TSINGHUA SCIENCE AND TECHNOLOGY
ISSN 1007-0214 03/22 pp34–45
DOI: 10.26599/TST.2023.9010042
Volume 29, Number 1, February 2024

Lesson Learned from COVID-19 Retrospective Study: An Entropy-Based Clinical-Interpretable Scorecard for Mortality Risk Control at ICU Admission

Chong Yao, Chonghui Huangqi, and Anpeng Huang*

Abstract: With severe acute respiratory syndrome coronavirus 2 spreading globally and causing 2019 coronavirus disease (COVID-19), a challenge that we unprepared for was about how to optimally plan and distribute limited top-medical resources for patients in need of urgent care. To address this challenge, physicians desperately needed a scientific tool to methodically differentiate between cases with varying severity. In this study, the unique data of COVID-19 intensive care unit (ICU) patients provided by the national medical team in Wuhan were classified into discrete and continuous variable types. All continuous data were discretized using an entropy-based method and transformed into serial information margins, in which each information margin is related to a specific symptom or clinical meaning. Finally, all these native and processed discrete data were used to configure a readable scorecard through logistic regression, which is the desired scientific tool aforementioned. A total of 322 ICU patients (age: [median: 64, interquartile range: 54–75], males: 178 [55.28%], and death: 72 [22.36%]) were included in the study. Probabilities of mortality in COVID-19 patients can be evaluated using a scorecard model (calibration slope: 1.343, Brier: 0.048, Dxy = 0.972, and population stability index = 0.071), with desired model performances (accuracy = 0.948, area under curve = 0.99, sensitivity = 1, and specificity = 0.939). This new model can interpret clinical meanings from complex data, and compare it with existing machine learning methods through a black-box mechanism. This new data-information model answers a critical question of how a computing algorithm produces clinically meaningful results that will help physicians logically allocate medical resources for COVID-19 patients. Notably, this tool has limitations, giving that this research is a retrospective study. Hopefully, this tool will be tested further and optimized for adaptation to similar clinical cases in the future.

Key words: COVID-19; scorecard; clinical-interpretable; machine learning; ICU admission control

1 Introduction

The 2019 coronavirus disease (COVID-19) is spreading globally and is known as the "killer" of public health. Until November 27, 2022, more than 645 million confirmed cases and over 6.6 million deaths worldwide have been recorded. How to maximize the utilization of intensive care unit (ICU) medical resources is a critical

• Chong Yao is with the Laboratory of Network Information Security, Beihang University, Beijing 100191, China. E-mail: 1592026301@qq.com.
• Chonghui Huangqi is with Andrew and Erna Viterbi School of Engineering, University of Southern California, Los Angles, CA 90089, USA. E-mail: huangqi@usc.edu.
• Anpeng Huang is with Beijing Goodwill Information and Technology Co., Ltd., and Mobile Health Laboratory, Peking University, Beijing 100871, China. E-mail: hapku@pku.edu.cn.
∗ To whom correspondence should be addressed.
Manuscript received: 2022-11-27; revised: 2023-03-27; accepted: 2023-05-09

issue, especially in areas with medical shortages. A related study will be useful for increasing the efficiency of treatment and resource allocation. More importantly, physicians need a tool to fight future pandemics when ICU admissions increase exponentially, thereby decreasing mortality rates. Driven by this motivation, a scientific tool is needed to help physicians make objective judgments, rather than subjective logical choices. This tool may include several essential components, namely, data preprocessing, computational methods, knowledge discovery, and clinical interpretation. In terms of data preprocessing, Selvan[1], Shang et al.[2], and Overmyer et al.[3] discussed data formalization and cleaning. Regarding survival analysis, related literature addressing logistic models and machine learning algorithms for risk scoring were examined by Zhang et al.[4], Tian et al.[5], and Liang et al.[6]. As for knowledge discovery, clinical information is extremely complex to be extracted in a linear style according to Yan et al.[7], Gao et al.[8], and Yadaw et al.[9]. Lastly, for clinical interpretation, the black-box data-information model of machine learning algorithms cannot be applied in clinical practice, as discussed by Knight et al.[10] and Razavian et al.[11] Therefore, this topic must be studied from a new perspective to efficiently translate data knowledge into clinical information. Taking this into consideration, we developed an entropy-based scorecard model to achieve clinical interpretability. This data-information model can be used to categorize patients depending on case severity with the use of a few key bioparameters. Arguably, this model helps in assigning limited medical resources to patients who possess a higher probability of survival and aids physicians in maximizing ICU resources.

2 Method

2.1 Data resource

This study is a retrospective research. The ICU patient cohort was from the national medical team in Wuhan and contained data from January 26, 2020, to April 6, 2020; the patients were confirmed by oropharyngeal swabs and positive antigen-antibody tests. The data included routine records at admission, comorbidities, and disease severity, along with laboratory test results such as nucleic acid tests, routine blood tests, biochemistry, etc. Each patient's information was recorded from ICU admission to final discharge. All data from patients' electronic medical records were formalized professionally and labeled by clinical experts. All data processing personnel were required to sign a data security and confidentiality agreement to protect patient-doctor confidentiality. For the assessment of this model's effectiveness in different situations, two more external cohorts were applied (one was from Yan et al.[7] study: https://www.nature.com/articles/s42256-020-0180-7; the other was a Peru cohort: https://figshare.com/articles/dataset/Database.xls/13869179).

2.2 Data preprocessing

For data quality control, 322 cases from 345 ICU patients were finally selected as study subjects, with the omission of 23 inconclusive data (i.e., 4 patients transferred out and 19 non-tracking documents). On the first day of admission, information on each patient was collected for clinical and blood biochemical samples, generating 49 separate categories of data entries. For the study of validity and data integrality, an entry may be deleted if its data distribution had low diversity, or more than 20% of patients lacked such an entry. As a result, 30 out of the 49 variables were selected as candidates and treated as input variables of our data-information model. To examine all potential dysfunction or bugs in our models, we selected 97 individuals (30% of patients) as the test data set, and the other 70% of subjects were grouped into the training dataset to optimize the model hyperparameters. The basic requirement was that the test dataset had no voids, and some nulls in the training dataset were filled in advance for data continuity. To avoid additional background noise in the process of null fillings, we applied K-Nearest-Neighbours (KNN) to enhance the resolution[12–15]. Furthermore, data classification processing methods were a part of our models and machine learning algorithms, but they likely relied on their statistical distribution features. For different data, including demographic, clinical, and blood biochemistry characteristics by survival status, the mean (Standard Deviation (SD)) and student t-test were applied to continuous variables following normal distribution; the median (interquartile range (IQR)) and Mann-Whitney U (U-test) were deployed if they did not show a normal distribution. Otherwise, the variables were considered categorical, frequency and chi-square tests were adopted for their distribution profile.

2.3 Model configuration

In this study, the main goal was to conceive a new data-informal model that can avoid the dilemma of

justifying the allocation of resources to patients who are more likely to survive. This dilemma decision cannot be made blindly by a data model without clinical explanation. Currently, most artificial intelligence data models run in a "black-box" style, which cannot present clinical explanations from computing results. To address this challenge above, we conceived an entropy-based scorecard model (Fig. 1) in this study and adopted logistic regression to compute data. Each component of this model is illuminated below.

2.3.1 Variable formalization

As discussed above, the first goal was to determine how to deal with mixed data with continuous and discrete variables. For any continuous variable, a clinical outcome is generally related to a range of data. For example, the level of systolic pressure can be interpreted as normal, hypertension, or hypotension among most people (90–140, above 140, and below 90 mmHg, respectively). Thus, these continuous variables need to be discretized into a sequence of data zones to match each one with a specific symptom or similar clinical outcome. An optimal solution is needed to group these continuous variables into a sequence of data zones. In the optimal solution, how the cutoff endpoints of these continuous variables can be determined for data zones must be resolved. To address this problem, we borrowed the entropy concept, which was originally from physics, and has been successfully applied in many areas, e.g., information theory, credit review for financial loans, binning computing, etc. Entropy, or the information content, is a way to calculate the information gain of different data segment solutions. The calculated information gain values was used to optimize the placement of data zones. This entropy method can avoid potential crosstalk-intermixing between data zones[16].

The function of entropy is listed below:

$$H(X) = -\sum p(k) \times \log p(k) \tag{1}$$

where X is an event (death or living), and $H(X)$ is the entropy value of variable X with the k^{th} data zone within K discrete data zones. $p(k)$ is the probability of the event X during the k^{th} data zone.

After discretization based on entropy, the continuous variable became the ordinal categorical data zone. To code these ordinal data zones, we applied the weight of evidence (WOE)[17, 18] to maintain logic linearity, which can satisfy the logic linearity hypothesis in logistic regression compared with other coding methods (one hot spot/virtual). The WOE of the i^{th} data zone of a categorical variable is defined as follows:

$$WOE_i = \ln \frac{N_{i(\text{nondeath})}/N_{tot(\text{nondeath})}}{N_{i(\text{death})}/N_{tot(\text{death})}} \tag{2}$$

$N_{i(\text{nondeath})}$: the number of patients who survived within the i^{th} data zone of the selected categorical variable; $N_{i(\text{death})}$: the number of deceased patients within the i^{th} data zone of the selected categorical variable; $N_{tot(\text{nondeath})}$: the total number of patients who survived; $N_{tot(\text{death})}$: the total number of deceased patients.

2.3.2 Variable selection

To evaluate the roles of each variable, we applied the information values (IV) formula[17].

$$IV = \sum_{i=1}^{m} \left(\frac{N_{attr(\text{nondeath})}}{N_{tot(\text{nondeath})}} - \frac{N_{attr(\text{death})}}{N_{tot(\text{death})}} \right) \times WOE_i \tag{3}$$

The IV of the i^{th} data zone is a measurement of how much information it can predict. All these data zones

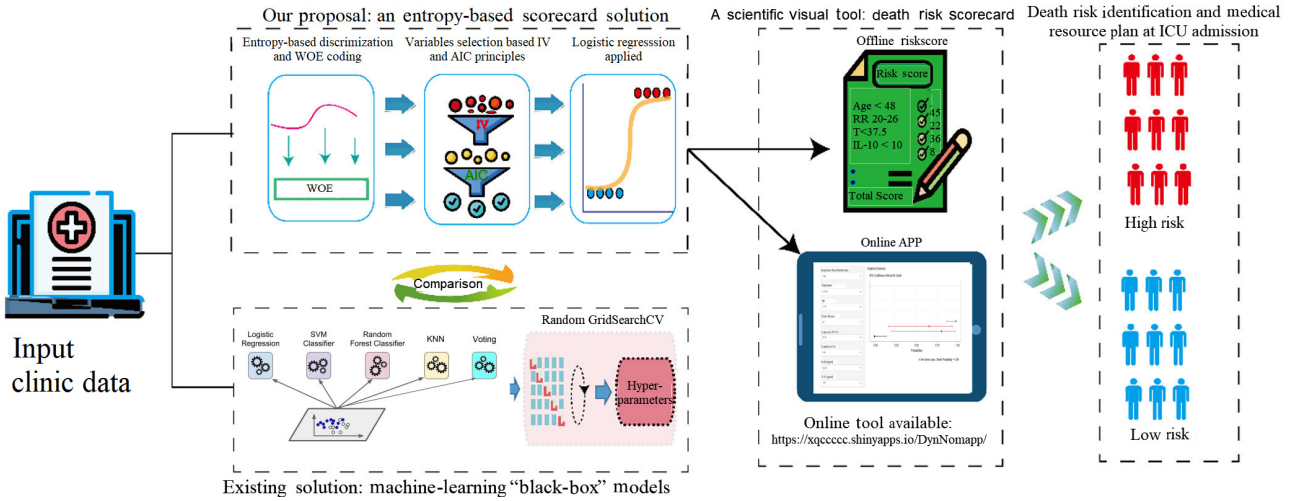

Fig. 1 Workflow of our new model.

were ranked using backward stepwise regression in accordance with the exact Akaike information criterion (AIC). In general, variables with $IV < 0.02$[19] were not considered in this model due to insignificant information predictability.

2.3.3 Model configuration

Based on multiple logistic regressions, the patient probability of mortality can be calculated as below:

$$d = \frac{1}{1 + \mathrm{e}^{-\theta^{\mathrm{T}} X}} \tag{4}$$

$$\ln \frac{d}{1-d} = \theta^{\mathrm{T}} X \tag{5}$$

$$\ln(odds) = W_0 + W_1 X_1 + \cdots + W_n X_n \tag{6}$$

d: the probability of mortality, θ: parameter vector $[W_0, W_1, \ldots, W_n]$, in which W_i is the i^{th} variable coefficient, X: variable vector $[1, X_1, \ldots, X_n]$, in which X_i is the i^{th} variable value.

To visualize and interpret this function for clinical application, which is widely used in financial credits, etc., we mapped its results to a scorecard. The scorecard formulas are listed in Eqs. (4) to (12).

$$score_{\mathrm{total}} = A + B \times \ln(odds) \tag{7}$$

A: intercept (the score points when probabilities of death and survival are equal); B: slope (a margin while scoring points against $\ln(odds)$(changing); $odds$: probability of death (d)/survival probability $(1-d)$.

To normalize the outcome results of the scorecard, we initialized the parameters as below. A specific point score of P_0 was set, with $odds = odds_0$. The point-to-double odds (PDO) refer to the difference when $odds = odds_0$ becomes $odds = \frac{odds_0}{2}$. Here, $P_0 = 600$, $odds_0 = \frac{1}{19}$, and PDO $= 50$. Thus, the odds (death probability/survival probability) of the patient at 600 score points was 1/19, and the value was reduced by half for every 50 points lost (Fig. 2).

These initialization factors were brought back into the following formula:

$$P_0 = A + B \times \ln(odds_0) \tag{8}$$

$$P_0 + \mathrm{PDO} = A + B \times \ln(odds_0/2) \tag{9}$$

Then,

$$B = -\frac{\mathrm{PDO}}{\ln 2} \tag{10}$$

$$A = P_0 - B \times \ln(odds_0) \tag{11}$$

Based on the settings above, the score function was deduced,

$$score_{\mathrm{total}} = A + BW_0 + BW_1 X_1^i + \cdots + BW_n X_n^i \tag{12}$$

where $A + BW_0$ is a base point, and $BW_n X_n^i$ is the mapping score of the i^{th} data zone of the n^{th} variable.

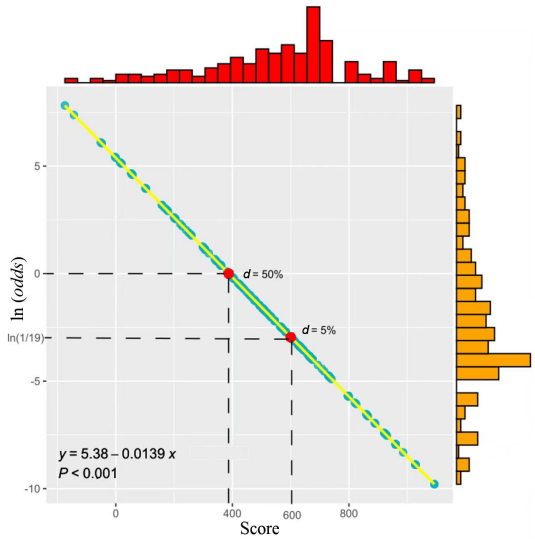

Fig. 2 Relationship between the score in the scorecard and probability.

2.4 Model evaluation

To evaluate the new model, we applied several key performance metrics.

First, whether the results of this new model matched real situations is significant. Here, Brier scores and calibration curves[20] were deployed to describe gaps between the model and statistical facts. The values of Brier scores were between 0 and 1, in which a smaller value indicates a better performance.

Second, training and test data sets should have no bias. The population stability index (PSI) is widely used to measure bias; it is normalized between 0 and 1, in which a small value indicates high stability. PSI is calculated as follows:

$$\mathrm{PSI} = \sum_{i=1}^{m} (l_i - q_i) \ln\left(\frac{l_i}{q_i}\right) / 100 \tag{13}$$

m: the number of groups, which are classified by a metric of equal score zone in the training data set; l_i: the patient proportion in the training data set when their scores are in the i^{th} group of m; q_i: the patient proportion in the test data set when their scores are in the i^{th} group of m, where PSI $= 0$; ideally, the patient proportions in training and test data sets must be equal in each group. Existing rules of thumb are as follows: PSI $<$ 0.10 means "little shift", 0.10 $<$ PSI $<$ 0.25 means "moderate shift", and PSI $>$ 0.25 means "significant shift, action required[21]".

Third, this model is needed to evaluate how much gain is valid. Here, a decision curve analysis[22] was applied to calculate "NET Benefits", which were referred to, was in two default cases (i.e., no model was applied in one

default case, and the other one applied the model for every case)[23, 24]. The formula for net benefits is shown below:

$$\text{NET Benefit} = \frac{\text{TPC}}{N} - \frac{\text{FPC}}{N} \times \frac{t}{1-t} \tag{14}$$

TPC: true positive count; FPC: false-positive count; t: threshold probability of a given event.

Fourth, this model's efficiency was compared with those of typical machine learning algorithms in the same test data set. Six machine learning algorithms were applied: penalty logistic regression (logistic elastic), KNN, support vector machine (SVM), extreme gradient boosting (XGBoost), random forests (RF), and a voting model (RF + XGBoost). To prevent prepositive bias from machine learning algorithms, we used the typical adaptive synthetic[25] strategy to balance the probabilities of both dead and survival samples in the training data set. Their hyperparameters were optimized by combining grid search and fivefold cross-validation.

Finally, the model was validated on the two other external datasets (one was from Yan et al.[7] study: https://www.nature.com/articles/s42256-020-0180-7, and the other was from a Peru cohort: https://figshare.com/articles/dataset/Database.xls/13869179).

Given that no same datasets had all identical variables, we checked the kernel performance of this model in part. Initially, we observed all common variables between our dataset and two external datasets. Second, our model was simplified by using these common variables as input, exclusively. Logically, our model can work well if its kernel performance is accepted in other different datasets with only common variables. All tests were two-tailed, and their results were statistically significance when P-value was less than 0.05. Python (3.9.1) was used for machine learning algorithms, including numpy, pandas, matplotlib, sklearn, imblearn, scipy, and XGBoost packages along with R (version 4.0.2), involving finalfit, glmnet, ggplot2, pROC, rmda, and tidyverse packages, were applied for the scorecard.

3 Result

3.1 Patient statistic

The 322 patients included in this study had a median (IQR) age of 64 (54–70.75) years old, and males accounted for 55.28%. A total of 139 (46.49%) patients had hypertension, 66 (22.54%) had diabetes, and 246 (88.17%) had symptoms of coughing. Among the 322 samples, 72 (22.4%) patients died, and their median (IQR) age was 68 (60.0–74.74) years old, which is higher than that of the survival group (68 versus 63). Each of these 72 deceased patients suffered from one or more symptoms of consciousness disorder, whose probability and seriousness levels were higher than those of the survival group (24% versus 14%). Men accounted for 69.01%, which was higher than the 50.41% observed in the survival group, as shown in Table 1. The distribution of blood oxygen saturation, respiratory rate (RR), temperature, and blood biochemical indicators, etc., deviated between samples of the deceased and surviving patients (Table 2). The dead patients suffered from higher RR, lower blood oxygen saturation, and higher temperature, accompanied by increased levels of leukocytes, neutrophils, high sensitive troponin-I (hsTnl), N-terminal pro-B-type natriuretic peptide (NT-proBNP), interleukin-2 (IL-2), interleukin-6 (IL-6), interleukin-10 (IL-10) procalcitonin, and decreased levels of lymphocytes along with platelets.

3.2 Our proposal: Scorecard model for death risk prediction

As aforementioned, 30 out of 49 variables were selected for this study (Table 3). Twelve items were excluded due to the data incompleteness issue, along with seven variables being deleted for the sake of low diversity. Among the selected 30 variables from 322 patients, two more items were removed as a consequence of the high correlation of neutrophils (cor [neutrophils, lymphocytes] = 0.92) and NT-proBNP (cor [TNI, NT-proBNP] = 0.82). Thus, only 28 variables were deemed suitable for the study.

To search for clinical meanings from these 28 study variables, we discretized the continuous variables by tree-like segmentation. In accordance with the clinical criteria, cutoff points were set based on their entropy values and adjusted if necessary. After data differentiation and entropy processing, 8 of the 28 items were discarded because their *IVs* were less than 0.02. The remaining 20 variables were explored by backward stepwise variable selection in accordance with the AIC. After the selection process, we observed that mortality risks were more significantly affected by eight factors: namely age, temperature, RR, disease cluster, leukocyte, lymphocytes, hsTnl, and IL-10. Their attributes are listed below: age (odds ratio(OR): 5.68; 95% confidence interval (CI): 1.74–24.1), RR (OR: 2.43; 95% CI: 1.12–5.59), temperature (OR: 2.88; 95% CI: 1.19–7.41), and disease cluster (OR: 1.78; 95% CI: 0.87–3.93), leukocyte (OR: 1.52; 95% CI: 0.90–2.77), lymphocytes (OR: 2.02;

Table 1 Clinical characteristics, comorbidities, and outcomes of 322 patients with COVID-19.

Characteristic		Missed heads	All patients	Survival group	Death group	t/X2	*P*-value
Age (IQR)		–	64 (54–70.75)	63.0 (51.25–69.0)	68.0 (60.0–74.75)	6181	<0.001
HR(IQR)		17	94 (84–104)	93.0 (83.75–105.0)	101.0 (85.0–114.0)	6526	0.0504
RR(IQR)		53	20 (20–24)	20.0 (20.0–23.0)	25.0 (20.0–32.0)	3764	<0.001
SpO2(IQR)		53	96 (92–98)	96.0 (94.0–98.0)	86.5 (78.0–94.75)	9139	<0.001
Temperature (IQR)		44	36.6 (36.2–37.08)	36.6 (36.2–37.0)	37.0 (36.5–37.85)	4547	<0.001
Gender	Male	–	178 (55.28%)	124 (51.20%)	50 (69.44%)	7.526	0.006
	Female	–	144 (44.72%)	122 (48.80%)	22 (39.56%)		
Coronary disease	Yes	–	46 (16.61%)	31 (14.42%)	15 (24.19%)	3.32	0.068
	No	–	231 (83.39%)	184 (85.58%)	47 (75.81%)		
Mental state	Normal	–	227 (85.35%)	185 (91.13%)	42 (66.67%)	16.19	0.003
	Slightly	–	29 (10.9%)	15 (7.4%)	14 (22.22%)		
	Severe	–	10 (2.75%)	3 (1.47%)	7 (11.11%)		
Diabetes	Yes	–	66 (22.54%)	49 (21.4%)	17 (26.56%)	0.765	0.381
	No	–	227 (77.47%)	180 (78.6%)	47 (73.44%)		
Hypertension	Yes	–	139 (46.49%)	127 (54.27%)	33 (50.77%)	0.251	0.616
	No	–	160 (53.51%)	107 (45.73%)	32 (49.23%)		
Cough	Yes	–	246 (88.17%)	188 (87.04%)	59 (92.19%)	1.26	0.262
	No	–	33 (11.83%)	28 (12.96%)	5 (7.81%)		

Table 2 Blood biochemistry of 322 patients with COVID-19

Characteristic	Missed head	Patient	Survival group	Death group	t/X2	*P*-value
Leukocyte (IQR)	5	6.26 (4.93–8.67)	5.7 (4.62–7.16)	9.67 (7.46–13.57)	2946.5	<0.001
Neutrophils (IQR)	5	71.1 (57.55–82.6)	65.7 (55.35–75.78)	89.52 (83.7–91.95)	1679.5	<0.001
Lymphocytes (IQR)	5	19.2 (9.65–29.15)	22.6 (14.08–30.15)	5.96 (3.16–8.85)	15519	<0.001
Platelets (IQR)	6	223 (165.38–286.5)	234.0 (185.25–292.75)	159.0 (105.71–225.75)	12299	<0.001
hsTnl (IQR)	23	6.1 (2.3–17.15)	3.6 (1.9–8.6)	35.92 (11.42–198.6)	1918.5	<0.001
NT-proBNP (IQR)	25	173 (49–503)	111.5 (35.75–296.88)	817.0 (383.0–2687.0)	2017	<0.001
PCT (IQR)	48	0.05 (0.03–0.16)	0.04 (0.02–0.09)	0.22 (0.13–0.94)	1687.5	<0.001
IL-6 (IQR)	40	14 (3–49.74)	9.02 (2.15–27.43)	69.77 (34.4–178.95)	1937.5	<0.001
IL-2 (IQR)	58	655 (399–1091)	570.75 (347.0–894.75)	1128.0 (816.0–1809.0)	2903.5	<0.001
IL-10 (IQR)	58	5 (4–5)	5.0 (5.0–5.9)	10.3 (6.8–20.7)	1968.5	<0.001

Table 3 Initial filtering of variables based on their missing data and variability.

Type	Parameter
Inclusion	Age, sex, hypertension, coronary heart disease
	Chronic nephropathy, history of diabetes, underlying lung disease, smoking history
	History of surgery/trauma/blood transfusion, disease cluster
	Heart rate (5.6%) (bpm), respiratory rate (16.1%), temperature (16.8%) (℃)
	Peripheral oxygen saturation (16.4%), night sweats, fever
	Mental status, dyspnea, cough, headache, chest pain
	Leukocyte (1.6%), neutrophils (1.6%), lymphocytes (1.6%)
	PCT (14.9%), IL-2 (18.3%), IL-6 (12.4%), IL-10 (18.3%)
	TNI (7.1%), NT-proBNP (7.5%), platelets (1.9%)
Exclusion	Weight (52.8%), height (54.3%), hepatitis B vaccine
	Disease of digestive tract, tumor, HIV
	Perinatal or lactation, contact histor of COVID-19 patients
	Contact history of fevrile personnel in Wuhan
	Pharyngal myalgia (41%), chilly (20.8%), fatigue (35.1%)
	Stroke, diarrhea (37.58%), C-reaction protein (72.46%)
	Albumin (72.17%), creatinine (72.75%), LDH (54.78%), D-dimer ration (27.54%)

95% CI: 1.48–2.84), hsTnl (OR: 1.60; 95% CI: 1.08–2.37), and IL-10 (OR: 1.60; 95%, CI: 1.08–2.37).

3.3 Model validation and comparison

As displayed in Fig. 3, the calibration curves matched the desired requirements: slope = 1343, Brier = 0.048, and Dxy = 0.972. In Fig. 4, the decision curves show that the scorecard model satisfied clinical utility in the test data set. In terms of computing effectiveness and efficiency, six typical algorithms (e.g., KNN, logistic-elastic, SVM, XGBoost, RF, and a voting model) were cited as alternates. According to Table 4, the RF achieved the best performance (ACC = 87.6%; AUC = 97.6%), followed by XGBoost (ACC = 85.6%; AUC = 96.1%), and logistic elastic (ACC = 86.6%; AUC = 94.7%). Compared with these traditional machine learning models, our entropy-based discretization scorecard model achieved a better performance (ACC = 94.8%; AUC = 98.6%).

The scorecard model is a visual quantitative tool for death risk in clinical applications (Table 5) and

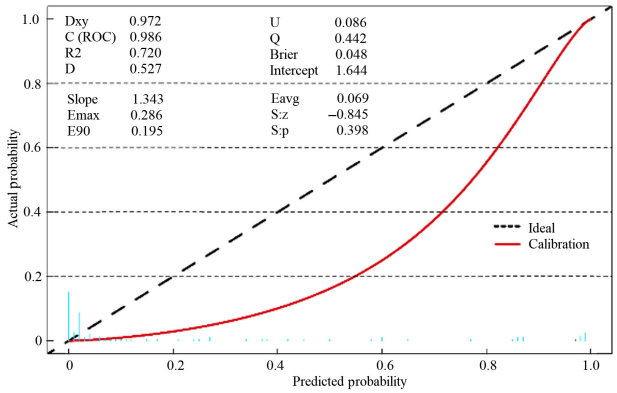

Fig. 3 Calibration curve, model results versus statistical facts regarding in-hospital mortality (red line: calibration curve).

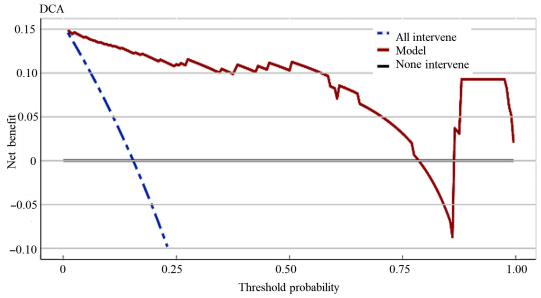

Fig. 4 Decision curve analysis (DCA) curves and red lines are standardized net benefits at different risk thresholds compared with two default cases (black line: no model applied; blue dash line: the model was used for all patients).

Table 4 Performance comparison of different models in the data test set.

Model	ACC (%)	AUC (%)	Sp	Se	PPV	NPV
KNN	82.5	76.3	0.878	0.533	0.444	0.911
SVM	85.6	88.5	0.914	0.533	0.530	0.915
XG	85.6	96.1	0.841	0.933	0.518	0.986
RF	87.6	97.6	0.866	0.933	0.560	0.986
Vote	87.6	97.6	0.866	0.933	0.560	0.986
Logistic	86.6	94.7	0.865	0.867	0.541	0.973
Scorecard	94.8	98.6	0.939	1.000	0.750	1.000

Note: ACC = accuracy; AUC = area under curve; Sp = specificity; Se = sensitivity; PPV = positive predictive value; NPV = negative predictive value.

Table 5 Visualization of risk scorecard for real applications.

Age (Risk coefficient)	Temperature (°C) (Risk coefficient)	Disease cluster (Risk coefficient)
<48 (291)	<37.3 (18)	No (–17)
48–72 (–5)	37.3–38.2 (–21)	Yes (51)
⩾72 (–62)	⩾38.2 (–135)	– (–)
Leukocyte (10^9/L) (Risk coefficient)	**hsTnl (pg/mL) (Risk coefficient)**	**IL-10 (pg/mL) (Risk coefficient)**
<10 (–106)	<16 (31)	<10 (42)
10–19 (37)	16–30 (6)	10–18 (–56)
⩾19 (99)	⩾30 (–70)	>18 (–161)

Note: pg/mL is the unit of drug dose.

reliable given the PSI = 0.071 between the training and test datasets (Fig. 5). As shown in Table 6, when the patient's death risk score was lower than a threshold of 205, the mortality rate was close to 100% (training set: 96.2%; test set: 100%). When the scores were between thresholds 205 to 459, the mortality rates decreased significantly, and when the score was higher than the score at our threshold of 459, the mortality rate was close to 0. In real scenarios, these score ranges can be adjusted based on local medical resource availability. Our new model is available for public sharing on the link (https://xqccccc.shinyapps.io/DynNomapp/), and it can interactively calculate the specific death probability and confidence interval of patients. Figure 6 shows its interactive windows.

In our dataset, the performance of our model with all eight factors showed ([ACC = 94.8%, AUC = 99.0%, sensitivity = 1, and specificity = 0.939]; the performance was adjusted to [ACC = 88.7%, AUC = 94.1%, sensitivity = 0.583, and specificity = 0.986]) when using the mentioned three common variables (age[6, 10, 26], temperature[27–29], and RR[19]). As shown in Fig. 7 and Table 7, the model performance in the other two datasets was listed as (Wuhan cohort

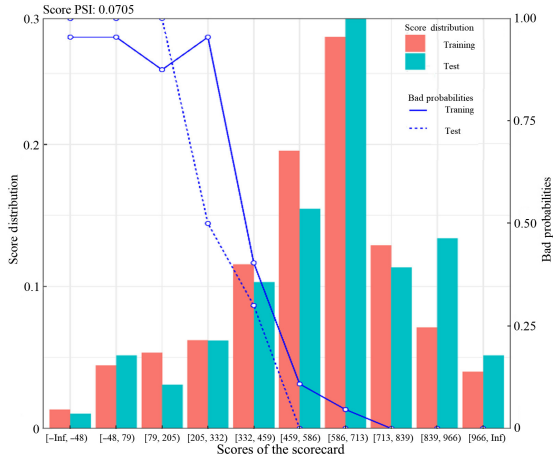

Fig. 5 Score distribution for a given PSI result.

https://www.nature.com/articles/s42256-020-0180-7; ACC = 82.0%, AUC = 92.4%, sensitivity = 0.886, and specificity = 0.785; Peru cohort: https://figshare.com/articles/dataset/Database.xls/13869179; ACC = 71.4%, AUC = 74.4%, sensitivity = 0.757, and specificity = 0.642). These kernel performance tests can mainly reflect the model's universal applicability in different datasets.

3.4 Illumination of model results

In this scorecard model, the roles of eight variables can be interpreted clinically and independently. Specifically, neutrophils are the main members of the leukocyte family. The increase in neutrophil levels introduces excessive reactive oxygen species and causes tissue damage, especially in elderly or frail people and indicates various types of bacterial infection[30, 31]. Existing studies[32, 33] explained the strong causation correlation between neutrophil growth and the increased death risk. The increased levels of hsTnl and IL-10 and decreased levels of lymphocytes are explicit risk factors for COVID-19 patient death[34–36]. These abnormalities indicate that COVID-19 infection may be associated with cellular immune deficiency, myocardial injury, and cytokine storm, which also reveals that systemic immune response is a critical factor related to the final outcomes of COVID-19 subjects. As for other variables, such as

Table 6 Scorecard performance evaluation regarding mortality rates and related metrics between cutoff points in the test set.

Number of cutoff points	Number of patients	ACC (%)	TP	TN	FP	FN	Se	Sp	PPV	NPV	F1	Death
48	1	85.6	1	82	0	14	0.07	1.00	1.00	0.85	0.13	1
79	6	90.7	6	82	0	9	0.40	1.00	1.00	0.90	0.57	6
205	9	93.8	9	82	0	6	0.6	1.00	1.00	0.93	0.75	9
332	15	93.8	12	79	3	3	0.80	0.96	0.80	0.96	0.80	12
459	27	87.6	15	70	12	0	1.00	0.85	0.56	1.00	0.71	15
586	40	74.2	15	57	25	0	1.00	0.70	0.38	1.00	0.55	15
713	68	45.4	15	29	53	0	1.00	0.35	0.22	1.00	0.36	15
839	79	34.0	15	18	64	0	1.00	0.22	0.19	1.00	0.32	15
966	92	20.6	15	5	77	0	1.00	0.06	0.16	1.00	0.28	15
1093	97	15.5	15	0	82	0	1.00	0.00	0.15	NA	0.27	15

Note: TP = true positive; TN = true negative; FP = false positive; FN = false negative; Se=sensitivity; Sp=specificity; PPV = positive predictive value; NPV = negative predictive value.

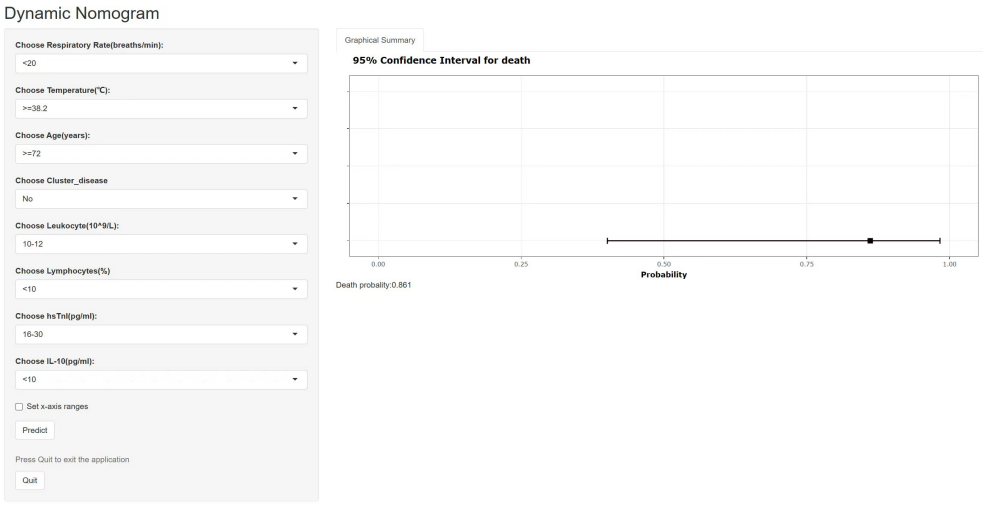

Fig. 6 The interactive window of a visualization tool for clinical treatment.

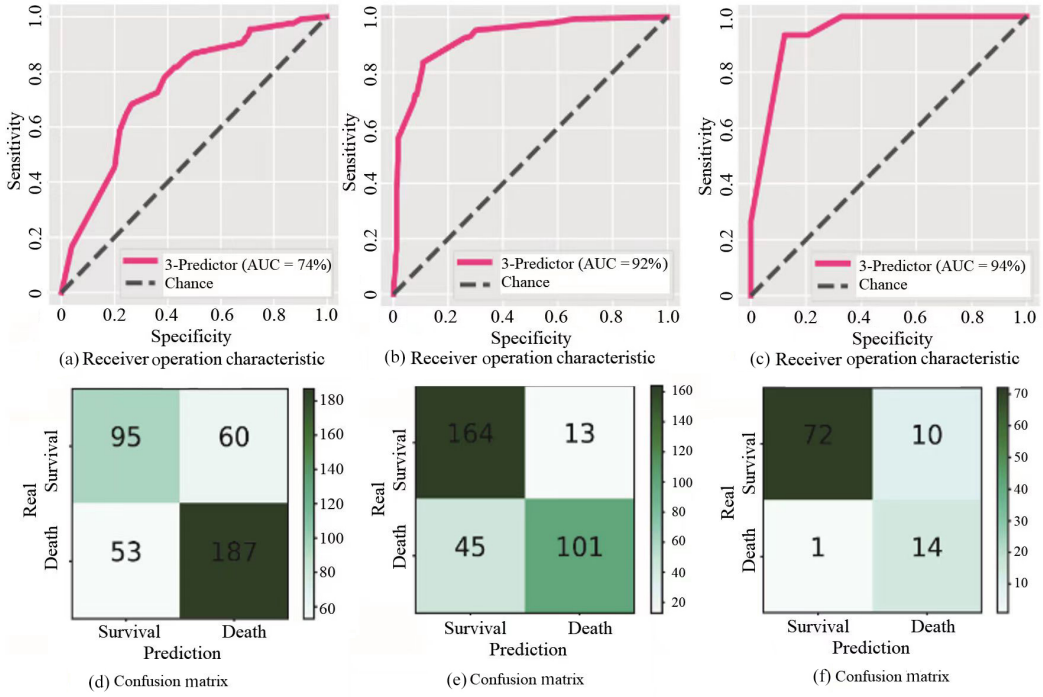

Fig. 7 Comparison between different datasets in terms of receiver operating characteristic (ROC) curves and confusion matrices for testing the new model's applicability (in the figure, ROC curves are in (a)–(c); confusion matrices are in (d)–(f), (a) and (d) stand for performance in the Peru dataset; (b) and (e) represent results in Ref. [7]; (c) and (f) mean results from our dataset of the national medical team).

Table 7 Performance comparison of the three-predictor model in different test sets.

Type	ACC (%)	AUC (%)	Sp	Se	PPV	NPV
Peru	71.4	74.4	0.642	0.757	0.613	0.779
Yan et al.'s data[7]	82.0	92.4	0.785	0.886	0.927	0.692
Our data	88.7	94.1	0.986	0.583	0.878	0.933

Note: ACC = accuracy; AUC = area under curve; Sp = specificity; Se = sensitivity; PPV = positive predictive value; NPV = negative predictive value.

age[6, 10, 26], temperature[27–29], and RR[19], their increase either means a high risk of hospitalization mortality for COVID-19 patients. These three factors are also common risk factors for other types of pneumonia[37–39]. Disease cluster is a special and extensively studied fact, and in such a case, a patient can infect those who are close to him or her in a relatively short period. These patients are likely to recover rapidly as they are determined by the scoreboard as positive for proper treatment as early as possible.

In this retrospective study, our main objective was to find an effective and efficient way that can provide a scientific tool to process complex data and help in processing unknown or unclear data knowledge to improve the medical quality of a new disease, for example, at the early stage of a disease such as the COVID-19 outbreak. The first issue was the removal of redundant noises from high-dimensional data variables using the entropy information theory. These variables were selected and processed layer by layer, to find key factors of COVID-19 death risk. Finally, the findings on these key factors were mapped to a simple scorecard, which is a digital tool for translating data into clinical information to allow doctors to control the death risks during ICU admission of COVID-19 patients. In applications, the thresholds of the scorecard can be adjusted based on the actual availability of local medical resources. To illustrate how the model works in different datasets, we created and applied a partial kernel performance test method. In this study, training data were all from Chinese patients, and the results' effectiveness may have some limitations when extended to other populations, including influence from differences in genes, environment, medical treatment methods, etc. This model should be further optimized for cases in different countries/regions. Thus, our model provides a handy digital assistant to physicians for handling new or unknown diseases for death risk control and medical resource management.

4 Conclusion

This study solved a critical problem of how to allow a computing algorithm to produce clinical explanation results, regardless of its black-box mechanism. To address this problem, we developed a visual scientific tool for clinical decision support during ICU admission of COVID-19 patients by refining some critical factors using an entropy-based method. Compared with traditional machine learning models, our scorecard model is a scientific clinical interpretation tool to be used by doctors for medical resource planning, especially in massive epidemics like the COVID-19 situation during the past three years globally. In this retrospective study, we assessed 322 adults (age: [median: 64; IQR: 54–75], males: 178 [55.28%], and death: 72 [22.36%]) who are the laboratory-confirmed COVID-19 ICU patients from the national medical team in Wuhan. These patients were randomly grouped into training (70%) and validation (30%) cohorts. In the training cohort, all kinds of continuous-type data were first discretized using an entropy-based method (tree-like segment) to obtain the cutoff endpoints of the information margin, and they were helpful in relating clinical explanation to complex data. Combining other discrete data, a logistic regression was adopted to configure a scorecard of death risk factors in visualization and compare it with machine learning algorithms. Logistic regression was utilized to identify the risk factors of patients with COVID-19, and a scorecard with the selected eight variables was built for clinical use. Calibration curves, PSI, and DCA were used to evaluate the performance of the scorecard in validation cohorts. In terms of survival probability analyses, this scorecard model can interpret clinical meanings from complex data with the desired computing performance (ACC = 94.8%; AUC = 99.0%, compared with several other typical machine learning methods, such as XGBoost (ACC = 85.6%; AUC = 96.0%), RF (ACC = 87.6%; AUC = 98.0%) and a voting classifier (XGBoost + RF) (ACC = 87.6%; AUC = 98.0%)). Notably, some limitations were noted in this study, such as how to set up a clinical cohort to validate this model in real-life, how to find clinical explanations from missing or unconsidered data, and how to design an online computing model to help in the treatment of any new diseases, etc. These topics are open for the next study.

Acknowledgment

This work was supported in part by the Scientific and Technological Innovation 2030-"New Generation Artificial Intelligence" Major Project (No. 2021ZD0140406), and the National Natural Science Foundation of China (No. 62041201). Qiang Ji contributed figure drawing and data analysis assistance to the article.

References

[1] M. Esai Selvan, Risk factors for death from COVID-19, *Nat. Rev. Immunol.*, vol. 20, no. 7, p. 407, 2020.

[2] Y. Shang, T. Liu, Y. Wei, J. Li, L. Shao, M. Liu, Y. Zhang, Z. Zhao, H. Xu, Z. Peng, et al., Scoring systems for predicting mortality for severe patients with COVID-19, *eClinicalMedicine*, vol. 24, p. 100426, 2020.

[3] K. A. Overmyer, E. Shishkova, I. J. Miller, J. Balnis, M. N. Bernstein, T. M. Peters-Clarke, J. G. Meyer, Q. Quan, L. K. Muehlbauer, E. A. Trujillo, et al., Large-scale multi-omic analysis of COVID-19 severity, *Clin. Transl. Discov.*, vol. 12, no. 1, pp. 23–40, 2021.

[4] G. Zhang, Y. An, L. Zhang, L. Xie, and X. Guo, Risk factors for in-hospital mortality in patients with cancer and COVID-19, *Lancet Oncol.*, vol. 21, no. 9, p. 407, 2020.

[5] J. Tian, X. Yuan, J. Xiao, Q. Zhong, C. Yang, B. Liu, Y. Cai, Z. Lu, J. Wang, Y. Wang, et al., Clinical characteristics and risk factors associated with COVID-19 disease severity in patients with cancer in Wuhan, China: A multicentre, retrospective, cohort study, *Lancet Oncol.*, vol. 21, no. 7, pp. 893–903, 2020.

[6] W. Liang, H. Liang, L. Ou, B. Chen, A. Chen, C. Li, Y. Li, W. Guan, L. Sang, J. Lu, et al., Development and validation of a clinical risk score to predict the occurrence of critical illness in hospitalized patients with COVID-19, *JAMA Intern. Med.*, vol. 180, no. 8, pp. 1081–1089, 2020.

[7] L. Yan, H. T. Zhang, J. Goncalves, Y. Xiao, M. Wang, Y. Guo, C. Sun, X. Tang, L. Jing, M. Zhang, et al., An interpretable mortality prediction model for COVID-19 patients, *Nat. Mach. Intell.*, vol. 2, no. 5, pp. 283–288, 2020.

[8] Y. Gao, G. Y. Cai, W. Fang, H. Y. Li, S. Y. Wang, L. Chen, Y. Yu, D. Liu, S. Xu, P. F. Cui, et al., Machine learning based early warning system enables accurate mortality risk prediction for COVID-19, *Nat. Commun.*, vol. 11, no. 1, p. 5033, 2020.

[9] A. S. Yadaw, Y. C. Li, S. Bose, R. Iyengar, S. Bunyavanich, and G. Pandey, Clinical features of COVID-19 mortality: Development and validation of a clinical prediction model, *Lancet Digit. Heath.*, vol. 2, no. 10, pp. 516–525, 2020.

[10] S. R. Knight, A. Ho, R. Pius, I. Buchan, G. Carson, T. M. Drake, J. Dunning, C. J. Fairfield, C. Gamble, C. A. Green, et al., Risk stratification of patients admitted to hospital with COVID-19 using the ISARIC WHO Clinical Characterisation Protocol: Development and validation of the 4C Mortality Score, *BMJ Clin. Res. Ed.*, vol. 370,

p. 3339, 2020.

[11] N. Razavian, V. J. Major, M. Sudarshan, J. Burk-Rafel, P. Stella, H. Randhawa, S. Bilaloglu, J. Chen, V. Nguy, W. Wang, et al., A validated, real-time prediction model for favorable outcomes in hospitalized COVID-19 patients, *NPJ Digit. Med.*, vol. 3, p. 130, 2020.

[12] O. Troyanskaya, M. Cantor, G. Sherlock, P. Brown, T. Hastie, R. Tibshirani, D. Botstein, and R. B. Altman, Missing value estimation methods for DNA microarrays, *Bioinform. Oxf. Engl.*, vol. 17, no. 6, pp. 520–525, 2001.

[13] J. Gupta, S. Paul, and A. Ghosh, A novel transfer learning-based missing value imputation on discipline diverse real test datasets—A comparative study with different machine learning algorithms, in *Advances in Intelligent Systems and Computing*, Singapore: Springer, 2019.

[14] Minakshi, Rajan Vohra, and Gimpy, Missing value imputation in multi attribute data set, *Int. J. Comput. Sci. Infor. Technol.*, vol. 5, no. 4, pp. 5315–5321, 2014.

[15] G. E. A. P. A. Batista and M. C. Monard, An analysis of four missing data treatment methods for supervised learning, *Appl. Artif. Intell.*, vol. 17, nos. 5–6, pp. 519–533, 2003.

[16] J. Dougherty, R. Kohavi, and Sahami M., Supervised and unsupervised discretization of continuous features, in *Proc. 12^{th} Int. Conf. Machine Learning*, Tahoe City, CA, USA: Morgan Kaufmann, 1995.

[17] G. Zeng, Metric divergence measures and information value in credit scoring, *J. Math.*, vol. 2013, pp. 1–10, 2013.

[18] M. Refaat, *Credit Risk Scorecards: Development and Implementation Using SAS*, Raleigh, NC, USA: LULU.COM, 2011.

[19] N. Siddiqi, *Credit Risk Scorecards: Developing and Implementing Intelligent Credit Scoring*, Hoboken, NJ, USA: John Wiley & Sons, Inc., 2012.

[20] E. W. Steyerberg, A. J. Vickers, N. R. Cook, T. Gerds, M. Gonen, N. Obuchowski, M. J. Pencina, and M. W. Kattan, Assessing the performance of prediction models: A framework for traditional and novel measures, *Epidemiol. Camb. Mass*, vol. 21, no. 1, pp. 128–138, 2010.

[21] R. Taplin and C. Hunt, The population accuracy index: A new measure of population stability for model monitoring, *Risks*, vol. 7, no. 2, p. 53, 2019.

[22] A. J. Vickers and E. B. Elkin, Decision curve analysis: A novel method for evaluating prediction models, medical decision making, *Med. Decis. Making*, vol. 26, no. 6, pp. 565–574, 2006.

[23] A. J. Vickers, B. van Calster, and E. W. Steyerberg, A simple, step-by-step guide to interpreting decision curve analysis, *Diagn. Progn. Res.*, vol. 3, p. 18, 2019.

[24] M. Fitzgerald, B. R. Saville, and R. J. Lewis, Decision curve analysis, *JAMA*, vol. 313, no. 4, p. 409, 2015.

[25] H. He, Y. Bai, E. A. Garcia, and S. Li, ADASYN: Adaptive synthetic sampling approach for imbalanced learning, in *Proc. 2008 IEEE Int. Joint Conf. Neural Networks (IEEE World Congress on Computational Intelligence)*, Hong Kong, China, 2008, pp. 1322–1328.

[26] F. Caramelo, N. Ferreira, and B. Oliveiros, Estimation of risk factors for COVID-19 mortality-preliminary results, medRxiv, https://europepmc.org/article/PPR/PPR114369, 2020.

[27] B. Zheng, Y. Cai, F. Zeng, M. Lin, J. Zheng, W. Chen, G. Qin, and Y. Guo, An interpretable model-based prediction of severity and crucial factors in patients with COVID-19, *BioMed Res. Int.*, vol. 2021, pp. 1–9, 2021.

[28] J. A. Kline, C. A. Camargo, D. M. Courtney, C. Kabrhel, K. E. Nordenholz, T. Aufderbeide, J. J. Baugh, D. G. Beiser, C. L. Bennett, J. Bledsoe, et al., Clinical prediction rule for SARS-CoV-2 infection from 116 U.S. emergency departments 2-22-2021, *PloS One*, vol. 16, no. 3, p. 0248438, 2021.

[29] K. B. Son, T. J. Lee, and S. S. Hwang, Disease severity classification and COVID-19 outcomes, Republic of Korea, *Bull. World Heath. Organ.*, vol. 99, no. 1, pp. 62–66, 2021.

[30] M. Laforge, C. Elbim, C. Frère, M. Hémadi, C. Massaad, P. Nuss, J. J. Benoliel, and C. Becker, Tissue damage from neutrophil-induced oxidative stress in COVID-19, *Nat. Rev. Immunol.*, vol. 20, no. 9, pp. 515–516, 2020.

[31] B. Kalyanaraman, Do free radical network and oxidative stress disparities in African Americans enhance their vulnerability to SARS-CoV-2 infection and COVID-19 severity? *Redox Biol.*, vol. 37, p. 101721, 2020.

[32] P. Pan, Y. Li, Y. Xiao, B. Han, L. Su, M. Su, Y. Li, S. Zhang, D. Jiang, X. Chen, et al., Prognostic assessment of COVID-19 in the intensive care unit by machine learning methods: Model development and validation, *J. Med. Internet Res.*, vol. 22, no. 11, p. 23128, 2020.

[33] A. Alnor, M. B. Sandberg, C. Gils, and P. J. Vinholt, Laboratory tests and outcome for patients with coronavirus disease 2019: A systematic review and meta-analysis, *J. Appl. Lab. Med.*, vol. 5, no. 5, pp. 1038–1049, 2020.

[34] X. Zhang, Y. Tan, Y. Ling, G. Lu, F. Liu, Z. Yi, X. Jia, M. Wu, B. Shi, S. Xu, et al., Viral and host factors related to the clinical outcome of COVID-19, *Nature*, vol. 583, no. 7816, pp. 437–440, 2020.

[35] D. Wang, B. Hu, C. Hu, F. Zhu, X. Liu, J. Zhang, B. Wang, H. Xiang, Z. Cheng, Y. Xiong, et al., Clinical characteristics of 138 hospitalized patients with 2019 novel coronavirus–infected pneumonia in Wuhan, China, *JAMA*, vol. 323, no. 11, p. 1061, 2020.

[36] F. He, Y. Quan, M. Lei, R. Liu, S. Qin, J. Zeng, Z. Zhao, N. Yu, L. Yang, and J. Cao, Clinical features and risk factors for ICU admission in COVID-19 patients with cardiovascular diseases, *Aging Dis.*, vol. 11, no. 4, p. 763, 2020.

[37] W. S. Lim, M. M. van der Eerden, R. Laing, W. G. Boersma, N. Karalus, G. I. Town, S. A. Lewis, and J. T. MacFarlane, Defining community acquired pneumonia severity on presentation to hospital: An international derivation and validation study, *Thorax*, vol. 58, no. 5, pp. 377–382, 2003.

[38] M. J. Fine, T. E. Auble, D. M. Yealy, B. H. Hanusa, L. A. Weissfeld, D. E. Singer, C. M. Coley, T. J. Marrie, and W. N. Kapoor, A prediction rule to identify low-risk patients with community-acquired pneumonia, *Dev. Camb. Engl.*, vol. 336, no. 4, pp. 243–250, 1997.

[39] J. L. Liu, F. Xu, H. Zhou, X. J. Wu, L. X. Shi, R. Q. Lu, A. Farcomeni, M. Venditti, Y. L. Zhao, S. Y. Luo, et al., Expanded CURB-65: A new score system predicts severity of community-acquired pneumonia with superior efficiency, *Sci. Rep.*, vol. 6, p. 22911, 2016.

Chong Yao received the MS degree from China University of Mining and Technology, Xuzhou, China, in 2020. Now, she is pursuing the PhD degree in Beihang University, China. Her research interests include medical information and blockchain.

Chonghui Huangqi received the BS degree from College of Agricultural and Environment Sciences at University of California, Davis, USA, in 2022. From August 2022 to May 2023, he was an employee at the Lab of Population Health and Reproduction, UC Davis, USA. He is currently a graduate student in University of Southern California, USA. His research interests include biomedical engineering, gene computing, medical imaging, and imaging information.

Anpeng Huang received the MS degree from the University of Electronic Science and Technology of China, Sichuan, China, in 2000, and the PhD degree from Peking University, Beijing, China, in 2003. From May 2004 to January 2005, he was a visiting scholar in the University of Waterloo, Canada. From February 2005 to March 2008, he was a postdoctoral researcher at the Department of Computer Science in the University of California, Davis, USA. Since November 2007, he has been a tenure-tracked associate professor of Peking University, China. He is also the chief scientist of National Health IT Program in China, and the chief scientist of Beijing Goodwill Information and Technology Co., Ltd. He has more than 75 journal papers and conference papers, He is the holder of 19 US patents and 74 Chinese patents, the advisor of "Best Student Paper Award" winner at 2012 14th IEEE HEALTHCOM conference, and the founder of Mobile Health Laboratory in PKU. His research interests include mobile health, medical big-data, artificial intelligence for health, etc.