

Research and Applications

Development and validation of a deep learning model to predict the survival of patients in ICU

Hai Tang^{1,2}, Zhuochen Jin³, Jiajun Deng^{1,2}, Yunlang She^{1,2}, Yifan Zhong^{1,2},
Weiyang Sun^{1,2}, Yijiu Ren^{1,2}, Nan Cao³, and Chang Chen ^{1,2}

¹Department of Thoracic Surgery, Shanghai Pulmonary Hospital, Tongji University School of Medicine, Shanghai, China, ²Shanghai Engineering Research Center of Lung Transplantation, Shanghai, China, and ³College of Design and Innovation, Tongji University, Shanghai, China

Hai Tang, Zhuochen Jin, and Jiajun Deng contributed equally to this work.

Chang Chen, Nan Cao, and Yijiu Ren were cosenior authors in this work.

Corresponding Author: Chang Chen, Department of Thoracic Surgery, Shanghai Pulmonary Hospital, Tongji University School of Medicine, Shanghai 200443, China; chenthoracic@163.com

Received 7 January 2022; Revised 23 May 2022; Editorial Decision 3 June 2022; Accepted 6 June 2022

ABSTRACT

Background: Patients in the intensive care unit (ICU) are often in critical condition and have a high mortality rate. Accurately predicting the survival probability of ICU patients is beneficial to timely care and prioritizing medical resources to improve the overall patient population survival. Models developed by deep learning (DL) algorithms show good performance on many models. However, few DL algorithms have been validated in the dimension of survival time or compared with traditional algorithms.

Methods: Variables from the Early Warning Score, Sequential Organ Failure Assessment Score, Simplified Acute Physiology Score II, Acute Physiology and Chronic Health Evaluation (APACHE) II, and APACHE IV models were selected for model development. The Cox regression, random survival forest (RSF), and DL methods were used to develop prediction models for the survival probability of ICU patients. The prediction performance was independently evaluated in the MIMIC-III Clinical Database (MIMIC-III), the eICU Collaborative Research Database (eICU), and Shanghai Pulmonary Hospital Database (SPH).

Results: Forty variables were collected in total for model development. 83 943 participants from 3 databases were included in the study. The New-DL model accurately stratified patients into different survival probability groups with a C-index of >0.7 in the MIMIC-III, eICU, and SPH, performing better than the other models. The calibration curves of the models at 3 and 10 days indicated that the prediction performance was good. A user-friendly interface was developed to enable the model's convenience.

Conclusions: Compared with traditional algorithms, DL algorithms are more accurate in predicting the survival probability during ICU hospitalization. This novel model can provide reliable, individualized survival probability prediction.

Key words: deep learning, intensive care unit, survival probability, model visualization

BACKGROUND

Patients in the intensive care unit (ICU) are often in critical condition and have a high mortality rate. The average in-hospital mortality of ICU patients can be as high as 11%–42%.^{1–6} In general hospitals, the ICU generally accounts for 10%–15% of hospital beds, but its operating cost accounts for 22% of the total cost of the hospital.⁷ In addition, limited by the development of medical care, the doctor-to-patient ratio is meager in most countries, especially developing countries with a large population base.⁸ Reducing the cost as much as possible and efficiently utilizing the resources of the ICU are the main problems that we are currently facing.^{9–11} Therefore, predicting ICU patients' survival probability is vital for identifying appropriate interventions and formulating health care policies.

The main characteristic of patients in ICU is that their vital signs are unstable and need to be monitored closely. Real-time monitoring data react to the patient's current condition. However, it is difficult for ICU doctors to quantify the survival probability of patients with the critical information collected from vast amounts of electronic medical records, which will lead to a low identification efficiency, treatment delay, and the condition deterioration of critical patients.

In the past 30 years, research on prediction models for ICU patients has been fruitful. The previous clinical practice mainly includes the Early Warning Score (EWS),¹² the Sequential Organ Failure Assessment Score (SOFA),¹³ the Simplified Acute Physiology Score (SAPS),^{14,15} and the Acute Physiology and Chronic Health Evaluation (APACHE) score.^{16–18} These models are based on previous research on mortality prediction, which plays a certain guiding role in the clinical evaluation of ICU patients.^{14–20} However, the results from different countries also suggested that the predicted mortality rate obtained by such scoring criteria is still generally overestimated.^{21,22} Moreover, although these scores distinguish between patients with expected death and expected survival, only the probability of survival during ICU hospitalization is given without considering the survival length so that the prediction effect is minimal.^{23,24}

In addition, inaccurate predictions are often due to the deviation in algorithm selection rather than variables selection. For example, the linear regression algorithm is unsuitable for complicated real-world scenes. Most traditional models only predict the survival probability during the entire hospitalization period and lack prediction in the time dimension. It is imprecise to determine the priority of medical resource allocation through the prediction results of these models.^{25,26} Models built around machine learning (ML) or deep learning (DL) algorithms have been widely used in the prediction of survival and prognosis of cancer patients and have achieved good results.^{27,28} At the same time, studies have shown that the DL model can be used for classification problems and processing the survival time.

Therefore, based on the variables of the existing EWS, SOFA, SAPS II, APACHE II, and APACHE IV prediction models, this study developed a new evaluation and prediction method based on DL algorithms to predict the survival probability of ICU patients during hospitalization accurately.

METHODS

Study population

This study was analyzed based on participants' information in the MIMIC-III Clinical Database (MIMIC-III),²⁹ approved after a strict deidentification process by the Harvard Medical School's Ethics Re-

view Board and the Massachusetts Institute of Technology. Then, we randomly divided the participants into a training set and a testing set in a 70%:30% ratio. At the same time, we also included participants in the eICU Collaborative Research Database (eICU) and the Shanghai Pulmonary Hospital Database (SPH) as the external testing set of the study. The eICU is released under the Health Insurance Portability and Accountability Act safe harbor provision. The institutional review board of Shanghai Pulmonary Hospital approved our study and waived the need for informed consent due to the retrospective nature of this study.

Patients aged ≥ 14 years were included, and the required variables were established according to the predictive model. Each ICU admission record of the patient was evaluated for multiple ICU admission records of the same patient. Patients with a missing proportion of the corresponding variables greater than 20% were excluded.

Data preprocessing

Forty variables were collected from the classic ICU patients' probability scores, including EWS, SOFA, APACHE II, APACHE IV, and SAPS II.^{14–20} We divided the variables into 4 categories: admission information, vital signs and arterial blood gas (ABG) analysis, history information, and laboratory results.

We obtained basic information about the participants (such as age and comorbidity) from their medical records at admission. We selected the first measurement results for laboratory results within 24 h after admission. For variables such as 24-h urine volume, we calculated the cumulative urine volume for 24 h since admission.

We manually screened the values of the included variables, and the screening details are shown in the [Supplementary eMethods](#). Subsequently, we filled in the missing data. We directly filled these values with the median data for variables missing less than 5% of values. For variables missing more than 5% of values, we filled in the data with the multiple imputation method.³⁰ The values range of variables in the database was shown in [Supplementary e-Table 1](#).

Study design

The research process, including the inclusion and exclusion of participants, variable selection, data extraction, model development, model validation, and evaluation, is shown in [Figure 1](#).

First, we integrated the existing variables among the classic ICU patients' scores and combined any duplicated or identical variables. Then, we used Cox regression to develop a prediction model for the survival probability of ICU patients. As a well-recognized regression model for survival probability prediction, we use it as a baseline for the models we developed. We introduced the random survival forest (RSF) and DL algorithms to build new prediction models with the same variables. The New-Cox model was developed by the survival R package.^{31–33} The relative scores of each variable were obtained by univariable and multivariable analyses. The randomForestSRC R package³⁴ was used to build the New-RSF model and obtained variable importance (VIMP) utilizing the VIMP method.³⁵ The New-DL model contained a core hierarchical structure with fully connected feed-forward neural networks with a single output node to calculate the survival probability $h\theta(x_i)$ of patient i using the negative log-partial likelihood function. More details about the New-DL model are described in the [Supplementary eMethods](#). To provide insights into the predictions made by the New-DL model, we provide local interpretable model-agnostic explanations (LIME).³⁶

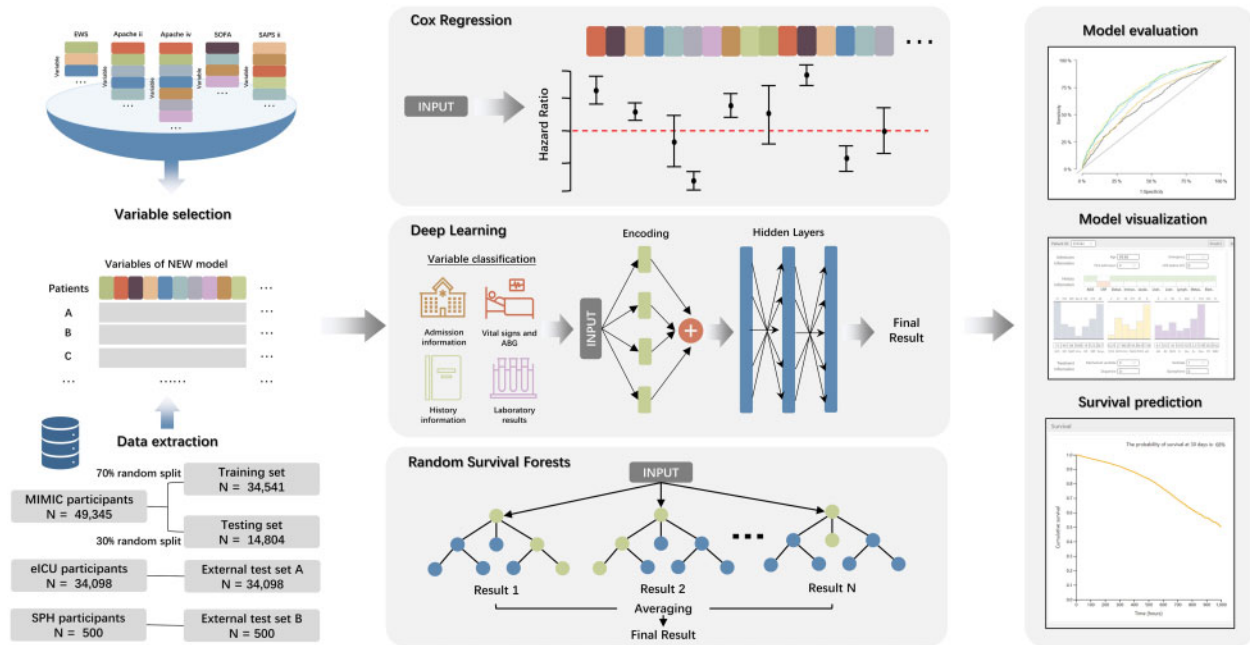


Figure 1. Schematic of the study design. A total of 49 345 participants in the MIMIC-III were randomly split into training ($n = 34\,541$) and testing ($n = 14\,804$) sets. Cox regression, the random survival forest (RSF), and the deep learning (DL) algorithm were used to develop models. All models were tested in the testing set of the MIMIC-III and externally tested among the participants from the eICU ($n = 34\,098$) and the SPH ($n = 500$).

A user-friendly tool was developed in Python for the New-DL model to facilitate survival probability predictions. The codes used in our study are available online (https://github.com/HuddTang/Deep-learning_ICU.git). The user interface consists of the information input interface and the survival probability prediction interface. The information input interface can help users input all entries regarding patient characteristics into the New-DL model. The user input view allows users to predict the survival probability based on specific patient information by clicking the “predict” button.

Statistical analysis

Differences in continuous variables were assessed with Student t test. Categorical variables were compared by the chi-squared test. Patient baseline information was compared with SPSS 22.0 (IBM Corporation). A 2-sided P value less than .05 was considered to be statistically significant. The ability of the prediction models was assessed by using the C-index (Hmisc R package)^{37,38} and the receiver operating characteristic curve (ROC) (ROCR and plotROC R package).³⁹ It is worth mentioning that the area under curve, which is often used to evaluate model efficiency, is generally defined as equal to the C-index for dichotomous variables. However, for analyses involving survival time variables, the C-index should be used for analysis. In addition, The C-index was compared with the compareC R package.⁴⁰ We plotted the calibration curve of the model for predicting survival with the pec R package⁴¹ and plotted the decision curve of the model for predicting survival with the code described in the [Supplementary eMethods](#).⁴²

Data availability

The datasets generated and analyzed during the current study are available in the MIMIC-III Clinical Database and the eICU Collaborative Research Database, <https://physionet.org/content/mimiciii/1.4>, <https://physionet.org/content/eicu-crd/2.0>. The other data sup-

porting this study’s findings are available from the corresponding author upon reasonable request.

RESULTS

Characteristics of the study participants

Forty variables, routinely available for patients admitted to the ICU, were included in the model development, consisting of 26 continuous variables and 14 categorical variables ([Supplementary e-Table 2](#)). Subsequently, the relevant data of participants in the MIMIC-III were extracted. After data preprocessing, 61 532 participants in the MIMIC-III were included in a dataset for this study. After removing the samples with missing values, 49 345 participants in the MIMIC-III were eventually included, with 34 541 and 14 804 participants in the training and testing sets. Subsequently, with the same inclusion and exclusion criteria, we selected 147 876 participants and included 34 098 participants from the eICU. Similarly, we randomly selected 500 participants from the SPH.

The characteristics of participants with complete data included in the analysis are shown in [Table 1](#). For the included participants, most of the participants in the MIMIC-III were admitted to the emergency department (85.8%), while medical ICU patients accounted for 39.6%. Most of the participants in the eICU were nonemergency patients (63.7%), and the majority were surgical ICU patients (60.9%). Most of the participants in the SPH were admitted to the emergency department (63.2%), and all participants in the SPH were medical ICU patients. 76.1% of the participants in the MIMIC-III received ventilation, compared with 43.4% of the participants in the eICU and 58.8% of the participants in the SPH. For other variables, especially laboratory variables, the data from the 3 databases were slightly different but generally consistent. After establishing the training set and the testing set in the MIMIC-III, we also added the external testing set, including participants from the eICU and SPH, to compare the patients’ basic information, predictor information, and prognosis. All variables were shown in [Supplementary e-Table 3](#).

Table 1. Demographic characteristics of the participants in MIMIC-III and the eICU included in the analysis

Characteristic	MIMIC-III	eICU	SPH
No. of participants	49 345	34 098	500
Age (years)	63.95 (52.67, 77.83)	63.83 (54, 76)	63.13 (55, 77)
Type of admission			
Elective	6992 (14.2%)	21 731 (63.7%)	184 (36.8%)
Emergency	42 353 (85.8%)	12 367 (36.3%)	316 (63.2%)
Type of ICU			
Medical	19 550 (39.6%)	3294 (9.7%)	500 (100%)
Medicosurgical	8027 (16.3%)	20 780 (60.9%)	0
Coronary and cardiac surgery	15 774 (32.0%)	8261 (24.2%)	0
Other (trauma surgical/neuro)	5994 (12.1%)	1763 (5.2%)	0
Temperature (°C)	36.7 (36.2, 37.2)	36.37 (36, 36.8)	36.63 (36.2, 37)
Arterial blood pH	7.38 (7.33, 7.44)	7.36 (7.31, 7.43)	7.38 (7.32, 7.44)
Heart rate (bpm)	88.05 (75, 100)	107.59 (95, 127)	94.38 (79, 110)
Mean arterial pressure (mmHg)	81.53 (70.74, 91)	85.31 (50, 127)	83.98 (65, 96.25)
Systolic arterial pressure (mmHg)	118.86 (106.94, 129.07)	118.7 (99, 136)	119.01 (106, 130)
Respiratory rate (cpm)	18.17 (14, 21)	27.56 (12, 38)	22.06 (15, 26)
Serum sodium (mmol/L)	138.72 (136, 141)	138.23 (135, 141)	138.21 (136, 141)
Serum potassium (mmol/L)	4.07 (3.7, 4.4)	4.13 (3.7, 4.5)	4.10 (3.7, 4.4)
Serum creatinine (mg/dL)	0.98 (0.7, 1.1)	1.39 (0.77, 1.72)	1.04 (0.7, 1.13)
Blood urea nitrogen (mmol/L)	21.26 (13, 26)	28.49 (15, 37)	24.21 (13.16, 30)
Albumin (g/dL)	3.06 (2.6, 3.5)	2.72 (2.2, 3.2)	3.03 (2.6, 3.6)
Bilirubin (μ mol/L)	0.78 (0.4, 1.07)	0.84 (0.4, 1)	0.78 (0.4, 1)
Blood glucose (mg/dL)	130.16 (103, 149)	162.77 (97, 207)	150.2 (106, 181)
Hematocrit (%)	31.4 (27.8, 34.7)	31.73 (26.7, 36.4)	31.93 (27.5, 35.9)
White blood cell count ($10^9/L$)	11.03 (7.6, 13.7)	12.85 (7.88, 16.7)	11.35 (7.42, 14.53)
Platelet ($10^9/L$)	212.01 (149, 263)	193.33 (133, 244)	205.1 (142, 259)
Ventilation, No. (%)	37 541 (76.1%)	17 525 (51.4%)	294 (58.8%)
Mechanical ventilation, No. (%)	17 434 (35.3%)	14 814 (43.4%)	170 (34.0%)
PaO ₂ (mmHg)	170.85 (81, 246)	126.21 (73, 146)	142.9 (74.9, 184.7)
PCO ₂ (mmHg)	41.01 (35, 46.45)	41.74 (33.4, 47)	40.84 (34.1, 45.3)
Bicarbonates (mmol/L)	23.86 (22, 26)	22.79 (19.4, 26)	23.95 (21.32, 26.7)
24-h urine output (mL)	1925.7 (1055, 2515)	1542.36 (582.34, 2242.72)	1708.39 (1002.28, 2194.67)
GCS score	13.78 (14, 15)	11.03 (8, 15)	13.14 (13, 15)
Chronic health condition			
Chronic renal failure, No. (%)	2776 (5.6%)	1375 (4.0%)	28 (5.6%)
Lymphoma	512 (1.0%)	191 (0.6%)	2 (0.4%)
Liver cirrhosis	2454 (5.0%)	886 (2.6%)	20 (4.0%)
Leukemia/myeloma	792 (1.6%)	291 (0.9%)	4 (0.8%)
Hepatic failure	2752 (5.6%)	727 (2.1%)	29 (5.8%)
Immunosuppression	1908 (3.9%)	1091 (3.2%)	8 (1.6%)
Metastatic carcinoma	2106 (4.3%)	712 (2.1%)	20 (4.0%)
AIDS	521 (1.1%)	57 (0.2%)	3 (0.6%)
Treatment			
Dopamine (mg)	1.87 (0, 0)	9.4 (0, 0)	11.84 (0, 0)
Epinephrine (mg)	0.02 (0, 0)	1.22 (0, 0)	0.46 (0, 0)
Pre-ICU length of stay (days)	1.04 (0, 0.58)	1.95 (0.05, 1.19)	1.20 (0, 0.57)

ICU: intensive care unit; PaO₂: arterial partial pressure of oxygen; PCO₂: arterial partial pressure of carbon dioxide; GCS: Glasgow Coma Score; AIDS: acquired immune deficiency syndrome.

Development and validation of the survival probability prediction model

A New-Cox model was developed and validated in the testing set. Variables from the classic ICU patients' risk scores, including the EWS, SOFA, and APACHE IV scores, were used. The C-index values of the EWS, SOFA, and APACHE IV model were 0.685 (95% CI, 0.668, 0.701), 0.737 (95% CI, 0.722, 0.753), and 0.785 (95% CI, 0.772, 0.798), respectively. The accuracy of prediction has been improved.

We integrated all variables to develop the New-Cox, New-RSF, and New-DL models, the C-index values of which were 0.802 (95% CI, 0.790, 0.815), 0.818 (95% CI, 0.807, 0.829), and 0.868 (95%

CI, 0.859, 0.878), respectively. The C-index for the prediction of patient death by the New-DL model was higher than that of the other models. To verify the model's stability further, we divided the ICU patients into subgroups based on the ICU type and age to verify the model performance. The results showed that the New-DL model showed good performance in different subgroups. We present the results in [Supplementary e-Figures 1 and 2](#).

We then applied the above models in the external testing set. The performance of the models in the external testing set A (eICU) and external testing set B (SPH) decreased compared with that in the testing set, but the C-index of each model was still within the acceptable range. The models developed by the DL method, C-index val-

Table 2. Univariate and multivariate analyses for the New-Cox model for survival probability

Classification	Variables	Univariate Cox regression		Multivariate Cox regression	
		Hazard ratio (95% CI)	P value	Hazard ratio (95% CI)	P value
Patient and admission information	Age	11 (9, 13)	<.0001	7.49 (6.06, 9.26)	<.0001
	Emergency	4 (3.4, 4.7)	<.0001	3.98 (3.21, 4.94)	<.0001
	First admission	0.96 (0.89, 1)	.2900	1.04 (0.97, 1.13)	.2358
	LOS before ICU	13 (0.1, 1700)	.3000	0.19 (0.00, 47.2)	.5588
History information	AIDS	0.94 (0.7, 1.3)	.6600	1.08 (0.73, 1.61)	.6782
	CRF	1.1 (0.99, 1.3)	.0650	0.94 (0.82, 1.06)	.3438
	Dobutamine	2.4 (2.1, 2.8)	<.0001	1.31 (1.14, 1.50)	.0001
	Immunosuppression	1.3 (1.2, 1.5)	<.0001	1.11 (0.83, 1.48)	.4600
	Leukemia	1.7 (1.4, 2)	<.0001	1.7 (1.28, 2.24)	.0002
	Liver cirrhosis	1.6 (1.4, 1.8)	<.0001	0.84 (0.64, 1.10)	.2148
	Liver failure	1.7 (1.5, 1.8)	<.0001	1.69 (1.30, 2.19)	.0001
	Lymphoma	1.5 (1.2, 1.8)	.0016	1.37 (0.96, 1.95)	.0763
	Metastatic cancer	2.2 (2, 2.4)	<.0001	2.38 (2.12, 2.67)	<.0001
	Elective surgery	0.29 (0.22, 0.38)	<.0001	1.05 (0.74, 1.49)	.7484
Vital signs and arterial blood gas	GCS	0.34 (0.3, 0.38)	<.0001	0.43 (0.39, 0.48)	<.0001
	HR	3.3 (2.6, 4.4)	<.0001	1.78 (1.33, 2.39)	.0001
	MAP	0.078 (0.055, 0.11)	<.0001	0.61 (0.41, 0.90)	.0143
	Urinary output	2.1e−08 (5.3e−09, 8.1e−08)	<.0001	8.81e−05 (2.27e−05, 3.42e−04)	<.0001
	RR	22 (16, 29)	<.0001	8.60 (6.13, 12.0)	<.0001
	SBP	0.017 (0.012, 0.026)	<.0001	0.13 (0.08, 0.20)	<.0001
	Temperature	0.33 (0.27, 0.4)	<.0001	0.49 (0.41, 0.60)	<.0008
	FiO ₂	2.2 (2, 2.5)	<.0001	1.81 (1.61, 2.02)	<.0001
	HCO ₃ [−]	0.053 (0.041, 0.069)	<.0001	0.38 (0.28, 0.51)	<.0005
	Hct	1.8 (1.3, 2.4)	.0002	5.26 (3.78, 7.31)	<.0001
	PaO ₂	0.58 (0.48, 0.71)	<.0001	0.70 (0.56, 0.87)	.0014
	PCO ₂	0.43 (0.33, 0.56)	<.0001	0.46 (0.33, 0.63)	<.0003
	pH	0.1 (0.079, 0.13)	<.0001	0.31 (0.23, 0.42)	<.0009
Laboratory results	Alb	0.1 (0.078, 0.13)	<.0001	0.54 (0.40, 0.72)	<.0002
	Bil	2.2 (1.9, 2.7)	<.0001	1.46 (1.20, 1.76)	.0001
	BUN	7.1 (5.9, 8.6)	<.0001	1.11 (0.88, 1.40)	.3538
	Cr	18 (14, 23)	<.0001	2.64 (1.96, 3.55)	<.0006
	Glu	3.5 (2.5, 4.9)	<.0001	1.86 (1.33, 2.58)	.0002
	K ⁺	3 (2.3, 3.9)	<.0001	1.99 (1.56, 2.55)	<.0004
	Na ⁺	2 (1.6, 2.5)	<.0001	1.41 (1.14, 1.74)	.0013
	Plt	0.57 (0.48, 0.67)	<.0001	0.67 (0.56, 0.81)	<.0001
	WBC	2 (1.6, 2.5)	<.0001	1.14 (0.92, 1.43)	.2210
	MV	1.3 (1.2, 1.4)	<.0001	1.42 (1.31, 1.54)	<.0001
	Ventilate	1.4 (1.3, 1.6)	<.0001	0.91 (0.82, 1.00)	.0599
Treatment information	Dopamine	90 (48, 170)	<.0001	13.0 (6.17, 27.6)	<.0007
	Epinephrine	2.0e+01 (2.3e−16, 1.8e+18)	.8800	1.53e+03 (8.27e−27, 2.82e+32)	.8311

Note: See Table 1 legend for expansion of other abbreviations.

LOS: length of stay; CRF: chronic renal failure; HR: heart rate; MAP: mean arterial pressure; Urinary Output: urinary output within 24 h after admission to the ICU; RR: respiratory rate; SBP: systolic blood pressure; FiO₂: fraction of inspiration oxygen; HCO₃[−]: bicarbonate; Hct: hematocrit; Alb: blood albumin; Bil: total bilirubin; BUN: blood urea nitrogen; Cr: serum creatinine; Glu: blood glucose; K⁺: serum potassium; Na⁺: serum sodium; Plt: platelet; WBC: white blood cell; MV: mechanical ventilate. Bold words indicate $P < .05$.

ues of which were 0.764 (95% CI, 0.756, 0.771), were still better than the other models (Supplementary e-Table 4).

For the New-Cox model, we evaluated the related risk ratio according to the training set data, shown in Table 2. The feature importance ranking of the New-RSF model is shown in Figure 2A. Insights into the variable importance of the New-DL model are provided in Figure 2B. Supplementary e-Figure 3 provides insights into the variable importance of the New-DL model for predicting the survival probability using local interpretation methods among a randomized sample of 300 participants in the testing set of the MIMIC-II and eICU. The feature component weightings in the New-DL model are listed in Supplementary e-Table 5.

According to the model prediction value and patient survival data information, we drew a ROC curve for each model, shown in Figure 3. We also drew decision curves of the models at 3 and 10 days after admission to the ICU, which indicated that the accuracy of the survival probability prediction of the modified model was excellent (Supplementary e-Figure 4A–F). The New-DL model performed well in both the testing and external testing sets. To test the calibration of the model, we ultimately drew the calibration curves of the models at 3 and 10 days after admission to the ICU, and most of the prediction and observation points were distributed directly on the 45° line (Figure 4A–F). The results show that the New-DL model performs the best, and its Brier Score is smaller than other models in

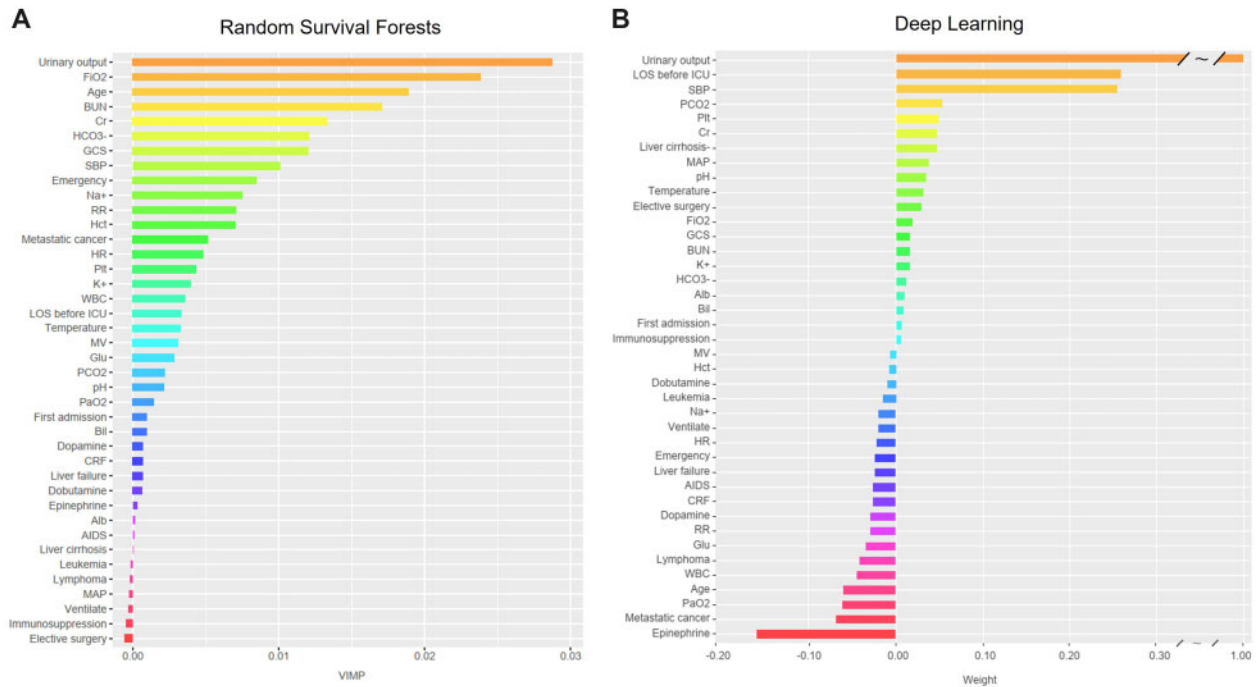


Figure 2. Variable importance (VIMP) based on the New-RSF model (A) and feature component weightings in the New-DL model trained in participants from the MIMIC-III (B).

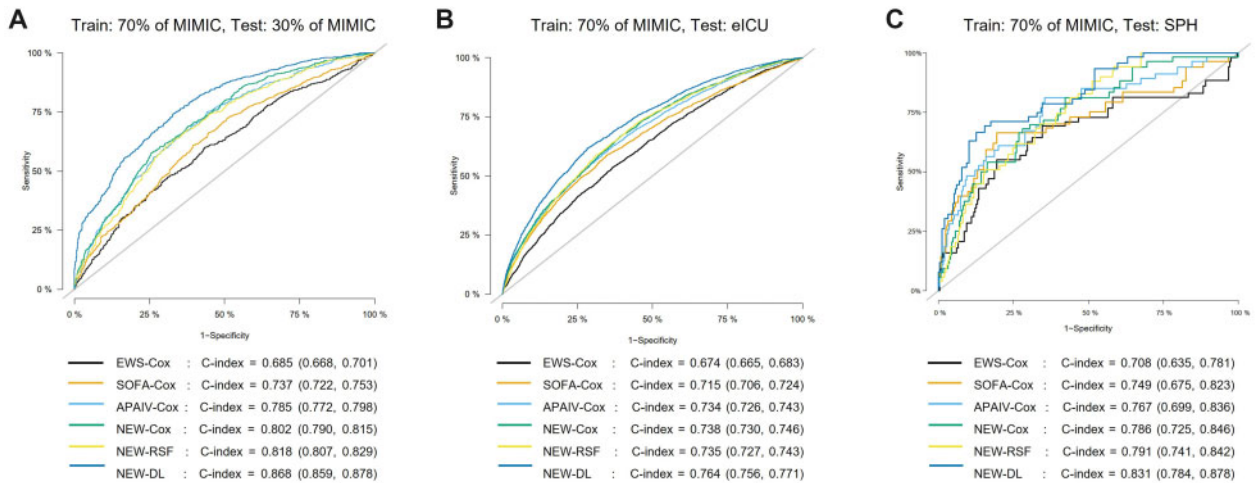


Figure 3. Receiver operating characteristic (ROC) curve and C-index comparing the models for survival probability in the testing set of the MIMIC-III (A), eICU (B), and SPH (C).

each dataset. We divided the included population from the MIMIC-III, eICU, and SPH into 3 groups according to the predicted value of ICU mortality risk (high, middle, and low). The Kaplan–Meier analysis was used to compare the survival of the patients in these groups (Figure 5). The model developed by DL algorithms can more accurately classify patients with different survival probability.

Model visualization

We developed the interface (https://github.com/HuddTang/Deep-learning_ICU.git) to facilitate the use of the model to explore the relative contribution of survival probability factors in ICU patients. In the prediction view, the system invokes a prediction model, and the New-DL

model predicts the patient’s survival probability. The analysis results are visualized in a graphic view as a survival curve, which indicates the survival probability of the patient input over time (Figure 6).

DISCUSSION

The estimation of the survival probability of critical patients is an essential reference for doctors to choose appropriate intervention times and allocate medical resources. The classic ICU patients’ scores mainly used the logistic regression model for prediction.^{14,17} With the advent of the era of big data, the application of DL algorithms has provided accurate and feasible methods for clinical prediction, which have been applied to the prognosis prediction of

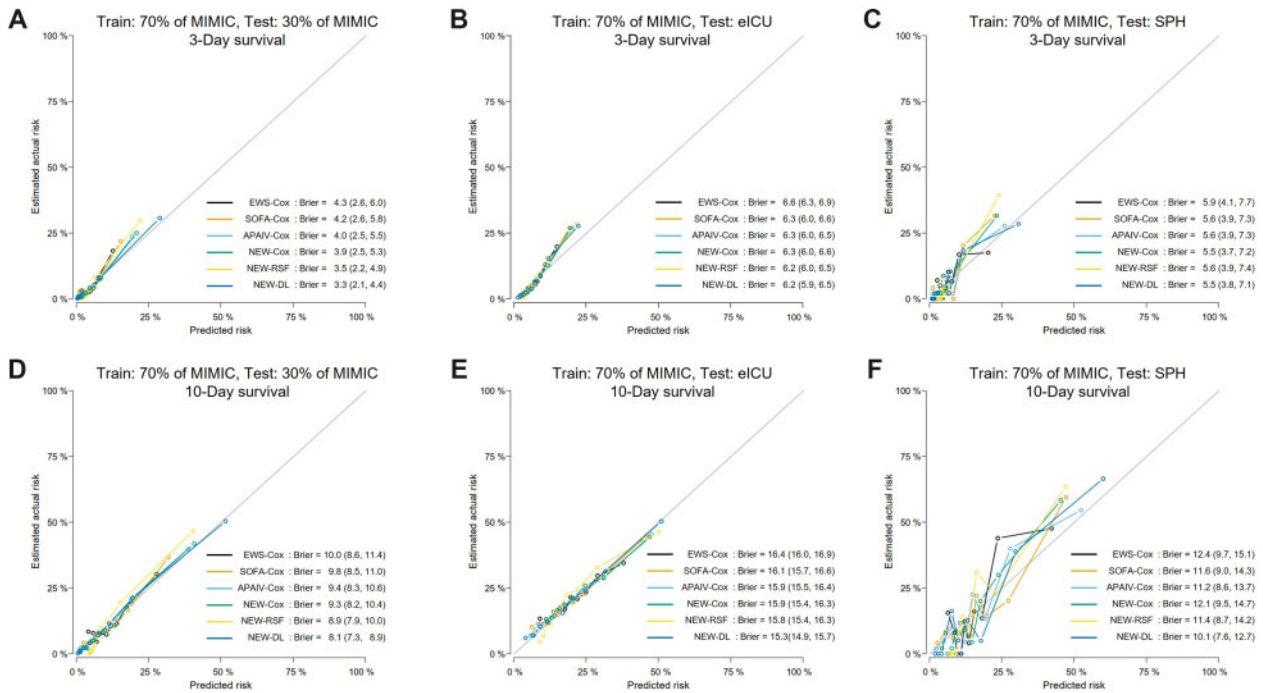


Figure 4. Calibration curves for the models for predicting survival probability at 3 days in the testing set of the MIMIC-III (A), eICU (B), and SPH (C) databases and 10 days in the testing set of the MIMIC-III (D), eICU (E), and SPH (F) databases.

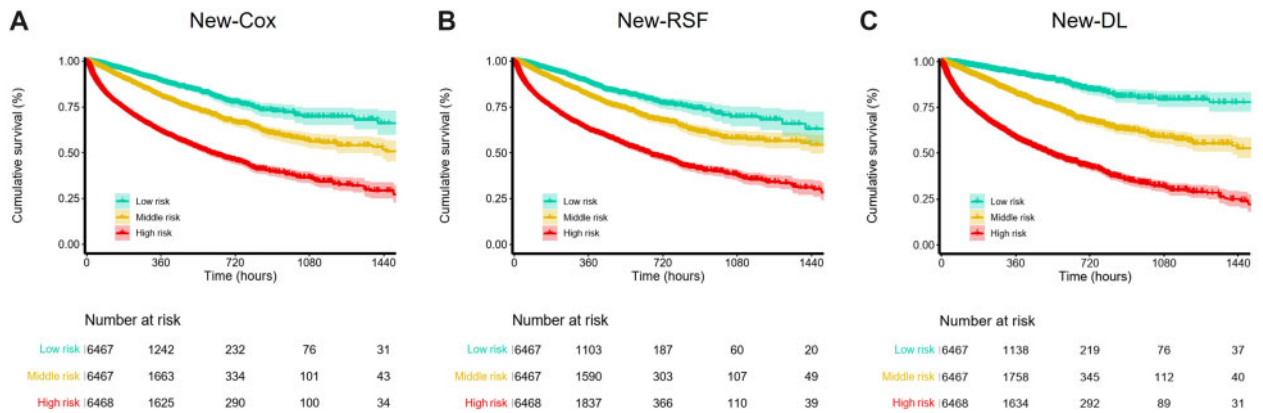


Figure 5. Kaplan-Meier analysis of participants in the pooled samples according to the survival probability predicted by the New-Cox, New-RSF, and New-DL model.

cancer patients.^{27,28,43} These studies show that the most significant advantage of DL algorithms, as discussed before, is that they try to learn high-level features from data in an incremental manner.

This study applied DL algorithms to clinical variables, including admission information, vital signs and ABG analysis, history information, and laboratory results. The prediction effect of the DL model was better than that of linear regression models and ML models (such as the RSF model). In addition, the model included not only patient death or survival outcomes during hospitalization but also the patient length of stay and survival. Therefore, this model can reflect the risk of events in each period after admission to the ICU. It is worth mentioning that we also used the currently relatively mature shallow algorithm. Given the need for studies to assess survival probability on time scales, we planned to use Support Vector

Machine (SVM) and RSF algorithms.^{44,45} However, the sample size involved in this study is vast, and the calculation amount of the SVM algorithm is difficult to support on the current platform, and it takes much time and cost in the process of parameter adjustment. Therefore, we finally chose RSF with similar performance as the shallow model. However, the results still show that complex architectures such as the DL model are better than the RSF model in predicting the survival probability of ICU patients.

Previously, a series of studies have been published on the prognostic risk of ICU patients.^{23,38} Most of the studies were based on the existing prediction models and incorporated more clinical predictive variables to make the prediction results of the model closer to real-world scenes. However, the methods used are always linear models, which are not fully applicable to complex clinical scenes, so

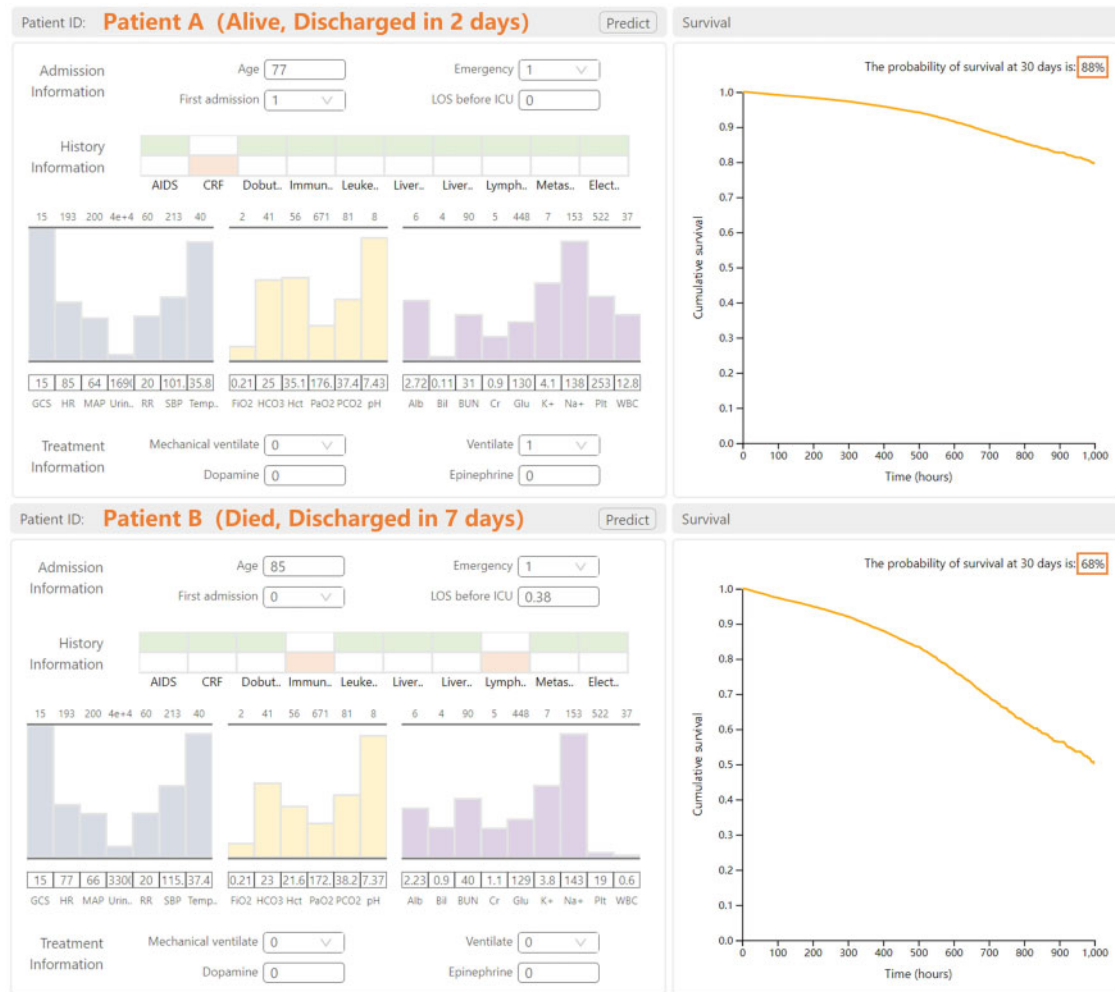


Figure 6. User-friendly interface of the New-DL model facilitating survival probability prediction.

the accuracy of the prediction results often presents bottlenecks. The input of big data also does not optimize algorithm performance. At the same time, the increasing number of predictive variables aggravates the complexity of the clinical operation, and the model efficiency cannot be significantly improved. Therefore, we propose a new model (the DL model) to address this bottleneck.

Most models only included patients' death or survival information in the ICU but did not evaluate patients' survival probability in the dimension of survival time.^{2,46,47} In other words, the lower the survival probability reflected in the prediction model was, the shorter the survival time of patients, and these patients should be treated sooner. To give a simple example, when using traditional models to assess the survival probability of 2 patients admitted to the ICU simultaneously. If the survival probability scores of the 2 patients (patient A and patient B) are the same, it means that the probability of death of the 2 patients throughout the hospitalization period is the same. Even though patient A may die after 3 days, patient B may die after 14 days. In these models, only survival/death is considered when the model is trained and the time is ignored. If the new model predicts the survival probability of these 2 patients, patient A, who may die after 3 days, will have a lower probability. Therefore, in the case of limited medical resources, prioritizing medical resources for patient A will help improve the overall prognosis of ICU patients.

To our knowledge, this study is the first to predict the survival probability of ICU patients using patient follow-up survival data. Compared with previous studies, this study has the following characteristics: (1) In addition to the death and survival information of patients, the survival time information of patients was included in the model development; therefore, the survival probability of patients admitted to the ICU can be predicted continuously in the dimension of survival time. (2) DL algorithms avoid the limitation that traditional linear regression models cannot reflect clinical reality. It is closer to real-world scenes. (3) A friendly interface was developed to realize the visualization of the model, which added great clinical application value to the model.

In terms of model evaluation, we evaluated the predictive efficacy of the Cox model, ML model, and DL model in the MIMIC-III testing set, and the DL model had advantages over the former 2 models. We applied the model to the external test (the eICU and the SPH), and although the evaluation accuracy decreased, the DL model still performed better than the other models. We believe that the decrease in the testing set may be related to the data heterogeneity in the 2 datasets, such as the median survival time, vital signs, and results of laboratory examinations. In our opinion, since the data were collected retrospectively, the ICU data of different medical centers were biased. For example, the conditions of laboratory examination instruments and the measurement methods of vital

signs were not restricted by the study, which was the main reason for the differences among the datasets. Ideally, multicenter ICU data should be adopted for training so that the prediction results can integrate multidimensional features to improve prediction accuracy.

In addition to the model's accuracy, we further showed the proportion of the impact efficiency of the prediction variables to confirm the rationality of the model. In the Cox model, there were statistically significant variable values, and the variables with the highest relative score were the dose of dobutamine, respiratory rate, age, hematocrit, and emergency admission. In the RSF model, the values with the highest weight for each variable in the model were urinary output within 24 h after admission to the ICU, the fraction of inspiration oxygen, age, blood nitrogen, and serum creatinine. The latter variables indicate that the most critical variables for the survival probability assessment of patients are correlated with the circulatory system, such as urine volume, urea nitrogen, and creatinine, which are associated with blood perfusion. At the same time, oxygenation index and age are mostly related to cardiopulmonary function. The model evaluation system of RSF is more combined with general clinical cognition.

Although our proposed model can improve the survival probability prediction for ICU patients, it is more important to translate the improved accuracy into better decision-making for clinicians and patients. For this purpose, we developed an interface that allows researchers and clinicians to explore the accuracy of our model's features in predicting the survival probability of ICU patients, making it easier for scientific and clinical applications. This prediction tool can help ICU physicians identify patients with a higher mortality risk ahead of time, enabling timely care and prioritizing medical resources to improve the overall patient population survival.

This study still has some limitations. It is a retrospective study based on ICU databases, and there may be data collection errors, missing data, and other problems in data collection and recording. In addition, because the DL algorithms have high requirements for data integrity, patients with more than 20% of missing variables were excluded. Therefore, there may be the risk that poses to generalizability and introducing bias. Ideally, a prospective cohort study should be carried out in the ICU patient population to verify the model prediction accuracy. In addition, the time span of the patient inclusion process in the MIMIC-III used in this study was 12 years, and there may be data drift. However, we cannot know the exact time due to the strict deidentification, so it cannot be verified. Future studies can also explore the application of the DL model in effectively screening high-risk groups to guide medical staff practice. The predictive ability of this DL model still needs to be further explored.

CONCLUSION

The results of this study indicate that the use of DL algorithms to predict the survival probability during ICU hospitalization has good accuracy and practicability. Compared with traditional linear models and ML models, the DL model is more accurate. Moreover, the user-friendly visual prediction tool developed based on this model can help clinicians make more accurate judgments and can compensate for the deficiency of previous experience judgments to make the treatment of ICU inpatients more technical with a more reasonable allocation of resources.

FUNDING

Supported by the projects from Shanghai Hospital Development Center (SHDC12017114), Shanghai Pulmonary Hospital Innovation Team

(FKCX1906, FKXY1902), and Shanghai Science and Technology Committee (20YF1441100, 20XD1403000, 18DZ2293400).

AUTHOR CONTRIBUTIONS

HT, ZJ, JD, YR, NC and CC conceived and designed the project. HT, ZJ, JD, YS, YZ and WS contributed to the data preparation, analysis, and interpretation. HT, ZJ and YR contributed to the design of the study and writing the manuscript. JD, YR, NC and CC drafted the manuscript. YR, NC and CC performed the quality assessment and revised the manuscript. All authors revised the manuscript together. All authors have read and approved the submitted version.

ETHICS APPROVAL AND CONSENT TO PARTICIPATE

The institutional review board (IRB) of Shanghai Pulmonary Hospital approved our study (L21-368) and waived the need for informed consent due to the retrospective nature of this study.

SUPPLEMENTARY MATERIAL

Supplementary material is available at *Journal of the American Medical Informatics Association* online.

CONFLICT OF INTEREST STATEMENT

The authors declare that there are no competing interests.

DECLARATIONS

This study was conducted in accordance with the provisions of the Declaration of Helsinki.

CONSENT FOR PUBLICATION

Not applicable.

AVAILABILITY OF DATA AND MATERIALS

The database was approved by the Harvard Medical School's Ethics Review Board and the Massachusetts Institute of Technology after a strict deidentification process. The source code of the proposed method can be found at https://github.com/HuddTang/Deep-learning_ICU.git.

REFERENCES

- 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 (2): R81–9.
- Pirracchio R, Petersen ML, Carone M, Rigon MR, Chevret S, van der Laan MJ. Mortality prediction in intensive care units with the Super ICU Learner Algorithm (SICULA): a population-based study. *Lancet Respir Med* 2015; 3 (1): 42–52.
- Galloway M, Hegarty A, McGill S, Arulkumaran N, Brett SJ, Harrison D. The effect of ICU out-of-hours admission on mortality: a systematic review and meta-analysis. *Crit Care Med* 2018; 46 (2): 290–9.
- Kaufmann M, Perren A, Cerutti B, Dysli C, Rothen HU; Swiss Society of Intensive Care Medicine. Severity-adjusted ICU mortality only tells half the truth—the impact of treatment limitation in a nationwide database. *Crit Care Med* 2020; 48 (12): e1242–e50.

5. Cavallazzi R, Marik PE, Hirani A, Pachinburavan M, Vasu TS, Leiby BE. Association between time of admission to the ICU and mortality: a systematic review and metaanalysis. *Chest* 2010; 138 (1): 68–75.
6. Kashiouris MG, Sessler CN, Qayyum R, et al. Near-simultaneous intensive care unit (ICU) admissions and all-cause mortality: a cohort study. *Intensive Care Med* 2019; 45 (11): 1559–69.
7. Halpern NA, Pastores SM. Critical care medicine in the United States 2000–2005: an analysis of bed numbers, occupancy rates, payer mix, and costs. *Crit Care Med* 2010; 38 (1): 65–71.
8. Halpern NA, Pastores SM. Critical care medicine beds, use, occupancy and costs in the United States: a methodological review. *Crit Care Med* 2015; 43 (11): 2452–9.
9. Murphy DJ, Ogbu OC, Coopersmith CM. ICU director data: using data to assess value, inform local change, and relate to the external world. *Chest* 2015; 147 (4): 1168–78.
10. Andre AS. The formation, elements of success, and challenges in managing a critical care program: Part II. *Crit Care Med* 2015; 43 (5): 1096–101.
11. Andre AS. The formation, elements of success, and challenges in managing a critical care program: Part I. *Crit Care Med* 2015; 43 (4): 874–9.
12. Churpek MM, Snyder A, Han X, et al. Quick sepsis-related organ failure assessment, systemic inflammatory response syndrome, and early warning scores for detecting clinical deterioration in infected patients outside the intensive care unit. *Am J Respir Crit Care Med* 2017; 195 (7): 906–11.
13. Ferreira FL, Bota DP, Bross A, Mélot C, Vincent JL. Serial evaluation of the SOFA score to predict outcome in critically ill patients. *JAMA* 2001; 286 (14): 1754–8.
14. Le Gall J-R, Lemeshow S, Saulnier F. A new simplified acute physiology score (SAPS II) based on a European/North American multicenter study. *JAMA* 1993; 270 (24): 2957–63.
15. Le Gall JR, Neumann A, Hemery F, et al. Mortality prediction using SAPS II: an update for French intensive care units. *Crit Care* 2005; 9 (6): R645–8.
16. Knaus WA, Draper EA, Wagner DP, Zimmerman JE. APACHE II: a severity of disease classification system. *Crit Care Med* 1985; 13 (10): 818–29.
17. Knaus WA, Wagner DP, Draper EA, et al. The APACHE III prognostic system: risk prediction of hospital mortality for critically III hospitalized adults. *Chest* 1991; 100 (6): 1619–36.
18. Zimmerman JE, Kramer AA, McNair DS, Malila FM. Acute Physiology and Chronic Health Evaluation (APACHE) IV: hospital mortality assessment for today's critically ill patients. *Crit Care Med* 2006; 34 (5): 1297–310.
19. McGaughey J, Alderdice F, Fowler R, Kapila A, Mayhew A, Moutray M. Outreach and Early Warning Systems (EWS) for the prevention of intensive care admission and death of critically ill adult patients on general hospital wards. *Cochrane Database Syst Rev* 2007; (3): CD005529.
20. Vincent JL, Moreno R, Takala J, et al. The SOFA (Sepsis-related Organ Failure Assessment) score to describe organ dysfunction/failure. On behalf of the Working Group on Sepsis-Related Problems of the European Society of Intensive Care Medicine. *Intensive Care Med* 1996; 22 (7): 707–10.
21. Poole D, Rossi C, Latronico N, Rossi G, Finazzi S, Bertolini G, GiViTI. Comparison between SAPS II and SAPS 3 in predicting hospital mortality in a cohort of 103 Italian ICUs. Is new always better? *Intensive Care Med* 2012; 38 (8): 1280–8.
22. Metnitz B, Schaden E, Moreno R, Le Gall J-R, Bauer P, Metnitz PG; ASDI Study Group. Austrian validation and customization of the SAPS 3 admission score. *Intensive Care Med* 2009; 35 (4): 616–22.
23. Beck DH, Smith GB, Pappachan JV, Millar B. External validation of the SAPS II, APACHE II and APACHE III prognostic models in South England: a multicentre study. *Intensive Care Med* 2003; 29 (2): 249–56.
24. Ledoux D, Canivet J-L, Preiser J-C, Lefrancq J, Damas P. SAPS 3 admission score: an external validation in a general intensive care population. *Intensive Care Med* 2008; 34 (10): 1873–7.
25. Moreno RP, Metnitz PG, Almeida E, et al.; SAPS 3 Investigators. SAPS 3—From evaluation of the patient to evaluation of the intensive care unit. Part 2: Development of a prognostic model for hospital mortality at ICU admission. *Intensive Care Med* 2005; 31 (10): 1345–55.
26. Liu VX, Lu Y, Carey KA, et al. Comparison of early warning scoring systems for hospitalized patients with and without infection at risk for in-hospital mortality and transfer to the intensive care unit. *JAMA Netw Open* 2020; 3 (5): e205191.
27. Manz CR, Chen J, Liu M, et al. Validation of a machine learning algorithm to predict 180-day mortality for outpatients with cancer. *JAMA Oncol* 2020; 6 (11): 1723–30.
28. She Y, Jin Z, Wu J, et al. Development and validation of a deep learning model for non-small cell lung cancer survival. *JAMA Netw Open* 2020; 3 (6): e205842.
29. Johnson AE, Pollard TJ, Shen L, et al. MIMIC-III, a freely accessible critical care database. *Sci Data* 2016; 3 (1): 1–9.
30. Allison PD. Multiple imputation for missing data: a cautionary tale. *Sociol Methods Res* 2000; 28 (3): 301–9.
31. Lin H, Zelterman D. Modeling survival data: extending the Cox model. *Technometrics* 2000; 44 (1): 85–6.
32. Andersen PK, Gill RD. Cox's regression model for counting processes: a large sample study. *Ann Stat* 1982; 10 (4): 1100–20.
33. Moll M, Qiao D, Regan EA, et al. Machine learning and prediction of all-cause mortality in COPD. *Chest* 2020; 158 (3): 952–64.
34. Ishwaran H, Kogalur UB, Blackstone EH, Lauer MS. Random survival forests. *Ann Appl Stat* 2008; 2 (3): 841–60.
35. Ishwaran H. Variable importance in binary regression trees and forests. *Electron J Stat* 2007; 1 (none): 519–37.
36. Ribeiro MT, Singh S, Guestrin C. “Why Should I Trust You?”: Explaining the Predictions of Any Classifier. In: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. San Francisco, CA: Association for Computing Machinery; 2016: 1135–44.
37. Pencina MJ, D'Agostino RB. Overall C as a measure of discrimination in survival analysis: model specific population value and confidence interval estimation. *Stat Med* 2004; 23 (13): 2109–23.
38. Harrell FE Jr, Lee KL, Mark DB. Multivariable prognostic models: issues in developing models, evaluating assumptions and adequacy, and measuring and reducing errors. *Stat Med* 1996; 15 (4): 361–87.
39. Sing T, Sander O, Beerenwinkel N, Lengauer T. ROCr: visualizing classifier performance in R. *Bioinformatics* 2005; 21 (20): 3940–1.
40. Kang L, Chen W, Petrick NA, Gallas BD. Comparing two correlated C indices with right-censored survival outcome: a one-shot nonparametric approach. *Stat Med* 2015; 34 (4): 685–703.
41. Mogensen UB, Ishwaran H, Gerds TA. Evaluating random forests for survival analysis using prediction error curves. *J Stat Softw* 2012; 50 (11): 1–23.
42. Roupêt M, Hupertan V, Seisen T, et al.; French National Database on Upper Tract Tumors. Prediction of cancer specific survival after radical nephroureterectomy for upper tract urothelial carcinoma: development of an optimized postoperative nomogram using decision curve analysis. *J Urol* 2013; 189 (5): 1662–9.
43. Matsuo K, Purushotham S, Jiang B, et al. Survival outcome prediction in cervical cancer: Cox models vs deep-learning model. *Am J Obstet Gynecol* 2019; 220 (4): 381.e1–81. e14.
44. Layeghian Javan S, Sepehri MM, Layeghian Javan M, Khatibi T. An intelligent warning model for early prediction of cardiac arrest in sepsis patients. *Comput Methods Programs Biomed* 2019; 178: 47–58.
45. Chen C, Zhou J, Yu H, et al. Identification of important risk factors for all-cause mortality of acquired long QT syndrome patients using random survival forests and non-negative matrix factorization. *Heart Rhythm* 2021; 18 (3): 426–33.
46. Hyland SL, Faltys M, Hüser M, et al. Early prediction of circulatory failure in the intensive care unit using machine learning. *Nat Med* 2020; 26 (3): 364–73.
47. Meyer A, Zverinski D, Pfahringer B, et al. Machine learning for real-time prediction of complications in critical care: a retrospective study. *Lancet Respir Med* 2018; 6 (12): 905–14.