IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

SPECIAL SECTION ON SMART HEALTH SENSING AND COMPUTATIONAL INTELLIGENCE: FROM BIG DATA TO BIG IMPACTS

Received June 20, 2019, accepted July 31, 2019, date of publication August 9, 2019, date of current version August 22, 2019.

Digital Object Identifier 10.1109/ACCESS.2019.2934166

Prediction of Length of Stay on the Intensive Care Unit Based on Least Absolute Shrinkage and Selection Operator

CHUNLING LI[1], LONGYI CHEN[1], JIE FENG[1], DUANPO WU[2], ZIMENG WANG[2], JUNBIAO LIU[3], AND WEIFENG XU[3]

[1]Sichuan Academy of Medical Sciences and Sichuan Provincial People's Hospital, Chengdu 610072, China
[2]School of Communication Engineering, Hangzhou Dianzi University, Hangzhou 310018, China
[3]Hangzhou Neuro Science and Technology Company Ltd., Hangzhou 310052, China

Corresponding author: Duanpo Wu (wuduanpo@hdu.edu.cn)

This work was supported in part by the Natural Science Foundation of Zhejiang Province under Grant LGG19F030013 and Grant LGF18F010007, in part by the Special Funds for Information Development in Shanghai under Grant 201801050, in part by the Scientific research project of Zhejiang Provincial Department of Education under Grant Y201533647, and in part by the National Natural Science Foundation of China under Grant 61602417.

ABSTRACT Length of stay (LoS) in the intensive care unit (ICU) is a common outcome measure used as an indicator of both quality of care and resource use. However, the existing analysis methods of LoS are poorly interpretable and extensible, and there is controversial for the predictive performance of LoS. In this paper, the study includes data from 1,214 unplanned ICU admissions to participate in the ICU of Sichuan Provincial People's Hospital between Dec. 11, 2015 and Dec. 6, 2018. On the basis of these data, this study creates a highly accurate and predictive model using advanced preprocessing techniques, exploratory data analysis (EDA) and least absolute shrinkage and selection operator (LASSO) algorithm. Next, this study evaluates the predictive performance of the proposed model by 10-fold cross validation and external validation method using the root mean square prediction error (RMSPE), mean absolute error (MAE), and coefficient of determination (R^2). The predictive performance of the proposed model is 0.88±0.13 day for RMSPE, 0.87±0.07 day for MAE and 0.35±0.09 for R^2. Experimental results show that the performance of the proposed method are competitive with the state-of-the-art methods and results. Furthermore, this study explores the risk factors for ICU LoS in survivors and non-survivors and compare their predictive performance.

INDEX TERMS Length of stay, intensive care unit, exploratory data analysis, least absolute shrinkage and selection operator.

I. INTRODUCTION

Intensive care unit (ICU) is an indispensable medical unit to modern hospitals, and it's also the last line of defense for the life of critically ill patients. However, the care provided by ICU is complex and expensive, so ICU is particularly interested in assessing, comparing and improving their quality of care and resource use [1]–[3]. Since costs are strongly related to length of stay (LoS) on the ICU, shorter LoS on the ICU generally means lower costs [4]. Conversely, longer LoS on the ICU is associated with increased stress and discomfort for patients and their families, as well as increased costs for patients, hospitals, and society [5]. Hence, an accurate tool to predict patient LoS on the ICU would facilitate efficient patient scheduling and maximise available capacity.

Conventional LoS analysis methods, such as APACHE II [6], [7], SAPS II [8] and APACHE IV [9], [10] have been proposed and widely used in the prediction of ICU LoS. However, the predictive performance for ICU LoS of these methods using patient features or ICU features is poor and little consensus exists on the best method for predicting the ICU LoS. As stated in the literature [9], [11], [12], these analysis methods may not be adequate predictors of ICU LoS. Recently, there are a variety of new approaches that have been used in order to improve the predictive performance of LoS on the ICU for patients [13]–[15]. Verburg *et al.* [13] compared the performance of 8 regression models, including 1) ordinary least squares regression (OLSR) on untransformed ICU LoS,

The associate editor coordinating the review of this article and approving it for publication was Guan Gui.

2) OLSR on LoS truncated at 30 days and log-transformed LoS, 3) a generalized linear model with a Gaussian distribution and a logarithmic link function, 4) Poisson regression, 5) negative binomial regression, 6) Gamma regression with a logarithmic link function, 7) the original APACHE IV model, 8) recalibrated APACHE IV model, which uses patient features at admission time only when predicting individual patient LoS in the ICU and concluded that it is difficult to predict ICU LoS. In 2016, Verburg *et al.* conducted a systematic review of the use of models to predict LoS on the ICU in [14]. They claimed that none of the models produced predictions with low bias and the R^2 was 0.05-0.28 across patients and 0.01-0.64 across ICU. In 2019, Daghistani *et al.* [15] developed a machine learning-based model method for predicting in-hospital LoS for cardiac patients. They divided the patients into three groups according to their LoS: short ($<$ 3 days), intermediate (3-5 days) and long ($>$ 5 days), and then using random forest model to obtain the optimal classification performance, but this could not obtain a quantitative LoS. Moreover, only a few studies have explored the risk factors for LoS in survivors and non-survivors, and the studies are not in-depth [9], [13], [16]. Hence, the aim of this paper is twofold. The first aim is to predict ICU LoS using machine learning techniques to plan the number of wards and members of staff required to fulfil demand for ICU care within a given hospital or geographical area. The second aim is to explore the risk factors for ICU LoS in survivors and non-survivors.

In this study, a least absolute shrinkage and selection operator (LASSO) regression algorithm is employed. The LASSO algorithm constructs a penalty function to make some regression coefficients with little influence on prediction performance to be 0, which can deal with biased estimates with complex collinear data on the one hand, and it can help improve the interpretability of the model on the other hand. Then, after data preprocessing, exploratory data analysis (EDA) technique is used to explore new data features. In addition, considering that LoS tends to be skewed, which leads to problems in LASSO regression algorithm, the Box-Cox technology is introduced to optimize the predictive performance of the LASSO algorithm. Furthermore, basing on the LASSO algorithm, the paper explores the risk factors for LoS in survivors and non-survivors, and compares their predictive performance.

The remainder of the paper is organized as follows. In Section II, the dataset used in this study and the prediction method of ICU LoS are described in detail, including data preprocessing, EDA, LASSO algorithm, and evaluation criterion. In Section III, the experimental results and the risk factors for LoS in survivors and non-survivors are explored and discussed. Finally, further discussion and conclusions are included in Section IV and Section V.

II. MATERIAL AND METHODS

This paper uses machine learning techniques to predict ICU LoS and explores LoS risk factors among different populations of survivors and non-survivors. Figure 1 shows the steps of proposed method and the following subsections describe the details of each step.

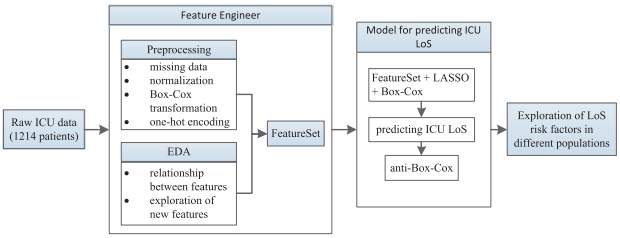

FIGURE 1. Schematic diagram of predicting ICU LoS.

A. DATA COLLECTION AND PRELIMINARY ANALYSIS

The dataset of this study includes ICU patients who were admitted between Dec. 11, 2015 and Dec. 6, 2018 in the ICU of Sichuan Provincial People's Hospital. A total of 1,214 ICU patients records and 72 features are extracted from medical records. These features include demographic information, ICU information, surgical information, drug information and laboratory parameters, and they are selected by experienced doctors and trained nurses. The collected features of this study are shown in Table 1. According to Table 1, the age of ICU patients ranged from 0.28 years old to 100 years old, LoS ranged from 0.04 day to 241 days. In general, previous studies [15], [17], [18] on ICU LoS often excluded data of patients who had been admitted for $<$ 4 hours, patients younger than 16 years old and patients readmitted to the ICU to improve ICU LoS prediction performance. However, this study includes all the ICU patients records so that the proposed model has better generalization performance. In other words, this study includes all patients who had been admitted to ICU. In term of LoS, Figure 2 shows that the distribution of LoS is markedly skewed, which will greatly affect the prediction performance of the regression model [14], [19]. Moreover, the bold font in Table 2 highlights the feature that the standard deviation is greater than mean (including LoS), which indicates that the feature is highly volatile. For white blood cell counts (WBC), its minimum value is $2.1 \times 10^9/L$ and its maximum value is $610.2 \times 10^9/L$. By examining the patient's medical records, it can be found that the patient had acute leukemia, which led to a sharp increase in WBC. Similarly, the maximum value of other features is associated with a disease. Although this may reduce the performance of the proposed model, it can greatly improve the scalability of the model due to the diversity of data.

1) DEFINITION OF LoS

This study defines ICU LoS as the period between ICU admission time and ICU discharge time. However, Figure 2 shows that the distribution of LoS is markedly skewed. Therefore, this study introduces Box-Cox transformation [20] in order to correct the skewed distribution to normal distribution. Eq.(1) shows the Box-Cox

TABLE 1. Dataset structure of 1,214 inpatients.

Category	Features (abbreviation)(unit or format)	Min	Max	Mean±std	Missing(%)	Types
Demographic information	gender	-	-	-	0.000%	binary
	age (year)	0.280	100.000	56.343±18.172	0.000%	numerical
ICU information	admission time (AT) (yy/mm/dd)	15/12/11	18/12/06	-	0.000%	datetime
	discharge time (DT) (yy/mm/dd)	16/01/01	18/12/21	-	0.000%	datetime
	admission department (AD)	-	-	-	0.000%	categorical
	discharge department (DD)	-	-	-	0.000%	categorical
	blood type (BT)	-	-	-	0.000%	categorical
	transfer or not (TON)	-	-	-	0.000%	binary
	resue or not (RON)	-	-	-	0.000%	binary
	initial diagnosis (IND)	-	-	-	25.900	text
	red blood cell in blood transfusion **(RBCT)**$(10^{12}/L)$	1.500	50.000	**5.262±7.571**	86.550	numerical
	platelet in blood transfusion (PTT) $(10^9/L)$)	1.000	4.000	1.903±1.091	97.752	numerical
	plasma in blood transfusion (PLAT) (mL)	150.000	2150.000	628.600±402.300	94.662	numerical
	total blood in blood transfusion (TBT) (mL)	-	-	-	99.892	numerical
	autologous blood in blood transfusion (ABT) (mL)	150.000	1450.000	396.400±371.800	98.292	numerical
	albumin in blood transfusion (ALT) (g)	10.000	280.000	77.300±70.900	97.223	numerical
	consciousness at admission(CA)	-	-	-	0.700	categorical
	consciousness before transfer (CBT)	-	-	-	44.392	categorical
	consciousness after transfer (CAT)	-	-	-	46.424	categorical
	consciousness after discharge (CAD)	-	-	-	3.100	categorical
	atrial fibrillation (AF)	-	-	-	0.000%	binary
	ventricular fibrillation (VF)	-	-	-	0.000%	binary
	diagnostic result (DR)	-	-	-	0.000%	text
	LoS	0.040	241.000	**16.900±17.500**	0.000%	numerical
Surgical information	surgical level (SL)	-	-	-	5.000	categorical
	type of anesthesia (TOA)	-	-	-	5.000	categorical
	healing of incision level (HIL)	-	-	-	5.000	categorical
	American society of anesthesiologists (ASA)	-	-	-	49.243	categorical
	date of operations (DAO) (yy/mm/dd)	15/12/21	18/12/16	-	47.100	datetime
	name of operations (NAO)	-	-	-	47.000	text
	duration of operations (DUO)(h)	1.200	22.800	2.909±2.200	50.900	numerical
	blood loss in operations **(BLO)** (mL)	50.000	8000.000	**346.433±606.523**	56.141	numerical
	red blood cell in operations (RBCO) $(10^{12}/L)$)	-	-	-	100.000	numerical
	platelet in operations (PLTO) $(10^9/L)$)	-	-	-	100.000	numerical
	plasma in operations (PLAO) (mL)	-	-	-	100.000	numerical
	total blood in operations (TBO) (mL)	400.000	800.000	600.000±282.800	99.800	numerical
	autologous blood in operations (ABTO) (mL)	-	-	-	100.000	numerical
	albumin in operations (ALO) (g)	-	-	-	100.000	numerical
	consciousness before operations (CBO)	-	-	-	48.663	categorical
	consciousness after operations (CAO)	-	-	-	51.800	categorical
Drug information	coagulant	-	-	-	0.000%	categorical
	coagulant medication time (CMT)	15/1/13	18/12/10	-	50.900	datetime
	anticoagulant	-	-	-	0.000%	categorical
	anticoagulant medication time (AMT)	15/12/22	18/9/27	-	23.400	datetime
Laboratory parameters	white blood cell counts **(WBC)** $(10^9/L)$	2.100	610.200	**14.300±20.200**	20.200	numerical
	neutrophil counts (NEU) $(10^9/L)$	1.200	128.100	12.100±6.900	20.200	numerical
	lymphocyte counts (LYM) $(10^9/L)$	0.009	18.312	1.100±1.000	20.200	numerical
	mononuclear cell counts **(MONO)** $(10^9/L)$	0.010	42.700	**0.600±1.400**	20.200	numerical
	eosinophil counts **(EOS)** $(10^9/L)$	0.004	0.900	**0.040±0.080**	20.200	numerical
	basophil counts (BASO) $(10^9/L)$	0.002	0.410	0.020±0.020	20.200	numerical
	red blood cell counts **(RBC)** $(10^{12}/L)$	0.002	131.000	**4.600±6.000**	20.200	numerical
	hemoglobin (HGB) (g/L)	0.180	249.000	128.200±28.700	20.200	numerical
	hematocrit **(HCT)**	0.040	147.000	**2.500±11.000**	20.200	numerical
	mean corpuscular volume (MCV) (fl)	0.300	118.800	89.900±9.800	20.200	numerical
	mean corpuscular hemoglobin content (MCH) (pg)	3.000	353.000	31.000±18.100	20.200	numerical
	mean corpuscular hemoglobin concentration (MCHC) (g/L)	27.200	3285.000	334.000±132.200	20.200	numerical
	red blood cell volume distribution width (RDW) (fl)	4.000	334.000	46.100±16.800	20.200	numerical
	platelet counts (PLT) $(10^9/L)$	0.252	581.000	171.600±79.200	20.500	numerical
	high-sensitivity C-reactive protein **(sCRP)** (mg/L)	0.500	458.000	**21.900±46.600**	74.200	numerical
	prothrombin time (PT) (s)	1.100	81.900	12.300±4.100	22.300	numerical
	PT(%)	6.200	190.200	87.900±24.900	22.300	numerical
	prothrombin international normalized ratio (PT-INR)	0.090	58.300	1.200±0.700	22.300	numerical
	activated partial thromboplastin time (APTT) (s)	0.800	116.400	27.700±8.600	23.200	numerical
	fibrinogen concentration (FIB) (g/L)	0.090	30.900	2.900±1.800	23.400	numerical
	thrombin time (TT) (s)	0.010	515.600	19.100±17.300	23.100	numerical
	D-Dimer$(mg/LFEU)$	0.030	340.000	**15.600±33.800**	25.700	numerical
	fibrinogen degradation product **(FDP)** (Ug/mL)	0.100	1662.000	**51.700±133.700**	25.500	numerical

TABLE 2. Table 1 continued.

Category	Features (abbreviation)(unit or format)	Min	Max	Mean±std	Missing(%)	Types
Laboratory parameters	neutrophil counts ratio **(NEU-R)** $(10^9/L)$	0.009	98.000	**4.500±17.400**	20.200	numerical
	basophil counts ratio **(BASO-R)** $(10^9/L)$	0.007	4.080	**0.010±0.150**	20.200	numerical
	eosinophil counts ratio **(EOS-R)** $(10^9/L)$	0.001	1.400	**0.010±0.080**	20.200	numerical
	mononuclear cell counts ratio **(MONO-R)** $(10^9/L)$	0.001	11.500	**0.200±0.900**	20.200	numerical
	lymphocyte counts ratio **(LYM-R)** $(10^9/L)$	0.007	26.200	**0.400±1.700**	20.200	numerical

std : it represents standard deviation.
Missing : it represents the proportion of missing values in all patients.

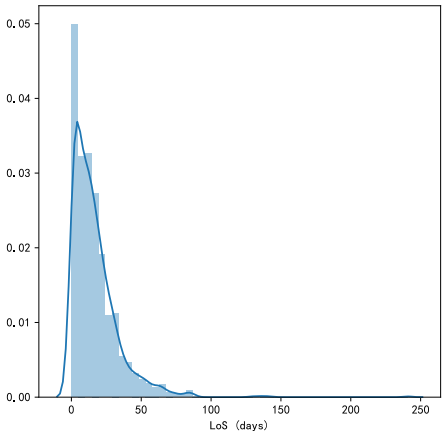

FIGURE 2. Distribution of LoS.

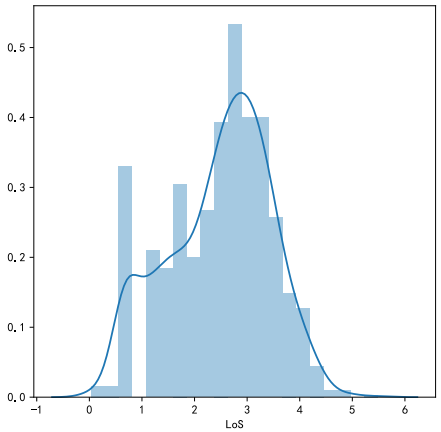

FIGURE 3. Distribution of LoS after Box-Cox transformation.

transformation method.

$$y = \begin{cases} \dfrac{x^{\alpha} - 1}{\lambda}, & \alpha \neq 0 \\ ln(x), & \alpha > 0 \end{cases} \tag{1}$$

x represents continuous positive data, y represents the transformed data, α represents transformation parameter, the transformation is a log transformation when $\alpha = 0$, a reciprocal transformation when $\alpha = -1$, and a square root transformation when $\alpha = 0.5$. There are two methods for estimating the parameter α in the Box-Cox transformation: (1) maximum likelihood estimation, and (2) Bayesian method. This paper implements this transformation by python package. Figure 3 shows the distribution of LoS after Box-Cox transformation, which approximates a normal distribution.

2) DEFINITION OF SURVIVORS AND NON-SURVIVORS

This study defines survivors and non-survivors by their consciousness after discharge from the ICU. If the consciousness after discharge is death, it is non-survivors. Moreover, the consciousness of patients after discharge is deep coma and patients whose family members signed waiver of informed consent are also classified as non-survivors according to clinical experience. Conversely, when the consciousness after discharge is drowsiness, awake, lethargy, light coma and unconscious, the study classifies them as survivors. As a result, the paper gets 925 cases of survivors and 289 cases of non-survivors.

B. FEATURE ENGINEER

Feature engineer is often mentioned as one of the most important steps in machine learning. It addresses the problem of attaining the most informative and compact set of features to improve the performance of machine learning models.

1) DATA PREPROCESSING

In the real world, data is usually incomplete and inconsistent, which can not be directly used for data mining, or the results are unsatisfactory [21]. In order to improve the quality of data mining, data preprocessing technology has been developed. There are many methods of data preprocessing: data cleaning, data integration, data transformation, data reduction and so on. These data processing technologies are used before data mining, which greatly improves the quality of data mining and reduces the time needed for actual mining [21].

1 Missing data
Most studies on ICU LoS prediction have not mentioned the handling of missing values [13], [22]–[24], even though the missing values processing is mentioned, most of them directly exclude patients with missing data [25]. But in this study, different methods are used according to the proportion of missing values. For features of SL (5.000%), TOA (5.000%), HIL (5.000%), since there are few missing values, this study uses the mode of this feature to fill in the missing values. For features of CBT (44.392%), CAT (46.424%), CBO (48.663%), CAO (51.800%), ASA (49.243%), BLO (56.141%), DAO (47.100%), CMT (50.900%)

and laboratory parameters (20.200%-25.700%) except sCRP (74.200%), the missing values are relatively large. In order to make full use of the collected data, the missing values are regarded as a feature in this study. For numerical features, 0 is used to fill in the missing values, and for categorical features,'None' is used to fill in the missing values. Particularly, considering that the higher the level of operation, the longer the duration of the operation, so this study fills in the missing values of DUO (50.900%) according to SL. Specifically, the DUO is divided into 5 groups according to the SL, and then the missing value of the DUO is filled with the average value of the corresponding subgroup. For features of sCRP (74.200%), RBCT (86.550%), PTT (97.752%), PLAT (94.662%), TBT (99.892%), ABT (98.292%), ALT (97.223%), RBCO (100%), PLTO (100.000%), PLAO (100.000%), TBO (99.800%), ABTO (100%) and ALO (100%), the paper excludes these 13 features that are excessive missing values to avoid negative impact on the actual data distribution. Moreover, for the text features, including IND (25.900%), DR, NAO (47.000%), coagulant and anticoagulant or datetime features, including AT, DT, DAO (47.100%), CMT (50.900%) and AMT (23.400%), these features are not applied to the the proposed model. Hence, there are still 49 features left.

2 Box-Cox transformation
Most of the numerical ICU data in this study presents skewed distribution, which would affect the final performance of the model. Therefore, Box-Cox technology is once again introduced to this study to correct it. In other words, Box-Cox transformation is performed on all features with skewness coefficient greater than 0.5, including MONO, PT-INR, BASO-R, TT, RBC, EOS-R, NLR, MCH, LYM-R, MONO-R, BASO, MCHC,P-FDP, LYM, NEU-R, D-DIMER, EOS I24H, NEU, RDW, FIB, PLR, PT, TC, NO and PLT.

3 Normalization
Considering that the dimension between numerical features (including age, WBC and NEU, etc.) is different and the range of ICU data varies greatly, it is necessary to normalize the original ICU data. The Z-score method shown in Eq.(2) is used to normalize the ICU data in this study. In this way, the obtained ICU data conforms to the standard normal distribution.

$$xp_i = \frac{x_i - \mu}{\sigma} \tag{2}$$

where xp_i represents normalized ICU data of the i-th feature and x_i represents the raw ICU data of the i-th feature, μ represents the mean of all ICU data, and σ represents the standard deviation of all ICU data.

4 One-hot encoding
It can be seen from Table 1 that the ICU data contained mixed data types including numerical measurements, binary flags, and text fields for admission diagnosis.

TABLE 3. One-hot encoding.

Original Blood Type	One-hot Encoding				
	'O'	'A'	'B'	'AB'	'Not checked'
'O'	1	0	0	0	0
'A'	0	1	0	0	0
'B'	0	0	1	0	0
'AB'	0	0	0	1	0
'Not checked'	0	0	0	0	1

To handle these mixed data types, one-hot encoding is used to encode a categorical feature with k possible values to k features, where the feature representing the corresponding categorical feature has a value of 1, and all other resulting features have values of 0. Table 3 shows the change of BT state using one-hot encoding. When the categorical feature BT is'O'-type blood, the feature is expressed as (1,0,0,0,0) after one-hot encoding and becomes a numerical feature available for the regression model. Moreover, the features are also expanded. For example, the categorical feature BT is extended to five binary blood type features, namely 'A'-type blood, 'B'-type blood, 'AB'-type blood, 'O'-type blood and 'Not checked'-type blood.

2) EXPLORATORY DATA ANALYSIS

Since exploratory data analysis (EDA) does not need to make assumptions of the ICU data, it explores the structure and pattern of the ICU data by mapping, tabulating, equation fitting, etc so that the structure and pattern of the data can be observed most realistically and directly [26].

1 Exploration of new features
In order to improve the prediction performance of ICU LoS, this study tries to explore some new features based on the original ICU data. According to the study of Yoldas *et al.* [27], neutrophil to lymphocyte ratio (NLR) and platelet to lymphocyte ratio (PLR) have significant effect in predicting LoS and mortality of ICU patients, so this study also extracts these two features. In addition, intuitively, this study also extracts the following features: admission season (AS), number of operations (NO), interval between admission time and operation time (IAOT), interval between operation 1 time and operation 2 time is more than 24 hours or not (I24H), operation or not (ON), types of coagulants (TC), types of anticoagulants (TA), take anticoagulants and coagulants at the same time or not (TAC). Table 4 presents some basic statistical information of the extracted new features. Until now, there are 59 original features.

2 Relationship between features
Because there may be correlation or collinearity between features, it will bring negative impact on regression model, and then affect the performance of the model. In this study, the Apriori algorithm [28] is used to analyze the associations between different categorical features, including gender, AS, SL, TOA, HIL, coagulant, anticoagulant and CAD, etc. Table 5 shows the

TABLE 4. Statistical information of extracted new features.

Features	Min	Max	Mean±std	Unique	Types
NLR	0.01	506.3	17.1±22.4	746	numerical
PLR	0.14	1500	223.7±171.4	740	numerical
AS	-	-	-	4	categorical
NO	0	3	0.6±0.7	4	numerical
IAOT	0	61	2±9	26	numerical
I24H	-	-	-	2	binary
ON	-	-	-	2	binary
TC	0	2	0.6±0.6	3	categorical
TA	0	3	1±0.8	4	categorical
TAC	-	-	-	2	binary

TABLE 5. Association rules between categorical features.

Rules	Confidence
TOA('general anesthesia')==>HIL('I/A')	0.865
SL('level 4')==>HIL('I/A')	0.883
CBO('coma')==>TOA('general anesthesia')	0.967
SL('level 4')==>TOA('general anesthesia')	0.992

==> : The left side of this symbol indicates antecedent of the association rule, and the right side indicates consequent of the association rule.

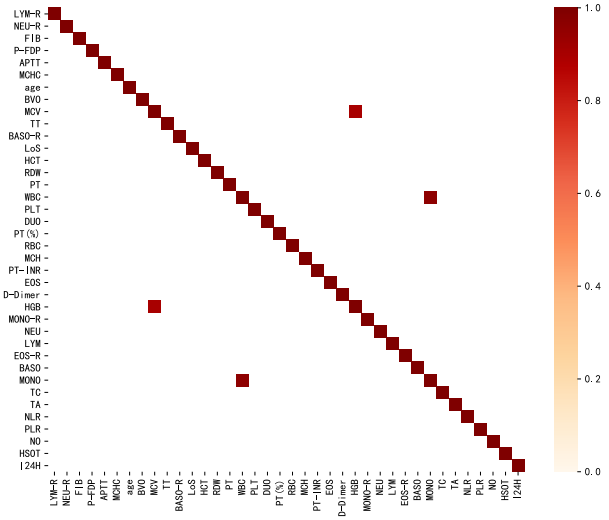

FIGURE 4. Pearson correlation between continuous features.

association rules for minimum support of 0.2 and minimum confidence of 0.8. According to Table 5, these association rules can give doctors some guidance from the data perspective. For example, general anesthesia may be a better option when a patient needs to have a level 4 operation or is in a coma before operation. However, It is worth emphasizing that these are all analyzed from the perspective of data, and more specific operations should be based on clinical practice. Moreover, the Pearson method [29] is used to calculate the correlation between continuous features, including age, WBC, and NEC, etc. Figure 4 visually shows the correlation between continuity features, and Pearson's coefficient over 0.9 is highlighted by deep colors, which are HGB and MCV, and WBC and MONO. In order to avoid the negative effects of collinearity on the model, this study removes MCV and WBC.

Moreover, considering that one of the aim of this study is to predict LoS, the paper removes relevant features after discharge, such as DD and CAD. Finally, the paper uses trial and error method to determine the 45 features used in the proposed model, namely age, gender, RON, TON, AD, CA, CBT, CAT, HIL, PLR, NLR, AS, NO, TOA, TC, TA, TAC, CBO, ASA, BD, PT, NEU, P-FDP, HGB, PT(%), RBC, EOS, FIB, MCH, BASO-R, APTT, MONO, NEU-R, TT, D-DIMER, MONO-R, BASO, MCHC, EOS-R, LYM-R, PT-INR, PLT, LYM, I24H and RDW as the candidate feature set. Furthermore, It is worth noting that one-hot encoding increases the number of features to 125.

C. MODEL FOR PREDICTING LENGTH OF STAY

To estimate expected LoS occurring between ICU admission and hospital discharge, this study developed and validated a LASSO-based model using ICU data for Sichuan Provincial People's Hospital.

1) LASSO ALGORITHM

LASSO is introduced in order to improve the prediction accuracy and interpretability of regression models by altering the model fitting process to select only a subset of the provided features for use in the final model rather than using all of them [30], [31]. Consider a dataset $D = \{(x_1, y_1), (x_2, y_2), ..., (x_n, y_n)\}$,$x \in \mathbb{R}^d$, $y \in \mathbb{R}$, n represents the number of patients in dataset D, and each patient consists of d features and a single outcome, namely LoS. Let y_i be the outcome and $x_i := (x_1, x_2, \ldots, x_d)^T$ be the feature vector for the i-th patient. Then the objective of LASSO is to solve:

$$\min_{w} \sum_{i=1}^{n} (y_i - w^T x_i)^2 + \lambda ||w||_1 \tag{3}$$

where w represents the weight of features, and the larger the weight, the greater the impact on LoS. λ represents the regularization parameter and is greater than 0. In general, the larger λ, the less likely the model is to overfit, but it may not be able to capture some useful information in the data. In fact, the parameter λ can effectively choose a simpler model that does not include those coefficients set to 0 by forcing the sum of the absolute value of the regression coefficients to be less than a fixed value, which is beneficial to enhance the interpretability of the model.

In this study, to develop and validate the LASSO-based model for predicting ICU LoS, the patient population is randomized into 2 groups. 70% of patients (849 patients) is used to create predictive models and cross validation, and the remaining 30% (365 patients) is used to externally validate the proposed model. Figure 5 shows a schematic diagram of the ICU LoS prediction model.

The proposed model includes the FeatureSet, LASSO algorithm and Box-Cox transformation. FeatureSet, a total of 125 discrete features are obtained by data preprocessing and EDA. LASSO algorithm, λ is determined as 1.0 and the maximum number of iterations is 1000 by grid

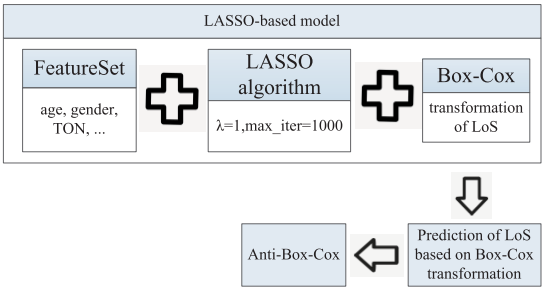

FIGURE 5. Schematic diagram of proposed model.

search method. Box-Cox transformation, which transforms LoS into Box-Cox form. In addition, anti-Box-Cox of the expected ICU LoS is applied when the performance of proposed model is assessed in the 10-fold cross validation and external validation method.

2) EVALUATION CRITERION

In order to evaluate proposed model, this paper uses 10-fold cross validation and external validation method. In 10-fold cross validation, the processed data are randomly partitioned into 10 similar-sized subsets, and each subset is mutually exclusive. Among the 10 subsets, a single subset is retained as the validation data for testing the model, and the remaining 9 subsets are used as training data. The cross validation process is then repeated 10 times, with each of the 10 subsets used exactly once as the validation data. The 10 results are then averaged to produce a single estimation. The advantage of this method over repeated random subsets is that all observations are used for both training and validation, and each observation is used for validation exactly once. In general, a main advantage of the 10-fold cross validation evaluation method is that it has a lower variance than a hold-out method. In external validation, the paper uses the data that does not appear in the training set to verify the performance of the model, which can better reflect the scalability of the proposed model.

Basing on the differences between the proposed model's predicted and observed ICU LoS, three measures of predictive performance are introduced to evaluate the proposed model's ability to predict ICU LoS.

1 Root mean squared prediction error (RMSPE)
RMSPE measures the average of the squares of the errors, that is, the average squared difference between the proposed model's predicted value and observed value [32]. It is always non-negative and value closer to 0 are better, and is computed as Eq.(4).

$$RMSPE = \sqrt{\frac{1}{q}\sum_{i=1}^{q}(y_i - \hat{y}_i)^2} \tag{4}$$

where q represents the number of data used for model testing, $\hat{y}_i$ is the value of the i-th sample proposed by predicted model and y_i is the corresponding observed value.

2 Mean absolute error (MAE)
Because of the skewness of the ICU LoS distribution, the RMSPE increases quickly if a long LoS are erroneously predicted to be short or vice versa. Therefore, this paper presents the MAE, which does not have this limitation. MAE is the average of the absolute value of the difference between the proposed model's predicted value and observed value [33]. It is always non-negative and value closer to 0 are better, and is computed as Eq.(5).

$$MAE = \frac{1}{q}\sum_{i=1}^{q}|y_i - \hat{y}_i| \tag{5}$$

3 Coefficient of determination (R^2)
R^2 provides a measure of how well future samples are likely to be predicted by the proposed model. It ranges from -1 to 1, where higher value correspond to better predictions [34]. Eq.(6) presents its calculation method.

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \overline{y_i})^2} \tag{6}$$

where n represents the number of all data and $\overline{y_i}$ represents the average LoS of all patients.

III. RESULTS

All statistical analyses and proposed methods are implemented using python 3.5 with a workstation: Intel(R) Core (TM) i5-8400 CPU @ 2.58 GHz and 8 GB of RAM.

TABLE 6. Top-15 features for predicting the ICU LoS.

Features	Coefficients
TON_No	-2.8e-01
TA	2.2e-01
NO	1.8e-01
TAC_No	7.2e-02
AD_NICU	7.1e-02
CA_lethargy	5.5e-02
CA_coma	-5.2e-02
CAT_deep coma	-4.9e-02
CAT_lethargy	4.8e-02
APTT	-4.7e-02
LYM	-4.6e-02
AD_rehabilitation	-4.2e-02
age	-3.5e-02
I24H	3.0e-02
TOA_No	2.7e-02

A. PATIENT FEATURES AND OUTCOMES

This study removed 87 features from the block of patient features by LASSO-based model and used the remaining 38 features to predict LoS. Table 6 presents the Top-15 patient features used in the proposed model and Figure 6 shows the order of importance of all features affecting LoS. Clearly, Tables 6 and Figure 6 show that TON, TA, NO, TAC, CA, and CAT have significant effects on LoS. For TON, if it is'No', it has a negative association with LoS. Otherwise, for TA, it has a positive association with LoS.

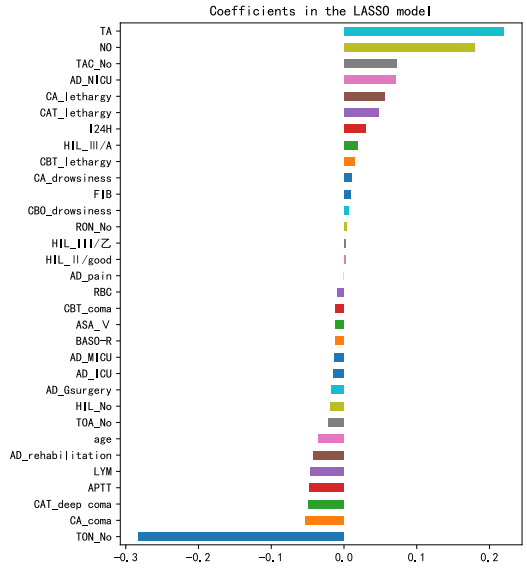

FIGURE 6. The order of importance of all features for predicting the ICU LoS.

TABLE 7. The Performance of proposed model evaluated using 10-fold cross validation and external validation method.

Populations	RMSPE	MAE	R^2
10-fold cross validation	0.88±0.13	0.87±0.07	0.35±0.09
External validation	0.74±0.12	0.82±0.08	0.33±0.12

Table 7 summarizes the performance of proposed model using 10-fold cross validation with 849 patients and external validation with 365 patients. The results obtained by the 10-fold cross validation method show that the average error between the observed values and the predicted values is within 1 day, which indicates that the proposed model has strong fitting ability for ICU patients. Moreover, the results obtained by the external validation method also show that the proposed model has good generalization performance, so the proposed method can be applied to clinical practice.

Figure 7 graphically displays the relationship between observed and predicted ICU LoS for the 365 patients in the external validation dataset. It can be found that when the LoS is less than 30 days, the proposed model prediction performance is better, but when the LoS is more than 30 days, the prediction performance of the proposed model decreases dramatically. This may be due to the insufficient long-term inpatients in the dataset, resulting in insufficient feature extraction for long-term patients.

B. COMPARISON OF DIFFERENT REGRESSION METHODS

In order to verify the validity and superiority of proposed model, this study compares it with other models commonly used in ICU LoS prediction [7], [13], [17], [35]–[38]. Table 8 presents a comparison on the results between the proposed model in this paper and other methods proposed in the previous literature. From Table 8, it can be found that most studies have lower R^2 and most of the validation methods are simple hold-out method, which makes the results random to some extent. In the previous literature, Niskanen [36] *et al.* obtained a best R^2 with 0.28, but this study obtained a better R^2 with 0.35. To the best knowledge of authors, this study achieved the best R^2 across patients. However, technically, R^2 obtained by the proposed model is not very high. In fact, when using cross-patient data to predict ICU LoS, R^2 is mostly concentrated in 0.05 to 0.28, while when using cross-ICU data to predict ICU LoS, R^2 is mostly concentrated in 0.43 to 0.64, which indicates that the data used in this study limits the further improvement of R^2. Moreover, Figure 7 also indicates that when the LoS is longer than 30 days, the predictive performance of the proposed model decreases dramatically, which also limits the improvement of R^2. However, the RMSPE and MAE obtained by the proposed model have significant promising results. RMSPE is about 4 days lower than the study of Verburg *et al.* [13], and MAE is about 2 days and 1 days lower than the studies of Verburg *et al.* [13] and Balkan and Subbian [39], respectively. It shows that the proposed model have substantial advantages over other LoS prediction models.

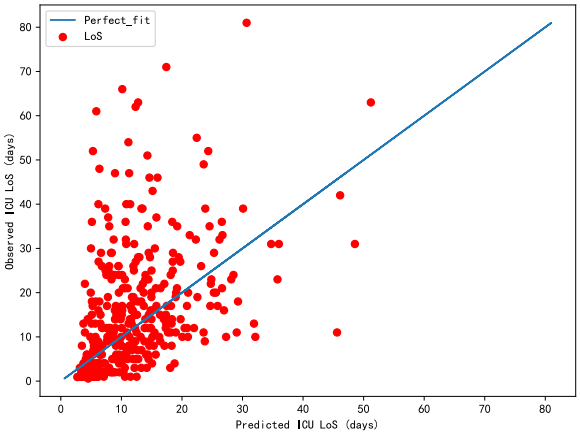

FIGURE 7. Calibration curve comparing observed and predicted ICU LoS for the 365 patients in the external validation dataset. Straight line indicates perfect predictive ability. The red dots represent ICU LoS.

C. EXPLORATION OF LENGTH OF STAY RISK FACTORS IN DIFFERENT POPULATIONS

In this paper, this study not only estimates the performance of the proposed model all patients, but also the subgroups of patients, namely, survivors and non-survivors. Moreover, this study also explores the risk factors for LoS in different subgroups. According to the definition of survivors and non-survivors, 925 survivors and 289 non-survivors are obtained. The survivors and non-survivors are used to construct corresponding models according Figure 5.

Table 9 presents the performance of separate models developed for survivors and non-survivors. In general, models built using ICU survivors and ICU non-survivors performed worse than models built using all patients in R^2, which may be limited by the amount of data we studied, especially in non-survivirs. For survivors, the value for RMSPE is about 0.87 day, for MAE about 0.86 day and for R^2 about 0.26.

TABLE 8. A Comparison of the performance obtained by proposed methods and others' works.

References	Model Validation			Model Performance		
	Models	Validation method	No. of patients	RMSPE	MAE	R^2
Zimmerman et al.(2006) [7]	APACHE IV	Simple random sample of 40%	46,517	-	-	0.22
Moran et al.(2008) [35]	random effects model	Random sample(pps) of 20% stratified by year of admission	44,625	-	-	0.18
Vasilevskis et al.(2009) [17]	APACHE IV, MPM III, SAPS II	Simple random sample of 40%	4,611	-	-	0.20/0.10/0.05
Niskanen et al.(2009) [36]	linear regression	Simple random sample of 40%	25,586	-	-	0.28
Kramer et al.(2010) [37]	multivariable regression model	Simple random sample of 50% and different time period	12,904	-	-	0.18
Verburg et al.(2014) [13]	OLS regression log(LoS), GLM, etc.	Bootstrap	32,667	5.15-8.74	3.00-4.10	0.09-0.21
Balkan et al.(2018) [39]	APACHEIV, APACHE IVa	Simple random sample of 33%	44,774	-	2.285/2.491	0.136/0.069
Proposed model.	LASSO	10-fold cross validation and simple random sample of 30%	1,214	**0.88±0.13**	**0.87±0.07**	**0.35±0.09**

APACHE: a recalibrated acute physiology and chronic health evaluation.
pps: probability proportional to size sampling.
MPM0 III: mortality probability model III at zero hours.
SAPS II: simplified acute physiology score II.
OLS: ordinary least square.
GLM: general linear model.

TABLE 9. The performance of proposed model evaluated using 10-fold cross validation in ICU survivors and ICU non-survivors.

Populations	RMSPE	MAE	R^2
Survivors	0.87±0.11	0.86±0.07	0.26±0.06
Non-survivors	0.87±0.16	0.90±0.08	0.26±0.10
All patients	0.88±0.13	0.87±0.07	0.35±0.09

For non-survivors, the value for RMSPE is about 0.87 day, for MAE about 0.90 day and for R^2 about 0.26. Compared survivors and non-survivors, it can be found that the volatility of evaluation parameters of non-survivors is greater than that of survivors. Similarly, it may be limited by the amount of non-survivors.

To explore this further, the study compares the risk factors affecting ICU LoS of survivors and non-survivors. The paper ranks the coefficients of LASSO-based model for survivors and find that the top-15 risk factors affecting survivors are TON_No, TA, AD_NICU, I24H, CAT_lethargy, AD_ICU, NO, ASA_IV, CA_lethargy, CBT_coma, CA_coma, CAT_deep coma, CBO_NO, TAC_NO, and HGB, while the top-15 risk factors affecting non-survivors are TA, NEU, HIL_No, NO, TT, TON_No, CA_deep coma, LYM, CBT_awake, PT, BASO, CA_coma, RBC, HIL_good and AD_NICU. Comparing the risk factors of survivors to non-survivors, it can be found that TA, NO, TON and CA have a significant impact on both survivors and non-survivors, but there are some differences in the influencing factors of the two sub-populations. For survivors, AD, I24H, CAT and ASA have a greater impact on them but have little effect on non-survivors. For non-survivors, NEU, TT, LYM and PT have a greater impact on them but have little effect on survivors.

IV. DISCUSSION

The main contribution of this study is to construct a highly accurate ICU LoS prediction model and to explore the risk factors of different population in depth. The prediction model in this study has several advantages. Firstly, the model is characterized by small number of features (only 22 original features), yet provides an accurate performance of the ICU LoS. Moreover, small number of features can improve interpretability of proposed models. Secondly, in order to make full use of the data obtained, this study does not restrict the data like other studies [15], [17], [18], which increases the scalability of the proposed models. Thirdly, Box-Cox technology is introduced to transform the significantly skewed ICU LoS into an approximate normal distribution, which greatly improves the prediction performance. Fourthly, new features such as TA and NO have a significant contribution to the prediction of ICU LoS. Whether TA and NO were added, the performance of proposed model are: RMSPE is 0.95±0.14 day vs 0.88±0.13 day, MAE is 0.98±0.08 day vs 0.87±0.07 day, R^2 is 0.28±0.10 vs 0.35±0.09. Clearly, the new features TA and NO can significantly improve the prediction of ICU LoS. Fifthly, this study uses different methods to deal with missing values according to the proportion of missing values to improve the predictive performance of the model, which avoids directly removing data in order to make full use of limited data. Sixthly, this study used the Apriori algorithm to analyze the association of categorical features. The results of the association analysis can be used as a guidance for clinicians. Furthermore, this study has a main strength over previous reviews of ICU LoS. Previously, most researchers have not conducted an in-depth study on the risk

factors of ICU LoS in different populations [40]. However, this study conducted an in-depth analysis of the survivors and non-survivors. For survivors, AD, I24H, CAT and ASA have a greater impact on them but have little effect on non-survivors. For non-survivors, NEU, TT, LYM and PT have a greater impact on them but have little effect on survivors. Moreover, this study also shows that using a new set of features may improve the predictive performance of ICU LoS which is confirmed by Table 8.

However, there are several limitations of the current study that need to be recognized. Firstly, this study deals with a single organization (Sichuan Provincial People's Hospital) instead of several organizations, which may not be reflective and may also restrict generalizability. This is because international differences in hospital and ICU structure, management, and patient care are likely to have an adverse impact on predictive accuracy. Moreover, compared with other studies [7], [13], [17], [35]–[38], the data used in this study is smaller. Secondly, this study has not incorporated ICU-related features such as the number of hospital/ICU beds, full-time equivalent ICU nurses and types of care protocol into the model, which may help improve the performance of the model and make it possible to plan ICU resources more rationally. Thirdly, we are uncertain about the impact of missing values on ICU LoS prediction. Fourthly, the proposed model does not achieve an ideal performance for predicting long-term LoS on the ICU according to Figure 7. In addition, Figure 2 also shows that the LoS in this study has a distribution of about 10% within one day, which improves the performance of the proposed model to some extent, because the proposed model predicts well in short-term LoS. To illustrate this, all patients with LoS for one day are removed and the proposed model obtained 0.58±0.05 day for RMSPE, 0.73±0.04 day for MAE, and 0.26±0.07 for R^2. Compared with the prediction model using all patients, although the performance of RMSPE and MAE is better, the performance of R^2 is poor, so the performance of the proposed model is improved to some extent by short-term ICU patients.

In fact, constructing a good model for ICU LoS requires rich data, specialized statistical and clinical knowledge. Although this study supports and extends the previous studies, it can still carry out further study in the following aspects in the future. Firstly, collecting more types and greater amount data, such as different hospitals, different regions and even different races, which can increase the reliability of the model's performance. Secondly, it may be possible to make more detailed ICU LoS prediction based on different types of disease or operation, which may further improve the predictive performance of the model. Thirdly, adding some ICU-related features such as the number of hospital/ICU beds, full-time equivalent ICU nurses and types of care protocol, or constructing new features based on clinical practice, which may also improve the performance of the model. Furthermore, this study is poor in predicting long-term LoS, and future work can focus on improving the predictive performance of long-term LoS.

V. CONCLUSION

This study undertakes an prediction of ICU LoS and exploration of risk factors for different populations in depth. On the one hand, this study has created a highly accurate predictive model of ICU LoS based on the LASSO algorithm, in which Box-Cox technology and new features TA and NO have made significant contributions to the proposed model. On the other hand, exploration of survivors and non-survivors also demonstrates that there are also differences in the risk factors affecting ICU LoS in survivors and non-survivors. Therefore, this study and the successful application of LASSO algorithm in real-world data shows that proposed model has the potential to improve the quality of care and resource use in ICU wards.

ETHICAL STANDARDS

This study has been approved by Sichuan Academy of Medical Science & Sichuan Provincial People's Hospital and registered in Chinese Clinical Trail Registry (ChiCTR1900020726). All patients gave their informed consent prior to their inclusion in the study.

ACKNOWLEDGMENT

(Chunling Li, Longyi Chen, and Jie Feng contributed equally to this work.)

REFERENCES

[1] M. O. Harhay, J. Wagner, S. J. Ratcliffe, R. S. Bronheim, A. Gopal, S. Green, E. Cooney, M. E. Mikkelsen, M. P. Kerlin, D. S. Small, and S. D. Halpern, "Outcomes and statistical power in adult critical care randomized trials," *Amer. J. Respiratory Crit. Care Med.*, vol. 189, no. 12, pp. 1469–1478, Jun. 2014. doi: 10.1164/rccm.201401-0056CP.

[2] N. A. Halpern and S. M. Pastores, "Critical care medicine beds, use, occupancy and costs in the united states: A methodological review," *Critical Care Med.*, vol. 43, no. 11, pp. 2452–2459, Nov. 2015. doi: 10.1097/CCM.0000000000001227.

[3] N. A. Halpern, S. M. Pastores, and R. J. Greenstein, "Critical care medicine in the United States 1985–2000: An analysis of bed numbers, use, and costs," *Crit. Care Med.*, vol. 32, no. 6, pp. 1254–1259, Jun. 2004. doi: 10.1097/01.ccm.0000128577.31689.4c.

[4] J. M. Kahn, G. D. Rubenfeld, J. Rohrbach, and B. D. Fuchs, "Cost savings attributable to reductions in intensive care unit length of stay for mechanically ventilated patients," *Med. Care*, vol. 6, no. 12, pp. 1226–1233, Dec. 2008. doi: 10.1097/mlr.0b013e31817d9342.

[5] W. Lin, S. D. Halpern, M. P. Kerlin, and D. S. Small, "A 'placement of death' approach for studies of treatment effects on ICU length of stay," *Stat. Methods Med. Res.*, vol. 26, no. 1, pp. 292–311, Feb. 2017. doi: 10.1177/0962280214545121.

[6] W. A. Knaus, E. A. Draper, D. P. Wagner, and J. E. Zimmerman, "APACHE II: A severity of disease classification system," *Crit. Care Med.*, vol. 13, no. 10, pp. 818–829, Oct. 1985. doi: 10.1097/00003246-198510000-00009.

[7] J. E. Zimmerman, A. A. Kramer, D. S. McNair, and F. M. Malila, "Acute physiology and chronic health evaluation (APACHE) IV: Hospital mortality assessment for today's critically ill patients," *Crit. Care Med.*, vol. 34, no. 5, pp. 1297–1310, May 2006. doi: 10.1097/01.ccm.0000215112.84523.f0.

[8] J.-R. Le Gall, S. Lemeshow, and F. Saulnier, "A new simplified acute physiology score (SAPS II) based on a European/North American multicenter study," *JAMA*, vol. 270, no. 24, pp. 2957–2963, Dec. 1993. doi: 10.1001/jama.1993.03510240069035.

[9] J. E. Zimmerman, A. A. Kramer, D. S. McNair, F. M. Malila, and V. L. Shaffer, "Intensive care unit length of stay: Benchmarking based on acute physiology and chronic health evaluation (APACHE) IV," *Crit. Care Med.*, vol. 34, no. 10, pp. 2517–2529, Oct. 2006. doi: 10.1097/01.ccm.0000240233.01711.d9.

[10] J. E. Zimmerman and A. A. Kramer, "Outcome prediction in critical care: The acute physiology and chronic health evaluation models," *Current Opinion Crit. care*, vol. 14, no. 5, pp. 491–497, Oct. 2008. doi: 10.1097/mcc.0b013e32830864c0.

[11] J. Rapoport, D. Teres, Y. Zhao, and S. Lemeshow, "Length of stay data as a guide to hospital economic performance for ICU patients," *Med. Care*, vol. 41, no. 3, pp. 386–397, Mar. 2003. doi: 10.1097/01.MLR.0000053021.93198.96.

[12] B. H. Nathanson, T. L. Higgins, D. Teres, W. S. Copes, A. Kramer, and M. Stark, "A revised method to assess intensive care unit clinical performance and resource utilization," *Crit. Care Med.*, vol. 35, no. 8, pp. 1853–1862, Aug. 2007. doi: 10.1097/01.ccm.0000275272.57237.53.

[13] I. W. M. Verburg, N. F. de Keizer, E. de Jonge, and N. Peek, "Comparison of regression methods for modeling intensive care length of stay," *PLoS ONE*, vol. 9, no. 10, Oct. 2014, Art. no. e109684. doi: 10.1371/journal.pone.0109684.

[14] I. W. M. Verburg, A. Atashi, S. Eslami, R. Holman, A. Abu-Hanna, E. de Jonge, N. Peek, and N. F. de Keizer, "Which models can i use to predict adult ICU length of stay? A systematic review," *Crit. Care Med.*, vol. 45, no. 2, pp. e222–e231, Feb. 2017. doi: 10.1097/CCM.0000000000002054.

[15] T. A. Daghistani, R. Elshawi, S. Sakr, A. M. Ahmed, A. Al-Thwayee, and M. H. Al-Mallah, "Predictors of in-hospital length of stay among cardiac patients: A machine learning approach," *Int. J. cardiology*, vol. 288, pp. 140–147, Aug. 2019. doi: 10.1016/j.ijcard.2019.01.046.

[16] O. V. Hein, J. Birnbaum, K. Wernecke, M. England, W. Konertz, and C. Spies, "Prolonged intensive care unit stay in cardiac surgery: Risk factors and long-term-survival," *Ann. Thoracic Surg.*, vol. 81, no. 3, pp. 880–885, Mar. 2006. doi: 10.1016/j.athoracsur.2005.09.077.

[17] E. E. Vasilevskis, M. W. Kuzniewicz, B. A. Cason, R. K. Lane, M. L. Dean, T. Clay, D. J. Rennie, E. Vittinghoff, and R. A. Dudley, "Mortality probability model III and simplified acute physiology score II: Assessing their value in predicting length of stay and comparison to APACHE IV," *Chest*, vol. 136, no. 1, pp. 89–101, Jul. 2009. doi: 10.1378/chest.08-2591.

[18] A. A. Kramer and J. E. Zimmerman, "The relationship between hospital and intensive care unit length of stay," *Crit. Care Med.*, vol. 39, no. 5, pp. 1015–1022, May 2011. doi: 10.1097/ccm.0b013e31820eabab.

[19] W. G. Manning, A. Basu, and J. Mullahy, "Generalized modeling approaches to risk adjustment of skewed outcomes data," *J. Health Econ.*, vol. 24, no. 3, pp. 465–488, May 2005. doi: 10.1016/j.jhealeco.2004.09.011.

[20] R. C. Langan, C.-C. Huang, W. R. Mao, K. Harris, W. Chapman, C. Fehring, K. Oza, P. G. Jackson, R. Jha, N. Haddad, J. Carroll, A. Parker, W. B. Al-Refaie, and L. B. Johnson, "Pancreaticoduodenectomy hospital resource utilization in octogenarians," *Amer. J. Surg.*, vol. 211, no. 1, pp. 70–75, Jan. 2016. doi: 10.1016/j.amjsurg.2015.04.014.

[21] M. Kantardzic, *Data Mining: Concepts, Models, Methods, and Algorithms*. Hoboken, NJ, USA: Wiley, 2011. doi: 10.1109/9780470544341.

[22] M. Najafi, H. Goodarzynejad, M. Sheikhfathollahi, and H. Adibi, "Role of surgeon in length of stay in ICU after cardiac bypass surgery," *J. Tehran Heart Center*, vol. 5, no. 1, pp. 9–13, 2010. [Online]. Available: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3466845/

[23] M. Najafi and H. Goodarzynejad, "Determinants of length of stay in surgical ward after coronary bypass surgery: Glycosylated hemoglobin as a predictor in all patients, diabetic or non-diabetic," *J. Tehran Univ. Heart Center*, vol. 7, no. 4, pp. 170–176, 2012. [Online]. Available: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3537203/

[24] K. Meadows, R. Gibbens, C. Gerrard, and A. Vuylsteke, "Prediction of patient length of stay on the intensive care unit following cardiac surgery: A logistic regression analysis based on the cardiac operative mortality risk calculator, EuroSCORE," *J. Cardiothoracic Vascular Anesthesia*, vol. 32, no. 6, pp. 2676–2682, Dec. 2018. doi: 10.1053/j.jvca.2018.03.007.

[25] T. L. Higgins, N. J. Starr, J.-C. Lee, G. J. Beck, and F. G. Estafanous, "Predicting prolonged intensive care unit length-of-stay following coronary artery bypass surgery," *Clin. Intensive Care*, vol. 10, no. 5, pp. 175–182, Oct. 1999. doi: 10.3109/tcic.10.5.175.182.

[26] S. Shekhar and H. Xiong, "Exploratory data analysis," in *Encyclopedia of GIS*, 1st ed. Boston, MA, USA: Springer, 2008, p. 297. doi: 10.1007/978-0-387-35973-1_394.

[27] H. Yoldas, I. Karagoz, M. N. Ogun, Y. Velioglu, I. Yildiz, M. Bilgi, and A. Demirhan, "Novel mortality markers for critically ILL patients," *J. Intensive Care Med.*, to be published. doi: 10.1177/0885066617753389.

[28] M. Tuchler, A. C. Singer, and R. Koetter, "Minimum mean squared error equalization using a priori information," *IEEE Trans. Signal Process.*, vol. 50, no. 3, pp. 673–683, Mar. 2002. doi: 10.1109/78.984761.

[29] J. Benesty, J. Chen, Y. Huang, and I. Cohen, "Pearson correlation coefficient," in *Noise Reduction in Speech Processing*. Berlin, Germany: Springer, 2009. doi: 10.1007/978-3-211-89836-9_1025.

[30] R. Tibshirani, "Regression shrinkage and selection via the lasso," *J. Roy. Stat. Soc., B (Methodol.)*, vol. 58, no. 1, pp. 267–288, 1996. doi: 10.1111/j.2517-6161.1996.tb02080.x.

[31] F. Santosa and W. W. Symes, "Linear inversion of band-limited reflection seismograms," *SIAM J. Sci. Stat. Comput.*, vol. 7, no. 4, pp. 1307–1330, Jul. 1986. doi: 10.1137/0907087.

[32] D. Harvey, S. Leybourne, and P. Newbold, "Testing the equality of prediction mean squared errors," *Int. J. Forecasting*, vol. 13, no. 2, pp. 281–291, Jun. 1997. doi: 10.1016/S0169-2070(96)00719-4.

[33] T. Chai and R. R. Draxler, "Root mean square error (RMSE) or mean absolute error (MAE)?—Arguments against avoiding RMSE in the literature," *Geosci. Model Develop.*, vol. 7, no. 3, pp. 1247–1250, Jan. 2014. doi: 10.5194/gmd-7-1247-2014.

[34] N. J. D. Nagelkerke, "A note on a general definition of the coefficient of determination," *Biometrika*, vol. 78, no. 3, pp. 691–692, 1991. doi: 10.2307/2337038.

[35] J. L. Moran, P. Bristow, P. J. Solomon, C. George, and G. K. Hart, "Mortality and length-of-stay outcomes, 1993–2003, in the binational Australian and New Zealand intensive care adult patient database," *Crit. Care Med.*, vol. 36, no. 1, pp. 46–61, Jan. 2008. doi: 10.1097/01.ccm.0000295313.08084.58.

[36] M. Niskanen, M. Reinikainen, and V. Pettilä, "Case-mix-adjusted length of stay and mortality in 23 finnish ICUs," *Intensive Care Med.*, vol. 35, no. 6, pp. 1060–1067, Jun. 2009. doi: 10.1007/s00134-008-1377-0.

[37] A. A. Kramer and J. E. Zimmerman, "A predictive model for the early identification of patients at risk for a prolonged intensive care unit length of stay," *BMC Med. Inf. Decis. Making*, vol. 10, no. 1, p. 27, May 2010. doi: 10.1186/1472-6947-10-27.

[38] M. Al Tehewy, M. El Houssinie, N. A. El Ezz, M. Abdelkhalik, and S. El Damaty, "Developing severity adjusted quality measures for intensive care units," *Int. J. Health Care Qual. Assurance*, vol. 23, no. 3, pp. 277–286, 2010. doi: 10.1108/09526861011029343.

[39] B. Balkan, P. Essay, and V. Subbian, "Evaluating ICU clinical severity scoring systems and machine learning applications: Apache IV/IVA case study," in *Proc. 40th Annu. Int. Conf. IEEE Eng. Med. Biol. Soc. (EMBC)*, Jul. 2018, pp. 4073–4076. doi: 10.1109/EMBC.2018.8513324.

[40] V. K. Moitra, C. Guerra, W. T. Linde-Zwirble, and H. Wunsch, "Relationship between ICU length of stay and long-term mortality for elderly ICU survivors," *Crit. Care Med.*, vol. 44, no. 4, pp. 655–662, Apr. 2016. doi: 10.1097/CCM.0000000000001480.

CHUNLING LI received the B.S. degree in clinical medicine and the M.S. degree in internal medicine from Harbin Medical University, in 1997 and 2006, respectively. After graduation, she was with Harbin 242 Hospital and Sichuan Provincial People's Hospital. She studied at the Beijing Union Medical College Hospital, Minster University of Germany. She is the Director of the ICU of neurosurgery with the Sichuan Provincial People's Hospital. She has participated in a medical monograph. She has published more than 20 papers in SCI and core journals in China.

LONGYI CHEN received the B.S. degree in clinical medicine from Chongqing Medical University, in 1987. After graduation, he was involved in neurosurgery with Sichuan Provincial People's Hospital. He studied at PLA General Hospital in clinical study of neurosurgery and involved in neurosurgery at Hospital of Vienna University, Austria. He has participated in the compilation of three monographs, and has published more than 60 articles in domestic and foreign medical journals.

JIE FENG received the B.S. degree in clinical medicine from Luzhou Medical University, in 2014, and the M.S. degree in surgery from Southwest Medical University, in 2017. After graduation, he was with the Neurosurgery Department, Sichuan Provincial people's Hospital.

DUANPO WU received the B.S. degree from the College of Electronics and Information, Hangzhou Dianzi University, in 2009, and the Ph.D. degree from the College of Information Science and Electronic Engineering, Zhejiang University, China, in 2014. Since 2014, he has been a Lecturer with Hangzhou Dianzi University, China. His research interests include intelligent signal processing, biologic data analysis, and machine learning.

ZIMENG WANG received the B.S. degree in electronic information engineering from the Qingdao University of Technology, in 2018. She is currently pursuing the master's degree with Hangzhou Dianzi University. Her research interests include automatic sleep staging, epileptic seizure detection, and prediction.

JUNBIAO LIU received the B.S. degree in electronic communication engineering from Zhejiang University, in 2004, the M.S. degree in control theory and control engineering from Hangzhou Dianzi University, in 2009, and the Ph.D. degree in circuits and systems from Zhejiang University, in 2015. Since 2015, he has been the CTO with Hangzhou Neuro Science and Technology Company Ltd., China. His research interests include signal processing, complex networks, and wireless sensor networks. He is currently the Chair of the National Natural Science Foundation Project.

WEIFENG XU received the B.S. degree from the College of Information Engineering, West Anhui University, in 2015, and the M.S. degree from the Key Laboratory of Advanced Design and Intelligent Computing, Dalian University, Ministry of Education, China, in 2018. Since 2018, he has been a Machine Learning Engineer at Hangzhou Neuro Science and Technology Company Ltd., China. His research interests include biologic data analysis, modeling, and machine learning.