Journal of Clinical Anesthesia 88 (2023) 111121

Contents lists available at ScienceDirect

Journal of Clinical Anesthesia

journal homepage: www.elsevier.com/locate/jclinane

Original contribution

Early prediction of delirium upon intensive care unit admission: Model development, validation, and deployment

Man-Ling Wang [a], Yu-Ting Kuo [a], Lu-Cheng Kuo [b], Hsin-Ping Liang [c], Yi-Wei Cheng [d,e], Yu-Chen Yeh [a], Ming-Tao Tsai [b], Wing-Sum Chan [f], Ching-Tang Chiu [a], Anne Chao [a], Nai-Kuan Chou [g], Yu-Chang Yeh [a,*], Shih-Chi Ku [b,**]

[a] Department of Anesthesiology, National Taiwan University Hospital, Taipei, Taiwan
[b] Department of Internal Medicine, National Taiwan University Hospital, Taipei, Taiwan
[c] School of Medicine and Public Health, Faculty of Health and Medicine, University of Newcastle, Callaghan, Australia
[d] Taiwan AI Labs, Taipei, Taiwan
[e] Department of Medical Education, National Taiwan University Hospital, Taipei, Taiwan
[f] Department of Anesthesiology, Far Eastern Memorial Hospital, New Taipei, Taiwan
[g] Department of Surgery, National Taiwan University Hospital, Taipei, Taiwan

HIGHLIGHTS

- A certain proportion of critically ill patients develop delirium within the first 24h after ICU admission.
- Traditional delirium prediction models make a prediction time of 24h after ICU admission cannot predict early delirium.
- Developing machine learning model to predict delirium upon ICU admission is crucial for early intervention.
- Make a second prediction at 24h after ICU admission can improve the prediction of delirium occurred after 24h.

ARTICLE INFO

Keywords:
Intensive care unit
Critically ill
Delirium
Prediction
Machine learning

ABSTRACT

Study objective: To develop, validate, and deploy models for predicting delirium in critically ill adult patients as early as upon intensive care unit (ICU) admission.
Design: Retrospective cohort study.
Setting: Single university teaching hospital in Taipei, Taiwan.
Patients: 6238 critically ill patients from August 2020 to August 2021.
Measurements: Data were extracted, pre-processed, and split into training and testing datasets based on the time period. Eligible variables included demographic characteristics, Glasgow Coma Scale, vital signs parameters, treatments, and laboratory data. The predicted outcome was delirium, defined as any positive result (a score ≥ 4) of the Intensive Care Delirium Screening Checklist that was assessed by primary care nurses in each 8-h shift within 48 h after ICU admission. We trained models to predict delirium upon ICU admission (ADM) and at 24 h (24H) after ICU admission by using logistic regression (LR), gradient boosted trees (GBT), and deep learning (DL) algorithms and compared the models' performance.
Main results: Eight features were extracted from the eligible features to train the ADM models, including age, body mass index, medical history of dementia, postoperative intensive monitoring, elective surgery, pre-ICU hospital stays, and GCS score and initial respiratory rate upon ICU admission. In the ADM testing dataset, the incidence of ICU delirium occurred within 24 h and 48 h was 32.9% and 36.2%, respectively. The area under the

Abbreviations: 24H, 24 h; ADM, admission; APACHE, acute physiologic assessment and chronic health evaluation; API, application programming interface; AUPRC, area under the precision-recall curve; AUROC, area under the receiver operating characteristic curve; CI, confidence interval; DL, deep learning; EHR, electronic health records; GCS, Glasgow coma scale; GBT, gradient boosted trees; HTTP, hypertext transfer protocol; ICDSC, the Intensive Care Delirium Screening Checklist; ICU, intensive care unit; LOS, length of stay; LR, logistic regression; SOFA, sequential organ failure assessment; TRIPOD, Transparent Reporting of a Multivariable Prediction Model for Individual Prognosis or Diagnosis.

* Correspondence to: Yu-Chang Yeh, Department of Anaesthesiology, National Taiwan University Hospital, No 7, Chung Shan South Road, Taipei, Taiwan.
** Correspondence to: Shih-Chi Ku, Department of Internal Medicine, National Taiwan University Hospital, No 7, Chung Shan South Road, Taipei, Taiwan.
E-mail addresses: tonyyeh@ntuh.gov.tw (Y.-C. Yeh), scku1015@ntuh.gov.tw (S.-C. Ku).

https://doi.org/10.1016/j.jclinane.2023.111121
Received 30 September 2022; Received in revised form 31 March 2023; Accepted 5 April 2023
Available online 12 April 2023

receiver operating characteristic curve (AUROC) (0.858, 95% CI 0.835–0.879) and area under the precision-recall curve (AUPRC) (0.814, 95% CI 0.780–0.844) for the ADM GBT model were the highest. The Brier scores of the ADM LR, GBT, and DL models were 0.149, 0.140, and 0.145, respectively. The AUROC (0.931, 95% CI 0.911–0.949) was the highest for the 24H DL model and the AUPRC (0.842, 95% CI 0.792–0.886) was the highest for the 24H LR model.

Conclusion: Our early prediction models based on data obtained upon ICU admission could achieve good performance in predicting delirium occurred within 48 h after ICU admission. Our 24-h models can improve delirium prediction for patients discharged >1 day after ICU admission.

1. Introduction

Delirium is an acute and fluctuating brain dysfunction affecting attention and cognition. Delirium is prevalent among intensive care unit (ICU) patients and is associated with increased mortality, length of stay (LOS), and hospital cost, and long-term cognitive decline. [1–3] Routine delirium assessment in the ICU using with validated bedside instruments, such as the Confusion Assessment Method for the ICU and the Intensive Care Delirium Screening Checklist (ICDSC), is recommended. [4–7] Many researchers have developed models that predicted delirium to reduce the incidence, duration, and severity of delirium. [8–13] The prediction time of a model is crucial to predict the outcome as early as possible to get more time to make decision or take action. In a systematic review of 12 delirium prediction models, most were based on logistic regression (LR) and the prediction time of these model were 24 h after ICU admission, and the highest area under the receiver operating characteristic curve (AUROC) was 0.94. [14] Only a multicenter cohort study used multiple LR analysis to predict delirium upon ICU admission and achieved an AUROC of 0.75. [13] The advantage of predicting delirium at 24 h after ICU admission is having more data to generate accurate prediction. However, a patient can develop delirium or be discharged within 24 h of ICU admission; therefore, a model that predicts delirium upon ICU admission is desirable and could assist in risk stratification, inform preventative strategies, and improve the allocation of medical resources. [15] Recently, machine-learning-based algorithms based on electronic health records (EHR) were employed to developing prediction models for ICU delirium. [16] The model in the Prediction of Intensive Care Unit Delirium study could predict delirium within the first 24 h after ICU admission by using data collected within the first 4 h after admission with an AUROC of 0.919. [17] However, that model used >60 features, and so many features complicate practical model deployment. Therefore, we developed and validated machine learning models using limited features to predict delirium within 48 h after ICU admission. We also tested the remote deployment of our optimal prediction model. We additionally trained machine learning models to predict ICU delirium 24 to 48 h after ICU admission from data obtained at 24 h after ICU admission and investigated its model performance.

2. Materials and methods

2.1. Study design and data source

We employed data from the anonymized database of a Taipei teaching hospital, to develop and validate machine learning models. Our data set comprised the data of adult patients from eight ICUs. The hospital's research ethics committee approved the use of data (REC number 202004016RINB). Written informed consent was waived because of the retrospective nature of the study and anonymization of health information in accordance with the Health Insurance Portability and Accountability Act Privacy Rules. [18] This study was conducted, and the results herein are reported in accordance with the Transparent Reporting of a Multivariable Prediction model for Individual Prognosis or Diagnosis (TRIPOD) statement [19] and the recommended guidance. [20] We calculated LOS in the ICU as the ICU discharge date minus ICU admission date.

2.2. Population and datasets

We excluded patients who were aged <20 years upon ICU admission and those whose delirium could not be assessed (e.g. Richmond Agitation Sedation Scale score < -3 during the whole first 48 h). We split the data by time into the training and testing dataset according to the recommendation of the TRIPOD statement [19]. It allowed for non-random variation between the training and testing datasets in real world. To develop the admission models (ADM models), we allocated the records of patients admitted between August 2020 and May 2021 to the ADM training dataset. To validate the ADM models, we allocated the records of patients admitted between June 2021 and August 2021 to the ADM testing dataset. When developing and validating the 24-h models (24H models), we excluded patients who were discharged within 1 day (on day 0 or 1) of ICU admission and allocated the remaining records into the 24H training and testing datasets.

2.3. Data processing

We extracted the following variables: demographic characteristics, premorbid status, treatments, Glasgow coma scale (GCS), vital signs measurements, and laboratory data obtained from the first 24 h after ICU admission. We used a frequency histogram to screen the data and deleted the outliers and erroneous data according to the definition of outliers in the data dictionary of Taiwan Center of Outcome and Resource Evaluation, in cooperation with the Australian and New Zealand Intensive Care Society Centre for Outcome and Resource Evaluation [21]. We used RapidMiner Studio (Version 9.10, RapidMiner, Boston, MA, USA) for additional data processing. [22–24] Missing data were imputed using the k-Nearest Neighbor model. Finally, the eligible feature data underwent *Z*-normalization.

2.4. Outcomes

The predicted outcome was any positive results (a score ≥ 4) of the delirium screening test using the Intensive Care Delirium Screening Checklist (ICDSC) that was assessed by primary care nurses in each 8-h shift within 48 h after ICU admission. The ICDSC has eight domains based on the Diagnostic and Statistical Manual of Mental Disorders (DSM-IV) criteria and evaluates features and severity of delirium: altered level of consciousness, inattention, disorientation, hallucinations or delusions, psychomotor activity, inappropriate speech or mood, sleep disturbance or fluctuation of symptoms [5]. Routine delirium assessment was promoted in our hospital since 2018, and the ICU nurses were trained and examined for using ICDSC to screen delirium. During the training periods, we had invited an experienced Psychiatrist to discuss and design the lectures, workshop, and training program together. The observational periods were 0–48 and 24–48 h after ICU admission for the ADM and 24H models, respectively.

2.5. Feature selection

We developed the models using RapidMiner Studio. When training the ADM models, only data obtained before or upon ICU admission could be used as features for training. Eligible features for training the

ADM models included age, sex, height, weight, medical history of coronary artery disease, congestive heart failure, dementia, diabetes, or end-stage renal disease, admission after elective surgery, admission from the emergency department, pre-ICU hospital stays, first GCS score, and initial vital signs parameters upon ICU admission. We extracted these features according to their descriptive statistics and clinical relevance. [25] We used the explain predictions operator of the RapidMiner Studio to calculate the feature weights and rank features. Multiple feature sets were used to training the ADM models. As many available features as possible were used for initial training. For further feasibility and practicality of model deployment, we aimed to limit the number of the features of the ADM models to <10. Several features with lower feature weights were removed and we checked whether the AUROC values of the new feature set would decrease a lot. If the AUROC values did not decrease a lot, we kept removing several features with lower feature weights. When the AUROC values decreased a lot, we stopped deleting feature and might add back one feature to maintain averagely high AUROC values. A feature set with a minimal limited number of features with an average high AUROC value was selected for model validation.

When training the 24H models, only data obtained ≤24 h after ICU admission could be used as features; these included demographic characteristics, premorbid status, GCS scores, vital signs measurements, diagnoses, treatments, and laboratory data. We extracted these features according to their descriptive statistics and clinical relevance. [25] We used the explain predictions operator of the RapidMiner Studio to calculate the feature weights and ranked. Multiple feature sets were used to training the 24H models. As many available features as possible were used for initial training. For further feasibility and practicality of model deployment, we aimed to limit the number of the features of the 24H models to <20. Several features with lower feature weights were removed and we checked whether the AUROC values of the new feature set would decrease a lot. If the AUROC values did not decrease a lot, we kept removing several features with lower feature weights. When the AUROC values decreased a lot, we stopped deleting feature and might add back one feature to maintain averagely high AUROC values. A feature set with a minimal limited number of features with an average high AUROC value was selected for model validation.

2.6. Model training and validation

The list of available algorithms in the RapidMiner Studio is presented in Fig. S1. We trained the models by using three common machine learning algorithms: logistic regression (LR), gradient boosted trees (GBT), and deep learning (DL). These three algorithms represented regression, tree-based, and neural networks techniques, respectively [26]. We performed 10-fold cross-validation method for model training. The Optimized Parameter Operator in RapidMiner was used to determine the optimal parameter set for training the GBT models. The ADM and 24H testing datasets were used for model validation. The models were interpreted on the basis of the ranking and weights of features. We generated a model simulator with a personalized risk assessment in RapidMiner Studio.

2.7. Remote model deployment

The ADM model with the highest AUROC was used in testing the feasibility of deploying a remote model with RapidMiner AI Hub for being accessed by multiple ICUs. We created a web server with a query function to communicate with the AI Hub using a web application programming interface (API). A web interface was designed to send hypertext transfer protocol (HTTP) request with the required features to the web server. The web server consolidated and converted the HTTP request into a web API request and then transmitted them to the ADM model deployed on AI Hub. The web server then received the web API response from the prediction results from AI Hub and converted and integrated the web API response into an HTTP response containing the prediction and related recommendations. Finally, the HTTP response was transmitted to the web interface to complete the query process. We tested two methods of model deployment: one employing a manual input query function for a single patient and one based on a CSV file upload for querying for multiple patients.

2.8. Statistical analysis

Categorical variables were compared using a $\chi 2$ test are expressed herein as numbers (percentages). Continuous variables were compared using the Mann–Whitney *U* test and are expressed as medians (interquartile ranges). A two-sided *p* value of <0.05 was considered statistically significant. We performed all statistical analyses using SPSS (SPSS 27; IBM SPSS, USA).

We analyzed the models' discrimination and calibration on the basis of their AUROC, area under the precision-recall curve (AUPRC), sensitivity, precision, specificity, Brier score, and calibration plot. Brier score is commonly used to evaluate accuracy of a set of probabilistic forecasts. Calibration plot is a visualization tool for assessing the agreement between predictions and observations for different percentiles of predicted values. Performance measures are reported with 95% confidence intervals (CIs). We compared AUROC and AUPRC by using Python (version 3.7.3, Python Software Foundation, DE, USA) and the scikit-learn package (version 0.23.1). The explain predictions operator of the RapidMiner Studio was used to calculate the feature weights and rank features. We interpreted the models on the basis of feature rankings and weights. We drew the calibration plots by using R (version 4.1.3, Foundation for Statistical Computing, Vienna, Austria).

3. Results

3.1. Dataset information

We used 4999 and 1239 admission records to respectively develop and validate the performance of the ADM models (Fig. 1). The clinical characteristics of the patients with and without delirium in the admission training and testing datasets are presented in Table 1. The incidence of delirium within 48 h in the ADM training and testing data sets was 32.5% and 36.2%, respectively. In the ADM testing dataset, the mortality rate was higher in the patients with delirium than in the patients without delirium (16.9% vs 2.4%, $p < 0.001$), and 32.9% of patients developed delirium within 24 h. The information of missing data in the ADM data sets are displayed in Table S1. We used 2773 and 763 admission records to develop and validate the performance of the 24H models, respectively (Fig. 1). The clinical characteristics of patients with and without delirium in the 24H training and testing datasets are presented in Table S2. In the 24H training and testing datasets, the incidence of delirium occurred at 24–48 h were 31.8% and 33.6%, respectively. In the 24H testing dataset, 84.4% of patients with delirium 24–48 h after ICU admission had developed delirium within 24 h of admission. The information of missing data in the 24H datasets are presented in Table S3.

3.2. Performance and feature importance of the ADM models

Multiple feature sets and their AUROC values of different ADM models during training are presented in Table S4. We selected a feature set with eight features for the validation of ADM models. These extracted features included age, body mass index, medical history of dementia, postoperative intensive monitoring, elective surgery, pre-ICU hospital stays, and GCS score and initial respiratory rate upon ICU admission. The model performance of the ADM models is illustrated in Fig. 2. The AUROC (0.858, 95% CI 0.835–0.879) and AUPRC (0.814, 95% CI 0.780–0.844) of the ADM GBT model were higher than those of the ADM LR and ADM DL models. The ranking and weight of the features of the

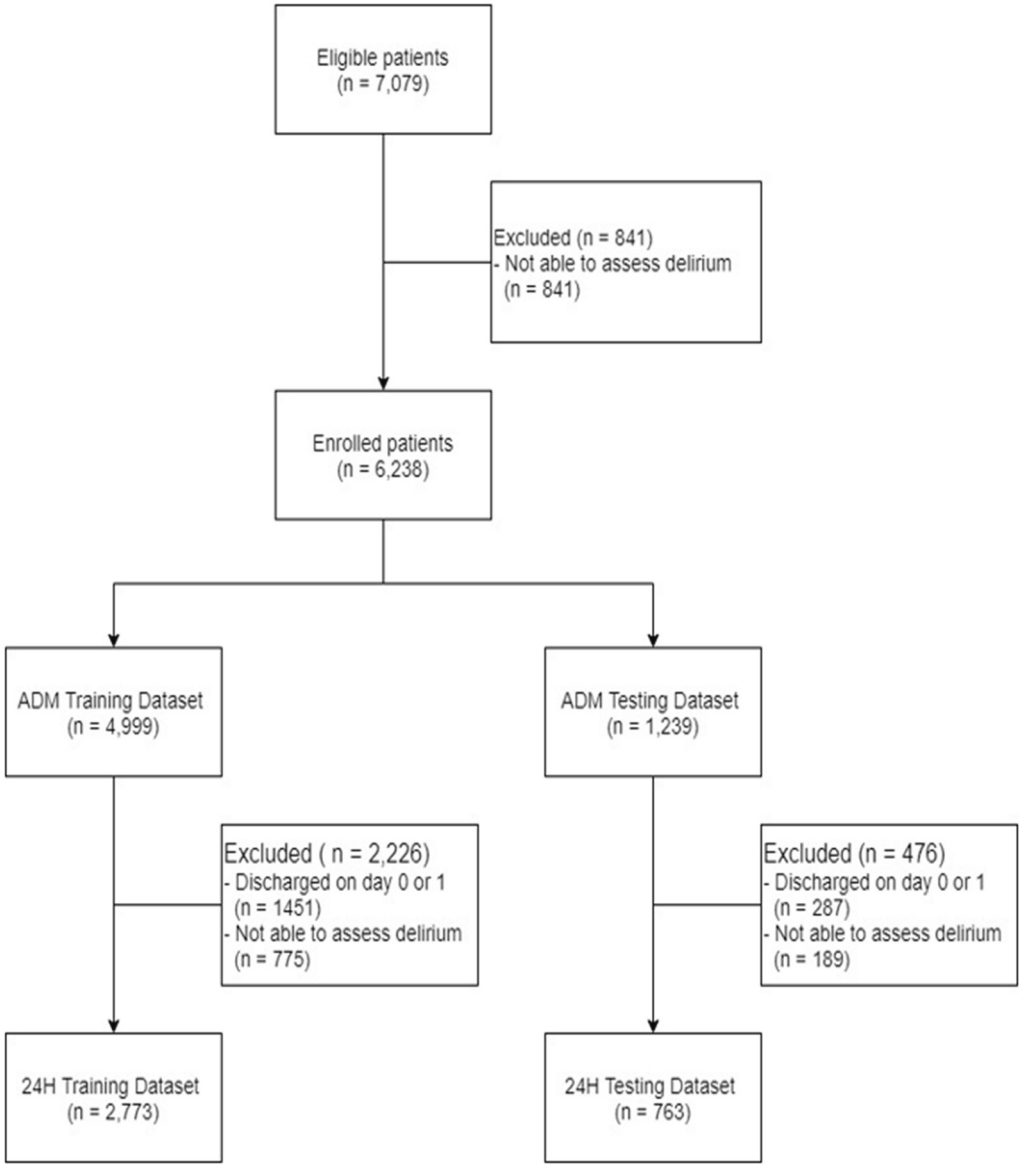

Fig. 1. Flowchart of patient inclusion criteria and allocation to training and testing data sets.
ADM = admission; 24H = 24 h.

ADM models are presented in Fig. S2. Postoperative intensive care, GCS score, and a history of elective surgery were influential features in the ADM models. The calibration plots of the ADM models are presented in Fig. 3. The Brier scores of the ADM LR, GBT, and DL models were 0.149, 0.140, and 0.145, respectively.

3.3. Performance and feature importance of the 24H models

Multiple feature sets and their AUROC values of different 24H models during training are presented in Table S5. We selected a feature set with 15 features for the validation of 24H models. The features for developing 24H models included the occurrence of delirium, age, sex, elective surgery, pre-ICU hospital stays, GCS score, use of invasive ventilator, mean arterial pressure, lowest mean arterial pressure, arterial oxygen tension, arterial carbon dioxide tension, lactate level, highest white cell count, blood urine nitrogen level, and lowest plasma sodium level. The model performance of the 24H models is presented in Fig. S3. The AUROC (0.931, 95% CI 0.911–0.949) was the highest for the 24H DL model and the AUPRC (0.842, 95% CI 0.792–0.886) was the highest for the 24H LR model. The ranking and weight of the features for the 24H models are presented in Fig. S4. The calibration plots for the 24H models are presented in Fig. S5. The Brier scores of the 24H LR, GBT, and DL models were 0.099, 0.105, and 0.098, respectively. Delirium within 24 h, GCS score, age, and intubation within 24 h were influential features of the 24H models.

3.4. Personalized risk assessment and model simulator

The ADM GBT admission model simulator with personalized risk assessment is presented in Fig. S6. In the demonstrated example, the predicted risk of ICU delirium was 34%, and the identified risk factors were tachypnea, long LOS before ICU admission, high body mass index, and older age.

3.5. Demonstration of remote model deployment

The workflow of the manual input and file uploading using the web interface, web server, API, and AI Hub of RapidMiner are presented in Fig. 4. Fig. 4 displays the web page for manual input of features and that for obtaining prediction results of the ADM GBT model. The web page

Table 1
Characteristics of patients from the ADM training and testing datasets.

	Training dataset			Testing dataset		
Delirium within 48 h	Negative	Positive	*P* values	Negative	Positive	P values
Patient number	3375	1624		790	449	
Age (years)	65 (55–74)	72 (62–82)	<0.001	65 (55–74)	74 (62–82)	<0.001
Female (%)	1260 (37.3%)	674 (41.5%)	0.005	317 (40.1%)	184 (41%)	0.81
Height (cm)	163 (156–169)	162 (155–168)	<0.001	162 (156–169)	161 (155–168)	0.014
Weight (cm)	62.5 (54.6–71.6)	59 (51.6–68.5)	<0.001	61.5 (53.5–71.4)	58.6 (50.2–68.4)	0.000
Body mass index	23.7 (21.3–26.4)	22.9 (20.3–25.8)	<0.001	23.4 (20.9–26.2)	22.8 (20.0–25.4)	0.001
Clinical conditions						
Hypertension	1705 (50.5%)	961 (59.2%)	<0.001	388 (49.1%)	257 (57.2%)	0.006
CAD	544 (16.1%)	294 (18.1%)	0.082	102 (12.9%)	72 (16%)	0.148
CHF	182 (5.4%)	140 (8.6%)	<0.001	36 (4.6%)	56 (12.5%)	<0.001
ESRD	171 (5.1%)	136 (8.4%)	<0.001	47 (5.9%)	27 (6%)	1.000
Diabetes	952 (28.2%)	598 (36.8%)	<0.001	240 (30.4%)	183 (40.8%)	<0.001
Dementia	30 (0.9%)	110 (6.8%)	<0.001	5 (0.6%)	38 (6.5%)	<0.001
Post-operation	1885 (55.9%)	472 (29.1%)	<0.001	427 (54.1%)	112 (24.9%)	<0.001
Elective operation	1536 (45.5%)	283 (17.4%)	<0.001	311 (39.4%)	57 (12.7%)	<0.001
Admitted from ED	851 (25.2%)	620 (38.2%)	<0.001	237 (30%)	196 (43.7%)	<0.001
Pre-ICU hospital stay	2 (2–5)	2 (1−11)	<0.001	2 (1–6)	2 (1–9)	0.616
Vital signs at ICU admission						
Glasgow Coma Scale	15 (14–15)	14 (10–15)	<0.001	15 (14–15)	13 (9–15)	<0.001
Body temperature (°C)	36.3 (35.7–36.7)	36.3 (35.9–36.8)	<0.001	36.3 (35.9–36.6)	36.3 (35.8–36.8)	0.254
Respiratory rate (bpm)	18 (15–22)	20 (17–24)	<0.001	19 (15–22)	20 (17–24)	<0.001
Heart rate (bpm)	88 (76–102)	98 (84–114)	<0.001	89 (77–104)	97 (80–115)	<0.001
SBP (mm Hg)	137 (118–156)	131 (112–152)	<0.001	135 (116–155)	129 (112–150)	0.001
DBP (mm Hg)	71 (62–83)	70 (60–83)	0.460	71 (61–84)	70 (59–84)	0.548
MAP (mm Hg)	93 (83–106)	92 (79–105)	<0.001	94 (81–107)	90 (79–106)	0.047
Disease severity						
SOFA*	3 (2–5)	7 (4–11)	<0.001	3 (2–5)	7 (4–11)	<0.001
APACHE II	14 (11–17)	21 (17–26)	<0.001	13 (11–17)	22 (17–27)	<0.001
Clinical outcomes						
LOS in ICU (days)	2 (1–4)	6 (3−13)	<0.001	2 (1–4)	6 (3–14)	<0.001
LOS in Hospital (days)	14 (7–27)	31 (16–60)	<0.001	18 (9–35)	31 (18–60)	<0.001
ICU mortality	92 (2.7%)	285 (17.5%)	<0.001	19 (2.4%)	76 (16.9%)	<0.001

Values are presented as numbers (percentage) or medians (interquartile ranges). ADM = admission; bpm = beats or breaths per minute for heart rate and respiratory rate = respectively; CAD = coronary artery disease; CHF = congestive heart failure; DBP = diastolic blood pressure; ESRD = end-stage renal disease; ICU = intensive care unit; LOS = length of stay; MAP = mean arterial pressure; SBP = systolic blood pressure; SOFA = sequential organ failure assessment.

for uploading a CSV file containing the features of many patients and the web page for the results of prediction are presented in Fig. S7.

4. Discussion

In this study, we revealed that the mortality rate was higher in the patients with delirium than in the patients without delirium in the ADM testing dataset, and 32.9% of ICU patients developed delirium within 24 h. Traditional delirium prediction models designed with a prediction time of 24 h after ICU admission cannot predict these cases. Upon ICU admission, our ADM GBT model could predict the delirium occurred within 48 h with good AUROC and AUPRC. Using updated information obtained within 24 h after ICU admission, the 24H DL model could improve the prediction of the delirium occurred at 24–48 h after ICU admission.

Delirium screening is an initial and indispensable step for preventing or ameliorating delirium. [27] Our results indicate that early delirium prediction, upon ICU admission, is feasible with good model performance. Because a conventional routine delirium prevention protocol is a bundle checklist of numerous assessments, non-pharmacologic measures, medication strategies, family companionship policy, and nursing care, we suggest that an early prediction model can help medical teams classify patients into different risk groups. Patients in the different risk groups may require different bundle of delirium screening frequencies and prevention measures. It may help to save time and reduce workload as it decreases the requirement to perform a full set of routine bundle care for every ICU patient. Additionally, saving time and reducing workload may allow medical teams to spend more time identifying and treating the causes of patients with delirium or taking intensive measures to prevent delirium in high-risk patients. Moreover, our prototype remote model application can automatically receive the data of many patients from an EHR system and send prediction results back to the EHR system. Further studies are warranted to investigate the effect of early delirium prediction model on the clinical outcomes of patients predicted to have a high risk of delirium.

Another clinical application of the prediction model is the analysis of the risk features for developing delirium. Several features were significant predictors for both the ADM and 24H models, including age, GSC score, admission after an elective surgery, and LOS before ICU admission. Some features are non-modified risk factors like age, elective surgery and LOS before ICU admission, [13,17,28–31] and some features are modified risk factors or associated with other clinical problems. Body mass index was valuable feature in our ADM models, because low body mass index might be associated with weight loss and frailty, which have been identified as significant risk factors for delirium and mortality. [32–35] Hypotension, tachycardia and invasive mechanical ventilation might be associated with sepsis, septic shock, respiratory failure, or circulatory failure. [36–44] Disturbances of serum sodium are common in older adults and associated with frailty, falls, hip fractures, and delirium. [45–47] Further study is required to investigate the effects of performing evidence-based interventions to modify these risk factors or treat the underlying diseases. In the 24H testing dataset, 84.4% patients with delirium at 24–48 h after ICU admission had developed delirium within the past 24 h. Further study is required to investigate the treatment and preventive strategies for these patients. Moreover, 11.8% patients without delirium at 24–48 h after ICU admission had developed delirium within the past 24 h. These findings showed that the disease progress of delirium might be fluctuated and reversible. We suggest that it is reasonable to obtain any clinically plausible and valuable information before a predictive model begins to predict features. The effect of

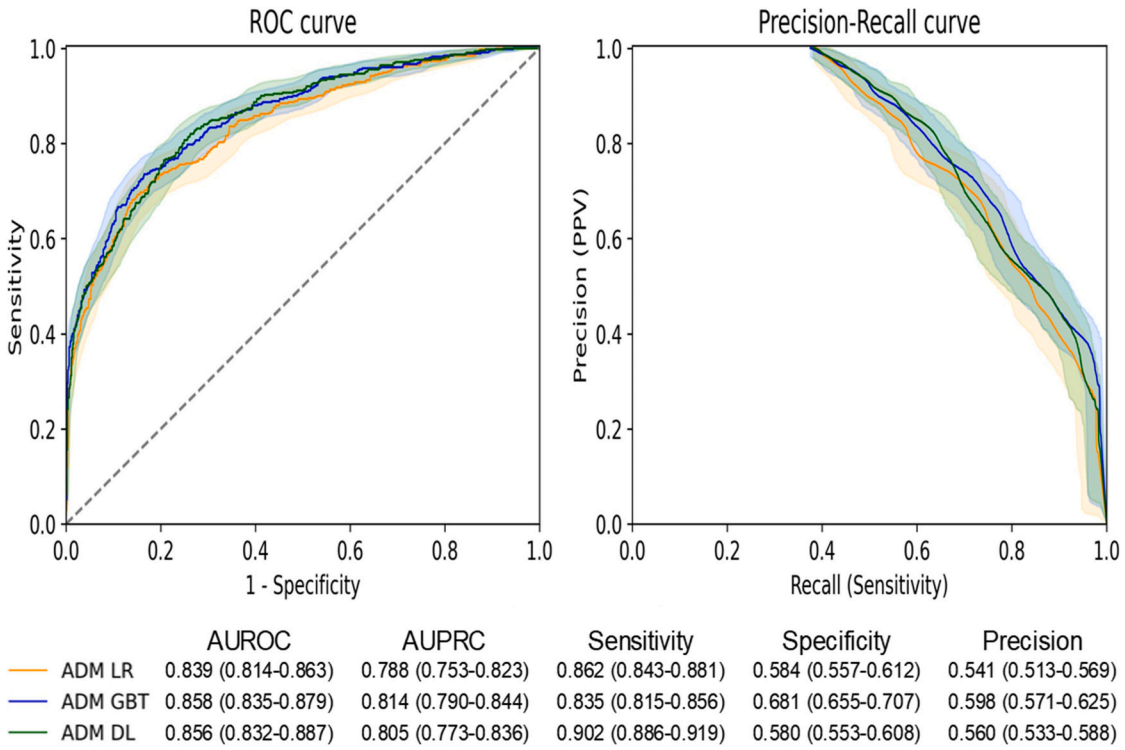

Fig. 2. Performance of the admission models on testing data sets.
Performance measures are presented as values with corresponding 95% confidence intervals. Cut-off value for sensitivity, precision, and specificity was 0.25 for all models. ADM = admission; AUPRC = area under the precision-recall curve; AUROC = area under the receiver operating characteristic curve; DL = deep learning; GBT = gradient-boosted trees; LR = logistic regression.

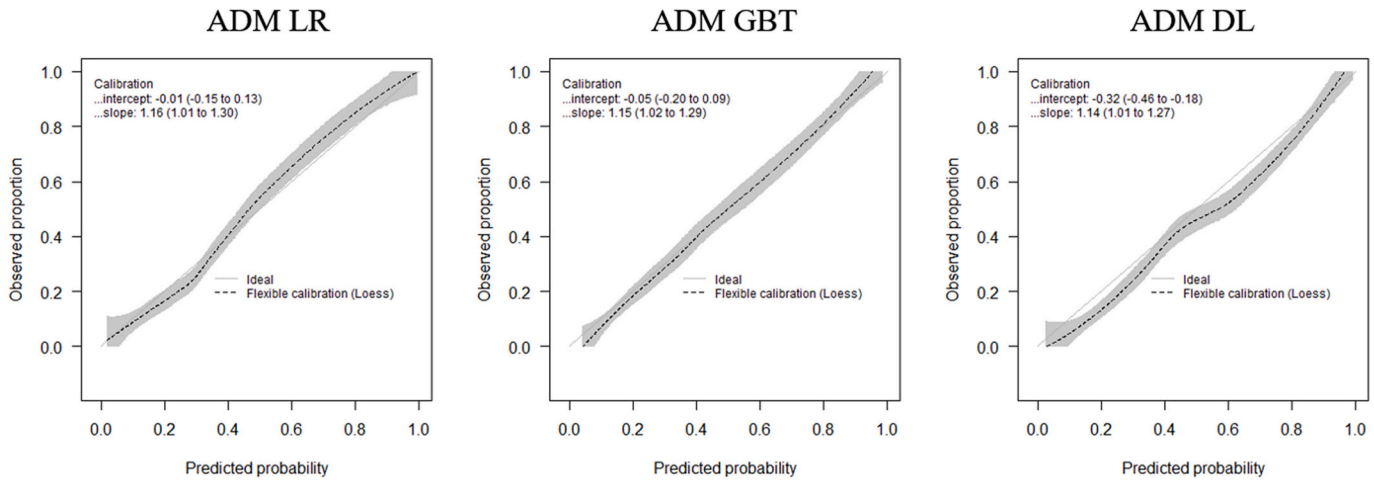

Fig. 3. Calibration plots for admission models.
For clinical utility, a good prediction model should have a predicted risk very close to the observed risk. ADM = admission; DL = deep learning; GBT = gradient-boosted trees; LR = logistic regression.

removing the feature of delirium occurred within 24 h on model performance of the 24H models was shown in Table S7.

Rank of feature importance of a model can help the clinical staffs to understand the model. Because of the different algorithms, we can observe that the ranks and weights of features differ between different models. In most research, a same feature set was used to train different models with different algorithms for easy comparison of performance among the models. In fact, an algorithm may have better model performance than another model when using a different feature set. Even a logistic regression model may have better performance than machine learning models in specific population. Beside good model performance, we suggest that determining a right model for deployment should consider its feasibility and practicability. A super model has an AUROC of 0.971 using 150 complex features may not be as feasible and practicable as a good model has an AUROC of 0.931 using 15 easily accessible features. Furthermore, regular testing of the performance and calibration of a deployed model with new data is crucial to determine the timing to retrain a model to maintain good or better performance. Regarding to missing data imputation, it requires clinical knowledge to correctly interpret the reason of missing and multiple techniques to make a good and reasonable imputation. In this study, k-Nearest Neighbor model imputation of missing data improved the performance

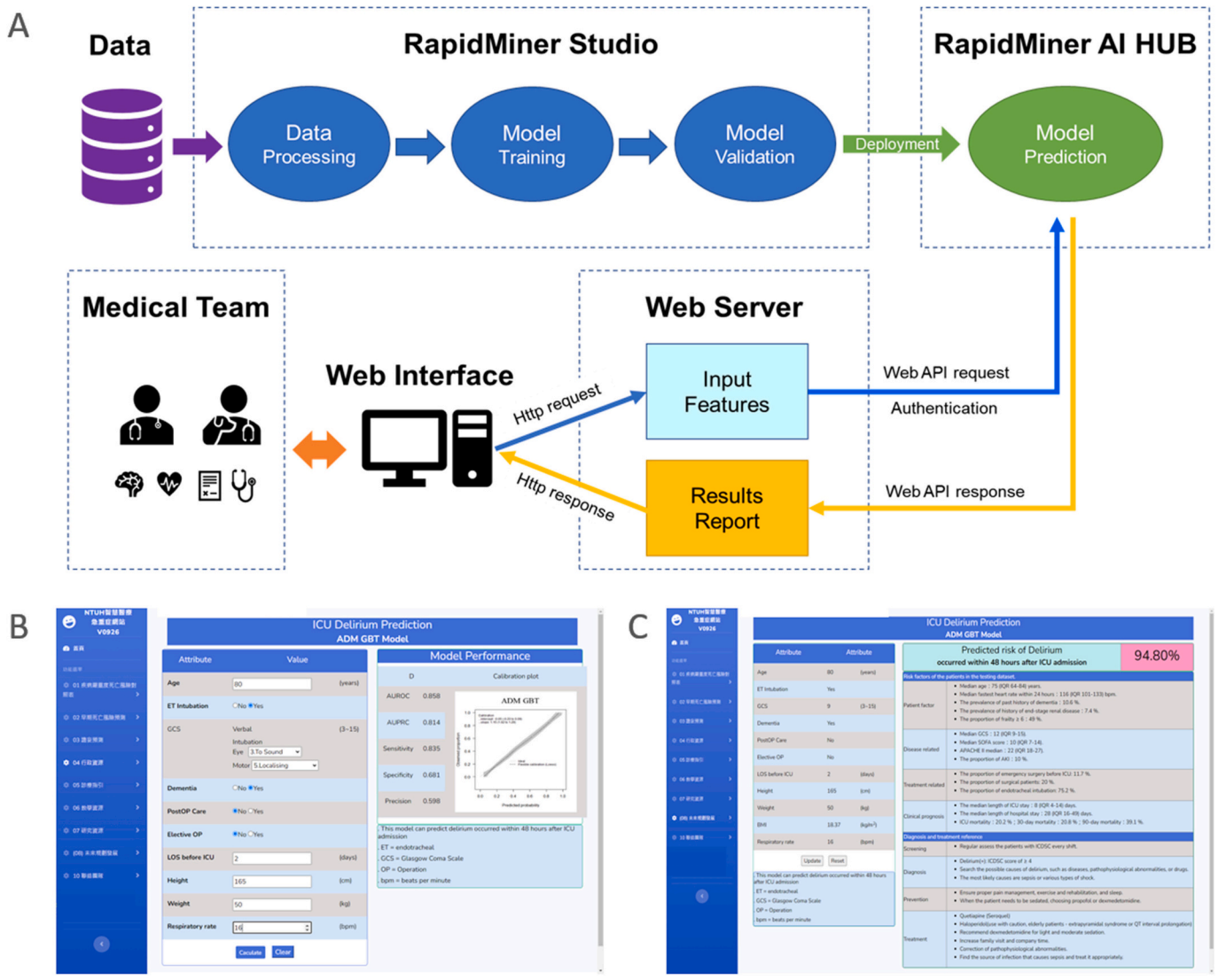

Fig. 4. Demonstration of remote model deployment.
Demonstration results and communications among web interface, web server, and the ADM GBT model deployed on RapidMiner AI HUB. (A) Workflow of model training, model deployment, web requests and responses. (B) Web page of manual input and description of model performance and calibration. (C) Web page of prediction result and recommendation. ADM, admission; GBT, gradient-boosted trees.

of ADM models (Table S8), but it did not affect the performance of 24H models. Other imputation methods can be considered to improve model performance in further study. [48,49]

Our study has several strengths. We developed a model that can predict delirium using data obtained soon after ICU admission. In addition, the training and testing data were split by time according the suggestion of the TRIPOD statement to allow the non-random variation between training data and real data in future. [19] Furthermore, delirium was defined on the routine clinical assessment of the ICDSC by ICU nurses, which was the real circumstance for model deployment, rather than specific assessment by study nurses only during the investigation periods. One reminder is that ICDSC is a screening checklist for finding out the patients at high risk of delirium. Further studies are required to develop a practical and accurate diagnostic checklist for delirium.

Nevertheless, our study has several limitations. First, this study focused on predicting early delirium within the first 48 h after ICU admission. More patients are required to train and validate prediction of late delirium. Second, this was a single-center study; therefore, our study lacks generalizability. However, this also meant that few data were missing for the selected features; thus, the models performed well for the local patients. [50] The key to multicenter collaboration is to obtain high-quality data and information on the data distribution of various institutions. Third, delirium is a syndrome of brain dysfunction. ICDSC and our model can help to screen and find out the patients with high risk of brain dysfunction, but more efforts are required to definitely diagnose delirium and treat the causes and symptoms of brain dysfunction. Fourth, extracted features of low GCS score and dementia are well-known risk factors of delirium. However, a proportion of Psychiatrists will not diagnose delirium in those patients with GCS < 15, and patients with dementia may have false positive results of ICDSC [5]. The effect of removing patients with dementia on the model performance of the ADM models was shown in Table S6. Finally, information on medications, procedures, the presence of family members, and environmental factors was unavailable in our data sets. Expanding the availability of data is crucial for improving the performance of machine learning models. Our models predicting delirium at two time points are just the beginning. We expect more dynamic delirium prediction models will be developed when real-time updating system for medical big data is clinically practical.

In conclusion, early delirium prediction upon ICU admission is feasible, and the model presented herein performs well. Crucially, early

prediction may facilitate early intervention. Furthermore, a second prediction at 24 h after ICU admission can improve the prediction of delirium occurred at 24–48 h after ICU admission.

Ethics approval

The work described has been carried out in accordance with the declaration of Helsinki (Code of Ethics of the World Medical Association for experiments involving humans). The data used in our study was approved by the Research Ethics Committee of the National Taiwan University Hospital (number 202004016RINB).

Funding

The database was funded by the National Taiwan University Hospital (NTUH 109–A150). The funding agencies had no role in the design and conduct of the study; collection, management, analysis and interpretation of the data; preparation, review, or approval of the manuscript; and decision to submit the manuscript for publication.

CRediT authorship contribution statement

Man-Ling Wang: Conceptualization, Methodology, Formal analysis, Writing – original draft. **Yu-Ting Kuo:** Methodology, Formal analysis, Writing – review & editing, Visualization. **Lu-Cheng Kuo:** Methodology, Writing – review & editing, Supervision. **Hsin-Ping Liang:** Methodology, Writing – review & editing. **Yi-Wei Cheng:** Formal analysis, Writing – review & editing, Visualization. **Yu-Chen Yeh:** Formal analysis, Writing – review & editing. **Ming-Tao Tsai:** Conceptualization, Methodology, Writing – review & editing. **Wing-Sum Chan:** Conceptualization, Formal analysis, Writing – original draft. **Ching-Tang Chiu:** Resources, Writing – review & editing. **Anne Chao:** Investigation, Resources, Writing – review & editing. **Nai-Kuan Chou:** Investigation, Resources, Writing – review & editing, Supervision. **Yu-Chang Yeh:** Conceptualization, Methodology, Formal analysis, Investigation, Data curation, Writing – original draft, Writing – review & editing, Visualization, Funding acquisition. **Shih-Chi Ku:** Conceptualization, Data curation, Writing – review & editing.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

The data used during the current study are available from the corresponding author on reasonable request after obtaining the agreement of Research Ethic Committee of National Taiwan University Hospital. All codes and processes used in the current study are available from the corresponding author on reasonable request.

Acknowledgement

The authors would like to thank the following collaborators: members of the National Taiwan University Hospital (NTUH) Center of Outcome and Resource Evaluation (CORE) database, and NTUH Smart Emergency and Critical Care (SECC) group, and Mrs. Hsiu-Yi Huang for data management. This manuscript was edited by Wallace Academic Editing.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.jclinane.2023.111121.

References

[1] Girard TD, Pandharipande PP, Ely EW. Delirium in the intensive care unit. Crit Care 2008;12(Suppl. 3):S3.
[2] Cavallazzi R, Saad M, Marik PE. Delirium in the ICU: an overview. Ann Intensive Care 2012;2(1):49.
[3] Stollings JL, Kotfis K, Chanques G, Pun BT, Pandharipande PP, Ely EW. Delirium in critical illness: clinical manifestations, outcomes, and management. Intensive Care Med 2021;47(10):1089–103.
[4] Gusmao-Flores D, Salluh JI, Chalhub RA, Quarantini LC. The confusion assessment method for the intensive care unit (CAM-ICU) and intensive care delirium screening checklist (ICDSC) for the diagnosis of delirium: a systematic review and meta-analysis of clinical studies. Crit Care 2012;16(4):R115.
[5] Bergeron N, Dubois MJ, Dumont M, Dial S, Skrobik Y. Intensive care delirium screening checklist: evaluation of a new screening tool. Intensive Care Med 2001;27(5):859–64.
[6] Ely EW, Inouye SK, Bernard GR, Gordon S, Francis J, May L, et al. Delirium in mechanically ventilated patients: validity and reliability of the confusion assessment method for the intensive care unit (CAM-ICU). JAMA. 2001;286(21):2703–10.
[7] Plaschke K, von Haken R, Scholz M, Engelhardt R, Brobeil A, Martin E, et al. Comparison of the confusion assessment method for the intensive care unit (CAM-ICU) with the intensive care delirium screening checklist (ICDSC) for delirium in critical care patients gives high agreement rate(s). Intensive Care Med 2008;34(3):431–6.
[8] Heesakkers H, Devlin JW, Slooter AJC, van den Boogaard M. Association between delirium prediction scores and days spent with delirium. J Crit Care 2020;58:6–9.
[9] Moon KJ, Jin Y, Jin T, Lee SM. Development and validation of an automated delirium risk assessment system (auto-DelRAS) implemented in the electronic health record system. Int J Nurs Stud 2018;77:46–53.
[10] Chen Y, Du H, Wei BH, Chang XN, Dong CM. Development and validation of risk-stratification delirium prediction model for critically ill patients: a prospective, observational, single-center study. Medicine (Baltimore) 2017;96(29):e7543.
[11] van den Boogaard M, Schoonhoven L, Maseda E, Plowright C, Jones C, Luetz A, et al. Recalibration of the delirium prediction model for ICU patients (PRE-DELIRIC): a multinational observational study. Intensive Care Med 2014;40(3):361–9.
[12] van den Boogaard M, Pickkers P, Slooter AJ, Kuiper MA, Spronk PE, van der Voort PH, et al. Development and validation of PRE-DELIRIC (PREdiction of DELIRium in ICu patients) delirium prediction model for intensive care patients: observational multicentre study. BMJ. 2012;344:e420.
[13] Wassenaar A, van den Boogaard M, van Achterberg T, Slooter AJ, Kuiper MA, Hoogendoorn ME, et al. Multinational development and validation of an early prediction model for delirium in ICU patients. Intensive Care Med 2015;41(6):1048–56.
[14] Ruppert MM, Lipori J, Patel S, Ingersent E, Cupka J, Ozrazgat-Baslanti T, et al. ICU delirium-prediction models: a systematic review. Crit Care Explor 2020;2(12):e0296.
[15] Tomasev N, Glorot X, Rae JW, Zielinski M, Askham H, Saraiva A, et al. A clinically applicable approach to continuous prediction of future acute kidney injury. Nature. 2019;572(7767):116–9.
[16] Lovejoy CA, Buch V, Maruthappu M. Artificial intelligence in the intensive care unit. Crit Care 2019;23(1):7.
[17] Hur S, Ko RE, Yoo J, Ha J, Cha WC, Chung CR. A machine learning-based algorithm for the prediction of intensive care unit delirium (PRIDE): retrospective study. JMIR Med Inform 2021;9(7):e23401.
[18] Gunn PP, Fremont AM, Bottrell M, Shugarman LR, Galegher J, Bikson T. The health insurance portability and accountability act privacy rule: a practical guide for researchers. Med Care 2004;42(4):321–7.
[19] Collins GS, Reitsma JB, Altman DG, Moons KG. Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD): the TRIPOD statement. Ann Intern Med 2015;162(1):55–63.
[20] Leisman DE, Harhay MO, Lederer DJ, Abramson M, Adjei AA, Bakker J, et al. Development and reporting of prediction models: guidance for authors from editors of respiratory, sleep, and critical care journals. Crit Care Med 2020;48(5):623–33.
[21] Paul E, Bailey M, Pilcher D. Risk prediction of hospital mortality for adult patients admitted to Australian and New Zealand intensive care units: development and validation of the Australian and New Zealand risk of death model. J Crit Care 2013;28(6):935–41.
[22] Chaudhry F, Hunt RJ, Hariharan P, Anand SK, Sanjay S, Kjoller EE, et al. Machine learning applications in the neuro ICU: a solution to big data mayhem? Front Neurol 2020;11:554633.
[23] Van Poucke S, Zhang Z, Schmitz M, Vukicevic M, Laenen MV, Celi LA, et al. Scalable predictive analysis in critically ill patients using a visual open data analysis platform. PloS One 2016;11(1):e0145791.
[24] Mierswa I, Klinkenberg R. RapidMiner studio (9.10). https://rapidminer.com/platform/educational/; 2018.
[25] Barrett TW, Martin AR, Storrow AB, Jenkins CA, Harrell Jr FE, Russ S, et al. A clinical prediction model to estimate risk for 30-day adverse events in emergency department patients with symptomatic atrial fibrillation. Ann Emerg Med 2011;57(1):1–12.
[26] Sanchez-Pinto LN, Luo Y, Churpek MM. Big data and data science in critical care. Chest. 2018;154(5):1239–48.
[27] Trogrlic Z, van der Jagt M, Bakker J, Balas MC, Ely EW, van der Voort PH, et al. A systematic review of implementation strategies for assessment, prevention, and

management of ICU delirium and their effect on clinical outcomes. Crit Care 2015; 19:157.
[28] Inouye SK. Delirium in older persons. N Engl J Med 2006;354(11):1157–65.
[29] Wong A, Young AT, Liang AS, Gonzales R, Douglas VC, Hadley D. Development and validation of an electronic health record-based machine learning model to estimate delirium risk in newly hospitalized patients without known cognitive impairment. JAMA Netw Open 2018;1(4):e181018.
[30] Jin Z, Hu J, Ma D. Postoperative delirium: perioperative assessment, risk reduction, and management. Br J Anaesth 2020;125(4):492–504.
[31] Ansaloni L, Catena F, Chattat R, Fortuna D, Franceschi C, Mascitti P, et al. Risk factors and incidence of postoperative delirium in elderly patients after elective and emergency surgery. Br J Surg 2010;97(2):273–80.
[32] Deng X, Qin P, Lin Y, Tao H, Liu F, Lin X, et al. The relationship between body mass index and postoperative delirium. Brain Behav 2022;12(4):e2534.
[33] Feinkohl I, Winterer G, Pischon T. Obesity and post-operative cognitive dysfunction: a systematic review and meta-analysis. Diabetes Metab Res Rev 2016; 32(6):643–51.
[34] Mazzola P, Tassistro E, Di Santo S, Rossi E, Andreano A, Valsecchi MG, et al. The relationship between frailty and delirium: insights from the 2017 delirium day study. Age Ageing 2021;50(5):1593–9.
[35] Xue QL. The frailty syndrome: definition and natural history. Clin Geriatr Med 2011;27(1):1–15.
[36] Maheshwari K, Ahuja S, Khanna AK, Mao G, Perez-Protto S, Farag E, et al. Association between perioperative hypotension and delirium in postoperative critically ill patients: a retrospective cohort analysis. Anesth Analg 2020;130(3): 636–43.
[37] Wachtendorf LJ, Azimaraghi O, Santer P, Linhardt FC, Blank M, Suleiman A, et al. Association between intraoperative arterial hypotension and postoperative delirium after noncardiac surgery: a retrospective multicenter cohort study. Anesth Analg 2022;134(4):822–33.
[38] Hirsch J, DePalma G, Tsai TT, Sands LP, Leung JM. Impact of intraoperative hypotension and blood pressure fluctuations on early postoperative delirium after non-cardiac surgery. Br J Anaesth 2015;115(3):418–26.
[39] 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(8):801–10.
[40] Roca O, Caralt B, Messika J, Samper M, Sztrymf B, Hernandez G, et al. An index combining respiratory rate and oxygenation to predict outcome of nasal high-flow therapy. Am J Respir Crit Care Med 2019;199(11):1368–76.
[41] Masip J, Peacock WF, Price S, Cullen L, Martin-Sanchez FJ, Seferovic P, et al. Indications and practical approach to non-invasive ventilation in acute heart failure. Eur Heart J 2018;39(1):17–25.
[42] Bradley BD, Green G, Ramsay T, Seely AJ. Impact of sedation and organ failure on continuous heart and respiratory rate variability monitoring in critically ill patients: a pilot study. Crit Care Med 2013;41(2):433–44.
[43] Zaal IJ, Devlin JW, Peelen LM, Slooter AJ. A systematic review of risk factors for delirium in the ICU. Crit Care Med 2015;43(1):40–7.
[44] Green C, Bonavia W, Toh C, Tiruvoipati R. Prediction of ICU delirium: validation of current delirium predictive models in routine clinical practice. Crit Care Med 2019; 47(3):428–35.
[45] Morley JE. Dehydration, hypernatremia, and hyponatremia. Clin Geriatr Med 2015;31(3):389–99.
[46] Wang LH, Xu DJ, Wei XJ, Chang HT, Xu GH. Electrolyte disorders and aging: risk factors for delirium in patients undergoing orthopedic surgeries. BMC Psychiatry 2016;16(1):418.
[47] Zieschang T, Wolf M, Vellappallil T, Uhlmann L, Oster P, Kopf D. The Association of Hyponatremia, risk of Confusional state, and mortality. Dtsch Arztebl Int 2016; 113(50):855–62.
[48] Li J, Yan XS, Chaudhary D, Avula V, Mudiganti S, Husby H, et al. Imputation of missing values for electronic health record laboratory data. NPJ Digit Med 2021;4 (1):147.
[49] Sterne JA, White IR, Carlin JB, Spratt M, Royston P, Kenward MG, et al. Multiple imputation for missing data in epidemiological and clinical research: potential and pitfalls. BMJ. 2009;338:b2393.
[50] Futoma J, Simons M, Panch T, Doshi-Velez F, Celi LA. The myth of generalisability in clinical research and machine learning in health care. Lancet Digit Health 2020; 2(9):e489–92.