexplore.ieee.org/document/8448335

14. Kaiming He GG, Piotr D, Ross G. *Mask R-CNN. IEEE International Conference on Computer Vision (ICCV)*. Piscataway, NJ: IEEE (2017).

15. Ren S, He K, Girshick R, Sun J. Faster R-CNN: towards real-time object detection with region proposal networks. *IEEE Trans Pattern Anal Mach Intell.* (2017) 39:1137–49. doi: 10.1109/TPAMI.2016.2577031

16. Lin T-Y, Dollar P, Girshick R, He K, Hariharan B, Belongie S. *Feature Pyramid Networks for Object Detection. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. Piscataway, NJ: IEEE (2017), p. 936–44.

17. Howard A, Sandler M, Chu G, Chen L-C, Chen B, Tan M, et al. Searching for MobileNetV3. In: *IEEE International Conference on Computer Vision (ICCV).* Seoul: IEEE (2019). p. 1314–24. doi: 10.1109/ICCV.2019.00140

18. Patil I. Visualizations with statistical details: the 'ggstatsplot' approach. *J Open Source Softw.* (2021) 6:3167. doi: 10.21105/joss.03167

19. Wickham H. *ggplot2: Elegant Graphics for Data Analysis*. New York, NY: Springer-Verlag New York (2016).

20. Collings S, Thompson O, Hirst E, Goossens L, George A, Weinkove R. *Non-invasive detection of anaemia using digital photographs of the conjunctiva*. *PloS ONE.* (2016) 11:286. doi: 10.1371/journal.pone.0153286

21. Jain P, Bauskar S, Gyanchandani M. Neural network based non-invasive method to detect anemia from images of eye conjunctiva. *Int J Imaging Syst Technol.* (2019) 30:112–25. doi: 10.1002/ima.22359

22. Muthalagu R. A Smart (phone) solution: an effective tool for screening anaemia - correlation with conjunctiva pallor and haemoglobin levels. *TAGA J.* (2018) 14:2611–21.

23. Saldivar-Espinoza B, Núñez-Fernández D, Porras-Barrientos F, Alva-Mantari A, Leslie LS, Zimic M. Portable system for the prediction of anemia based on the ocular conjunctiva using Artificial Intelligence. *arXiv [Preprint].* (2019). arXiv: 1910.12399. Available online at: https://arxiv.org/pdf/1910.12399.pdf

frontiers | Frontiers in Medicine

TYPE Perspective
PUBLISHED 30 August 2023
DOI 10.3389/fmed.2023.1241508

OPEN ACCESS

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

REVIEWED BY
Harpreet Singh Grewal,
Radiology Associates of Florida, United States
Akshatha Kamath,
Bayhealth Medical Center, United States

*CORRESPONDENCE
Kwadwo Kyeremanteng
✉ kkyeremanteng@toh.ca

RECEIVED 16 June 2023
ACCEPTED 18 August 2023
PUBLISHED 30 August 2023

CITATION
Hryciw BN, Fortin Z, Ghossein J and Kyeremanteng K (2023) Doctor-patient interactions in the age of AI: navigating innovation and expertise.
Front. Med. 10:1241508.
doi: 10.3389/fmed.2023.1241508

Doctor-patient interactions in the age of AI: navigating innovation and expertise

Brett N. Hryciw[1], Zanna Fortin[2], Jamie Ghossein[3] and Kwadwo Kyeremanteng[1,4,5]*

[1]Department of Medicine, Division of Critical Care, University of Ottawa, Ottawa, ON, Canada, [2]Gemeinschaftspraxis im Bayerwald, Bavaria, Germany, [3]Department of Medicine, University of Ottawa, Ottawa, ON, Canada, [4]Clinical Epidemiology, Ottawa Hospital Research Institute, University of Ottawa, Ottawa, ON, Canada, [5]Institute du Savoir Montfort, Ottawa, ON, Canada

The integration of artificial intelligence (AI) in healthcare has the capacity to transform medical practice. Despite its revolutionary potential, the influence of AI may affect the physician-patient interaction and presents ethical challenges that will need to be carefully considered. This article discusses how patients may interact with this technology, considers how emerging technologies may alter the dynamics of the physician-patient relationship, and reviews some of the limitations that continue to exist. We identify potential challenges that may arise with the integration of AI into medical settings and propose solutions to help mitigate these issues.

KEYWORDS

artificial intelligence, healthcare, decision-making, patient care, ethics, medical practice

Introduction

The adoption of artificial intelligence (AI) in healthcare has the potential to revolutionize medical practice, improving diagnostics, treatment planning, and overall patient care (1). However, the integration of AI into clinical settings also presents new challenges for doctor-patient interactions, as well as ethical concerns that must be carefully considered. In this article, we will explore the complexities of patients and families introducing AI-generated medical opinions into doctor-patient relationships and discuss strategies for effectively navigating these challenges.

Patients as technology consumers

With the development of the internet, patients have been increasingly empowered to become informed about their health, allowing them to access a wealth of medical information from various sources. As a result, patients can take an increasingly active role in their healthcare decision-making (2). This has both positive and negative implications for doctor-patient relationships. For instance, when patients bring internet-generated opinions to their medical appointments, it can promote informed discussions and better decision-making. On the other hand, patients may develop rigid beliefs about optimal medical management based on internet advice, which may not align with the doctor's professional opinion, potentially straining the therapeutic alliance (3).

Addressing the shift in dynamics

Similarly, as AI-generated medical opinions become more accessible and reliable, patients may turn to AI software to provide advice regarding their healthcare. Consequently, the rise of AI-generated medical opinions will likely lead to a further shift in dynamics between physicians, who historically held all the knowledge and expertise, and patients or family members, who can now access AI-generated opinions with increasing sophistication and accuracy through large language model (LLM) chatbots such as OpenAI's GPT-4 or Google's Bard. A medicine specific consumer technology, Glass AI, is a GPT-4 based technology where users present a clinical scenario and subsequently a differential diagnosis or clinical management plan is generated. As these technologies become more mainstream, it seems likely that patients will arrive to clinical encounters with specific expectations for next steps in their care. The advantage of AI lies in its ability to process large volumes of data and identify patterns that may not be readily apparent to human clinicians, which has revolutionary potential in the age of AI innovation (4, 5). Already these technologies have been applied to a range of clinical scenarios and, in select circumstances, may be able to recognize biological signatures in patient data that is beyond human interpretation. In fact, an automated deep learning model of retinal fundus photographs from a UK database was able to reliably predict a patient's reported sex which is beyond human capabilities (6).

While empowerment of patients and families to participate in health care decision is undeniably important, those who choose to seek AI-generated medical opinions could strain doctor-patient relationships if the physician feels threatened or if families do not accept the current limitations of these tools and believe that the AI-generated opinions are superior. This possibility necessitates a more collaborative approach in doctor-patient relationships, emphasizing partnership and shared decision-making with an openness to discussing AI-generated opinions. Physicians should be encouraged to embrace this shift and actively engage patients and families as partners in the decision-making process, acknowledging the value of AI-generated insights while maintaining their unique role as human experts (7).

Identifying ethical concerns

Amidst the evolving domain of AI ethics, its implications in medicine raise concerns of informed consent, training biases, and transparency among others. Firstly, informed consent is a core component of medical ethics but has the potential to be compromised by providing misinformation. AI algorithms are not infallible and can produce false or misleading information, known as AI hallucinations. These errors can arise from biases in the training data or limitations in the AI's understanding of complex medical scenarios (8, 9). Overreliance on AI-generated opinions by patients may in fact lead to suboptimal healthcare decisions and outcomes when the uniqueness of individual patients, the broader clinical context, and the expertise of human clinicians are not appropriately considered. Further, with the potential for unrecognized AI hallucinations, knowledge provided to the patient and families has the potential to bias and misinform patients, in turn clouding judgement. This is particularly relevant as patients receive and place increasing value on AI-generated advice without understanding its limitations. This may in fact compromise patient autonomy and lead to ill-informed decision-making (7, 10). Additionally, LLMs are limited by their training data set. Inherent biases can arise when AI is trained on non-representative patient data, potentially leading to less accurate predictions for underrepresented populations or diseases (8, 11). Unfortunately, underrepresentation biases often further disadvantage marginalized populations. Consequently, physicians may be obliged to educate patients while relying on their expertise and judgment to interpret AI advice in the context of the individual patient's condition and needs, helping to mitigate potential biases. Transparency is a cornerstone of the evolving physician-patient relationship in the era of AI-driven healthcare. As AI systems can sometimes be perceived as "black boxes" with their complex decision-making processes, physicians must highlight that while AI can provide useful information, it may not yet consider all relevant factors or nuances of a patient's unique circumstances that are considered by a human physician. Lastly, the question of who bears the responsibility when AI-based decisions lead to poor patient choices or adverse patient outcomes remains an ongoing debate. Clear guidelines on responsibility attribution and informed consent procedures are needed to address this issue.

Involving patients in decision-making processes

Patient involvement in the decision-making process is paramount for promoting responsible AI integration in healthcare. By engaging patients with AI-generated insights, physicians can ensure that these are considered alongside human expertise and experience, as well as the patient's preferences and unique circumstances (12). Patient-centered care models, which focus on active collaboration between patients, families, and healthcare providers, can help achieve this goal (13). By fostering a patient-centered approach, healthcare providers can maintain the human element of care while leveraging the benefits of AI-generated medical opinions.

Healthcare providers must also be educated about AI's capabilities and limitations, enabling them to effectively explain AI-generated opinions to their patients (14). This can be achieved through targeted training programs and patient education initiatives, promoting ethical AI adoption and informed decision-making. By enhancing patient awareness of AI's capabilities and limitations, healthcare providers can help to ensure that patients make well-informed decisions based on a combination of human expertise and AI-generated insights (10). Nevertheless, when an AI-generated opinion that resonates with patients or families differs from a doctor's recommendation, this discrepancy may deter patients and family members from accepting the medical opinion. Clinicians must be prepared to explain their reasoning and engage in open conversations with patients to address potential concerns. Transparent communication is essential in maintaining trust and fostering collaborative decision-making in the age of AI.

Future directions

It seems increasingly inevitable that AI, much like the internet previously, will permeate many aspects of society. Almost certainly,

patients be among AI consumers who will turn to these technologies to illicit medical advice when access to healthcare is not readily available. As AI tools become increasingly reliable, validated consumer tools should be trained and validated using data that encompasses the local diversity of the patient population they are meant to service. This approach could reduce potential underrepresentation biases and disparities in healthcare outcomes and enhance the tool's relevance and effectiveness. One of the most compelling advantages of AI is its potential to alleviate limitations in healthcare access. For example, AI tools may eventually be capable of triaging patient concerns, identifying those requiring immediate attention and those suitable for virtual consultations. This would not only enhance resource allocation but also extend the reach of healthcare to underserved populations.

Concurrently, the importance of AI training programs for medical professionals cannot be overstated. As we transition into an era of AI-augmented healthcare, it's essential that our doctors, nurses, and other healthcare workers are equipped with the necessary skills to navigate this new landscape. They need to understand how to integrate AI-generated advice into their practice and communicate these insights effectively to patients. This training will not only augment their ability to provide care but also bolster their confidence as they navigate this new frontier in medicine. Patient education and engagement are equally vital. Patients, now more than ever, are active participants in their healthcare journey. As such, they must be equipped with a basic understanding of AI's strengths and limitations. Educational resources or initiatives could help patients make sense of AI-generated insights, promoting informed discussions and decision-making while promoting trust in their healthcare providers.

Additionally, as we grapple with the ethical and logistical aspects of AI deployment, longitudinal studies can offer much-needed insight into AI's real-world impact over time. Concurrently, a cost–benefit analysis is crucial. While AI's immense potential cannot be understated, the cost associated with integrating AI into healthcare systems must be justifiable.

Conclusion

The integration of AI into healthcare is inevitable and offers many benefits for patient care. However, it is crucial to address potential challenges in doctor-patient interactions and maintain trust in the face of AI-generated medical opinions. By fostering open communication, recognizing AI's limitations, and valuing human expertise, clinicians can successfully navigate the evolving landscape of healthcare and ensure the best possible care for their patients. The education for healthcare providers and involving patients in decision-making processes are essential strategies for the responsible integration of AI in healthcare. As we move forward with integrating AI into healthcare, it's paramount that we do so with a thoughtful and comprehensive approach. Ensuring effective regulation, standardization, and education will pave the way for a healthcare landscape where AI is not just a tool for doctors, but an ally for patients as well.

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

KK and BH contributed to the conceptualization of the paper. AI-generated content from structured and refined input data using GPT-4 was overseen by BH. BH, ZF, and JG monitored content generation iteratively and were responsible for editing, revising and validating all AI-generated content to ensure reliability and accuracy. KK was responsible for overseeing project completion. 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

References

1. Jiang F, Jiang Y, Zhi H, Dong Y, Li H, Ma S, et al. Artificial intelligence in healthcare: past, present and future. *Stroke Vasc Neurol*. (2017) 2:230–43. doi: 10.1136/svn-2017-000101

2. Tang G, Goh K, Stine R, Zavala E, Gupta A. The impact of the internet on patient empowerment: a systematic review. *Health Inform*. (2016) 22:202–20. doi: 10.1177/1460458214563275

3. Laranjo L, Dunn AG, Tong HL, Kocaballi AB, Chen J, Bashir R, et al. Conversational agents in healthcare: a systematic review. *J Am Med Inform Assoc*. (2018) 25:1248–58. doi: 10.1093/jamia/ocy072

4. Esteva A, Kuprel B, Novoa RA, Ko J, Swetter SM, Blau HM, et al. Dermatologist-level classification of skin cancer with deep neural networks. *Nature*. (2017) 542:115–8. doi: 10.1038/nature21056

5. Rajpurkar P, Irvin J, Zhu K, Yang B, Mehta H, Duan T, et al. CheXNet: Radiologist-level pneumonia detection on chest X-rays with deep learning. *arXiv preprint*. (2017) arXiv:1711.05225. doi: 10.48550/arXiv.1711.05225

6. Korot E, Pontikos N, Liu X, Wagner SK, Faes L, Huemer J, et al. Predicting sex from retinal fundus photographs using automated deep learning. *Sci Rep*. (2021) 11:10286. doi: 10.1038/s41598-021-89743-x

7. Leslie D (2019), Understanding artificial intelligence ethics and safety: A guide for the responsible design and implementation of AI Systems in the Public Sector. Social Science Research Network, Available at: https://ssrn.com/abstract=3403301.

8. 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

9. Char DS, Shah NH, Magnus D. Implementing machine learning in health care—addressing ethical challenges. *N Engl J Med*. (2018) 378:981–3. doi: 10.1056/NEJMp1714229

10. Ji Z, Lee N, Frieske R, Yu T, Su D, Xu Y, et al. Survey of hallucination in natural language generation. *ACM Comput Surv*. (2022) 55:1–38. doi: 10.1145/3571730

11. Rajkomar A, Hardt M, Howell MD, Corrado G, Chin MH. Ensuring fairness in machine learning to advance health equity. *Ann Intern Med*. (2018) 169:866–72. doi: 10.7326/M18-1990

12. Aminololama-Shakeri S, López JE. The doctor-patient relationship with artificial intelligence. *Am J Roentgenol*. (2019) 212:308–10. doi: 10.2214/AJR.18.20509

13. Stacey D, Légaré F, Lewis K, Barry MJ, Bennett CL, Eden KB, et al. Decision aids for people facing health treatment or screening decisions. *Cochrane Database Syst Rev*. (2017) 2017:CD001431. doi: 10.1002/14651858.CD001431.pub5

14. Topol EJ. High-performance medicine: the convergence of human and artificial intelligence. *Nat Med*. (2019) 25:44–56. doi: 10.1038/s41591-018-0300-7

frontiers | Frontiers in Medicine

TYPE Original Research
PUBLISHED 04 October 2023
DOI 10.3389/fmed.2023.1089087

OPEN ACCESS

EDITED BY
Gulzar H. Shah,
Georgia Southern University, United States

REVIEWED BY
Kristie Cason Waterfield,
Georgia Southern University, United States
Hong Qin,
University of Tennessee at Chattanooga, United States

*CORRESPONDENCE
Rahul Kashyap
✉ kashyapmd@gmail.com

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

RECEIVED 20 April 2023
ACCEPTED 14 September 2023
PUBLISHED 04 October 2023

CITATION
Valencia Morales DJ, Bansal V, Heavner SF, Castro JC, Sharma M, Tekin A, Bogojevic M, Zec S, Sharma N, Cartin-Ceba R, Nanchal RS, Sanghavi DK, La Nou AT, Khan SA, Belden KA, Chen J-T, Melamed RR, Sayed IA, Reilkoff RA, Herasevich V, Domecq Garces JP, Walkey AJ, Boman K, Kumar VK and Kashyap R (2023) Validation of automated data abstraction for SCCM discovery VIRUS COVID-19 registry: practical EHR export pathways (VIRUS-PEEP). *Front. Med.* 10:1089087.
doi: 10.3389/fmed.2023.1089087

Validation of automated data abstraction for SCCM discovery VIRUS COVID-19 registry: practical EHR export pathways (VIRUS-PEEP)

Diana J. Valencia Morales[1†], Vikas Bansal[2†], Smith F. Heavner[3], Janna C. Castro[4], Mayank Sharma[1], Aysun Tekin[1], Marija Bogojevic[1], Simon Zec[1], Nikhil Sharma[2], Rodrigo Cartin-Ceba[5], Rahul S. Nanchal[6], Devang K. Sanghavi[7], Abigail T. La Nou[8], Syed A. Khan[9], Katherine A. Belden[10], Jen-Ting Chen[11], Roman R. Melamed[12], Imran A. Sayed[13], Ronald A. Reilkoff[14], Vitaly Herasevich[1], Juan Pablo Domecq Garces[2], Allan J. Walkey[15], Karen Boman[16], Vishakha K. Kumar[16] and Rahul Kashyap[1*] on behalf of Society of Critical Care Medicine's Discovery, the Critical Care Research Network

[1]Division of Critical Care Medicine, Department of Anesthesiology and Perioperative Care, Mayo Clinic, Rochester, MN, United States, [2]Division of Nephrology and Critical Care Medicine, Department of Internal Medicine, Mayo Clinic, Rochester, MN, United States, [3]CURE Drug Repurposing Collaboratory, Critical Path Institute, Tucson, AZ, United States, [4]Department of Information Technology, Mayo Clinic, Scottsdale, AZ, United States, [5]Division of Critical Care Medicine, Department of Pulmonary Medicine, Mayo Clinic, Scottsdale, AZ, United States, [6]Division of Pulmonary and Critical Care Medicine, Department of Internal Medicine, Medical College of Wisconsin, Milwaukee, WI, United States, [7]Department of Critical Care Medicine, Mayo Clinic Florida, Jacksonville, FL, United States, [8]Department of Critical Care Medicine, Mayo Clinic Health System, Eau Claire, WI, United States, [9]Department of Critical Care Medicine, Mayo Clinic Health System, Mankato, MN, United States, [10]Division of Infectious Diseases, Sidney Kimmel Medical College at Thomas Jefferson University, Philadelphia, PA, United States, [11]Division of Critical Care Medicine, Department of Internal Medicine, Montefiore Medical Center, Albert Einstein College of Medicine, Bronx, NY, United States, [12]Department of Critical Care Medicine, Abbott Northwestern Hospital, Allina Health, Minneapolis, MN, United States, [13]Department of Pediatrics, Children's Hospital of Colorado, University of Colorado Anschutz Medical Campus, Colorado Springs, CO, United States, [14]Division of Pulmonary, Allergy, Critical Care and Sleep Medicine, Department of Internal Medicine, University of Minnesota Medical School, Edina, MN, United States, [15]Division of Pulmonary, Allergy, Critical Care and Sleep Medicine, Department of Medicine, Evans Center of Implementation and Improvement Sciences, Boston University School of Medicine, Boston, MA, United States, [16]Society of Critical Care Medicine, Mount Prospect, IL, United States

Background: The gold standard for gathering data from electronic health records (EHR) has been manual data extraction; however, this requires vast resources and personnel. Automation of this process reduces resource burdens and expands research opportunities.

Objective: This study aimed to determine the feasibility and reliability of automated data extraction in a large registry of adult COVID-19 patients.

Materials and methods: This observational study included data from sites participating in the SCCM Discovery VIRUS COVID-19 registry. Important demographic, comorbidity, and outcome variables were chosen for manual and automated extraction for the feasibility dataset. We quantified the degree of

agreement with Cohen's kappa statistics for categorical variables. The sensitivity and specificity were also assessed. Correlations for continuous variables were assessed with Pearson's correlation coefficient and Bland–Altman plots. The strength of agreement was defined as almost perfect (0.81–1.00), substantial (0.61–0.80), and moderate (0.41–0.60) based on kappa statistics. Pearson correlations were classified as trivial (0.00–0.30), low (0.30–0.50), moderate (0.50–0.70), high (0.70–0.90), and extremely high (0.90–1.00).

Measurements and main results: The cohort included 652 patients from 11 sites. The agreement between manual and automated extraction for categorical variables was almost perfect in 13 (72.2%) variables (Race, Ethnicity, Sex, Coronary Artery Disease, Hypertension, Congestive Heart Failure, Asthma, Diabetes Mellitus, ICU admission rate, IMV rate, HFNC rate, ICU and Hospital Discharge Status), and substantial in five (27.8%) (COPD, CKD, Dyslipidemia/Hyperlipidemia, NIMV, and ECMO rate). The correlations were extremely high in three (42.9%) variables (age, weight, and hospital LOS) and high in four (57.1%) of the continuous variables (Height, Days to ICU admission, ICU LOS, and IMV days). The average sensitivity and specificity for the categorical data were 90.7 and 96.9%.

Conclusion and relevance: Our study confirms the feasibility and validity of an automated process to gather data from the EHR.

KEYWORDS

validation, data automation, electronic health records, COVID-19, VIRUS COVID-19 registry

Introduction

The pandemic of the coronavirus disease 2019 (COVID-19) has created a need to develop research resources rapidly (1). In response to the global demand for robust multicenter clinical data regarding patient care and outcomes, the Society of Critical Care Medicine (SCCM) Discovery Viral Infection and Respiratory Illness Universal Study (VIRUS) COVID-19 registry was created early in the pandemic (2–4).

Due to the surging nature of pandemic waves, and the subsequent workload and staffing burdens, clinical researchers have encountered difficulty in engaging in rapid, reliable manual data extraction from the electronic health record (EHR) (5). Manual chart review is the gold standard method for gathering data for retrospective research studies (6, 7). This process, however, is time consuming and necessitates personnel resources not widely available at all institutions (8, 9). Prior to the pandemic, automated data extraction from the EHR utilizing direct database queries was shown to be faster and less error-pone than manual data extraction (8, 10). Nonetheless, data quality challenges related to high complexity or fragmentation of data across many EHR systems make automated extraction vulnerable (11–14). Both manual and automatic extraction rely on the EHR, which is an artifact with its own biases, mistakes, and subjectivity (15–20).

Although previous research has looked at these notions, the best methods for obtaining data from EHR systems for research still need to be discovered. In response, we sought to assess the feasibility, reliability, and validity of an automated data extraction process using data for the VIRUS COVID-19 registry.

Methods

VIRUS COVID-19 registry

The SCCM Discovery VIRUS COVID-19 registry (Clinical Trials registration number: NCT04323787) is a multicenter, international database with over 80,000 patients from 306 health sites across 28 countries (21). VIRUS COVID-19 registry is an ongoing prospective observational study that aims at real-time data gathering and analytics with a feedback loop to disseminate treatment and outcome knowledge to improve COVID-19 patient care (3). The Mayo Clinic Institutional Review Board authorized the SCCM Discovery VIRUS COVID-19 registry as exempt on March 23, 2020 (IRB number: 20–002610). No informed consent was deemed necessary for the study subjects. The procedures were followed in accordance with the Helsinki Declaration of 2013 (22). Among the participating sites, 30 individual centers are collaborating to rapidly develop tools and resources to optimize EHR data collection. These efforts are led by the VIRUS Practical EHR Export Pathways group (VIRUS-PEEP).

Abbreviations: CAD, Coronary artery disease; CHF, Congestive heart failure; CI, Confidence interval; CKD, Chronic kidney disease; COPD, Chronic obstructive pulmonary disease; CRF, Case report forms; DM, Diabetes mellitus; ECMO, Extracorporeal membrane oxygenation; EHR, Electronic health records; HFNC, High flow nasal canula; HTN, Hypertension; ICU, Intensive care unit; IMV, Invasive mechanical ventilation; IRB, Institutional review boards; LOS, Length of stay; NIMV, Non-invasive mechanical ventilation; PCC, Pearson interclass correlation coefficient; REDCap, Research electronic data capture software; SCCM, Society of critical care medicine; SD, Standard deviations; SE, Standard error; SFTP, Secure file transfer platform; SOP, Standard operating procedure; SQL, Sequential query language; VIRUS, Viral Infection and Respiratory Illness Universal Study; VIRUS-PEEP, VIRUS Practical EHR Export Pathways group; WHO, World Health Organization; WHO-ISARIC, World Health Organization- International Severe Acute Respiratory And Emerging Infection Consortium.

Data collection

The VIRUS COVID-19 registry has over 500 variables which represents the pandemic registry common data standards for critically ill patients adapted from the World Health Organization- International Severe Acute Respiratory and Emerging Infection Consortium (WHO-ISARIC) COVID-19 CRF v1.3 24 February 2020 (23). The VIRUS-PEEP validation cohort was developed in an iterative, consensus process by a group of VIRUS: COVID-19 registry primary investigators to explore the feasibility of an automation process at each site. The Validation cohort variable was internally validated with seven core VIRUS COVID-19 investigators and subsequently validated from VIRUS-PEEP leads site's principal investigators. Because of the timeline, the cohort could not be externally validated. A purposeful representative sample of the 25 most clinically relevant variables from each category (Baseline demographic and clinical characteristics of patient and ICU and Hospital-related outcomes) were selected and prioritized for this study (4). We focused on demographic data (age, sex, race, ethnicity, height, weight), comorbidities (coronary artery disease (CAD), hypertension (HTN), congestive heart failure (CHF), chronic obstructive pulmonary disease (COPD), asthma, chronic kidney disease (CKD), diabetes mellitus (DM), dyslipidemia/hyperlipidemia), and clinical outcomes (intensive care unit (ICU) admission, days to ICU admission, ICU length of stay (LOS), type to oxygenation requirement, extracorporeal membrane oxygenation (ECMO), ICU discharge status, hospital LOS, and in-hospital mortality).

To avoid data extraction errors, we utilized precise variable definitions [VIRUS COVID-19 registry code book, cases report form (CRF), and Standard Operating Procedure (SOP)], which were already implemented in the registry and during the pilot phase of the automation implementation. Additionally, all manual and automation data extraction personnel were educated regarding the definitions and procedures needed to collect and report the data.

System description

De-identified data were collected through Research Electronic Data Capture software (REDCap, version 8.11.11, Vanderbilt University, Nashville, Tennessee) at Mayo Clinic, Rochester, MN, United States (24). The REDCap electronic data capture system is a secure, web-based application for research data capture that includes an intuitive interface for validated data entry; audit trails for tracking data manipulation and export procedures; automated export procedures for seamless data downloads to standard statistical packages; and provide a secure platform for importing data from external sources.

Manual abstraction

The VIRUS PEEP group has implemented a comprehensive process for data extraction, which involves training manual data extractors. These data extractors are trained to identify, abstract, and collect patient data according to the project's SOP. During a patient's hospitalization, extractors follow them until discharge, ensuring that all relevant information is collected. The CRF used in this process includes two main sections: demographics and outcomes, composed of categorical and continuous variables. Extractors answer a mix of binary ("yes" or "no") and checkbox ("check all that apply") questions in the nominal variable portions of the CRF. They are instructed to avoid free text and use the prespecified units for continuous variables. In any disagreement, a trainer is always available for guidance and correction. It's important to note that the manual extractors are unaware of the automated data extraction results.

Automated extraction

A package of sequential query language (SQL) scripts for the "Epic Clarity" database was developed at one institution and shared through the SCCM's Secure File Transfer Platform (SFTP) with participating sites. A second site offered peer coaching on the development and utility of end-user Epic™ reporting functions and how to adapt and modify the SQL scripts according to their EHR environment and security firewall. Other tools included R-Studio™ scripts, Microsoft Excel™ macros, STATA 16, and REDCap calculators for data quality checks at participating sites before data upload to VIRUS Registry REDCap. These tools were designed to aid in data extraction, data cleaning, and adherence to data quality rules as provided in VIRUS COVID-19 Registry SOPs. Institutions participated in weekly conference calls to discuss challenges and share successes in implementing automated data abstraction; additionally, lessons learned from adapting the SQL scripts and other data quality tools to their EHR environments were shared between individual sites and members of the VIRUS PEEP group.

Statistical analysis

We summarized continuous variables of manual and automation process data using mean ± SD and calculated mean difference and SE by matched pair analysis. Pearson correlation coefficient (PCCs) and 95% confidence intervals (CI) were generated for continuous data as a measure of inter-class dependability (25). Pearson correlations were classified as trivial (0.00–0.30), low (0.30–0.50), moderate (0.50–0.70), high (0.70–0.90), and extremely high (0.90–1.00) (26). Bland–Altman mean-difference plots for continuous variables were also provided to aid in the understanding of agreement (27).

Percent agreements were determined for the data collected using each of the two extraction techniques in a categorical variable:

$$\frac{\textit{Number of patients categorized identically by both sources}}{\textit{Total number of cases examined by both sources}}$$

The total number of agreeing outcomes divided by the total number of results is the summary agreement for each variable. For categorical variables we used Cohen's kappa coefficient (28). We used the scale created by Landis et al. to establish the degree of agreement (29). This scale is divided by almost perfect ($\varkappa$ =0.81–1.00), substantial ($\varkappa$ = 0.61–0.80), moderate ($\varkappa$ = 0.41–0.60), fair ($\varkappa$ = 0.21–0.40), slight ($\varkappa$ = 0.00–0.20), and poor ($\varkappa$ < 0.00). Additionally, the sensitivity and specificity were calculated by comparing the results of the automated data extractions method to the results of manual data extraction method (gold standard). The 95% confidence intervals were calculated using an exact test for proportions. We used JMP statistical software version 16.2 for all data analysis.

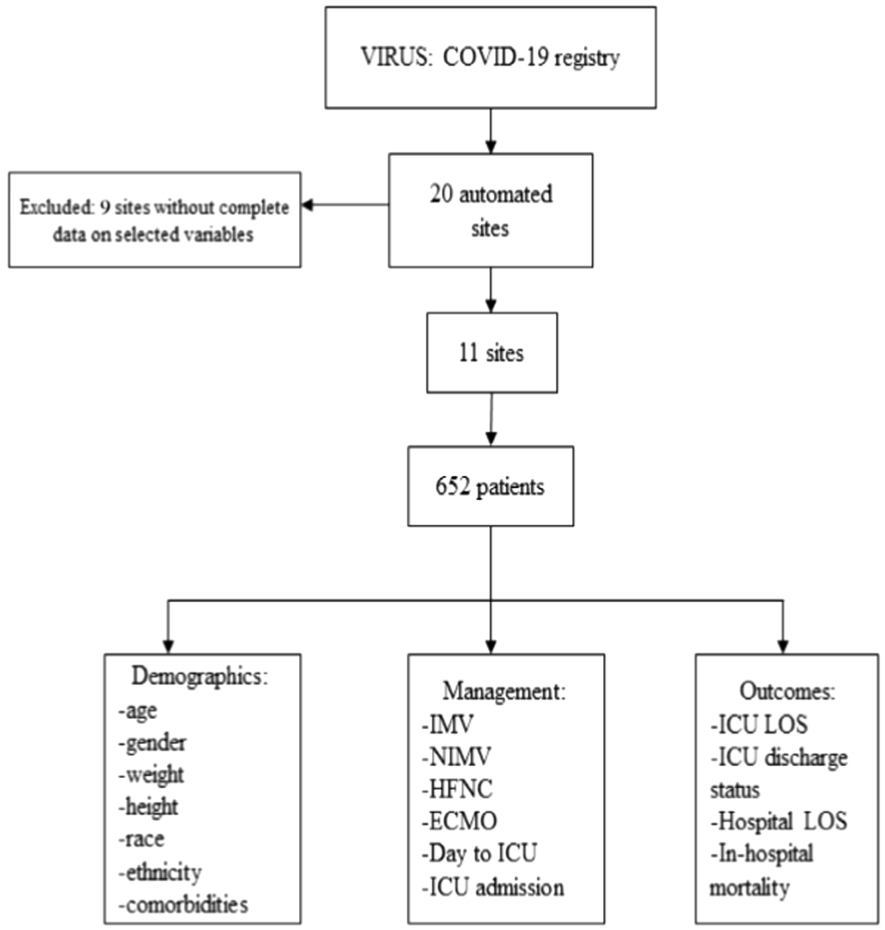

FIGURE 1
Study flowchart.

Results

Our cohort consisted of data from 652 patients from 11 sites (Figure 1). A total of 25 variables were collected for each patient for manual and automated methods. Of these 25 variables, 16 (64.0%) were nominal, 7 (28.0%) were continuous, and 2 (8.0%) were categorical variables.

Table 1 summarizes the continuous variables. The automated results for three variables (age, weight and hospital LOS) agreed "extremely high" (>90%) to the manual extraction results. The agreement was "high" (70–90%) for height, days to ICU admission, ICU LOS, and IMV days. Figure 2 presents the Bland–Altman plots for seven continuous variables.

Tables 2, 3 describe the ordinal and nominal variables. The agreement between manual and automated extraction was almost perfect in 13 (72.2%) of the studied variables, and substantial in five (27.8%). The comorbidity "dyslipidemia/hyperlipidemia" had the lowest degree of agreement (moderate 0.61); however, overall percent agreement was high (86.9%). The only variable that showed a Kappa Coefficient equal to 1 was "ICU-discharge status." The average Kappa Coefficient was 0.81 for the eight comorbidities collected and was 0.86 for outcomes variables, considered almost perfect. The automated electronic search strategy achieved an average sensitivity of 90.7% and a specificity of 96.9%. The sensitivity and specificity of each data-extraction method for all variables are presented in Table 3.

Discussion

The automated search strategy for EHR data extraction was highly feasible and reliable. Our investigation observed substantial and almost perfect agreement between automated and manual data extraction. There was almost perfect agreement in two-thirds of the categorical variables, and all continuous variables showed Extremely High or High agreement.

The results of our validation study are similar to other studies that validated and evaluated automated data (30–33). Singh et al. (31) developed several algorithm queries to identify every component of the Charlson Comorbidity Index and found median sensitivity and specificity of 98–100% and 98–100%, respectively. In the validation cohort, the sensitivity of the automated digital algorithm ranged from 91 to 100%, and the specificity ranged from 98 to 100% compared to

ICD-9 codes. These results are comparable to our study as the comorbidities analyzed presented a sensitivity and specificity of 90.2 and 96.8%, respectively. Our results are superior to the results of Schaerfer et al. (34), who found a sensitivity of 72% and a specificity of 95% for comorbidities (CHF, cerebral vascular disease, CKD, cancer, DM, human immunodeficiency virus, HTN) in patients with

TABLE 1 Comparison of patients in automated versus manual reviews and measures of agreement for individual responses for continuous variables.

Variable name	Automation (Mean, SD)	Manual (Mean, SD)	Mean difference (SE)	Pearson interclass correlation coefficient (PCC), 95% CI	Strength of agreement based on PCC
Age, *N* = 652	57.9 (21.9)	58.5 (19.9)	−0.5 (0.3)	0.95 (0.94–0.96)	Extremely High
Height, *N* = 632	167.6 (15.6)	167 (17.2)	0.6 (0.3)	0.89 (0.87–0.90)	High
Weight, *N* = 632	87.2 (27)	88.4 (28.5)	−1.2 (0.4)	0.94 (0.93–0.95)	Extremely High
Hospital LOS, *N* = 540	9.0 (9.1)	9.0 (9)	0.1 (0.1)	0.97 (0.96–0.97)	Extremely High
Days to ICU admission, *N* = 176	1.3 (3.3)	1.1 (2.6)	0.2 (0.1)	0.80 (0.74–0.85)	High
ICU LOS, *N* = 168	7.5 (9.3)	9.0 (10.5)	−1.5 (0.4)	0.88 (0.85–0.91)	High
IMV Days, *N* = 71	9.7 (9.6)	11.6 (11.1)	−1.9 (0.6)	0.88 (0.81–0.92)	High

CI, Confidence interval; ICU, Intensive Care Unit; IMV, Invasive Mechanical Ventilation; LOS, Length of stay; PCC, Pearson Interclass Correlation Coefficient; SD, Standard deviation; SE, Standard error.

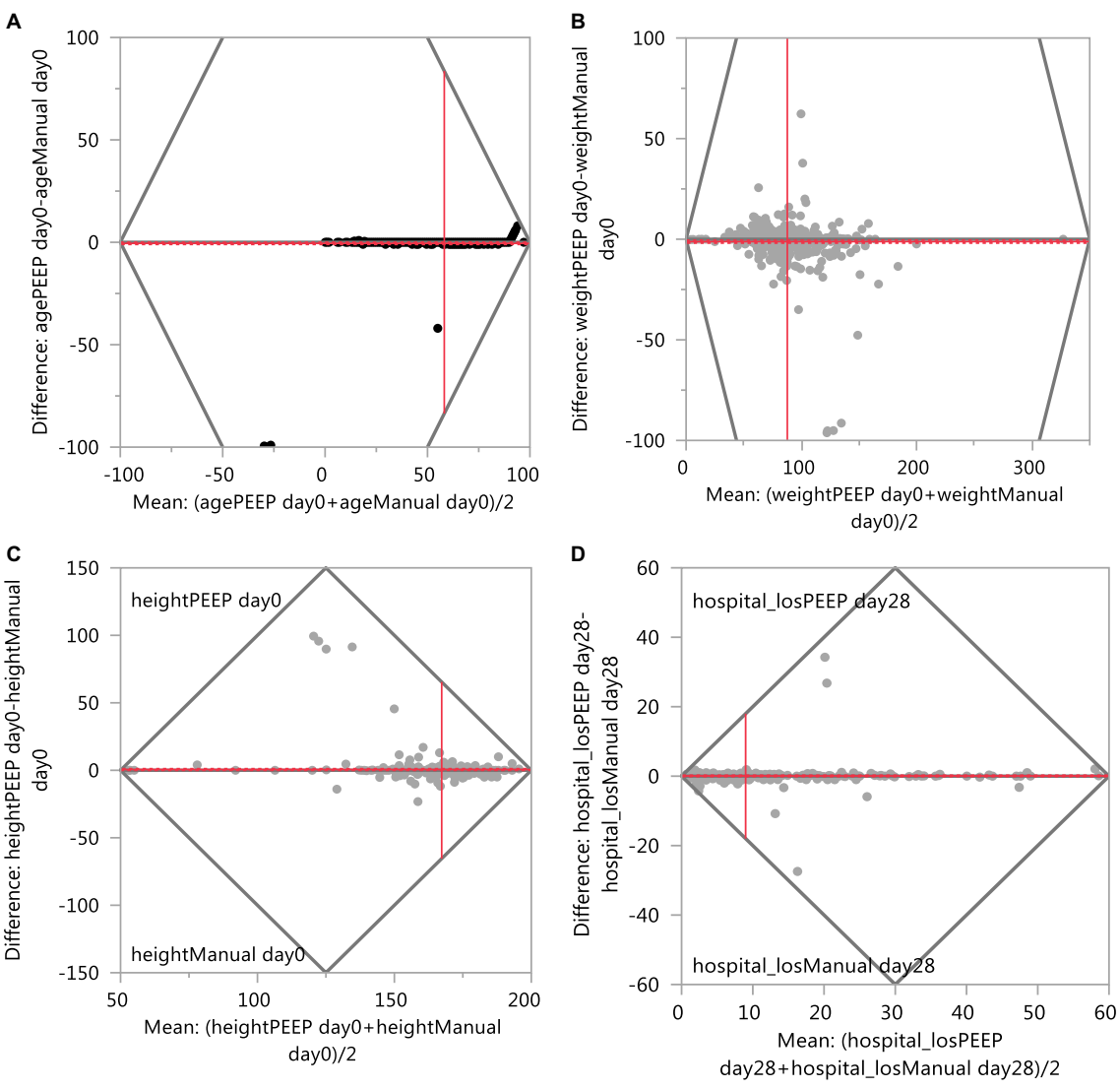

FIGURE 2 (Continued)

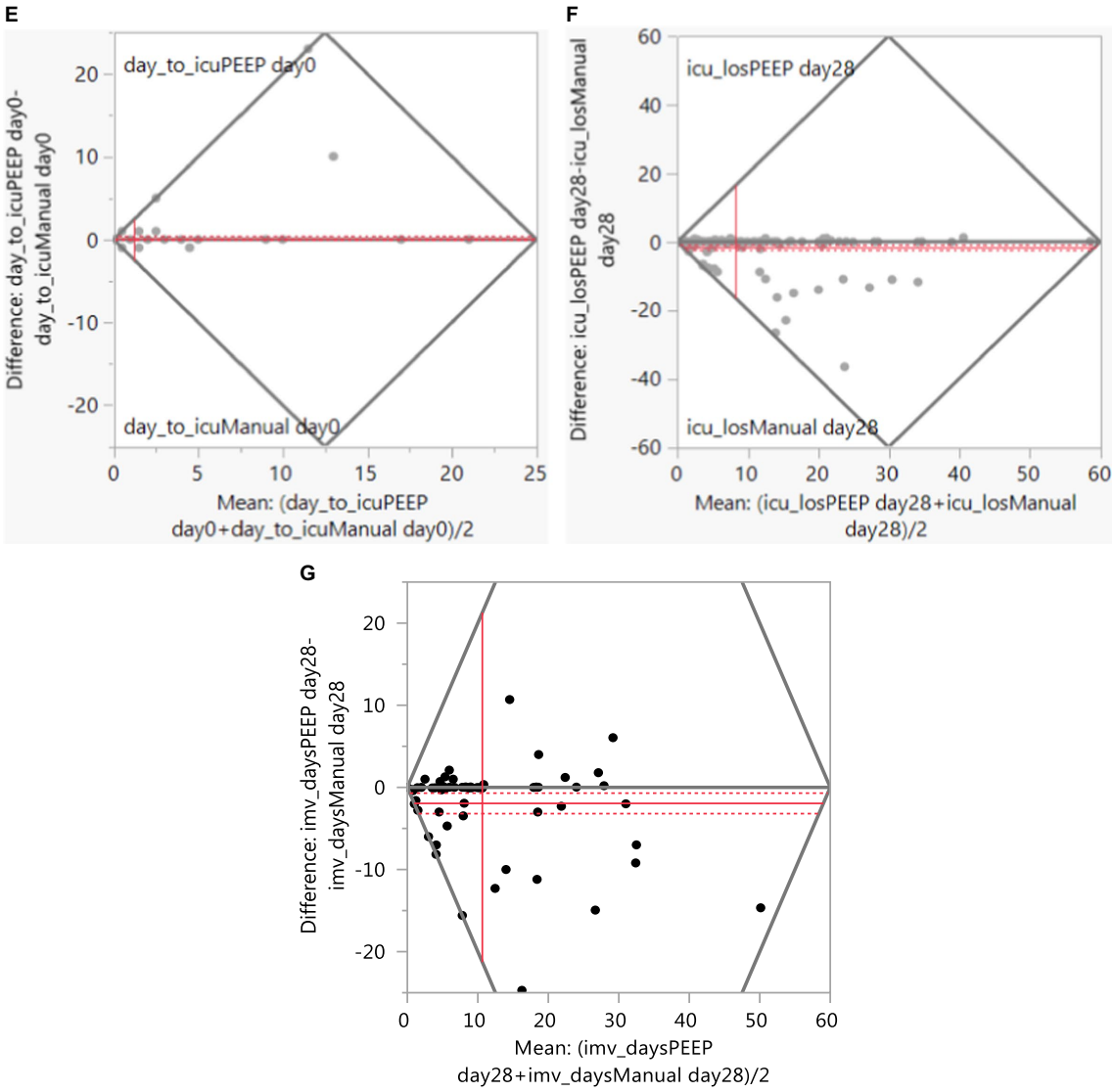

FIGURE 2
Agreement between manual and PEEP (Bland–Altman plot). **(A)** Age. **(B)** Weight. **(C)** Height. **(D)** Hospital Length of Stay. **(E)** Days to ICU admission. **(F)** ICU Length of Stay. **(G)** IMV Days.

TABLE 2 Comparison of patients in automated versus manual reviews and measures of agreement for individual responses for categorical (ordinal) variables.

Variable name	Automated vs. manual, percent agreement	Kappa coefficient (95% CI, SE)	Strength of agreement based on Kappa coefficient
Race, *N* = 652		0.91 (0.88–0.93, 0.01)	Almost perfect
White Caucasian	365/372 (98.1)		
Black or African American	138/139 (99.3)		
Others	111/141 (78.7)		
Total	614/652 (94.2)		
Ethnicity, *N* = 652		0.88 (0.84–0.93, 0.02)	Almost perfect
Non-Hispanic	506/512 (98.8)		
Hispanic	97/105 (92.4)		
Unknown/Not applicable	23/35 (65.7)		
Total	626/652 (96)		

CI, Confidence interval; SE, Standard error.

10.3389/fmed.2023.1089087

TABLE 3 Comparison of patients in automated versus manual reviews and measures of agreement for individual responses for categorical (nominal) variables.

Variable name	Percent agreement, automated vs. manual	Sensitivity	Specificity	Kappa coefficient (95% CI, SE)	Strength of agreement based on Kappa coefficient
Sex, *N* = 652		99.7	99.7	0.99 (0.99–1.0, 0)	Almost perfect
Male	359/360 (99.7)				
Female	291/292 (99.7)				
Total	650/652 (99.7)				
Coronary artery disease, *N* = 540		98.6	97.4	0.90 (0.85–0.96, 0.03)	Almost perfect
Yes	73/74 (98.6)				
No	454/466 (97.4)				
Total	527/540 (97.6)				
Hypertension, *N* = 540		92.0	93.5	0.85 (0.80–0.89, 0.02)	Almost perfect
Yes	298/324 (92.0)				
No	202/216 (93.5)				
Total	500/540 (92.6)				
Congestive heart failure, *N* = 540		88.0	97.8	0.82 (0.74–0.90, 0.04)	Almost perfect
Yes	44/50 (88)				
No	479/490 (97.8)				
Total	523/540 (96.7)				
Chronic obstructive pulmonary disease, *N* = 540		92.7	96.3	0.80 (0.72–0.88, 0.04)	Substantial
Yes	51/55 (92.7)				
No	467/485 (96.3)				
Total	518/540 (95.9)				
Asthma, *N* = 540		93.7	95.8	0.81 (0.73–0.88, 0.04)	Almost perfect
Yes	59/63 (93.7)				
No	457/477 (95.8)				
Total	516/540 (95.6)				
Chronic kidney disease, *N* = 540		81.2	96.2	0.79 (0.72–0.85, 0.03)	Substantial
Yes	95/117 (81.2)				
No	407/423 (96.2)				
Total	502/540 (93)				
Diabetes mellitus, *N* = 540		92.1	96.3	0.89 (0.85–0.93, 0.02)	Almost perfect
Yes	176/191 (92.1)				
No	336/349 (96.3)				
Total	512/540 (94.8)				
Dyslipidemia/Hyperlipidemia, *N* = 540		88.9	86.4	0.61 (0.53–0.69, 0.04)	Substantial
Yes	80/90 (88.9)				
No	389/450 (86.4)				
Total	469/540 (86.9)				
ICU admission rate, *N* = 611		90.3	95.2	0.86 (0.82–0.90, 0.02)	Almost perfect
Yes	215/238 (90.3)				
No	355/373 (95.2)				
Total	570/611 (93.3)				

(Continued)

TABLE 3 (Continued)

Variable name	Percent agreement, automated vs. manual	Sensitivity	Specificity	Kappa coefficient (95% CI, SE)	Strength of agreement based on Kappa coefficient
IMV rate, *N* = 582		87.7	98	0.85 (0.79–0.92, 0.03)	Almost perfect
Yes	64/73 (87.7)				
No	499/509 (98)				
Total	563/582 (96.7)				
NIMV rate, *N* = 581		83.3	99.3	0.80 (0.66–0.95, 0.07)	Substantial
Yes	15/18 (83.3)				
No	559/563 (99.3)				
Total	574/581 (98.3)				
HFNC rate, *N* = 581		100	98.9	0.86 (0.75–0.97, 0.06)	Almost perfect
Yes	19/19 (100)				
No	556/562 (98.9)				
Total	575/581 (99)				
ECMO rate, *N* = 581		72.7	99.3	0.69 (0.47–0.91, 0.11)	Substantial
Yes	8/11 (72.7)				
No	566/570 (99.3)				
Total	574/581 (98.8)				
ICU discharge status, *N* = 172		100	100	1.0 (1–1, 0.0)	Almost perfect
Death	9/9 (100)				
Alive	163/163 (100)				
Total	172/172 (100)				
Hospital discharge status, *N* = 541		90	100	0.94 (0.88–1, 0.03)	Almost perfect
Death	27/30 (90)				
Alive	511/511 (100)				
Total	538/541 (99.4)				

CI, Confidence interval; ECMO, Extracorporeal membrane oxygenation; HFNC, High Flow Nasal Canula; ICU, Intensive Care Unit; IMV, Invasive Mechanical Ventilation; LOS, Length of stay; NIMV; Non-Invasive Mechanical Ventilation; PCC, Pearson Interclass Correlation Coefficient; SE, Standard Error.

COVID-19 pneumonia using ICD-10 base-data comparing to manual data collection. We also successfully compared seven continuous variables with three extremely high agreement and four high agreement in comparison to Brazeal et al. (35), who compared two variables (age and BMI) for manual versus automation in a study population comprised of patients with histologically confirmed advanced adenomatous colorectal polyp.

Manual data extractors can overcome diverse interface issues, read and analyze free text, and provide clinical judgment when retrieving and interpreting data; however, manual data extraction is limited to human resources and is prone to human error (7, 32, 36). In addition to requiring considerable amount of time, manual data extraction also necessitates qualified personnel (30, 33). During the COVID-19 pandemic, where real-time data is paramount, automated data has proven validity and efficacy, and may divert personnel to patient care and other vital tasks. Nonetheless, automated data is not flawless. A significant limitation is finding a unique algorithm that can be applied to every center. Variables collected as free text fields are another challenge for such validations. The automated VIRUS COVID-19 sites had reported over a large majority of variables collected using this method. Currently, more than 60,000 patients and their data variables in the registry had been collected through efforts of the VIRUS-PEEP group, which has allowed for updates and complete data in the shortest possible time.

Challenges in automation

The environment for data collection is often a shared environment within an institution, and there are limitations on how much data may be extracted and processed in one job and how much post-abstraction processing is necessary. Microsoft SQL and TSQL solutions process substantial amounts of data from many different tables and can take a long time to run on large populations. There are clinical documentation differences between the various sites requiring additional coding when applying the data requirements and rules. Establishing logic for data elements within a given EHR can be time consuming up front, requiring close collaboration between clinician and analytics teams. Data may be stored differently between multiple medical centers in one institution, requiring processing to comply with data requirements for standardization. While sites can share coding experience in data abstraction between similar data storage structure, variable coding schemes pose challenges for direct

translation between sites. Lastly, one information technology employee often works on such projects with competing priorities.

Strengths and limitations

To our knowledge this is first multicenter study to evaluate the feasibility of automation process during COVID-19 pandemic. This automation process should be applicable to any EHR vendor (EHR type agnostic), and these purposeful sampled representative data points would be relevant to any other clinical study/trial, which is a major strength of this study. Nonparticipation of 19 sites out of 30 sites in the VIRUS-PEEP group, which leads to a possibility of selection bias, is a major limitation. The time constraints in the ongoing pandemic at participating sites were the reason behind this non-participation in the validation process. However, extracting data across 11 different centers is one of the strengths of this study; it could also highlight the variations in staff, procedures, and patients at these institutions. Although the SQL queries could be applicable in most sites, some sites required a new SQL tailored to their data architecture. One key limitation for our group was that all sites found a portion of data extraction that could not be automated, including variables which are described in narrative, such as, patient symptoms, estimated duration of onset of symptoms, and imaging interpretations. Another limitation is a notable discrepancy between manual and EMR extraction for important outcomes like ICU LOS and IMV days. The automation process relies on procedure order date (intubation/extubation) and ADT (hospital/ICU admission discharge transfer) order date and time and discontinuation date in EHR; however the manual extractor look for first-time documented ICU or IMV in her, which probably could account for such notable discrepancy in outcomes like ICU LOS and IMV days. Transferring a patient to a location that was not a usual ICU due to COVID-19 surge may be another possible explanation for the observed lower sensitivity of ICU admission rate. Variation in creation of make-shift ICUs at different institution may have caused this discrepancy in automation of ICU admissions documentation. It partially explains the lower sensitivity and high specificity of ICU admission, IMV, NIMV, and ECMO rates by automation process. Another noticeable issue was that the manual data extraction was done in real time and automation was done when the patient was discharged and mainly relied on billing codes and manually verified data available in EHR.

Future direction

Future research on this topic could involve a thorough comparison of all patient records extracted using two methods: manual extraction and automated SQL queries. The data comparison could be done by aligning data points across a wide range of variables for each data extraction method and then statistically analyzing their consistency and discrepancies. This detailed comparison would verify the reliability of automated data extraction and provide insights into areas that could be improved for greater accuracy.

Conclusion

This study confirms the feasibility, reliability, and validity of an automated process to gather data from the EHR. The use of automated data is comparable to the gold standard. The utilization of automated data extraction provides additional solutions when a rapid and large volume of patient data needs to be extracted.

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.

Ethics statement

The studies involving humans were approved by Mayo Clinic Institutional Review Board. The studies were conducted in accordance with the local legislation and institutional requirements. The ethics committee/institutional review board waived the requirement of written informed consent for participation from the participants or the participants' legal guardians/next of kin because The Mayo Clinic Institutional Review Board authorized the SCCM Discovery VIRUS COVID-19 registry as exempt on March 23, 2020 (IRB number: 20–002610). No informed consent was deemed necessary for the study subjects. The procedures were followed in accordance with the Helsinki Declaration of 2013.

Author contributions

DV and VB contributed equally in the defining the study outline and manuscript writing. VB, SH, JC, MS, AT, MB, SZ, NS, RC-C, RN, DS, AN, SK, KAB, J-TC, RM, IS, RR, and KB did the data review and collection. DV, VB, and SH did the statistical analysis. VH, JD, AW, VK, and RK did the study design and critical review. DV, VB, SH, and RK were guarantor of the manuscript and took 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

The VIRUS: COVID-19 Registry was supported, in part, by the Gordon and Betty Moore Foundation, and Janssen Research & Development, LLC. They have no role in data gathering, analysis, interpretation, and writing.

Acknowledgments

Data from this study was submitted and presented as an abstract format for the Chest 2023 Conferences at Hawai'i Convention Center, Honolulu, Hawai'i.

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. Wang C, Horby PW, Hayden FG, Gao GF. A novel coronavirus outbreak of global health concern. *Lancet*. (2020) 395:470–3. doi: 10.1016/S0140-6736(20)30185-9

2. Domecq JP, Lal A, Sheldrick CR, Kumar VK, Boman K, Bolesta S, et al. Outcomes of patients with coronavirus disease 2019 receiving organ support therapies: the international viral infection and respiratory illness universal study registry. *Crit Care Med*. (2021) 49:437–48. doi: 10.1097/CCM.0000000000004879

3. Walkey AJ, Kumar VK, Harhay MO, Bolesta S, Bansal V, Gajic O, et al. The viral infection and respiratory illness universal study (VIRUS): an international registry of coronavirus 2019-related critical illness. *Crit Care Explor*. (2020) 2:e0113. doi: 10.1097/CCE.0000000000000113

4. Walkey AJ, Sheldrick RC, Kashyap R, Kumar VK, Boman K, Bolesta S, et al. Guiding principles for the conduct of observational critical care research for coronavirus disease 2019 pandemics and beyond: the Society of Critical Care Medicine discovery viral infection and respiratory illness universal study registry. *Crit Care Med*. (2020) 48:e1038–44. doi: 10.1097/CCM.0000000000004572

5. Grimm AG Hospitals Reported That the COVID-19 Pandemic Has Significantly Strained Health Care Delivery Results of a National Pulse Survey. USA: U.S. Department of Health and Human Services Office of Inspector General. (2021). Available at: https://oig.hhs.gov/oei/reports/OEI-09-21-00140.pdf

6. Vassar M, Holzmann M. The retrospective chart review: important methodological considerations. *J Educ Eval Health Prof*. (2013) 10:12. doi: 10.3352/jeehp.2013.10.12

7. Yin AL, Guo WL, Sholle ET, Rajan M, Alshak MN, Choi JJ, et al. Comparing automated vs. manual data collection for COVID-specific medications from electronic health records. *Int J Med Inform*. (2022) 157:104622. doi: 10.1016/j.ijmedinf.2021.104622

8. Byrne MD, Jordan TR, Welle T. Comparison of manual versus automated data collection method for an evidence-based nursing practice study. *Appl Clin Inform*. (2013) 4:61–74. doi: 10.4338/ACI-2012-09-RA-0037

9. Lan H, Thongprayoon C, Ahmed A, Herasevich V, Sampathkumar P, Gajic O, et al. Automating quality metrics in the era of electronic medical records: digital signatures for ventilator bundle compliance. *Biomed Res Int*. (2015) 2015:396508:1–6. doi: 10.1155/2015/396508

10. Brundin-Mather R, Soo A, Zuege DJ, Niven DJ, Fiest K, Doig CJ, et al. Secondary EMR data for quality improvement and research: a comparison of manual and electronic data collection from an integrated critical care electronic medical record system. *J Crit Care*. (2018) 47:295–301. doi: 10.1016/j.jcrc.2018.07.021

11. Hersh WR, Cimino J, Payne PR, Embi P, Logan J, Weiner M, et al. Recommendations for the use of operational electronic health record data in comparative effectiveness research. *EGEMS*. (2013) 1:1018. doi: 10.13063/2327-9214.1018

12. Hersh WR, Weiner MG, Embi PJ, Logan JR, Payne PR, Bernstam EV, et al. Caveats for the use of operational electronic health record data in comparative effectiveness research. *Med Care*. (2013) 51:S30–7. doi: 10.1097/MLR.0b013e31829b1dbd

13. Kahn MG, Callahan TJ, Barnard J, Bauck AE, Brown J, Davidson BN, et al. A harmonized data quality assessment terminology and framework for the secondary use of electronic health record data. *EGEMS*. (2016) 4:1244. doi: 10.13063/2327-9214.1244

14. Wei WQ, Leibson CL, Ransom JE, Kho AN, Caraballo PJ, Chai HS, et al. Impact of data fragmentation across healthcare centers on the accuracy of a high-throughput clinical phenotyping algorithm for specifying subjects with type 2 diabetes mellitus. *J Am Med Inform Assoc*. (2012) 19:219–24. doi: 10.1136/amiajnl-2011-000597

15. Botsis T, Hartvigsen G, Chen F, Weng C. Secondary use of EHR: data quality issues and informatics opportunities. *Summit Transl Bioinform*. (2010) 2010:1–5.

16. Hripcsak G, Albers DJ. Next-generation phenotyping of electronic health records. *J Am Med Inform Assoc*. (2013) 20:117–21. doi: 10.1136/amiajnl-2012-001145

17. Prokosch HU, Ganslandt T. Perspectives for medical informatics. Reusing the electronic medical record for clinical research. *Methods Inf Med*. (2009) 48:38–44. doi: 10.3414/ME9132

18. Weiskopf NG, Hripcsak G, Swaminathan S, Weng C. Defining and measuring completeness of electronic health records for secondary use. *J Biomed Inform*. (2013) 46:830–6. doi: 10.1016/j.jbi.2013.06.010

19. Weiskopf NG, Cohen AM, Hannan J, Jarmon T, Dorr DA. Towards augmenting structured EHR data: a comparison of manual chart review and patient self-report. *AMIA Annu Symp Proc*. (2019) 2019:903–12.

20. Kern LM, Malhotra S, Barrón Y, Quaresimo J, Dhopeshwarkar R, Pichardo M, et al. Accuracy of electronically reported "meaningful use" clinical quality measures: a cross-sectional study. *Ann Intern Med*. (2013) 158:77–83. doi: 10.7326/0003-4819-158-2-201301150-00001

21. The Society of Critical Care Medicine, Lyntek Medical Technologies Inc. *VIRUS COVID-19 registry dashboard: a COVID-19 registry of current ICU and hospital care patterns USA2020*. (2021). Available at: https://sccmcovid19.org/.

22. General Assembly of the World Medical Association. World medical association declaration of Helsinki: ethical principles for medical research involving human subjects. *J Am Coll Dent*. (2014) 81:14–8.

23. World Health Organization-International Severe Acute Respiratory and Emerging Infection Consortium (WHO-ISARIC). *Clinical data collection – the COVID-19 case report forms (CRFs)* (2020). Available at: https://media.tghn.org/medialibrary/2020/03/ISARIC_COVID-19_CRF_V1.3_24Feb2020.pdf

24. Harris PA, Taylor R, Thielke R, Payne J, Gonzalez N, Conde JG. Research electronic data capture (REDCap)--a metadata-driven methodology and workflow process for providing translational research informatics support. *J Biomed Inform*. (2009) 42:377–81. doi: 10.1016/j.jbi.2008.08.010

25. Wang J. Pearson correlation coefficient In: W Dubitzky, O Wolkenhauer, K-H Cho and H Yokota, editors. *Encyclopedia of systems biology*. New York, NY: Springer (2013). 1671.

26. Mukaka MM. Statistics corner: a guide to appropriate use of correlation coefficient in medical research. *Malawi Med J*. (2012) 24:69–71.

27. Altman DG, Bland JM. Measurement in Medicine - the analysis of method comparison studies. *J Roy Stat Soc D-Sta*. (1983) 32:307–17. doi: 10.2307/2987937

28. Sun S. Meta-analysis of Cohen's kappa. *Health Serv Outc Res Methodol*. (2011) 11:145–63. doi: 10.1007/s10742-011-0077-3

29. Landis JR, Koch GG. The measurement of observer agreement for categorical data. *Biometrics*. (1977) 33:159–74. doi: 10.2307/2529310

30. Alsara A, Warner DO, Li G, Herasevich V, Gajic O, Kor DJ. Derivation and validation of automated electronic search strategies to identify pertinent risk factors for postoperative acute lung injury. *Mayo Clin Proc*. (2011) 86:382–8. doi: 10.4065/mcp.2010.0802

31. Singh B, Singh A, Ahmed A, Wilson GA, Pickering BW, Herasevich V, et al. Derivation and validation of automated electronic search strategies to extract Charlson comorbidities from electronic medical records. *Mayo Clin Proc*. (2012) 87:817–24. doi: 10.1016/j.mayocp.2012.04.015

32. Rishi MA, Kashyap R, Wilson G, Hocker S. Retrospective derivation and validation of a search algorithm to identify extubation failure in the intensive care unit. *BMC Anesthesiol*. (2014) 14:41. doi: 10.1186/1471-2253-14-41

33. Smischney NJ, Velagapudi VM, Onigkeit JA, Pickering BW, Herasevich V, Kashyap R. Retrospective derivation and validation of a search algorithm to identify emergent endotracheal intubations in the intensive care unit. *Appl Clin Inform*. (2013) 4:419–27. doi: 10.4338/ACI-2013-05-RA-0033

34. Schaefer JW, Riley JM, Li M, Cheney-Peters DR, Venkataraman CM, Li CJ, et al. Comparing reliability of ICD-10-based COVID-19 comorbidity data to manual chart review, a retrospective cross-sectional study. *J Med Virol*. (2022) 94:1550–7. doi: 10.1002/jmv.27492

35. Brazeal JG, Alekseyenko AV, Li H, Fugal M, Kirchoff K, Marsh C, et al. Assessing quality and agreement of structured data in automatic versus manual abstraction of the electronic health record for a clinical epidemiology study. *Res Methods Med Health Sci*. (2021) 2:168–78. doi: 10.1177/26320843211061287

36. Wu L, Ashton CM. Chart review. A need for reappraisal. *Eval Health Prof*. (1997) 20:146–63. doi: 10.1177/016327879702000203

frontiers | Frontiers in Medicine

TYPE Policy and Practice Reviews
PUBLISHED 08 January 2024
DOI 10.3389/fmed.2023.1174429

OPEN ACCESS

EDITED BY
Md. Mohaimenul Islam,
The Ohio State University, United States

REVIEWED BY
Yu-wen Chen,
Chinese Academy of Sciences (CAS), China
Tim Hulsen,
Philips Research, Netherlands

*CORRESPONDENCE
Yun Long
ly_icu@aliyun.com

†These authors have contributed equally to this work

RECEIVED 26 February 2023
ACCEPTED 09 November 2023
PUBLISHED 08 January 2024

CITATION
Su L, Liu S, Long Y, Chen C, Chen K, Chen M, Chen Y, Cheng Y, Cui Y, Ding Q, Ding R, Duan M, Gao T, Gu X, He H, He J, Hu B, Hu C, Huang R, Huang X, Jiang H, Jiang J, Lan Y, Li J, Li L, Li L, Li W, Li Y, Lin J, Luo X, Lyu F, Mao Z, Miao H, Shang X, Shang X, Shang Y, Shen Y, Shi Y, Sun Q, Sun W, Tang Z, Wang B, Wang H, Wang H, Wang L, Wang L, Wang S, Wang Z, Wang Z, Wei D, Wu J, Wu Q, Xing X, Yang J, Yang X, Yu J, Yu W, Yu Y, Yuan H, Zhai Q, Zhang H, Zhang L, Zhang M, Zhang Z, Zhao C, Zheng R, Zhong L, Zhou F and Zhu W (2024) Chinese experts' consensus on the application of intensive care big data.
Front. Med. 10:1174429.
doi: 10.3389/fmed.2023.1174429

Chinese experts' consensus on the application of intensive care big data

Longxiang Su[1†], Shengjun Liu[1†], Yun Long[1*], Chaodong Chen[2], Kai Chen[3], Ming Chen[4], Yaolong Chen[5], Yisong Cheng[6], Yating Cui[7], Qi Ding[2], Renyu Ding[8], Meili Duan[9], Tao Gao[4], Xiaohua Gu[10], Hongli He[11], Jiawei He[9], Bo Hu[12], Chang Hu[12], Rui Huang[13], Xiaobo Huang[11], Huizhen Jiang[14], Jing Jiang[15], Yunping Lan[11], Jun Li[3], Linfeng Li[16], Lu Li[12], Wenxiong Li[2], Yongzai Li[17], Jin Lin[9], Xufei Luo[5], Feng Lyu[18], Zhi Mao[7], He Miao[8], Xiaopu Shang[19], Xiuling Shang[3], You Shang[20], Yuwen Shen[21], Yinghuan Shi[22], Qihang Sun[23], Weijun Sun[24], Zhiyun Tang[25], Bo Wang[6], Haijun Wang[26], Hongliang Wang[13], Li Wang[27], Luhao Wang[28], Sicong Wang[13], Zhanwen Wang[29,30,31], Zhong Wang[8], Dong Wei[22], Jianfeng Wu[29], Qin Wu[6], Xuezhong Xing[27], Jin Yang[15], Xianghong Yang[26], Jiangquan Yu[10], Wenkui Yu[4], Yuan Yu[21], Hao Yuan[28], Qian Zhai[22], Hao Zhang[26], Lina Zhang[29,30,31], Meng Zhang[15], Zhongheng Zhang[32], Chunguang Zhao[29,30,31], Ruiqiang Zheng[10], Lei Zhong[26], Feihu Zhou[7], Weiguo Zhu[33]

[1]Department of Critical Care Medicine, State Key Laboratory of Complex Severe and Rare Diseases, Peking Union Medical College Hospital, Peking Union Medical College, Chinese Academy of Medical Sciences, Beijing, China, [2]Department of Surgical Intensive Critical Unit, Beijing Chao-yang Hospital, Capital Medical University, Beijing, China, [3]Department of Critical Care Medicine, Fujian Provincial Key Laboratory of Critical Care Medicine, Shengli Clinical Medical College of Fujian Medical University, Fujian Provincial Hospital, Fujian Provincial Center for Critical Care Medicine, Fuzhou, Fujian, China, [4]Department of Critical Care Medicine, Nanjing Drum Tower Hospital, The Affiliated Hospital of Nanjing University Medical School, Nanjing, Jiangsu, China, [5]Evidence-based Medicine Center, School of Basic Medical Sciences, Lanzhou University, Lanzhou, China, [6]Department of Critical Care Medicine, West China Hospital of Sichuan University, Chengdu, China, [7]Department of Critical Care Medicine, The First Medical Center, Chinese PLA General Hospital, Beijing, China, [8]Department of Intensive Care Unit, The First Hospital of China Medical University, Shenyang, Liaoning, China, [9]Department of Critical Care Medicine, Beijing Friendship Hospital, Capital Medical University, Beijing, China, [10]Department of Critical Care Medicine, Northern Jiangsu People's Hospital; Clinical Medical College, Yangzhou University, Yangzhou, China, [11]Intensive Care Unit, Sichuan Academy of Medical Sciences & Sichuan Provincial People's Hospital, School of Medicine of University of Electronic Science and Technology, Chengdu, China, [12]Department of Critical Care Medicine, Zhongnan Hospital of Wuhan University, Wuhan, Hubei, China, [13]Department of Critical Care Medicine, The Second Affiliated Hospital of Harbin Medical University, Harbin, Heilongjiang, China, [14]Department of Information Center, Peking Union Medical College Hospital, Peking Union Medical College, Chinese Academy of Medical Sciences, Beijing, China, [15]Department of Critical Care Medicine, Chongqing General Hospital, Chongqing, China, [16]Medical Data Research Institute, Chongqing Medical University, Chongqing, China, [17]Information Network Center, QiLu Hospital, ShanDong University, Jinan, China, [18]Department of Computer Science and Engineering, Central South University, Changsha, China, [19]Department of Information Management, Beijing Jiaotong University, Beijing, China, [20]Department of Critical Care Medicine, Union Hospital, Tongji Medical College, Huazhong University of Science and Technology, Wuhan, China, [21]Intensive Care Unit of Cardiovascular Surgery Department, Qilu Hospital of Shandong University, Jinan, China, [22]National Institute of Healthcare Data Science, Nanjing University, Nanjing, China, [23]British Chinese Society of Health Informatics, Beijing, China, [24]Faculty of Automation, Guangdong University of Technology, Guangzhou, China, [25]Department of Intensive Care Unit, Zhejiang Provincial People's Hospital, Affiliated

People's Hospital, Emergency and Intensive Care Unit Center, Hangzhou Medical College, Hangzhou, Zhejiang, China, [26]Department of Intensive Care Unit, National Cancer Center/National Clinical Research Center, Cancer Hospital, Chinese Academy of Medical Sciences and Peking Union Medical College, Beijing, China, [27]Department of Epidemiology and Biostatistics, Institute of Basic Medical Sciences Chinese Academy of Medical Sciences; School of Basic Medicine Peking Union Medical College, Beijing, China, [28]Department of Critical Care Medicine, Sun Yat-Sen University First Affiliated Hospital, Guangzhou, China, [29]Intensive Care Unit, XiangYa Hospital, Central South University, Changsha, China, [30]National Clinical Research Center for Geriatric Disorders, Xiang Ya Hospital, Central South University, Changsha, China, [31]Hunan Provincial Clinical Research Center for Critical Care Medicine, Xiang Ya Hospital, Central South University, Changsha, China, [32]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, [33]Department of General Medicine, Peking Union Medical College Hospital, Peking Union Medical College, Chinese Academy of Medical Sciences, Beijing, China

The development of intensive care medicine is inseparable from the diversified monitoring data. Intensive care medicine has been closely integrated with data since its birth. Critical care research requires an integrative approach that embraces the complexity of critical illness and the computational technology and algorithms that can make it possible. Considering the need of standardization of application of big data in intensive care, Intensive Care Medicine Branch of China Health Information and Health Care Big Data Society, Standard Committee has convened expert group, secretary group and the external audit expert group to formulate Chinese Experts' Consensus on the Application of Intensive Care Big Data (2022). This consensus makes 29 recommendations on the following five parts: Concept of intensive care big data, Important scientific issues, Standards and principles of database, Methodology in solving big data problems, Clinical application and safety consideration of intensive care big data. The consensus group believes this consensus is the starting step of application big data in the field of intensive care. More explorations and big data based retrospective research should be carried out in order to enhance safety and reliability of big data based models of critical care field.

KEYWORDS

machine learning, intensive care medicine, big data, critical care medicine, consensus

Introduction

The development of intensive care medicine is inseparable from the diversified monitoring data, which specifically presents the clinical manifestations of patients with critical symptoms. These data illustrate a certain clinical phenomena, and represents the nature of disease behind the phenomenon. Intensive care medicine has been closely integrated with data since its birth. The complexity of critical illness makes the traditional reductionist approach to medical research insufficient (1). Critical care research requires an integrative approach that embraces the complexity of critical illness and the computational technology and algorithms that can make it possible (2). Hence, the organic combination of artificial intelligence and critically ill patient data can provide significant assistance for clinical diagnosis and treatment (3). Pirracchio et al. summarize the current application of machine learning for predictive analytics and decision support in the ICU and propose online learning in the future (4). Sanchez-Pinto et al. review the definitions, types of algorithms, applications, challenges, and future of Big Data and data science in critical care (5). There are no concenus of application of big data on the field of intensive care in China so far. Specifically, the conception, clinical research site, standard of dataset, methodology and limitation are not fully exhibited. In this experts consensus, we would like to summarize the problem above and give recommendations based on evidence.

Consensus formation

This consensus is initiated and formulated by Intensive Care Medicine Branch of China Health Information and Health Care Big Data Society, Standard Committee, and is methodically supported by the Health Data Sciences and Research Institute of Lanzhou University/Research Innovation Unit of Evidence-based Evaluation and Guidelines of Chinese Academy of Medical Sciences/Guidelines for Implementation and Knowledge Transformation Cooperation Center of the World Health Organization. This consensus has been registered on the International Practice Guide Registration Platform (Practice

guideline registration for transparency, PREPARE[1]) with the registration number being PREPARE-2022CN566.

The consensus development group consists of the consensus expert group, secretary group, working group and external audit expert group. The work flow of formation of consensus are shown in Figure 1. The enrollment criteria and obligation of these groups are shown in Supplementary material S1.

For each recommendation, external audit expert used the Likert scale (Score range: 1–6) to evaluate recommendation degree. Specifically, 6 points for total agree, 5 points for general agreement, 4 points for uncertainty, 3 points for a little disagree, 2 points for disagree and 1 point for total disagree. For each recommendation, if more than 70% external audit expert grade no less than 6 points, the consensus is reached. In this formula, 31 recommendations were put forward. Except for the two recommendations on missing value and outlier value, the remaining 29 recommendations were finalized. The degree of expert recommendation was marked with "Consensus Degree," which equal to the total number of experts no less than 6 points/total number of experts×100%.

Consensus text

The concept, significance and necessity of intensive care data

Recommendation 1 (5–8)**: The concept of intensive care big data (97% consensus)**

Intensive care big data refers to the datasets with logical connotations formulated by various indicators which are large-scale, multi-heterogeneous, variably dynamic, high-speed and real-time acquisition, low-value density and difficult to analyze traditionally in the whole process of diagnosis and treatment of patients or potential ones with critical symptoms.

Recommendation 2: The intensive care big data is multi-modal, massive, dynamic, continuous, and objective, and its correct acquisition can provide auxiliary evidential support for diagnosis of critical illnesses and early warning. (98% consensus)

Background and Evidence:

The monitoring methods used in the intensive care unit are abundant, and the data obtained by the combined use of multiple monitoring equipment have a multimodal characteristic (9–11). According to the needs, multi-parameter sampling can be performed at different levels and time to obtain a large amount of continuous data. Therefore, the intensive care data has the characteristics of abundance (9), dynamics, continuation, and accuracy (5, 12). Correct and effective data processing has a guiding and early-warning role in the diagnosis and treatment of critical illnesses (8). Recently, Epimed Monitor System®, a cloud-based ICU management system that includes data of more than 2.5 million hospitalization in the ICU of Brazil, has been deployed to predict the duration of ICU stays, provide guidance for risk assessment of patients becoming long-term ones in the ICU, and help to plan the use of hospital beds (13). Komorowski et al. (14) used reinforced learning techniques to guide patients with sepsis to use fluid or vasoactive medication, and external validation showed that the model made better choices for treatment than intensive care physicians. In the aspect of building predictive models by using data mining techniques, Nemati et al. (15) demonstrated that "AI sepsis experts" can be used for real-time data processing to predict new sepsis within 4–12 h. Although big data research has shown broad prospects for application, at this stage, the number of random clinical trials is small, and various technical models need to be testing prospectively in the clinic to prove their effectiveness and safety (15). In view of the characteristics of individualized and differential conditions in patients with critical symptoms (1), at this stage, the intensive care big data cannot provide maturely clinical guidance and can be used as an auxiliary support tool.

Recommendation 3: The establishment of a large database for intensive care in China should follow the principles of multiple center, multiple disease and automatic capture, and provide reliable and accurate data support for the application of big data and the development of artificial intelligence. (92% consensus)

1 http: //www.guidelines-registry.org

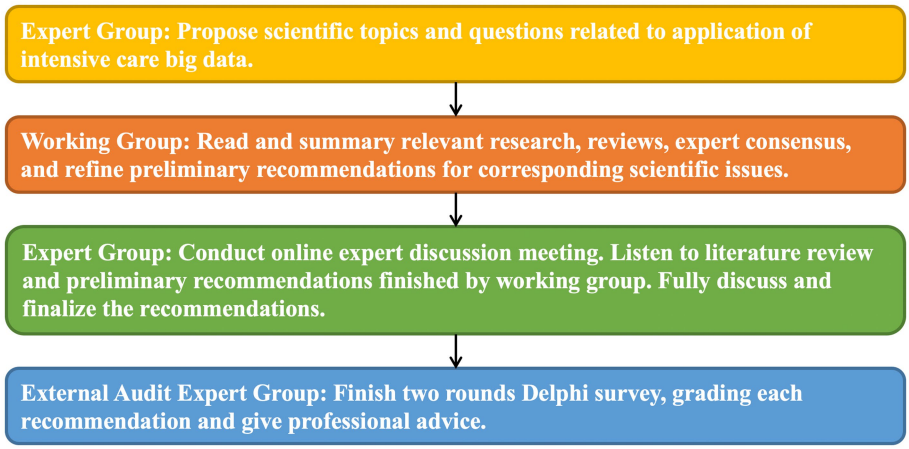

FIGURE 1
Work flow of formation of consensus.

Background and Evidence:

The establishment of a large database for intensive care in China is in the preliminary exploration. By drawing on the experience of existing databases at home and abroad and summarizing the deficiencies of the existing databases, the database can provide a basis for clinical decision-making in China, precise medicine implementation and formulation of medical policies in China. At present, a number of large databases for intensive care have been established abroad, such as the latest version of the Intensive Care Information Database (Medical Information Mart for Intensive Care-IV, MIMIC-IV) (16), the eICU Collaborative Research Database (eICU-CRD) (17), high time resolution ICU dataset (HiRID) (18) and Amsterdam University Medical Centers Database (Amsterdam University Medical Centers Database, AmsterdamUMCdb) (19), etc., mainly based on European and American races. The volume of data is large and the types of data are abundant, but the vital signs are regularly monitored, which are not fully automatically captured, and the scoring system for critical illnesses does not have functions of automatic data collection and integration (20), there is a general lack of software embedding of preliminary data analysis online. The large database for intensive care abroad needs improving in terms of real-time data and availability. Based on the MIMIC database, the researchers conducted in-depth mining of big data to build clinical models. With the help of artificial intelligence and machine learning, artificial intelligence physicians can be used to assist clinical decision-making and provide personalized, clinical and optimal treatment for patients with critical symptoms and improve the prognosis (14).

In recent years, China has been exploring large databases for intensive care, and has successively established a database of ICU infected patients (21), a pediatric intensive care database (PIC) (22), and HeartFailure database (23), etc. The existing large databases for intensive care started late, and their development is not yet mature. They are all single-center databases, with a single type of disease or population. They are limited to the initial collection of early data, and they do not have functions of automatic data capture and data analysis, and the overall quality of data is relatively low and The utilization efficiency is not high. It has not been integrated into clinically artificial intelligence and application technology of big data (24).

Table 1 shows the brief information comparison of major foreign intensive care databases. It can be seen that the existing databases at home and abroad are mainly single-center, and various illnesses may develop into critical one and require admission to ICU for treatment, so it is significant to improve the comprehensiveness of the data. Therefore, the existing single-center or multi-center databases established for certain diseases obviously cannot meet the needs of the vast majority of ICU patients. As we all know, the most widely used database such as MIMIC-III database records vital signs every hour, but for patients with critical symptoms who need continuous dynamic monitoring, this temporal resolution ratio is far from satisfactory. HiRID has a higher temporal resolution ratio than other published datasets, and data storage processes every 2 min (18), which is not yet possible for other databases. To sum up, the establishment of a large database for intensive care in China should follow the principles of multiple centers, multiple diseases and automatic capture to provide data support for the development and application of artificial intelligence.

Recommendation 4: Build a large database of patients with critical symptoms in China for their condition monitoring, the research and development of clinical drug and clinical trials can provide the standardized and individualized treatment for patients with critical symptoms. (97% consensus)

Background and Evidence:

Understanding the relationship between intensive care big data and critical clinic is crucial. The relationship between intensive care big data and the clinic is that: data integration can provide clinicians with manageable, interpretable, operational and treatment plan data, give certain reference to clinical treatment. Data management can provide better personalized and accurate medical guarantee through predicted and prognostic model, It can also use supervised and unsupervised learning algorithm to provide clinical researchers with handy, highly-credible and highly-utilizable database, provide scientific data support for drug development and exploration process, and finally promote the development of intensive care medicine. At present, the application of intensive care big data in clinical practice is gradually increasing, but it is mainly limited to mechanical data collection and manual data processing. The expert group believes that machine learning modeling and multi-disciplinary combination can be used to warn, track and summarize different clinical problems, so as to summarize past experience, warn current decisions and predict future progress.

The first is the application of intensive care big data in clinical decision-making. An RCT study conducted in two community hospitals in 2010 pointed out that remote data algorithms could effectively improve the medical quality of patients with critically symptoms (25). Meanwhile, a review in 2015 showed strategies for the application of big data in the use of antibiotics in patients with critically symptoms. They proposed the concept of AutoKinetics to provide decision support for clinical dosing. And through direct interaction with electronic medical records, they broadened the way to use big data and provided the right dose for each patient at the right

TABLE 1 Brief information comparison of major foreign databases of critical illnesses.

	MIMIC-III	MIMIC-IV	eICU	HiRID	Amsterdam UMCdb
Sources of included population	Single Center, Large Sample, Beth Israel Deaconess Medical Center in MIT		Multi-center, mainly small and medium-sized hospitals, organized by non-intensive specialists, with patients in 335 ICUs in the United States	Single-center, ICU patients at the University Hospital of Bern, Switzerland	Multi-center, with 20,109 ICU patients in Europe
Country/Region	USA	USA	USA	Switzerland	Europe
Time	2001~2012	2008~2019	2014~2015	2008.1~2016.6	2013~2016
Number of patients	46,520	383,220	139,367	36,098	20,109

time (26). Kindle et al. (27) and Carra et al. (8) summarized the developmental results of all remote algorithms and concluded that machine learning algorithms have important implications for sepsis detection, sepsis management, mechanical ventilation, reduction of false alarms, and prognosis in ICU. In addition, intensive care big data is also of great significance for decision making of clinical care. In 2022, the Stanford medical team developed an unsupervised process mining algorithm to evaluate the quality of care. The final result of the patient cohort had an average compliance score of 0.36. The highest was 0.64, and the lowest was 0.20. The results demonstrated the reliability of big data algorithms for data mining of electronic medical records, and the scheme could also be used to evaluate the quality of care in other diseases (28). In 2022, Jens Michael Boss et al. (29) proposed "ICU Cockpit," an integration platform of algorithmic model, which pointed out the early warning effect of severe big data on clinical decision-making. Since 2016, the platform has processed over 89 billion data points (979 patients) from 200 signals and laboratory in the analysis, and an infrastructure-based framework has been proposed for deploying and validating intensive care algorithms. It allows algorithms to seamlessly integrate into real-time data streams to generate clinical decision support and predictions in clinical practice (29). The second is the guidance of intensive care big data for clinical research. Taglang and Jackson (30) and Xu et al. (21) expounded the importance of big data to explore clinical trials systematically and, respectively. In the exploration of big data in the past 2 years, a number of studies have carried out analysis of individualized computational models constructed through big data, pointing out risk factors for high mortality in patients with critical symptoms (31–33). Finally, in terms of the relationship between clinical drug R&D and big data, we have not seen any evidence that relevant big data is used in drug R&D in the field of critical care medicine. However, due to the considerable progress of application in drug R&D and big data during recent years (34), we recommend that big data can also be combined with drug development in intensive care medicine. Therefore, the expert group recommends intensive care big data be used to detect changes in clinical practice, but more databases and algorithms and large-scale RCT experiments are needed to jointly promote the development of this field, which is also the future path of clinical practice. We point out that the multidisciplinary and interactive development of intensive care big data can build a large database of critical diseases in China, and ultimately guide the standardized treatment of patients with critical symptoms.

Clinical scientific issues concerned by intensive care big data in clinical research

Recommendation 5: It is recommended to use machine learning method to build modeling to make early warning of sepsis, acute kidney injury (AKI), and acute respiratory distress syndrome (ARDS). (94% consensus)

Background and Evidence:

Research on early warning models for sepsis, AKI, and ARDS is increasing, and most models can provide early warning with good sensitivity and specificity. The ability of different models to predict and popularize needs to be further verified. The expert group believes that machine learning method modeling can be used in the early prediction of the risk of sepsis, AKI and ARDS in ICU patients, so as to reduce the possibility, improve early coping ability, and possibly improve prognosis.

The sepsis early warning model compared with manual screening and scoring, made early and accurate predictions, and achieved external validation. A meta-analysis of sepsis prediction models in 2020 showed that a single machine learning model can be an accurately early prediction of sepsis (AUROC 0.68–0.99) and could replace traditional scores, but heterogeneity between studies limited the evaluation of results (35). A study in 2022 (36) developed a sepsis screening tool by using a learning approach to gradient-boosted supervision that was more sensitive (84.6% vs. 80.4%) and more accurate (28.8% vs. 11.4%) than traditional scoring. A controlled study in 2021 (37) developed an algorithm that accurately predicted sepsis 12 h in advance (AUC 0.94, sensitivity 0.87, specificity 0.87). A multi-center study in 2021 (38) showed the use of a transfer-learning algorithm to enable the validity of the external validation datasets in sepsis.

Early warning models for AKI patients with critical symptoms can make early and accurate predictions, but few models have external validation, clinical interpretability, and high predictive performance in one (39). Studies have shown (40) that the early warning model of AKI has an AUC of 88%, which can predict AKI 6 h in advance. A multi-center study in 2020 showed that the AKI early warning model could predict AKI 48 h in advance, and performed well in both internal and external validation (AUC of 0.86 and 0.85, 0.86 respectively) (41). A 2020 study (42) established a continuous prediction model based on the data of electronic medical record, which could predict AKI in real time during hospitalization, and its performance was significantly better than the one-time prediction model (AUC of 0.724 vs. 0.653).

The ARDS early warning model can make early prediction of ARDS efficiently, and some models can achieve external validation, and some incorporate variables of iconography. A study in 2020 (43) using the XGBoost gradient boosting tree model could accurately predict ARDS 48 h in advance (AUROC of 79.0%). A study in 2020 (44) performed a secondary analysis of prospective study data using the text of radiology reports to build a model that performed well and achieved external validation (C-statistic, 0.78; 95% CI, 0.72–0.84). The diagnosis of ARDS is strongly dependent on iconography, which is, however, not necessarily available at the time of diagnosis or there is uncertainty in its interpretation. This information is called privileged information and uncertainty labels, but the model incorporating variables of iconography is closer to clinical practice. A study in 2021 (44) used a transfer-learning algorithm based on radiographs to build a predictive model that performed well and had external validation (AUROC of 92 and 88%). A study in 2021 (45) successfully used privileged information and a learning model with uncertainty labels to predict ARDS (AUC of 85.78 and 87.01%).

Recommendation 6: The prediction model based on machine learning can effectively predict the risk of patients at high risk of potential organ damage in the ICU. (89% consensus)

Background and Evidence:

The proposed early warning scoring system enables medical staff to better identify potential patients with critical symptoms and achieve the purpose of early identification and intervention to improve patient prognosis. However, this scoring system may fail to identify patients until significant deterioration occurs. A systematic review in 2019 (46)

found that the early warning score using statistical modeling was more accurate in identifying high-risk patients than weighted early warning (mean AUC of 0.80 vs. 0.73), with one true finding of positive case being 4.9 and 7.1 alarm events required. A similar 2021 systematic review (47) also showed that an early warning system for clinical deterioration based on machine learning could more accurately predict the risk of patient with lower survival rate in the ICU, with an area under the model ranging from 0.57 to 0.97.

Specifically, in addition to the progression of the primary disease, patients with critical symptoms may develop a variety of life-threatening comorbidities. The common ones include failure of circulatory function. In 2020, a study Hyland et al. (18) independently established an early warning system for circulatory failure, which could identify patients at risk of circulatory failure more than 2 h in advance, and successfully conducted external validation in an independent patient cohort. There was also a study by Broch Porcar et al. (48) and they considered that by using data mining, modeling, machine learning and other techniques to generate predictions, risk quantification methods could be developed to predict QTc interval prolongation. The QTc interval risk score showed good predictive performance, with good sensitivity (74% high risk, 67% intermediate risk), specificity (77% high risk, 88% intermediate risk), positive (79% high risk, 55% intermediate) and predictive value of being negative (high risk 76%, intermediate risk 88%). In addition to circulatory function and ECG function, water and electrolyte disturbances are also risk factors for patients with critical symptoms. The Spanish researchers Broch Porcar et al. (48) developed a Spanish national algorithm by reviewing the management of hyponatremia in ICU patients to improve the standardized diagnosis and treatment of hyponatremia. There was also a study (49) that the analysis group of machine learning and the analysis library of collaborative data which were based on the intensive care information system were used to know the area under the curve could be greater than 0.80 when gastrointestinal bleeding in patients was after 5 h, and it had good predictability. In addition to bleeding risk, ICU patients are also at risk of embolism. Deep vein thrombosis (DVT) is associated with high morbidity, mortality, and increased healthcare costs. Researchers (50) developed gradient boosting machine learning algorithms to predict the risk of DVT in patients 12 and 24 h before onset. The area under the curve for the diagnosis of in-hospital DVT obtained by machine learning predictors was 0.83 and 0.85, respectively.

Recommendation 7: It is recommended to use machine learning method to build modeling to conduct early screening of hospitalized patients, so as to provide help for clinicians intervene early and reduce the severity of the disease. (88% consensus)

Background and Evidence:

Compared with ordinary patients, patients with critical symptoms often undergo longer hospitalization time, more expense, and poorer prognosis. Early detection of the change of patients' condition and timely intervention are of significance for preventing the progression of the disease. Machine learning methods can facilitate early screening of diseases and timely treatment of diseases. However, for different subjects, attention should be paid to the correction of heterogeneity before the model is applied, otherwise it will easily lead to wrong clinical guidance. Experts suggest using machine learning method to building modeling for early screening of patients with critical symptoms, so as to help clinicians intervene early and reduce the severity of the disease.

A study published in 2020 evaluated several machine learning methods by using 5-fold cross validation, and applied the XGBoost algorithm to make a AI prediction model for sepsis. The validation results showed that its accuracy = 82% ± 1%; sensitivity = 65% ± 5%; specificity = 88% ± 2%; area under the receiver operating characteristic curve (AUROC) was approximately 0.89, significantly better than the SOFA score (AUROC = 0.596), which might help clinicians deploy appropriate therapeutic regimen, so early and precise responses to this AI algorithm will reduce costs, improve outcomes, and benefit healthcare systems, medical staff, and patients (51). For example, a multi-center and real-world data study in 2020 confirmed that after applying the early warning model in the clinical setting, the in-hospital mortality rate of patients with sepsis decreased by an average of 39.5%, the length of hospital stay decreased by 32.3%, and the 30-day readmission rate of sepsis-related hospitalization decreased by 22.7% (52). In addition to sepsis, machine learning methods have also been used in early screening of other critical illnesses, and a study published in 2021 used a model built with four machine learning methods (Random Forest, XGBoost, GLM-Boost, and LASSO-GLM) to predict pediatric multiple organ dysfunction (MOD). The results showed that the early prediction model of all methods achieved an AUROC of 0.91, and early prediction through risk-based patient monitoring could provide more than 22 h of lead time for the occurrence of MOD, which would play an important role in improving the prognosis of patients (53). However, there were also articles that suggest that clinicians should first calibrate the model according to the heterogeneity of patients before applying the relevant model, so as to avoid misjudgment that might affect clinical decision-making (35). However, in clinical work, first-line clinical staff should pay more attention to the existing scoring system and supervise the actual application, otherwise it will be futile to simply improve the performance of the model without improving the clinical application and response speed (54).

[Diagnosis]

Recommendation 8: It is recommended that the image data of patients with critical symptoms be included in the intensive care database to provide more comprehensive, accurate and timely diagnostic information, so as to guide clinical decision-making through relevant algorithms. (92% consensus)

Background and Evidence:

There have long been studies using AI in the screening and diagnosis of tumors and the images of infectious foci, and have confirmed its advantages in rapidly processing a large amount of image data, moved the diagnostic "gateway" forward, and avoided missed diagnosis and misdiagnosis (55, 56). The disease state and imaging manifestations of patients with critical symptoms are more complex and diverse, and the optimal timing and scenarios for using artificial intelligence for imaging diagnosis need to be more verified. The expert group believes that AI-assisted imaging diagnosis of ICU patients has good application prospects, and recommends devoting to relevant exploration to improve the efficiency and accuracy of diagnosis and provide reference for clinical decision-making.

A study of 3,078 chest radiographs from 500 ICU patients at Michigan Hospital used directional measurements and deep machine learning features to model ARDS with an accuracy of 83% and an AUC value of 0.79 (57). Cerebellar model arithmetic computer analyzed the supine chest radiograph: the AUC values for the diagnosis of pneumonia and pleural effusion were 0.737 and 0.740,

respectively, which were similar to those of imaging experts (AUC values are 0.779 and 0.698) (58). In the outbreak of COVID-19, AI-assisted imaging diagnosis has performed well. Various machine learning methods could not only quickly identify the CT images of COVID-19 (AUC values were between 0.951 and 0.980) from a large number of lung CT images, but also It could predict severe transformation in patients (AUC value was 0.848) (59). The machine learning method combining classical imaging processing and deep learning analyzed CT images of 110 patients with severe subdural hematoma, and showed that the sample recall rate and precision rate were 78.61 and 76.12%, respectively, and the specificity judged based on the severity of the hematoma volume was 92.31%, which could help physicians save decision-making time (60).

In addition to radiological imaging, AI has also been applied in other ICU bedside imaging diagnosis. One study in 2019 showed that the neural network model could detect bedside lung ultrasound B-lines with a sensitivity and specificity of 0.871 and 0.930 (61); two studies in 2021 showed that the neural network model used ultrasound images to diagnose patients with Sepsis early and the accuracy and sensitivity of developing AKI are higher than those of professional radiologists (62, 63). Electrical impedance tomography (EIT) can only roughly show the distribution of ventilation and blood flow in various regions of the lung, but it cannot be quantified as a bedside monitoring index. The neural network model trained by deep learning can calculate information such as lung volume, air flow rate, normalized airway pressure and even transpulmonary pressure from the EIT signal, and AI can also optimize the output image of EIT and even reconstruct the chest image (64).

Recommendation 9: It is recommended to divide patients with sepsis, acute kidney injury, and acute adult respiratory distress syndrome into phenotypes with different clinical outcomes and treatment responses by means of cluster analysis, and identify patients who are most likely to benefit from specific treatment strategies. (91% consensus)

Background and Evidence:

Cluster analysis can identify relatively homogeneous groups within heterogeneous populations. Some treatments are only effective in certain groups of people. Clustering techniques were used to classify patients with critical symptoms into distinct phenotypes by significant differences in comorbidities, laboratory indicators, vital signs, clinical outcomes, and treatment responsiveness, identifying groups that benefit from specific therapies. At present, the identification of phenotypes has made research progress in sepsis, AKI, and ARDS, but the accuracy and generalizability of phenotypes still need further verification. The expert group recommends that patients with critically symptoms be divided into different phenotype by cluster analysis to identify those most likely to benefit from specific treatment strategies.

Clinical and/or host response data and machine learning (e.g., latent class analysis and K-means clustering) were used to segment critically-ill patients with sepsis, AKI, ARDS, etc. into distinct phenotypes (65–68). A RCT study in 2021 identified 4 coagulation-based sepsis phenotypes by K-means clustering and used a machine learning means to determine which phenotype would benefit from rhTM (69); another RCT study by Cluster analysis identified 4 clinical phenotypes of sepsis. These phenotypes differed in demographic characteristics, laboratory abnormalities, patterns of organ dysfunction, and were not homologous to traditional patient groups such as site of infection, pattern of organ dysfunction, or disease severity (70); a latent class analysis of an AKI cohort in 2020 identified two phenotypes of sepsis acute kidney injury with distinct clinical outcomes (71); a prospective observational cohort research through unsupervised consensus clustering and machine learning analyzed expression profiles of the whole blood RNA and identified 4 sepsis endophenotypes (Mars 1–4), of which Mars 1 was significantly associated with 28-day mortality. To facilitate clinical application, the study also extracted accurate classification biomarkers for each phenotype (72). Two different ARDS phenotypes have been identified by the LCA method using data from randomized controlled trials of ARDS. These phenotypes had different clinical outcomes. And different treatment responses to positive end-expiratory pressure strategies (73), fluid therapy (74), and simvastatin (75) have been identified.

[Treatment]

Recommendation 10: In specific clinical scenarios, such as decision making for tracheal intubation and intensive care drug decision, it is recommended to build a decision-making model that can be used for clinical treatment based on machine learning algorithms. (74% consensus)

Background and Evidence:

The condition of ICU patients is usually difficult and critical. Electronic medical record systems, monitors, ventilators and other instruments and equipment can generate massive amounts of vital information data, which far exceeds the ability of ICU doctors to continuously process and correctly interpret them, and affects the effectiveness of clinical decision-making and responsiveness. Artificial intelligence (AI) models can continuously clear, categorize, classify, calculate, and correlate a large amount of data, and make predictions for patients, thereby assisting clinical decision-making and improving the quality and efficiency of critical care.

Several studies have evaluated the clinical impact of applying artificial intelligence techniques such as machine learning to make treatment decisions. In 2018, Komorowski et al. applied reinforcement learning to the sepsis population, and AI clinicians could optimize fluid and vasoactive drug treatment and reduce the fatality rate (14). In 2019, a study established a model to predict urine output in patients with AKI. Compared with the traditional Logistic regression model, the XGBoost model could better distinguish whether patients had volume responsiveness (76).

AI technology has been tried to be applied to clinical situations such as extubation decision-making and optimization of drug treatment for patients. A 2018 retrospective study used machine learning to identify patients requiring prolonged mechanical ventilation (PMV) and those with high risk of tracheostomy (77). In 2021, Fabregat et al. compared three classification learning methods (Logistic regression, XGBoost, and support vector machines) to predict extubation outcomes, which may potentially reduce extubation failure rates (about 9%) (78). Another study in 2021 established a predictive model for accidental extubation through a machine learning algorithm, in which the random forest algorithm obtained the best AUROC of 0.787 (79).

The application of machine learning to optimize the therapeutic effects of anticoagulation, anti-infection and sedation in patients with critical symptoms is still in the exploratory stage. Chen et al. (80), Su et al. (81), Li et al. (82) compared different machine learning methods to predict the therapeutic effect of anticoagulant drugs

(citrate, heparin). The scores are overall better than the other models. A single-center retrospective study in 2022 used machine learning and cluster analysis to provide guidance on antibiotic management in patients with critical symptoms (83). Another study in 2022 based on self-attention and residual structure of convolutional neural network (CNN) had a good predictive effect on anesthesia depth monitoring (84). The examples above illustrate the potential role of AI in guiding critical decisions in patients with critical symptoms. But the vast majority of developed ICU-AI models are still in the testing or prototyping stage, and only a few have actually been evaluated in clinical practice. Van de Sande et al. found no studies suggesting the results of integrating AI models in routine clinical practice (85). Research on AI used to guide clinical decision-making is mostly calculated from retrospective and observational datasets. Therefore, in order to have AI directly guide clinical decision-making, it is necessary to conduct a comprehensive analysis of the suggested sequences or strategies derived from such AI systems with more high-quality and prospective studies to be designed.

[Prognosis and follow-up]

Recommendation 11: It is recommended to use machine learning methods to predict the prognosis of patients with critical symptoms. (85% consensus)

Background and Evidence:

There are more and more predictive models for mortality in ICU patients. Many data models are better at than disease prediction than clinical scoring systems. The sensitivity and specificity of some predictive models still rely on the assistance of clinical scoring systems. AI models in intensive care medicine are mainly generated by retrospective data, with small sample sizes and low reproducibility of conclusions, which are lack of sufficient external validation or prospective evaluation.

There are various machine learning models and algorithms, such as: support vector machines (SVM), Gradient Boosting Decision Tree (GBDT), Logistic regression (LR), adjacent algorithms (KNN, K-Nearest Neighbor), and Random Forest (RF). Studies have shown that the SVM model is a useful tool for early prediction of patients with a higher risk of death upon admission to the ICU. Compared with the early warning score of the SAPS II score, it was better at predicting 7-day mortality. However, the sensitivity and specificity of the SVM model without SAPS II significantly decreased (86). The prediction performance of the machine learning method and the traditional scoring system was further compared according to different diseases. The results were as follows: (1) Sepsis; The results in 2021 showed that GBDT is more accurate than other models (GBDT, LR, KNN, RF, and SVM) in predicting death in patients with sepsis (87). García-Gallo et al. used an assembly algorithm such as SGB to generate a sepsis model that was more accurate in predicting 1-year mortality than traditional scoring systems such as SAPS II, SOFA or OASIS (88). (2) Intracebral Hemorrhage (ICH); Nie et al. (89) indicated that RF was the best model for predicting mortality in ICH patients treated in the ICU, and all machine learning algorithms used to predict mortality in the ICU showed better results compared to the APACHE-II score. (3) Severe acute pancreatitis (SAP); Halonen et al. (90) established an artificial neural network (ANN) model for predicting the severity of acute severe pancreatitis, and the results were better than the Rason score, Glasgow-imrie, APACHE-II, and SOFA scores. The article by Ding et al. (91) also showed that the ANN model could easily screen patients with high risk of death in the early stages of acute pancreatitis.

Finally, it is important to note that the study by Niven et al. (92) showed that a minority of critical care practices with research published in high-profile journals were evaluated for reproducibility; less than half had reproducible effects. This question highlighted the importance of accurate labeling and precise reporting methods, including data preprocessing and functionalization.

[Auxiliary decision-making system changes the clinical path]

Recommendation 12: A clinical decision support system (CDSS) can be used to improve compliance with guidelines for diagnosis and treatment of patients with critical symptoms and the implementation of clinical pathways. (86% consensus)

Background and Evidence:

Evidence-based clinical diagnosis and treatment guidelines provide standardized and homogeneous diagnosis and treatment strategies for the treatment of patients with critical symptoms. However, compliance with clinical guidelines is not high in routine ICU care, resulting in an increase in avoidable patient mortality (93, 94). A clinical decision support system (CDSS) is a computer program that helps health care workers make decisions. With the clinical application of CDSS, most studies have shown that the application of CDSS can assist ICU physicians in decision making, improve compliance with diagnosis and treatment guidelines, and improve outcomes of patient. However, there are many types of CDSSs. One CDSS is aimed at a certain disease, and the development cost is high. The CDSS based on big data has been applied to clinical decision-making, but it has not been used to change guideline compliance. Moreover, CDSS needs to be integrated with the patient electronic health record system. Due to the different electronic health record systems adopted by different regions or hospitals, the promotion and application of CDSS in different hospitals are limited. Therefore, the expert group believes that CDSS can be used to improve the compliance with the guidelines for diagnosis and treatment of patients with critical symptoms, but CDSS based on big data is still in the stage of research and development. It is recommended that qualified hospitals take the development and clinical application of CDSS based on big data into consideration to improve compliance with guidelines.

As early as in 2011, CDSS, such as a "flow sheet," can monitor various parameters of patients in real time at the bedside, screen patients with sepsis early and make a series of mandatory treatment measures according to SSC guidelines. The application of CDSS can significantly improve the compliance with SSC guideline of resuscitation bundle strategy, shorten the duration of antibiotic use (90), and reduce hospital mortality (95). In the clinical implementation of lung protective ventilation with low tidal volume, by using CDSS to guide medical staff to set the ventilator mode and support level, the compliance with lung protective ventilation improved, and the level of tidal volume increased significantly after CDSS was discontinued (96). In a study of delirium management, the duration of delirium episodes was significantly reduced, followed the adoption of the tailored ICU delirium guideline CDSS and the duration of coma was reduced, with the brain function improved (97). In another prospective observational study assessing the compliance with AKI guidelines, the CDSS for AKI was integrated into the intensive care information system in the ICU and found the proportion of patients with worse condition from stage 1 AKI, and the proportion of inappropriate use of enoxaparin dose as well as that of morbidity rate

of patients with AKI was significantly reduced (98). It can be seen that CDSS can improve guideline compliance. However, there is currently no big data-based CDSS application in clinical practice to improve guideline compliance, which needs to be confirmed by further research in the future.

Establishment, standards and principles of a large database for intensive care

Recommendation 13: It is recommended to build a intensive care medicine database and data analysis platform. (98% consensus)

Background and Evidence:

The intensive care database can provide a good data foundation and new ideas for clinical medical research, which in turn can improve the understanding of diseases. For example, in Sepsis 1.0 (99), sepsis was defined as a systemic inflammatory response syndrome (SIRS) caused by infection. Although various diagnostic indicators were more complete in Sepsis 2.0 (100), it still continued the standard of Sepsis 1.0. However, the diagnostic criteria of infection and SIRS cannot accurately describe the disease characteristics of patients, such as different primary diseases, different symptoms and mortality of patients. In 2016, Sepsis 3.0, which was mainly based on big data analysis, was born (101), which defined sepsis as life-threatening organ failure caused by the body's uncontrolled response to infection, i.e., infection and organ function diagnosis. Patterns, making the definition of Sepsis more adaptable to pathophysiology and easier to implement in clinical practice. It can be said that the intensive care medicine databases that have been constructed abroad, such as the Medical Information Mart for Intensive Care (MIMIC) and the eICU Collaborative Research Database (eICU-CRD) (17), are used in clinical practice. The role played in diagnosis and treatment has gradually become prominent. At present, the pace of establishing a intensive care big data platform has also been accelerated in China, but most of them are limited to individual databases in each hospital, and there are still some deficiencies in data exchange and influence. Therefore, we recommend building a intensive care medicine database and data analysis for Chinese people platform to strengthen discipline construction and improve the level of treatment for patients with critical symptoms.

Recommendation 14: It is recommended to form a standard normative intensive care dataset. (97% consensus)

Background and Evidence:

Standard and normalized datasets are the basis for big data applications and facilitate data collaboration between research centers in different regions. There is a lot of information obtained by ICU equipment and instruments, and the data can be included in a reasonable and standardized manner and classified, so that they can be used more fully and conveniently. At present, there are many big data information systems for intensive care medicine at home and abroad. These information systems divide clinical data into different data elements according to specific classification standards, and then use specific data collection methods to acquire and analyze data. Referring to basic structure and data standard of the national electronic medical record (102), Beijing local standard - intensive care medicine dataset and the intensive care medicine database widely used in the field of medical research (103), the recommended standard data set should include the following data sets: (1) Basic information data of patients; (2) Diagnostic information data of patients; (3) Monitoring data of Patients; (4) Drug use data of patients; (5) Laboratory information data of patients; (6) In and out data of patients; (7) Imaging data of patients; (8) Etiology data of patients. See Table 2 for details.

It is also recommended that adjustments can be made in combination with actual conditions such as hospital disease conditions, information centers, laboratory testing items and other objective conditions. For example, based on acute respiratory distress syndrome, sepsis, acute kidney injury and other common diseases in intensive care medicine to build a special disease database, which is necessary to strengthen the sampling frequency and categories of intensive care information related to special diseases. For example, the acute respiratory distress syndrome database needs to further collection of biomarkers, etc.; The sepsis database requires further collection of vasoactive drugs, etiology collection, organ function assessment, etc.

TABLE 2 Standard datasets.

Basic information data of patients	Time information on patient admission and discharge, demographic information, source of admission, ICU category, time of death, etc.
Diagnostic information data of patients	All disease diagnosis information during the patient's stay in the ICU; the main diagnosis needs to be distinguished from the secondary diagnosis
Monitoring data of Patients	Routine vital signs, ventilator parameter information, blood purification parameter information, aortic balloon counter pulsation parameter information, the mental state, the score information, etc.
Treatment data of patients	The route of administration, use time and drug dose of all drugs during the patient's stay in the ICU; the name, time and related information of the operation; the name, time and related information of the treatment operation, etc.
Laboratory information data of patients	Laboratory examination information during the patient's stay in the ICU, such as sampling time, specimen type, test items, reference range of normal values, etc.
In and out data of patients	Data of all fluids entering and expelling from the body during the patient's stay in the ICU, including fluid type, entry and exit route, time, etc.
Imaging data of patients	Text reports related to radiographic imaging during patient stay in the ICU
Etiology data of patients	The etiological data collected during the patient's stay in the ICU, including sampling time, specimen type, etiological name, etiological drug susceptibility, etc.

Recommendation 15: It is recommended to select automatic collection for objective data first. For data that cannot be automatically collected for the time being, targeted collection should be carried out in combination with research needs, data sources and data types. (92% consensus)

Background and Evidence:

The data collection process should follow the principles of comprehensiveness, multi-dimensionality, efficiency and timeliness. In view of the many data sources and rich data structures in the ICU, it is recommended to use automated data collection technology to realize the data collection process so as to avoid human errors affecting the use of subsequent data.

Data in the ICU can be broadly classified into "phenotypic data" and "physiologic data." Phenotypic data include demographics, age, sex, laboratory values, and physician and nursing records. Phenotypic data collection can be queried and extracted from electronic medical records (EMRs). Relevant content can be obtained through Python or API, and the required attribute content can be extracted from it. Physiological data include vital signs (blood pressure, heart rate, respiratory rate, core temperature) and other parameters (intracranial pressure, EEG) generated by bedside monitoring equipment. If the data interface of the device can be obtained through various software manufacturers, data collection and aggregation can be realized through the interface docking method. If some devices cannot obtain the data interface, collecting all the data generated by the target device can be tried to acquire the underlying data exchange of the system, the network package between the client and the database, which can convert the data into with restructure and output to new database, based on underlying IO request and network analysis technologies.

Alarms in the ICU, such as ECG leads, blood pressure cuff detachment from patients, completion of infusion pump or air bubbles in tubing, high airway pressure, air leak, or apnea in mechanical ventilation ventilators, etc., which can be classified into the type of physiological data. This part of the data can be collected by collecting logs from log sources of various devices. Continuous waveform data is more complicated to acquire due to its continuous nature and high sampling rate. In recent years, many studies have used time series databases and unstructured databases such as InfluxDB, MongoDB, etc. to explore the writing, storage, and query processes of various continuous-time signals, which can solve the storage-transmission-exchange-exploitation problem (104). For image data, since most of the images are currently stored in the PACS system, it is necessary to clarify whether to collect from the equipment (CT machine, ultrasound machine, etc.) or through the PACS docking port (105).

Recommendation 16: It is recommended to optimize standard system for intensive care big data, standardize multi-center source data, and constrain standard codes, measurement units, field standards, as well as naming dictionaries to ensure the homogeneity and standardization of the use of the large database for intensive care. (95% consensus)

Background and Evidence:

"Information integration, standards first" (106), the construction of large databases for intensive care must be implemented in accordance with the corresponding norms and standards, the standard codes, measurement units, field standards, and naming dictionaries, and it is constrained by standard norms to ensure the subsequent modeling and application process. The consistency of data processing ensures the standardized production of data from the source, and lays the foundation for the construction, data integration, data exchange and data sharing of large databases for intensive care. Intensive care big data are multi-modal data with high privacy and diverse sources, and have the characteristics of multiple data dimensions, good timeliness, high value density and high data quality. The "phenotypic data" and "physiological data" in the ICU can be classified into structured discrete data, time series data, and unstructured text data, image data, and audio-video data (107). The main contents are as follows: (1) Discrete data: basic information and routine data of patients' physical sign, including a series of discrete data such as gender, age, blood type, height, weight, etc., which are mainly characterized data. These data volumes are small and stable. (2) Time series data: mainly physiological data, including time series data of various vital sign parameters such as blood oxygen, heart rate, and ECG. These data are closely related to the real-time symptoms of patients, with high real-time performance, strong continuity, and large datasets. (3) Image data: mainly physiological data, including a large amount of image data such as ultrasound and radiation. These image data are large in volume and are important reference data for diagnosis and operation. (4) Text data: a large amount of text data about patient medical records and diagnostic results, mainly for representation data, including electronic medical records, surgical records, inspection reports, etc. Among all data types of critical diseases, time series data, image data and text data have high information value density and play an important role in clinical diagnosis, treatment and decision making.

Due to the uneven level of informatization in each center and a wide range of coverage, the above-mentioned data formats for intensive care are complicated and difficult to integrate. After negotiation, multiple centers have formulated unified data fields, contents and formats for the big database for intensive care, and established a standard system. For example, for the standardization of image data, the level of imaging departments in different hospitals varies, and multiple centers need to negotiate the image quality standards for uploading compressed original images. For different types of data, in order to ensure the standardization of large databases for intensive care, data governance rules for different types of data can be formulated, and the system will automatically clean the data when it enters the database, supplemented by manual review if necessary to ensure data quality. For the quality assessment of inbound data, it can be measured from normative (the extent to which the data conforms to data standards, data models, business rules, metadata or authoritative reference data), integrity (the extent to which data elements are assigned values according to data rules), accuracy (the degree to which the data accurately represents the true value of the real entity, "real object" that it describes), consistency (the degree to which the data does not contradict the data used in other specific contexts), timeliness (the degree to which the data is correct over time), and accessibility (the degree to which data can be accessed), which are six aspects to manage and evaluate (108).

Recommendation 17: It is recommended to establish a data security system to ensure the security of data storage, processing, sharing and use. (98% consensus)

Background and Evidence:

The information security system in China mainly includes five technical tasks: risk assessment and grade protection, monitoring system, cryptography and network trust system, emergency response system, and disaster preparedness. The security level of

information system is divided into five levels, and the levels from one to five are gradually increased. Centering on the "Network Security Law," "Data Security Law" and "Personal Information Protection Law," China carries out the construction of data classification system. In terms of data security, the security of the data itself (using modern cryptographic algorithms to actively protect data) and security of data protection (active protection of data using modern information storage methods) must both be paid attention to. New security issues need to be addressed in an environment of big data, including balancing privacy and utility, analyzing and governing encrypted data, and verifying authenticated and anonymous users. With the continuous expansion of the application scope of intensive care big data, the content is becoming richer and more valuable with a large amount of sensitive personal information. A security system and a safety management responsibility system for intensive care big data should be established to ensure the security in data storage, opening, and processing (109, 110).

When storing data, system security reinforcement as well as software and hardware architecture design in a distributed environment (such as Apache Hadoop) should be done well. Strict fine-grained access control and risk registration management strategies should be set for static data, and privacy-related data storage should realize classified isolation data encryption (such as AES, RSA, SHA-256 and other encryption methods) and other security technical means, dynamic data classification and identification of important sensitive data should be through encryption and dynamic audit capabilities, using TLS (transport layer security technology) to communicate between cluster nodes and maintain confidentiality during transmission, and enabling unified management across platforms (endpoints, mobile devices, networks, and storage systems) (106).

During data processing, the software architecture and network configuration should be designed according to the database volume and access method, especially for multi-center, and the appropriate hardware architecture should be designed according to the software architecture. And policy configuration such as network security should be done to ensure data security. After the data is authorized to be processed by other parties, the most important question is whether there is misuse and malicious restoration of sensitive data during the processing, whether it complies with laws and regulations, and whether it complies with the privacy clauses agreed by both parties (106). In multi-party computation, data leakage is avoided through system policy design such as data desensitization (111) and federated learning (112).

When sharing data, measures such as data desensitization, rights management, and log auditing should be taken to ensure data security. Data cannot be unconditionally open to the public or third parties. Consideration should be given to the fact that sensitive information can be easily restored after a single information is desensitized through multi-source collisions which may lead to security risks, therefore, only point-to-point sharing, or multilateral transactions based on certain special constraints, such as sharing health records, patient medication information, medical images and other information about intensive care big data. Whether the data sharing is justified or not should be comprehensively weighed on the occasions of the data and the subject's right to know.

Ways and methods to solve big data problems in intensive care medicine

[Type of data]

Recommendation 18: It is recommended to use processing methods of digital signals such as filters to preprocess time series data, deep learning to process image data, use Natural Language Processing (NLP) technology to process unstructured text data. (93% consensus)

Background and Evidence:

From the perspective of machine model building, intensive care data can be roughly divided into four categories: numerical time series data, numerical non-series data, text data, and image data. Among them, numerical data can be divided into two categories according to the collection density: (1) time series data, or "streaming data," including electrocardiogram, arterial and intracranial pressure, hemodynamic monitoring, ventilator data, brain waves and other data with relatively high collection frequency; (2) non-sequential data, or "sparse data," including blood gas analysis, laboratory test results, medical history and other data with relatively low collection frequency. Different types of data can be combined to improve the accuracy of AI prediction models (113), provide decision support under complex and uncertain diagnostic conditions (114), and better adapt to the clinical real-time data environment.

For time series data, before further pattern recognition or other processing through different algorithms, processing methods of digital signals such as filters are usually used for preprocessing. The main purpose is to use various mathematical methods to strip components of different frequencies in the signal for targeted treatment. For example, in electrocardiogram (ECG) data processing, a five-minute moving average is often used for low-pass and high-pass filtering (29, 115, 116), and when building an EEG signal model, Narula et al. also used a band-pass filter to remove baseline drift and high-frequency interference (117).

For non-series data, the processing skills are mainly reflected in solving the problems of data (parameter) outliers and missing values, screening and dimensionality reduction according to different algorithm models. After the corresponding preprocessing of the data, whether it is a simple algorithm such as linear regression and logistic regression, or a sophisticated algorithm such as lifting algorithm and reinforcement learning (14, 118), it can achieve good results in the corresponding scene. So no special recommendation is made.

For image data, such as CT, pathological slices, ultrasound images, etc., most of them are processed by deep learning (such as convolutional neural network CNN, etc.) and other tasks (119–122). In particular, Walsh et al. believed that deep learning methods can directly extract important features from images, which could help to generate novel biomarkers and more accurate image analysis tools (123).

For unstructured text data, such as narrative text in EMR, as well as radiology, pathology reports, etc., the content can be mined and processed through natural language processing technology to obtain pathological information, social environment information, etc., which can be combined with the existing expert knowledge base (such as the unified medical language system, etc.) as a supplement to improve the accuracy of related prediction models, and show a speed and accuracy that exceeds manual processing (124–126). In particular, natural language processing for Chinese, ICTCLAS system, THULAC toolkit,

etc. are all good auxiliary tools, but the Sinicization of knowledge bases such as UMLS (or other Chinese medical knowledge bases) needs to be demonstrated in the literature.

[Data preprocessing]

Recommendation 19: It is recommended to use resampling methods to deal with unbalanced datasets. (78% consensus)

Background and Evidence:

In intensive care medical datasets, unbalanced data is very common. Unbalanced data refers to the uneven distribution of the number of samples among each category in the classification task, and there will be a particularly large gap, which will greatly affect the final performance of the prediction model. For example, a small number of death samples in intensive care medicine datasets carry important information about mortality prediction, but are ignored because the model is insensitive to data imbalances. In response to the phenomenon of data imbalance, the expert group recommends using resampling methods to process imbalanced data, which are mainly divided into three types: undersampling, oversampling and synthetic oversampling. Undersampling is the random sampling of fewer samples from most classes so that the data tends to be balanced. Edited Nearest Neighbors (ENN) is the most typical undersampling method. Oversampling is to generate more labeled samples according to the sample rules with fewer sample labels so that the data tend to be balanced. Synthetic Minority Over-sampling Technique (SMOTE) is an oversampling technique that generates synthetic samples for the minority class. In order to reduce the fitting problem caused by oversampling and undersampling, a method combining oversampling and undersampling is extended to deal with data imbalance on this basis, such as SMOTEENN, SMOTETomek, etc. In the study of using machine learning to predict atrial fibrillation, Tiwari et al. used a variety of sampling methods to deal with the imbalance problem that the data in the control group was much more than that in the experimental group, and compared the data under different sampling methods, and finally chose the random oversampling method according to the classifier effect (127). Papp et al. used SMOTE sampling to synthesize samples from the minority class for the class-imbalance problem and analyzed the synthesized new data results through cross validation and confusion matrix (128).

Recommendation 20: It is recommended to convert original categorical variables and numerical variables into variables that can be directly processed by machine learning algorithms through one-hot encoding, sequential encoding, etc. (83% consensus)

Background and Evidence:

The function of variable category transformation is to convert the original category of intensive care medical data containing the above information into a form suitable for data mining and easy for model understanding. The transformation of variable categories makes the original data more tidy and consistent through operations such as encoding. It is recommended to use methods such as one-hot encoding and sequential encoding. One-hot encoding is a common numerical processing method for unordered categorical variables, with "1" to indicate that it belongs to this category, and "0" to indicate that it does not belong to this category. One-hot encoding will add new variables to the original variables. The number of new variables being added is the number of types. Ordinal coding is a common numerical processing method for ordinal categorical variables. This coding makes numerical one according to the different degrees represented by the ordinal variables, such as scores about a patient's health status from 0 to 5.

Recommendation 21: It is recommended to use dimensionality reduction methods such as principal component analysis to perform variable screening of high-dimensional features in intensive care datasets. (90% consensus)

Background and Evidence:

In most research problems of intensive care big data, the datasets used usually have high-dimensional feature variables, which can easily lead to overfitting problems and increase training costs. Therefore, it is necessary to extract important features through variable screening to achieve the purpose of data dimensionality reduction. Experts recommend principal component analysis, variance selection, univariate feature selection, regularization models, feature ranking based on machine learning models, and recursive feature elimination methods.

Principal Component Analysis (PCA) is a popular general feature dimensionality reduction method, which can be used to reduce the dimensionality of various types of data such as numerical values, texts, and images. Essentially, multiple variables are synthesized into a few independent components, and each component can reflect the information of the original variable, which can improve the learning speed and reduce the training cost. Variance selection is a simple feature selection method that filters features by removing features with low variance. Univariate feature selection usually uses statistical test methods such as chi-square test and F test, or measures such as Pearson correlation coefficient and distance correlation coefficient to determine the relationship between variables. The regularization model is mainly divided into L1 regularization and L2 regularization. By adding additional constraints or penalty terms to the loss function of the existing model, it can prevent overfitting and improve the generalization ability of the model. L2 regularization is more stable than L1 regularization and is more favorable for the understanding of features. Regularization models are often used in feature selection of medical data. In the study on early triage of COVID-19 patients with critical symptoms, Liang et al. selected 10 statistically significant variables as predictors by the Lasso method (129). Many machine learning methods can achieve feature scoring, such as feature ranking by measuring feature importance. Therefore, it is recommended to use the selected machine learning model to complete feature selection, including SVM, random forest, decision tree, XGBoost, LGBM and other models. By adjusting the calculation parameters of feature importance, the feature ranking of different methods can be obtained. This method is convenient, effective and easy to understand the relationship between the model and features, but it is needed to verify the model fitting effect by means of cross-validation. In addition, recursive feature elimination methods can be considered to screen the features of intensive care medical data.

[Model Construction]

Recommendation 22: It is recommended to select supervised learning, unsupervised learning and reinforcement learning models for critical disease prediction and identification according to different scenarios and different data types. (97% consensus)

Background and Evidence:

The intensive care unit monitoring system collects a large number of the patients' respiratory, hemodynamic, neurological and clinical data, and its electronic medical record system also records the patient's

clinical treatment and medication information in detail. The data types include types of text, digital and image. Through the processing and analysis capabilities of big data by machine learning algorithms, key features of the data can be mined to assist in diagnosis and decision making. Machine learning algorithms can be classified into supervised learning, unsupervised learning, and reinforcement learning, depending on whether the dataset has labels. Among them, supervised learning can learn and summarize models, including decision trees, support vector machines, random forests, naive Bayesian models, artificial neural networks and other models; Unsupervised learning models can discover hidden patterns without manual annotation or data grouping, which can find potential similarities and differences in the data. Common algorithms include k-means, principal component analysis, hierarchical clustering, etc.; Reinforcement learning can learn the best behavior or mode that should be taken from experience. The model type should be selected according to the data type and medical task. Among them, for numerical data and clinical prediction problems, supervised learning models can be used; For text data, natural language processing models and unsupervised learning models can be used; For image data, Convolutional Neural Networks (CNN), Recurrent Neural Network (RNN) can be used for medical image recognition and segmentation; For clinical auxiliary decision-making tasks, reinforcement learning models can be used. According to the literature survey, the usage scenarios of three different learning methods include: (1) Supervised learning: prognosis prediction, phenotype classification, analgesic and sedation strategy selection, mortality risk prediction, disease severity prediction, prediction for length of stay in the ICU, etc. (2) Unsupervised learning: disease pattern mining and representation based on electronic health records (EHR). (3) Reinforcement learning: decision making of treatment plan, recommendation of fluid volume, robot-assisted surgery, etc.

Specifically, examples of the usages and indications for the three types of learning are as follows:

1 Supervised learning: Prognosis prediction and dose recommendation for heparin patients (82); monitoring and adjustment of Local citric acid anticoagulation (80); prediction of in-hospital mortality risk in patients with critical symptoms (124, 130) prediction of mortality risk in patients with candidemia (125), prediction of the severity of lung ultrasound in ICU patients (126), etc.
2 Unsupervised learning: Phenotype classification and sedation strategy selection in mechanically ventilated patients (131); temperature pattern recognition in patients with critical symptoms (67), blood pressure pattern recognition (132); subtype of diseases extracted from electronic health record data (133, 134).
3 Reinforcement learning: Dynamically provide optimal treatment plan and select intravenous fluids and vasopressor doses for patients in the ICU (135).

Recommendation 23: It is recommended to use a causal inference model to explore and discover causal relationships in the intensive care field. (89% consensus)

Background and Evidence:

The model system of causal inference is built on the basis of causal-heuristic learning and reasoning. It conducts in-depth mining of relevant data to extract causal structure, and conducts causal-heuristic estimation. It studies the influence of intervention variables on prognosis and obtains the key index of prognosis evaluation. The directions involved include causal discovery, causal structure learning, causal inference, causal deep learning, etc. In response to the need for poor ICU prognosis or poor survival rate, as well as the need to accurately determine the influencing factors of prognosis, it is advisable to use the frameworks of DoWhy, CDT and CausalML and establish a causal-heuristic learning inference and decision-making evaluation system based on the database of specialized diseases and multi-center of intensive care.

First, implement big data-driven causal structure identification, mine causal relationships, and conduct feature analysis, effect analysis, and interpretability analysis. Richens et al. (136) proposed a counterfactual diagnostic strategy for expected failure and expected adequacy, breaking the traditional diagnosis method of diseases based on symptoms and narrowing the scope of possible conditions by using counterfactual questions. Wei et al. (137) described the causal relationship between some variables in the recommended system from the perspective of causal inference and solved the influence of popularity bias on the model from the perspective of counterfactual inference. Goudet et al. (138) used deep learning methods to propose a causal generative neural network (CausalGNN), which exploited conditional independence and distribution asymmetry to discover bivariate and multivariate causal structures, and learned functional causal models from observational data to figure out a causal roadmap between clinicopathological features.

In addition, the causal effect was further estimated on the basis of the causal relationship, and machine learning methods such as generalized random forest (GRF) (139) were used to calculate CATE and HTE to predict the difference in prognosis under different ICU intervention methods and research the degree of impact on prognosis by intervention variables. Tan et al. (140) used an approach like adversarial training to give an interpretable means for recommended systems. The advantages of these methods are that the data can be used to reason about the source characteristics of heterogeneity to estimate a series of estimators, which also apply to high-dimensional data and missing data and have good interpretability. Through techniques based on causal discovery and estimation, learning the most discriminative characterization, discovering diagnostic basis and key characteristic indexes, judging prognosis accurately, and providing effective interventions for clinical treatment can be realized.

[Verification of the model]

Recommendation 24: It is recommended to add external validation to internal validation of the model. (94% consensus)

Background and Evidence:

Model validation is the process of evaluating the predictive performance of a model after it has been constructed. The importance of model validation is reflected in measuring the prediction accuracy of the prediction model, feeding back the model building process, and adjusting the model building ideas if necessary. The model verification idea is relatively mature at present, and there is a relatively consistent method consensus. In practice, model verification is mainly divided into internal verification and external verification. The expert group believes that the following methods can be used to evaluate the model validation process.

Internal verification: In general, verification based on their own data (internal verification) is required. That is, a part of the data (like

80% of the total) are randomly selected as the training set for building the prediction model, and the rest of the data are used as the test set to evaluate the performance of the model. In order to verify that the model has good performance on newly generated clinical data, "spatio-temporal division" can be added to the random division, which is the data of the latest period specially divided as an independent validation set (141). In order to improve the estimation robustness of the evaluation indicators, K-fold cross-validation can be used (18). Divide the data set into K parts (such as 10 parts), use K-1 data to build a prediction model, use the remaining data for verification, repeat K times, and take the average of the K times of model prediction evaluation indicators as the accuracy index of the final model. The implementation of internal verification is relatively simple, but since the training set and test set are both derived from the same data, the model extrapolation ability (i.e., "generalization" ability) is relatively weak.

External verification: Different regions and hospitals may have differences in data distribution due to differences in population, disease characteristics, and diagnosis and treatment habits. In order to verify that the model has good extrapolation, it is recommended to perform external validation on multi-center data from different regions and different hospitals.

Recommendation 25: It is recommended to use indicators such as sensitivity, specificity, F1 score, and AUC to evaluate the performance of classification models, and indexes such as R^2, MSE, RMSE, and MAE to evaluate the performance of regression models. (91% consensus)

Background and Evidence:

During model validation, a series of evaluation metrics should be used to measure model performance (i.e., predictive effect). For classification model and regression model, different indicators are used for evaluation.

Performance evaluation indicators of classification model: For classification models (models whose predicted values are categorical variables), sensitivity (also known as recall), specificity, F1 score, precision, AUC (Area Under Curve) and other metrics to evaluate the performance are generally used (3). Among them, the F1 score is the harmonic value of sensitivity and positive accuracy rate, and the larger the value, the better the model performance is. AUC is the area under the ROC curve drawn by "1-specificity" and "sensitivity." The larger the value, the better the model performance is. When the sample categories are not balanced, it is recommended to use the area under the PR curve, AUPRC, to evaluate model performance.

Performance evaluation indicators of regression model: For regression models (models whose predicted values are continuous variables), R^2 (R squared, coefficient of determination, coefficient of determination), Mean Squared Error (MSE), Root Mean Squared Error (RMSE), Mean Absolute Error (MAE), and other indicators are generally used to evaluate the performance. The closer the determinant coefficient, R^2, is to 1, the better the model performance is. The closer MSE, RMSE and MAE are to 0, the better the model performance is.

[**Model interpretability**]

Recommendation 26: It is recommended to explore the interpretability of the model to facilitate the clinical transformation of complex machine learning models. The recommended model interpretation methods include Feature Importance, LIME, and Shapley. (91% consensus)

Background and Evidence:

AI models based on big data training in intensive care medicine are often complex, and their complexity is mainly reflected in the large number of parameters and the complex functional relationship between various parameters. Such complex models are often not conducive to clinicians to analyze the pathophysiological mechanisms, and it is difficult to determine the causal relationship between variables, which seriously hinders the clinical transformation of AI research results. The interpretability of the model is considered as an effective way to solve the above problems. Understanding characteristics, classification, and prediction of indicators, and then understanding why a machine learning model makes such a decision, and what features play the most important role in the decision allow us to judge whether the model is in line with common sense. For example, an AI doctor trained by a reinforcement learning model is used to treat septic shock (14). The AI prompts the need to increase norepinephrine while appropriately limiting fluid replacement. Understanding the mechanism behind such an algorithm is critical for the reliability of the model. If the algorithm tells you that you need to increase the dose of norepinephrine for the patient because their main contradiction is peripheral vasodilation, rather than fluid deficiency, it can greatly enhance the confidence of the physician in the use of this model, because the diagnosis and treatment made by AI decision-making is consistent with clinical pathophysiological changes.

In addition, several other methods are also used for model interpretability exploration (142). Feature importance can be used. Its main working principle is to change the arrangement of the data in a certain column of the data table and keep the rest of the features unchanged to see how much it affects the prediction accuracy.

Locally Interpretable Agnostic Modeling (LIME) is an algorithm (143) that provides a novel technique to interpret the results of any predictive model in an interpretable and trustworthy way. It works by training an interpretable model locally around the predictions it wants to explain. In layman's terms, select a sample and a point near the sample, and then train a simple model to fit. Although the simple model cannot be effective on the complete data set, it is at least effective near this point. The characteristics of this simple model are human-analyzable, and the trained weights can also represent feature importance.

The Shapley value was proposed by Loyd Shapley, a professor at the University of California, Los Angeles, USA, to solve the problem of contribution and profit distribution in cooperative games. In cooperation of N persons, the contribution of individual member is different, and the distribution of income should also be different. The ideal distribution method is: contribution = income; Is there a quantifiable method for the distribution of contribution and income? The Shapley method is one such method, where the Shapley value of a feature is the average marginal contribution of that feature across all feature sequences.

Clinical application of intensive care big data

Recommendation 27: It is recommended to transform and promote early warning tools that meet critical needs. (91% consensus)

Background and Evidence:

The construction of early warning tools can make early predictions for the risk of various adverse events in the ICU, thus helping clinicians to take timely measures to prevent problems before they occur, reduce the incidence of adverse events in patients effectively, and improve early response capabilities. At present, although early warning models have been constructed and verified for the occurrence and prognosis of a variety of critical diseases at home and abroad, there is still insufficient research to truly conduct large-scale clinical trials to evaluate their application value. A Big-data clinical trial (BCT) of an early warning tool was implemented in terms of injury and disease deterioration. However, there are still differences in the predictive performance of different early warning tools in different application scenarios, and further promotion and verification are needed. So far, no mature disease-targeted early warning tools have been launched at home and abroad. The expert group believes that it is possible to use AI technology to provide early warning for various adverse events in the ICU. At the same time, it is necessary to carry out BCT research to further verify the clinical practical value of early warning tools, so as to achieve early detection, diagnosis and treatment of diseases.

For the early warning of sepsis, Shimabukuro et al. (144) conducted a BCT study in 2017 and found that patients who used the early warning tool for sepsis shortened the length of hospital stay significantly (10.3 days vs. 13.0 days, $p = 0.042$), and the in-hospital mortality rate was reduced significantly (8.96% vs. 21.3%, $p = 0.018$). However, a single-center BCT study conducted by Semler et al. (145) found that the application of a sepsis electronic warning system neither improved the completion of the 6-h bundle of sepsis ($p = 0.159$), nor improved clinical outcomes (including ICU fatality rate, days in ICU, days of vasoactive drug use).

For the early warning of acute kidney injury (AKI), a large multi-center BCT study in the United States in 2021 (146) found that the AKI early warning system did not improve disease progression in patients ($p = 0.67$). However, BCT evidence from the United Kingdom (147) and China (148) found that although the AKI early warning system could not improve the mortality rate of patients, it could significantly improve the early identification rate of AKI (RR: 1.12, 95% CI: 1.03–1.22, $p < 0.01$) and AKI diagnosis rate (7.9% vs. 2.7%, $p = 0.001$). Another BCT study from the United States found that the AKI electronic automatic alert system did not improve the composite outcome (maximum creatinine change, the need for dialysis or death) within 7 days of patients ($p = 0.88$) (149).

For the early warning of disease deterioration, the Escobar et al. (150) conducted a multi-center BCT study in 2020 that included a total of 43,949 people (15,487 people in the intervention group and 28,462 people in the control group). And it found that early warning tool for disease progression can significantly reduce patient mortality rate (adjusted RR: 0.84, 95% CI: 0.78–0.90, $p < 0.001$).

Recommendation 28: It is recommended to use the information system for intensive care as a carrier to access real-time data and output recommendations for decision making. (91% consensus)

Background and Evidence:

The condition of patients with critical symptoms is complex and fast-changing, and ICU equipment and instruments have a large amount of information, so the data dimension and the update frequency is high. The application carrier should be effectively integrated with the hospital information system, which can obtain high-dimensional information in real time, and be equipped with a prediction model. Based on Hadoop distributed processing technology, Xia et al. (33) designed a big data analysis system for intensive care medicine, and conducted a performance test through the "Study on the Effect of Xuebijing on AKI-related Sepsis" (33). The information system of intensive care big data can integrate ICU high-dimensional information, obtain analysis data in real time, and use it as a carrier for results of intensive care big data such as prediction models and scores (151). Boss et al. (29) developed an online real-time ICU decision support platform that could be used to collect multimodal waveform data and AI-based computational disease modeling, calling it "ICU Cockpit." In the cohort of 979 patients admitted to this 12-bed neurocritical care unit since 2016, the total number of data points processed and stored by the "ICU Cockpit" platform has been approximately 88.9 billion (29). Based on the intensive care information system, Zhang Suzhen et al. used the XGBoost model to integrate relevant parameters and performed machine learning to predict the risk of AKI in patients with septic shock. The sensitivity of the prediction results was 73.3%, the specificity was 71.7%, and the accuracy was 72.5%. Compared with the traditional score, it was significantly improved (152).

When there is no information system of intensive care medicine, the intensive care big data can also be equipped with a online prediction tool of web page, APP, applet, or bedside form and other carriers. Flechet et al. developed a prediction model for acute kidney injury, AKI predictor, and conducted a multicenter prospective cohort study to verify the prediction effect of clinicians and AKI Predictor. The performance of the two at ICU admission was: AUROC was 0.80 [0.69–0.92] and 0.75 [0.62–0.88] ($n = 120$, $p = 0.25$), the net benefit ranges were 0–26% and 0–74%. The machine learning-based AKI predictor achieved similar discriminative performance to physicians in predicting AKI-2 and AKI-3, with a higher overall net benefit because physicians overestimated the risk of AKI. This showed that AKI Predictor has added value to the doctor's prediction. The study also came with an online version of the predictive model[2] (153).

With the development of Internet of Things technology, 5G technology, database technology, etc., the carrier to realize the application of intensive care big data in the future should focus more on the "dynamic holographic prediction system" that obtains ICU information in a comprehensive and real-time way, analyzes the data and makes real-time predictions.

Recommendation 29: It is suggested that the current practice of intensive care diagnosis and treatment should still be led by clinicians with the use of big data technology to coordinate to improve medical efficiency and ensure medical quality and safety. (98% consensus)

Background and Evidence:

In recent years, the development of applications of intensive care big data has made rapid progress, and a large number of articles have been published, including prediction of diseases, early warning of risks, and real-time guidance of clinical medication. In the foreseeable future, big data applications can assist ICU clinical diagnosis and treatment activities. However, at the same time, applications of big data still have problems such as lack of clinical integration, lack of high-quality verification, poor interpretability, few application

2 https://www.akipredictor.com/en/

TABLE 3 Summary of recommendations on the application of intensive care big data.

Part	No.	Recommendation contents	Consensus degree
1. The concept, significance and necessity of intensive care data	1	The concept of intensive care big data: Intensive care big data refers to the datasets with logical connotations formulated by various indicators which are large-scale, multi-heterogeneous, variably dynamic, high-speed and real-time acquisition, low-value density and difficult to analyze traditionally in the whole process of diagnosis and treatment of patients or potential ones with critical symptoms	97
	2	The intensive care big data is multi-modal, massive, dynamic, continuous, and objective, and its correct acquisition can provide auxiliary evidential support for diagnosis of critical illnesses and early warning	98
	3	The establishment of a large database for intensive care in China should follow the principles of multiple center, multiple disease and automatic capture, and provide reliable and accurate data support for the application of big data and the development of artificial intelligence	92
	4	Building a large database of patients with critical symptoms in China for their condition monitoring, the research and development of clinical drug and clinical trials can provide the standardized and individualized treatment for patients with critical symptoms	97
2. Clinical scientific issues concerned by intensive care big data in clinical research	5	It is recommended to use machine learning method to build modeling to make early warning of sepsis, acute kidney injury (AKI), and acute respiratory distress syndrome (ARDS)	94
	6	The prediction model based on machine learning can effectively predict the risk of patients at high risk of potential organ damage in the ICU	89
	7	It is recommended to use machine learning method to build modeling to conduct early screening of hospitalized patients, so as to provide help for clinicians intervene early and reduce the severity of the disease	88
	8	It is recommended that the image data of patients with critical symptoms be included in the intensive care database to provide more comprehensive, accurate and timely diagnostic information, so as to guide clinical decision-making through relevant algorithms	92
	9	It is recommended to divide patients with sepsis, acute kidney injury, and acute adult respiratory distress syndrome into phenotypes with different clinical outcomes and treatment responses by means of cluster analysis, and identify patients who are most likely to benefit from specific treatment strategies	91
	10	In specific clinical scenarios, such as decision making for tracheal intubation and intensive care drug decision, it is recommended to build a decision-making model that can be used for clinical treatment based on machine learning algorithms	74
	11	It is recommended to use machine learning methods to predict the prognosis of patients with critical symptoms	85
	12	A clinical decision support system (CDSS) can be used to improve compliance with guidelines for diagnosis and treatment of patients with critical symptoms and the implementation of clinical pathways	86
3. Establishment, standards and principles of a large database for intensive care	13	It is recommended to build a intensive care medicine database and data analysis platform	98
	14	It is recommended to form a standard normative intensive care dataset	97
	15	It is recommended to select automatic collection for objective data first. For data that cannot be automatically collected for the time being, targeted collection should be carried out in combination with research needs, data sources and data types	92
	16	It is recommended to establish a standard system for intensive care big data, standardize multi-center source data, and constrain standard codes, measurement units, field standards, as well as naming dictionaries to ensure the homogeneity and standardization of the use of the large database for intensive care	95
	17	It is recommended to establish a data security system to ensure the security of data storage, processing, sharing and use	98

(Continued)

TABLE 3 (Continued)

Part	No.	Recommendation contents	Consensus degree
4. Ways and methods to solve big data problems in intensive care medicine	18	It is recommended to use processing methods of digital signals such as filters to preprocess time series data, deep learning to process image data, use Natural Language Processing (NLP) technology to process unstructured text data	93
	19	It is recommended to use resampling methods to deal with unbalanced datasets	78
	20	It is recommended to convert original categorical variables and numerical variables into variables that can be directly processed by machine learning algorithms through one-hot encoding, sequential encoding, etc.	83
	21	It is recommended to use dimensionality reduction methods such as principal component analysis to perform variable screening of high-dimensional features in intensive care datasets	90
	22	It is recommended to select supervised learning, unsupervised learning and reinforcement learning models for critical disease prediction and identification according to different scenarios and different data types	97
	23	It is recommended to use a causal inference model to explore and discover causal relationships in the intensive care field	89
	24	It is recommended to add external validation to internal validation of the model	94
	25	It is recommended to use indicators such as sensitivity, specificity, F1 score, and AUC to evaluate the performance of classification models, and indexes such as R^2, MSE, RMSE, and MAE to evaluate the performance of regression models	91
	26	It is recommended to explore the interpretability of the model to facilitate the clinical transformation of complex machine learning models. The recommended model interpretation methods include Feature Importance, LIME, and Shapley	91
5. Clinical application of intensive care big data	27	It is recommended to transform and promote early warning tools that meet critical needs	91
	28	It is recommended to use the information system for intensive care as a carrier to access real-time data and output recommendations for decision making	91
	29	It is suggested that the current practice of intensive care diagnosis and treatment should still be led by clinicians with the use of big data technology to coordinate to improve medical efficiency and ensure medical quality and safety	98

scenarios, and ethics. Therefore, this consensus believes that because of the current developmental level of big data applications, it is advisable to be guided by existing evidence and clinical experience to assist the diagnosis and treatment, and improve the quality and efficiency of medical care with the help of big data technology.

Big data models produce seemingly accurate results through complex computations, but often fail to provide end users with the logic behind them (154). AI is weak in determining causality, at least its interpretability does not meet current clinical needs. Models developed based on intensive care big data are often more accurate in predictions when validating data from the same population, but the results may be unreliable when tested in external populations (155). In clinical practice, the diagnosis and treatment process is often highly subjective, especially for patients with critically complex symptoms, and their plans for diagnosis and treatment also have large individual heterogeneity, resulting in low reliability of the ICU model (156). In summary, most of the current research is still in the exploited phase and lacks effective external validation (157). Therefore, unnecessary interventions or changes in treatment strategies that are not supported by scientific evidence may lead to medical safety issues such as overmedication or treatment failure.

When these algorithms are developed into intelligent assistance systems deployed as alerting tools, they must be concise and accurate enough to prevent alert fatigue and thus avoid delays in clinical decision-making (158, 159). Considering the scientific preciseness, the maturity and stability of AI-driven models are less convincing for clinical practice to a certain extent, and indiscriminate development and use of data models may lead to overdiagnosis and waste of resources. In addition, the clinical application of intensive care big data also faces ethical issues. At present, the hidden dangers of big data applications in terms of patient privacy and safety responsibility cannot be ignored. First of all, the establishment of the database will inevitably involve data of patient privacy, and protecting patient privacy has become a problem that must be solved in the development of intensive care big data. It is not appropriate to develop a medical database at full speed without guaranteeing privacy and security. Secondly, in terms of application security, in the process of big data-assisted clinical diagnosis and treatment practice, if a medical safety accident occurs, computer algorithms cannot be responsible for clinical decision-making with the current developmental level of ethics and AI. In order to avoid mistakes and abuses in the big data system for diagnosis and treatment, the clinician must act as the person in charge of clinical decision-making to "be responsible for" big data applications.

Discussion

With the increase of computing power and data scale, the emergence of large models has enabled AI systems to handle more complex and

massive tasks, improving the model's performance and generalization ability, which also brings new opportunities for critical big data applications. As a "double-edged sword," the application of big data science in intensive care has pros and cons. This consensus reach a consensus on five parts: conception, important scientific issues, standards and principles of database, methodology in solving big data problems, clinical application and safety consideration of intensive care big data. All recommendations has been summarized in Table 3. Actually, this is the starting step of application big data in the field of intensive care. In order to ensure data security and ensure the professionalism of the model, the medical industry needs a medical vertical domain big language model based on professional mapping knowledge domain and high-quality data. More explorations and big data based retrospective research should be carried out in order to enhance safety and reliability of big data based models of critical care field.

Author contributions

LS and SL: write and translate the Chinese version of this consensus and overall planning of the discussion and finalize the draft. YLo: organize and supervise the writing of this article. CC, KC, MC, YCheng, YCui, QD, TG, XG, HH, JH, CH, RH, HJ, JJ, YLa, JuL, LuL, JiL, XL, ZM, HM, YuS, QS, WS, ZT, HaW, LuW, SW, ZhaW, ZhoW, DW, QW, JYu, YY, HY, HZ, MZ, CZ, RZ, and LeZ: literature review and write the origin version of the Chinese version of this consensus. YChen, RD, MD, BH, XH, LiL, WL, YLi, FL, XiaS, XiuS, YoS, YiS, BW, HoW, LiW, JW, XX, JYa, XY, WY, QZ, LiZ, ZZ, FZ, and WZ: supervise and provide consulting of manuscript. All authors contributed to the article and approved the submitted version.

Funding

The author(s) declare financial support was received for the research, authorship, and/or publication of this article. National High-Level Hospital Clinical Research Fouding (2022-PUMCH-B-115, 2022-PUMCH-D-005); CAMS Innovation Fund for Medical Sciences (2021-I2M-1-056); Beijing Municipal Natural Science Foundation (7222134); Capital's Funds for Health Improvement and Research (2020-2-40111); Excellence Program of Key Clinical Specialty of Beijing in 2020 for Critical Care Medicine (ZK128001); Beijing Municipal Science and Technology Commission (Z201100005520051); 5G Network Construction of Remote Intensive care and Application of Critical Infection Standard Diagnosis and Treatment Cloud Platform.

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.2023.1174429/full#supplementary-material

References

1. Celi LA, Mark RG, Stone DJ, Montgomery RA. Big data in the intensive care unit. Closing the data loop. *Am J Respir Crit Care Med*. (2013) 187:1157–60. doi: 10.1164/rccm.201212-2311ED

2. Buchman TG, Billiar TR, Elster E, Kirk AD, Rimawi RH, Vodovotz Y, et al. Precision medicine for critical illness and injury. *Crit Care Med*. (2016) 44:1635–8. doi: 10.1097/CCM.0000000000002028

3. Jung Y, Hu J. A K-fold averaging cross-validation procedure. *J Nonparamet Stat*. (2015) 27:167–79. doi: 10.1080/10485252.2015.1010532

4. Pirracchio R, Cohen MJ, Malenica I, Cohen J, Chambaz A, Cannesson M, et al. Big data and targeted machine learning in action to assist medical decision in the ICU. *Anaesth Crit Care Pain Med*. (2019) 38:377–84. doi: 10.1016/j.accpm.2018.09.008

5. 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

6. Rumsfeld JS, Joynt KE, Maddox TM. Big data analytics to improve cardiovascular care: promise and challenges. *Nat Rev Cardiol*. (2016) 13:350–9. doi: 10.1038/nrcardio.2016.42

7. Yang S, Stansbury LG, Rock P, Scalea T, Hu PF. Linking big data and prediction strategies: tools, pitfalls, and lessons learned. *Crit Care Med*. (2019) 47:840–8. doi: 10.1097/CCM.0000000000003739

8. Carra G, Salluh JIF, da Silva Ramos FJ, Meyfroidt G. Data-driven ICU management: using big data and algorithms to improve outcomes. *J Crit Care*. (2020) 60:300–4. doi: 10.1016/j.jcrc.2020.09.002

9. Le Roux P, Menon DK, Citerio G, Vespa P, Bader MK, Brophy GM. Consensus summary statement of the international multidisciplinary consensus conference on multimodality monitoring in Neurocritical care: a statement for healthcare professionals from the Neurocritical care society and the European Society of Intensive Care Medicine. *Neurocrit Care*. (2014) 21:S1–S26. doi: 10.1007/s12028-014-0041-5

10. Citerio G, Park S, Schmidt JM, Moberg R, Suarez JI, Le Roux PD. Data collection and interpretation. *Neurocrit Care*. (2015) 22:360–8. doi: 10.1007/s12028-015-0139-4

11. Schmidt JM, De Georgia M. Multimodality monitoring: informatics, integration data display and analysis. *Neurocrit Care*. (2014) 21:S229–38. doi: 10.1007/s12028-014-0037-1

12. Docherty A. B and N I lone, exploiting big data for critical care research. *Curr Opin Crit Care*. (2015) 21:467–72. doi: 10.1097/MCC.0000000000000228

13. Zampieri FG, Soares M, Borges LP, Salluh JIF, Ranzani OT. The Epimed monitor ICU database®: a cloud-based national registry for adult intensive care unit patients in Brazil. *Revista Brasileira de Terapia Intensiva*. (2017) 29:418–26. doi: 10.5935/0103-507X.20170062

14. Komorowski M, Celi LA, Badawi O, Gordon AC, Faisal AA. The artificial intelligence clinician learns optimal treatment strategies for sepsis in intensive care. *Nat Med*. (2018) 24:1716–20. doi: 10.1038/s41591-018-0213-5

15. 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

16. Johnson ABL, Pollard T, Horng S, Celi LA, Mark R. MIMIC-IV (version 2.0). *PhysioNet*. (2022)

17. 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

18. Hyland SL, Faltys M, Huser M, Lyu X, Gumbsch T, Esteban C. 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

19. Thoral PJ, Peppink JM, Driessen RH, Sijbrands EJG, Kompanje EJO, Kaplan L, et al. Sharing ICU patient data responsibly under the Society of Critical Care Medicine/European Society of Intensive Care Medicine Joint Data Science Collaboration: the Amsterdam university medical centers database (AmsterdamUMCdb) example. *Crit Care Med*. (2021) 49:e563–77. doi: 10.1097/ccm.0000000000004916

20. Qi S, Mao Z, Hu X, Liu C, Kang H, Zhou F. Introduction of critical care database based on specialized information systems: a model of critical care medicine database in large level III grade a hospital. *Zhonghua Wei Zhong Bing Ji Jiu Yi Xue*. (2020) 32:743–9. doi: 10.3760/cma.j.cn121430-20200520-00393

21. 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

22. Zeng X, Yu G, Lu Y, Tan L, Wu X, Shi S, et al. PIC, a paediatric-specific intensive care database. *Sci Data*. (2020) 7:14. doi: 10.1038/s41597-020-0355-4

23. Zhang Z, Cao L, Chen R, Zhao Y, Lv L, Xu Z, et al. Electronic healthcare records and external outcome data for hospitalized patients with heart failure. *Sci Data*. (2021) 8:46. doi: 10.1038/s41597-021-00835-9

24. Xu J, Zhou Y, Xia X. Construction and application of big data platform for intensive care medicine. *Chin J Emerg Med*. 31:129–132. doi: 10.3760/cma.j.issn.1671-0282.2022.01.028

25. Morrison JL, Cai Q, Davis N, Yan Y, Berbaum ML, Ries M, et al. Clinical and economic outcomes of the electronic intensive care unit: results from two community hospitals. *Crit Care Med*. (2010) 38:2–8. doi: 10.1097/CCM.0b013e3181b78fa8

26. Elbers PW, Girbes A, Malbrain ML, Bosman R. Right dose, right now: using big data to optimize antibiotic dosing in the critically ill. *Anaesth Intens Ther*. (2015) 47:457–63. doi: 10.5603/AIT.a2015.0061

27. Kindle RD, Badawi O, Celi LA, Sturland S. Intensive care unit telemedicine in the era of big data, artificial intelligence, and computer clinical decision support systems. *Crit Care Clin*. (2019) 35:483–95. doi: 10.1016/j.ccc.2019.02.005

28. Noshad M, Rose CC, Chen JH. Signal from the noise: a mixed graphical and quantitative process mining approach to evaluate care pathways applied to emergency stroke care. *J Biomed Inform*. (2022) 127:104004. doi: 10.1016/j.jbi.2022.104004

29. Boss JM, Narula G, Straessle C, Willms J, Azzati J, Brodbeck D, et al. ICU cockpit: a platform for collecting multimodal waveform data, AI-based computational disease modeling and real-time decision support in the intensive care unit. *J Am Med Inform Assoc*. (2022) 29:1286–91. doi: 10.1093/jamia/ocac064

30. Taglang G, Jackson DB. Use of "big data" in drug discovery and clinical trials. *Gynecol Oncol*. (2016) 141:17–23. doi: 10.1016/j.ygyno.2016.02.022

31. Zhu Y, Yin H, Zhang R, Ye X, Wei J. The effect of dobutamine vs milrinone in sepsis: a big data, real-world study. *Int J Clin Pract*. (2021) 75:e14689. doi: 10.1111/ijcp.14689

32. Huang X, Shan S, Khan YA, Salem S, Mohamed A, Attia EA. Risk assessment of ICU patients through deep learning technique: a big data approach. *J Glob Health*. (2022) 12:04044. doi: 10.7189/jogh.12.04044

33. Xia Y, Wang X, Wu W, Shi H. Rehabilitation of Sepsis patients with acute kidney injury based on intelligent medical big data. *J Healthc Eng*. (2022) 2022:8414135–10. doi: 10.1155/2022/8414135

34. Vergetis V, Skaltsas D, Gorgoulis VG, Tsirigos A. Assessing drug development risk using big data and machine learning. *Cancer Res*. (2021) 81:816–9. doi: 10.1158/0008-5472.Can-20-0866

35. 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. doi: 10.1007/s00134-019-05872-y

36. Delahanty RJ, Alvarez J, Flynn LM, Sherwin RL, Jones SS. Development and evaluation of a machine learning model for the early identification of patients at risk for Sepsis. *Ann Emerg Med*. (2019) 73:334–44. doi: 10.1016/j.annemergmed.2018.11.036

37. Goh KH, Wang L, Yeow AYK, Poh H, Li K, Yeow JJL, et al. Artificial intelligence in sepsis early prediction and diagnosis using unstructured data in healthcare. *Nat Commun*. (2021) 12:711. doi: 10.1038/s41467-021-20910-4

38. Wardi G, Carlile M, Holder A, Shashikumar S, Hayden SR, Nemati S. Predicting progression to septic shock in the emergency department using an externally generalizable machine-learning algorithm. *Ann Emerg Med*. (2021) 77:395–406. doi: 10.1016/j.annemergmed.2020.11.007

39. Ozrazgat-Baslanti T, Loftus TJ, Ren Y, Ruppert MM, Bihorac A. Advances in artificial intelligence and deep learning systems in ICU-related acute kidney injury. *Curr Opin Crit Care*. (2021) 27:560–72. doi: 10.1097/MCC.0000000000000887

40. Xiao Z, Huang Q, Yang Y, Liu M, Chen Q, Huang J, et al. Emerging early diagnostic methods for acute kidney injury. *Theranostics*. (2022) 12:2963–86. doi: 10.7150/thno.71064

41. Churpek MM, Carey KA, Edelson DP, Singh T, Astor BC, Gilbert ER, et al. Internal and external validation of a machine learning risk score for acute kidney injury. *JAMA Netw Open*. (2020) 3:e2012892. doi: 10.1001/jamanetworkopen.2020.12892

42. Kate RJ, Pearce N, Mazumdar D, Nilakantan V. A continual prediction model for inpatient acute kidney injury. *Comput Biol Med*. (2020) 116:103580. doi: 10.1016/j.compbiomed.2019.103580

43. Le S, Pellegrini E, Green-Saxena A, Summers C, Hoffman J, Calvert J, et al. Supervised machine learning for the early prediction of acute respiratory distress syndrome (ARDS). *J Crit Care*. (2020) 60:96–102. doi: 10.1016/j.jcrc.2020.07.019

44. Mayampurath A, Churpek MM, Su X, Shah S, Munroe E, Patel B, et al. External validation of an acute respiratory distress syndrome prediction model using radiology reports. *Crit Care Med*. (2020) 48:e791–8. doi: 10.1097/CCM.0000000000004468

45. Jabbour S, Fouhey D, Kazerooni E, Wiens J, Sjoding MW. Combining chest X-rays and electronic health record (EHR) data using machine learning to diagnose acute respiratory failure. *J Am Med Informat Assoc*. (2022) 29:1060–8. doi: 10.1093/jamia/ocac030

46. Linnen DT, Escobar GJ, Hu X, Scruth E, Liu V, Stephens C. Statistical modeling and aggregate-weighted scoring Systems in Prediction of mortality and ICU transfer: a systematic review. *J Hosp Med*. (2019) 14:161–9. doi: 10.12788/jhm.3151

47. 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

48. Broch Porcar MJ, Cubillo BR, Domínguez-Roldán JM, Rocha LÁ, Sanz MÁB, Montes MC, et al. Practical document on the management of hyponatremia in critically ill patients. *Med Intensiva (Engl Ed)*. (2019) 43:302–16. doi: 10.1016/j.medin.2018.12.002

49. Levi R, Carli F, Arevalo AR, Altinel Y, Stein DJ, Naldini MM, et al. Artificial intelligence-based prediction of transfusion in the intensive care unit in patients with gastrointestinal bleeding. BMJ health care. *Inform*. (2021) 28:e100245. doi: 10.1136/bmjhci-2020-100245

50. Ryan L, Mataraso S, Siefkas A, Pellegrini E, Barnes G, Green-Saxena A, et al. A machine learning approach to predict deep venous thrombosis among hospitalized patients. *Clin Appl Thromb Hemost*. (2021) 27:1076029621991185. doi: 10.1177/1076029621991185

51. Yuan KC, Tsai LW, Lee KH, Cheng YW, Hsu SC, Lo YS, et al. The development an artificial intelligence algorithm for early sepsis diagnosis in the intensive care unit. *Int J Med Inform*. (2020) 141:104176. doi: 10.1016/j.ijmedinf.2020.104176

52. Burdick H, Pino E, Gabel-Comeau D, McCoy A, Gu C, Roberts J, et al. Effect of a sepsis prediction algorithm on patient mortality, length of stay and readmission: a prospective multicentre clinical outcomes evaluation of real-world patient data from US hospitals. *BMJ Health Care Inform*. (2020) 27:e100109. doi: 10.1136/bmjhci-2019-100109

53. Bose SN, Greenstein JL, Fackler JC, Sarma SV, Winslow RL, Bembea MM. Early prediction of multiple organ dysfunction in the pediatric intensive care unit. *Front Pediatr*. (2021) 9:711104. doi: 10.3389/fped.2021.711104

54. Bedoya AD, Clement ME, Phelan M, Steorts RC, O'Brien C, Goldstein BA. Minimal impact of implemented early warning score and best practice alert for patient deterioration. *Crit Care Med*. (2019) 47:49–55. doi: 10.1097/CCM.0000000000003439

55. Ardila D, Kiraly AP, Bharadwaj S, Choi B, Reicher JJ, Peng L, et al. End-to-end lung cancer screening with three-dimensional deep learning on low-dose chest computed tomography. *Nat Med*. (2019) 25:954–61. doi: 10.1038/s41591-019-0447-x

56. Qin ZZ, Sander MS, Rai B, Titahong CN, Sudrungrot S, Laah SN, et al. Using artificial intelligence to read chest radiographs for tuberculosis detection: a multi-site evaluation of the diagnostic accuracy of three deep learning systems. *Sci Rep*. (2019) 9:15000. doi: 10.1038/s41598-019-51503-3

57. Reamaroon N, Sjoding MW, Gryak J, Athey BD, Najarian K, Derksen H. Automated detection of acute respiratory distress syndrome from chest X-rays using directionality measure and deep learning features. *Comput Biol Med*. (2021) 134:104463. doi: 10.1016/j.compbiomed.2021.104463

58. Rueckel J, Kunz WG, Hoppe BF, Patzig M, Notohamiprodjo M, Meinel FG, et al. Artificial intelligence algorithm detecting lung infection in supine chest radiographs of critically ill patients with a diagnostic accuracy similar to board-certified radiologists. *Crit Care Med*. (2020) 48:e574–83. doi: 10.1097/CCM.0000000000004397

59. Zhang K, Liu X, Shen J, Li Z, Sang Y, Wu X, et al. Clinically applicable AI system for accurate diagnosis, quantitative measurements, and prognosis of COVID-19 pneumonia using computed tomography. *Cells*. (2020) 181:1423–1433.e11. doi: 10.1016/j.cell.2020.04.045

60. Farzaneh N, Williamson CA, Jiang C, Srinivasan A, Bapuraj JR, Gryak J, et al. Automated segmentation and severity analysis of subdural hematoma for patients with traumatic brain injuries. *Diagnostics (Basel, Switzerland)*. (2020) 10:773. doi: 10.3390/diagnostics10100773

61. van Sloun RJG, Demi L. Localizing B-lines in lung ultrasonography by weakly supervised deep learning, *in-vivo* results. *IEEE J Biomed Health Inform*. (2020) 24:957–64. doi: 10.1109/JBHI.2019.2936151

62. Lv Y, Huang Z. Account of deep learning-based ultrasonic image feature in the diagnosis of severe Sepsis complicated with acute kidney injury. *Comput Math Methods Med*. (2022) 2022:8158634–9. doi: 10.1155/2022/8158634

63. Ying F, Chen S, Pan G, He Z. Artificial intelligence pulse coupled neural network algorithm in the diagnosis and treatment of severe Sepsis complicated with acute kidney injury under ultrasound image. *J Healthc Eng*. (2021) 2021:1–8. doi: 10.1155/2021/6761364

64. Strodthoff N, Strodthoff C, Becher T, Weiler N, Frerichs I. Inferring respiratory and circulatory parameters from electrical impedance tomography with deep recurrent models. *IEEE J Biomed Health Inform*. (2021) 25:3105–11. doi: 10.1109/JBHI.2021.3059016

65. Bos LD, Schouten LR, van Vught LA, Wiewel MA, Ong DSY, Cremer O, et al. Identification and validation of distinct biological phenotypes in patients with acute respiratory distress syndrome by cluster analysis. *Thorax*. (2017) 72:876–83. doi: 10.1136/thoraxjnl-2016-209719

66. Chaudhary K, Vaid A, Duffy Á, 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

67. Bhavani SV, Carey KA, Gilbert ER, Afshar M, Verhoef PA, Churpek MM. Identifying novel Sepsis subphenotypes using temperature trajectories. *Am J Respir Crit Care Med*. (2019) 200:327–35. doi: 10.1164/rccm.201806-1197OC

68. Sinha P, Churpek MM, Calfee CS. Machine learning classifier models can identify acute respiratory distress syndrome phenotypes using readily available clinical data. *Am J Respir Crit Care Med*. (2020) 202:996–1004. doi: 10.1164/rccm.202002-0347OC

69. Kudo D, Goto T, Uchimido R, Hayakawa M, Yamakawa K, Abe T, et al. Coagulation phenotypes in sepsis and effects of recombinant human thrombomodulin: an analysis of three multicentre observational studies. *Crit Care*. (2021) 25:114. doi: 10.1186/s13054-021-03541-5

70. Seymour CW, Kennedy JN, Wang S, Chang CCH, 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

71. 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

72. Scicluna BP, van Vught LA, Zwinderman AH, Wiewel MA, Davenport EE, Burnham KL, et al. Classification of patients with sepsis according to blood genomic endotype: a prospective cohort study. *Lancet Respir Med*. (2017) 5:816–26. doi: 10.1016/s2213-2600(17)30294-1

73. Calfee CS, Delucchi K, Parsons PE, Thompson BT, Ware LB, Matthay MA, et al. Subphenotypes in acute respiratory distress syndrome: latent class analysis of data from two randomised controlled trials. *Lancet Respir Med*. (2014) 2:611–20. doi: 10.1016/S2213-2600(14)70097-9

74. Famous KR, Delucchi K, Ware LB, Kangelaris KN, Liu KD, Thompson BT, et al. Acute respiratory distress syndrome subphenotypes respond differently to randomized fluid management strategy. *Am J Respir Crit Care Med*. (2017) 195:331–8. doi: 10.1164/rccm.201603-0645OC

75. Calfee CS, Delucchi KL, Sinha P, Matthay MA, Hackett J, Shankar-Hari M, et al. Acute respiratory distress syndrome subphenotypes and differential response to simvastatin: secondary analysis of a randomised controlled trial. *Lancet Respir Med*. (2018) 6:691–8. doi: 10.1016/s2213-2600(18)30177-2

76. 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

77. Parreco J, Hidalgo A, Parks JJ, Kozol R, Rattan R. Using artificial intelligence to predict prolonged mechanical ventilation and tracheostomy placement. *J Surg Res*. (2018) 228:179–87. doi: 10.1016/j.jss.2018.03.028

78. Fabregat A, Magret M, Ferré JA, Vernet A, Guasch N, Rodríguez A, et al. A machine learning decision-making tool for extubation in intensive care unit patients. *Comput Methods Prog Biomed*. (2021) 200:105869. doi: 10.1016/j.cmpb.2020.105869

79. Hur S, Min JY, Yoo J, Kim K, Chung CR, Dykes PC, et al. Development and validation of unplanned Extubation prediction models using intensive care unit data: retrospective, comparative, machine learning study. *J Med Internet Res*. (2021) 23:e23508. doi: 10.2196/23508

80. Chen H, Ma Y, Hong N, Wang H, Su L, Liu C, et al. Early warning of citric acid overdose and timely adjustment of regional citrate anticoagulation based on machine learning methods. *BMC Med Inform Decis Mak*. (2021) 21:126. doi: 10.1186/s12911-021-01489-8

81. Su L, Liu C, Li D, He J, Zheng F, Jiang H, et al. Toward optimal heparin dosing by comparing multiple machine learning methods: retrospective study. *JMIR Med Inform*. (2020) 8:e17648. doi: 10.2196/17648

82. Li D, Gao J, Hong N, Wang H, Su L, Liu C, et al. A clinical prediction model to predict heparin treatment outcomes and provide dosage recommendations: development and validation study. *J Med Internet Res*. (2021) 23:e27118. doi: 10.2196/27118

83. Maviglia R, Michi T, Passaro D, Raggi V, Bocci MG, Piervincenzi E, et al. Machine learning and antibiotic management. *Antibiotics (Basel)*. (2022) 11:304. doi: 10.3390/antibiotics11030304

84. Wang Y, Zhang H, Fan Y, Ying P, Li J, Xie C, et al. Propofol anesthesia depth monitoring based on self-attention and residual structure convolutional neural network. *Comput Math Methods Med*. (2022) 2022:8501948–13. doi: 10.1155/2022/8501948

85. van de Sande D, van Genderen ME, Huiskens J, et al. Moving from bytes to bedside: a systematic review on the use of artificial intelligence in the intensive care unit[J]. *Intensive Care Med*. (2021) 47:750–760. doi: 10.1007/s00134-021-06446-7

86. Barchitta M, Maugeri A, Favara G, Riela P, Gallo G, Mura I, et al. Early prediction of seven-day mortality in intensive care unit using a machine learning model: results from the SPIN-UTI project. *J Clin Med*. (2021) 10:992. doi: 10.3390/jcm10050992

87. Li K, Shi Q, Liu S, Xie Y, Liu J. Predicting in-hospital mortality in ICU patients with sepsis using gradient boosting decision tree. *Medicine*. (2021) 100:e25813. doi: 10.1097/md.0000000000025813

88. Garcia-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 (Engl Ed)*. (2020) 44:160–70. doi: 10.1016/j.medin.2018.07.016

89. Nie X, Cai Y, Liu J, Liu X, Zhao J, Yang Z, et al. Mortality prediction in cerebral hemorrhage patients using machine learning algorithms in intensive care units. *Front Neurol*. (2020) 11:610531. doi: 10.3389/fneur.2020.610531

90. Halonen K, Leppäniemi A, Lundin J, Puolakkainen PA, Kemppainen EA, Haapiainen RK. Predicting fatal outcome in the early phase of severe acute pancreatitis by using novel prognostic models. *Pancreatology*. (2003) 3:309–15. doi: 10.1159/000071769

91. Ding N, Guo C, Li C, Zhou Y, Chai X. An artificial neural networks model for early predicting in-hospital mortality in acute pancreatitis in MIMIC-III. *Biomed Res Int*. (2021) 2021:6638919–8. doi: 10.1155/2021/6638919

92. Niven DJ, McCormick TJ, Straus SE, Hemmelgarn BR, Jeffs L, Barnes TRM, et al. Reproducibility of clinical research in critical care: a scoping review. *BMC Med*. (2018) 16:26. doi: 10.1186/s12916-018-1018-6

93. Weiss CH. Why do we fail to deliver evidence-based practice in critical care medicine? *Curr Opin Crit Care*. (2017) 23:400–5. doi: 10.1097/MCC.0000000000000436

94. Rosa RG, Teixeira C, Sjoding M. Novel approaches to facilitate the implementation of guidelines in the ICU. *J Crit Care*. (2020) 60:1–5. doi: 10.1016/j.jcrc.2020.07.014

95. Liu VX, Morehouse JW, Marelich GP, Soule J, Russell T, Skeath M, et al. Multicenter implementation of a treatment bundle for patients with Sepsis and intermediate lactate values. *Am J Respir Crit Care Med*. (2016) 193:1264–70. doi: 10.1164/rccm.201507-1489OC

96. Eslami S, Abu-Hanna A, Schultz MJ, de Jonge E, de Keizer NF. Evaluation of consulting and critiquing decision support systems: effect on adherence to a lower tidal volume mechanical ventilation strategy. *J Crit Care*. (2012) 27:425.e1. doi: 10.1016/j.jcrc.2011.07.082

97. Trogrlic Z, van der Jagt M, Lingsma H, Gommers D, Ponssen HH, Schoonderbeek JFJ, et al. Improved guideline adherence and reduced brain dysfunction after a multicenter multifaceted implementation of ICU delirium guidelines in 3,930 patients. *Crit Care Med*. (2019) 47:419–27. doi: 10.1097/CCM.0000000000003596

98. Bourdeaux C, Ghosh E, Atallah L, Palanisamy K, Patel P, Thomas M, et al. Impact of a computerized decision support tool deployed in two intensive care units on acute kidney injury progression and guideline compliance: a prospective observational study. *Crit Care*. (2020) 24:656. doi: 10.1186/s13054-020-03343-1

99. Bone RC, Balk RA, Cerra FB. American College of Chest Physicians/Society of Critical Care Medicine consensus conference: definitions for sepsis and organ failure and guidelines for the use of innovative therapies in sepsis. *Crit Care Med*. (1992) 20:864–74. doi: 10.1097/00003246-199206000-00025

100. Levy MM, Fink MP, Marshall JC, Abraham E, Angus D, Cook D, et al. SCCM/ESICM/ACCP/ATS/SIS international Sepsis definitions conference. *Crit Care Med*. (2001) 31:1250–6. doi: 10.1097/01.Ccm.0000050454.01978.3b

101. 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

102. (中华人民共和国卫生部, 电子病历基本规范(试行)(basic specifications for electronic medical records (trial implementation). (2010).

103. 北京市市场监督管理局 重症医学数据集·患者数据*(Intensive Care Medical Dataset Patient Data) DB11/T 1866-2021*. 2021.

104. 全国信息技术标准化技术委员会, 工业物联网—数据采集结构化描述规范 *(National Information Technology Standardization Technical Committee, industrial internet of things - structured description specification for data collection)*. (2020).

105. 全国通信标准化技术委员会, 智慧城市 数据融合 第3部分:数据采集规范 *(National Communications Standardization Technical Committee, Smart City data fusion part 3: Data collection specification)*. (2022).

106. 全国信息技术标准化技术委员会大数据标准工作组, 大数据标准化白皮书(2020版)(big data standards working Group of the National Information Technology Standardization Technical Committee, *White Paper on Big Data Standardization*) (2020).

107. 朱志勇 李, 大数据技术在医疗急重症领域的应用(application of big data technology in the field of medical emergency and critical care). 邮电设计技术 (2016).

108. 家市场监督管理总局与中国国家标准化委员会, 信息技术—数据质量评价指标*GB/T36344-2018 ICS 35.24.01. (state Administration for Market Regulation and China National Standardization Commission, information technology - data quality evaluation indicators)*. (2018).

109. 美国国家标准技术研究院, *NIST*大数据互操作框架:第4册 安全与隐私 *NIST big data interoperability framework: Volume 4, Security and Privacy*. (2013).

110. 中国人民共和国国家质量监督检验检疫总局与中国国家标准化管理委员会. 信息安全技术—大数据服务安全能力要求(information security technology - security capability requirements for big data services). *GB/T 35274-2017*. (2016)

111. 国家市场监督管理总局与国家标准化管理委员会, 信息安全技术—政务信息共享数据信息安全技术要求*(requirements for government information sharing data information security technology). GB/T 39477-2020*. (2020).

112. Rieke N, Hancox J, Li W, Milletarì F, Roth HR, Albarqouni S, et al. The future of digital health with federated learning. *NPJ Digit Med*. (2020) 3:119. doi: 10.1038/s41746-020-00323-1

113. Mahmoudi E, Kamdar N, Kim N, Gonzales G, Singh K, Waljee AK. Use of electronic medical records in development and validation of risk prediction models of hospital readmission: systematic review. *BMJ*. (2020) 369:m958. doi: 10.1136/bmj.m958

114. Liang H, Tsui BY, Ni H, Valentim CCS, Baxter SL, Liu G, et al. Evaluation and accurate diagnoses of pediatric diseases using artificial intelligence. *Nat Med*. (2019) 25:433–8. doi: 10.1038/s41591-018-0335-9

115. Rusin CG, Acosta SI, Vu EL, Ahmed M, Brady KM, Penny DJ. Automated prediction of cardiorespiratory deterioration in patients with single ventricle. *J Am Coll Cardiol*. (2021) 77:3184–92. doi: 10.1016/j.jacc.2021.04.072

116. Hamilton PS, Tompkins WJ. Quantitative investigation of QRS detection rules using the MIT/BIH arrhythmia database. *IEEE Trans Biomed Eng*. (1986) BME-33:1157–65. doi: 10.1109/tbme.1986.325695

117. Narula G, Haeberlin M, Balsiger J, Strässle C, Imbach LL, Keller E. Detection of EEG burst-suppression in neurocritical care patients using an unsupervised machine learning algorithm. *Clin Neurophysiol*. (2021) 132:2485–92. doi: 10.1016/j.clinph.2021.07.018

118. Maddali MV, Churpek M, Pham T, Rezoagli E, Zhuo H, Zhao W, et al. Validation and utility of ARDS subphenotypes identified by machine-learning models using clinical data: an observational, multicohort, retrospective analysis. Lancet. *Respir Med*. (2022) 10:367–77. doi: 10.1016/s2213-2600(21)00461-6

119. Walsh SLF, Calandriello L, Silva M, Sverzellati N. Deep learning for classifying fibrotic lung disease on high-resolution computed tomography: a case-cohort study. The lancet. *Respir Med*. (2018) 6:837–45. doi: 10.1016/s2213-2600(18)30286-8

120. Niazi MKK, Parwani AV, Gurcan MN. Digital pathology and artificial intelligence. *Lancet Oncol*. (2019) 20:e253–61. doi: 10.1016/s1470-2045(19)30154-8

121. Massion PP, Antic S, Ather S, Arteta C, Brabec J, Chen H, et al. Assessing the accuracy of a deep learning method to risk stratify indeterminate pulmonary nodules. *Am J Respir Crit Care Med*. (2020) 202:241–9. doi: 10.1164/rccm.201903-0505OC

122. Cheng N, Ren Y, Zhou J, Zhang Y, Wang D, Zhang X, et al. Deep learning-based classification of hepatocellular nodular lesions on whole-slide histopathologic images. *Gastroenterology*. (2022) 162:1948–1961.e7. doi: 10.1053/j.gastro.2022.02.025

123. Walsh SLF, Humphries SM, Wells AU, Brown KK. Imaging research in fibrotic lung disease; applying deep learning to unsolved problems. *Lancet Respir Med*. (2020) 8:1144–53. doi: 10.1016/S2213-2600(20)30003-5

124. Winslow CJ, Edelson DP, Churpek MM, Taneja M, Shah NS, Datta A, et al. The impact of a machine learning early warning score on hospital mortality: a multicenter clinical intervention trial. *Crit Care Med*. (2022) 50:1339–47. doi: 10.1097/CCM.0000000000005492

125. Yuan S, Sun Y, Xiao X, Long Y, He H. Using machine learning algorithms to predict Candidaemia in ICU patients with new-onset systemic inflammatory response syndrome. *Front Med*. (2021) 8:720926. doi: 10.3389/fmed.2021.720926

126. Dastider AG, Sadik F, Fattah SA. An integrated autoencoder-based hybrid CNN-LSTM model for COVID-19 severity prediction from lung ultrasound. *Comput Biol Med*. (2021) 132:104296. doi: 10.1016/j.compbiomed.2021.104296

127. Tiwari P, Colborn KL, Smith DE, Xing F, Ghosh D, Rosenberg MA. Assessment of a machine learning model applied to harmonized electronic health record data for the prediction of incident atrial fibrillation. *JAMA Netw Open*. (2020) 3:e1919396. doi: 10.1001/jamanetworkopen.2019.19396

128. Papp L, Spielvogel CP, Grubmüller B, Grahovac M, Krajnc D, Ecsedi B, et al. Supervised machine learning enables non-invasive lesion characterization in primary prostate cancer with [68Ga]Ga-PSMA-11 PET/MRI. *Eur J Nucl Med Mol Imaging*. (2021) 48:1795–805. doi: 10.1007/s00259-020-05140-y

129. Liang W, Yao J, Chen A, Lv Q, Zanin M, Liu J, et al. Early triage of critically ill COVID-19 patients using deep learning. *Nat Commun*. (2020) 11:1–7. doi: 10.1038/s41467-020-17280-8

130. 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) 8:664966. doi: 10.3389/fmed.2021.664966

131. Su L, Zhang Z, Zheng F, Pan P, Hong N, Liu C, et al. Five novel clinical phenotypes for critically ill patients with mechanical ventilation in intensive care units: a retrospective and multi database study. *Respir Res*. (2020) 21:325. doi: 10.1186/s12931-020-01588-6

132. Liu S, Su L, Liu X, Zhang X, Chen Z, Liu C, et al. Recognizing blood pressure patterns in sedated critically ill patients on mechanical ventilation by spectral clustering. *Ann Transl Med*. (2021) 9:1404. doi: 10.21037/atm-21-2806

133. Landi I, Glicksberg BS, Lee H-C, Cherng S, Landi G, Danieletto M, et al. Deep representation learning of electronic health records to unlock patient stratification at scale. *NPJ Digit Med*. (2020) 3:96. doi: 10.1038/s41746-020-0301-z

134. Hyun S, Kaewprag P, Cooper C, Hixon B, Moffatt-Bruce S. Exploration of critical care data by using unsupervised machine learning. *Comput Methods Prog Biomed*. (2020) 194:105507. doi: 10.1016/j.cmpb.2020.105507

135. Greco M, Caruso PF, Cecconi M. Artificial intelligence in the intensive care unit. *Semin Respir Crit Care Med*. (2021) 42:002–9. doi: 10.1055/s-0040-1719037

136. Richens JG, Lee CM, Johri S. Improving the accuracy of medical diagnosis with causal machine learning. *Nat Commun*. (2020) 11:1–9. doi: 10.1038/s41467-020-17419-7

137. Wei T, Feng F, Chen J, Wu Z, Yi J, He X. Model-agnostic counterfactual reasoning for eliminating popularity bias in recommender system In *Proceedings of the 27th ACM SIGKDD conference on Knowledge Discovery & Data Mining* (2021)

138. Goudet O, Kalainathan D, Caillou P, Guyon I, Lopez-Paz D, Sebag M. Learning functional causal models with generative neural networks In *Explainable and interpretable models in computer vision and machine learning*: Springer (2018). 39–80.

139. Athey S, Tibshirani J, Wager S. Generalized random forests. *Ann Stat*. (2019) 47:1148–78. doi: 10.1214/18-AOS1709

140. Tan J, Xu S, Ge Y, Li Y, Chen X, Zhang Y. Counterfactual explainable recommendation In *Proceedings of the 30th ACM international conference on Information & Knowledge Management* (2021)

141. Tharwat A. Classification assessment methods. *Appl Comput Informat*. (2021) 17:168–92. doi: 10.1016/j.aci.2018.08.003

142. Hong ZZLPY. Predictive analytics with ensemble modeling in laparoscopic surgery: a technical note. *Laparoscopic, Endoscopic and Robotic Surgery*. (2022) 5:25–34. doi: 10.1016/j.lers.2021.12.003

143. Zhang Z, Beck MW, Winkler DA, Huang B, Sibanda W, Goyal H, et al. Opening the black box of neural networks: methods for interpreting neural network models in clinical applications. *Ann Trans Med*. (2018) 6:216. doi: 10.21037/atm.2018.05.32

144. Shimabukuro DW, Barton CW, Feldman MD, Mataraso SJ, das R. Effect of a machine learning-based severe sepsis prediction algorithm on patient survival and hospital length of stay: a randomised clinical trial. *BMJ Open Respir Res*. (2017) 4:e000234. doi: 10.1136/bmjresp-2017-000234

145. Semler MW, Weavind L, Hooper MH, Rice TW, Gowda SS, Nadas A, et al. An electronic tool for the evaluation and treatment of Sepsis in the ICU: a randomized controlled trial. *Crit Care Med*. (2015) 43:1595–602. doi: 10.1097/CCM.0000000000001020

146. Wilson FP, Martin M, Yamamoto Y, Partridge C, Moreira E, Arora T, et al. Electronic health record alerts for acute kidney injury: multicenter, randomized clinical trial. *BMJ*. (2021) 372:m4786. doi: 10.1136/bmj.m4786

147. 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

148. Wu Y, Chen Y, Li S, Dong W, Liang H, Deng M, et al. Value of electronic alerts for acute kidney injury in high-risk wards: a pilot randomized controlled trial. *Int Urol Nephrol*. (2018) 50:1483–8. doi: 10.1007/s11255-018-1836-7

149. Wilson FP, Shashaty M, Testani J, Aqeel I, Borovskiy Y, Ellenberg SS, et al. Automated, electronic alerts for acute kidney injury: a single-blind, parallel-group, randomised controlled trial. *Lancet*. (2015) 385:1966–74. doi: 10.1016/S0140-6736(15)60266-5

150. Escobar GJ, Liu VX, Schuler A, Lawson B, Greene JD, Kipnis P. Automated identification of adults at risk for in-hospital clinical deterioration. *N Engl J Med*. (2020) 383:1951–60. doi: 10.1056/NEJMsa2001090

151. 齐霜, 毛智, 胡新, et al., 基于专科信息系统建立的重症医学数据库:大型三甲医院重症医学数据库的模式*(a critical care medicine database established based on specialized information systems: A model of critical care medicine database in large third class hospitals)*. 中华危重病急救医学: p. 743–749.

152. 张素珍,唐素娟,戎珊等, 基于机器学习的重症监护病房脓毒性休克患者早期发生急性肾损伤风险的预测模型构建(construction of a machine learning based predictive model for the risk of early acute kidney injury in patients with septic shock in intensive care units). 中华危重病急救医学. (2022) 34:255–9.

153. Flechet M, Falini S, Bonetti C, Güiza F, Schetz M, van den Berghe G, et al. Machine learning versus physicians' prediction of acute kidney injury in critically ill adults: a prospective evaluation of the AKIpredictor. *Crit Care*. (2019) 23:282. doi: 10.1186/s13054-019-2563-x

154. Thorsen-Meyer H-C, 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

155. Gutierrez G. Artificial intelligence in the intensive care unit. *Crit Care*. (2020) 24:101. doi: 10.1186/s13054-020-2785-y

156. Sjoding MW, Hofer TP, Co I, Courey A, Cooke CR, Iwashyna TJ. Interobserver reliability of the Berlin ARDS definition and strategies to improve the reliability of ARDS diagnosis. *Chest*. (2018) 153:361–7. doi: 10.1016/j.chest.2017.11.037

157. Fleuren LM, Thoral P, Shillan D, Ercole A, Elbers PWGRight Data Right Now Collaborators. Machine learning in intensive care medicine: ready for take-off? *Intensive Care Med*. (2020) 46:1486–8. doi: 10.1007/s00134-020-06045-y

158. Hravnak M, Pellathy T, Chen L, Dubrawski A, Wertz A, Clermont G, et al. A call to alarms: current state and future directions in the battle against alarm fatigue. *J Electrocardiol*. (2018) 51:S44–8. doi: 10.1016/j.jelectrocard.2018.07.024

159. Yoon JH, Jeanselme V, Dubrawski A, Hravnak M, Pinsky MR, Clermont G. Prediction of hypotension events with physiologic vital sign signatures in the intensive care unit. *Critical Care London, England*. (2020) 24:661. doi: 10.1186/s13054-020-03379-3

Glossary

AES	Advanced encryption standard
AI	Artificial intelligence
AKI	Acute kidney injury
ANN	Artificial neural network
APACHE-II	Acute physiology and chronic health evaluation-II
API	Application programming interface
ARDS	Acute respiratory distress syndrome
AUC	Area under curve
AUPRC	Area under precision recall curve
AUROC	Area under receiver operating characteristic curve
BCT	Big-data clinical trial
CATE	Conditional average treatment effect
CDSS	Clinical decision support system
CNN	Convolutional neural network
COVID-19	Corona virus disease 2019
CT	Computed tomography
DVT	Deep vein thrombosis
ECG	Electrocardiograph
EEG	Electroencephalogram
EHR	Electronic health records
eICU-CRD	eICU collaborative research database
EIT	Electrical impedance tomography
EMR	Electronic medical records
ENN	Edited nearest neighbors
GBDT	Gradient boosting decision tree
GRF	Generalized random forest
ICH	Intracerebral hemorrhage
ICTCLAS	Institute of Computing Technology, Chinese Lexical Analysis System
ICU	Intensive care unit
KNN	K-nearest neighbor
LASSO-GLM	Least absolute shrinkage and selection operator-generalized linear models

LGBM	Light gradient boosting machine
LIME	Locally interpretable agnostic modeling
LR	Logistic regression
MAE	Mean absolute error
MIMIC	Medical information mart for intensive care
MOD	Multiple organ dysfunction
MSE	Mean squared error
NLP	Natural language processing
PACS	Picture archiving and communication systems
PCA	Principal component analysis
PIC	Pediatric intensive care
PMV	Prolonged mechanical ventilation
RCT	Randomized controlled trial
RF	Random forest
RMSE	Root mean squared error
RNA	Ribonucleic acid
RNN	Recurrent neural network
RR	Risk ratio
RSA	Ron Rivest, Adi Shamir and Leonard Adleman Algorithm
SAP	Severe acute pancreatitis
SAPS	Simplified acute physiology scores
SIRS	Systemic inflammatory response syndrome
SMOTE	Synthetic minority over-sampling technique
SOFA	Sequential organ failure assessment
SSC	Surviving sepsis campaign
SVM	Support vector machine
THE	Heterogenous treatment effects
THULAC	Thu lexical analyzer for Chinese
TLS	Transport layer security technology
UMLS	Unified medical language system
USA	United States of America

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