



An Early Warning Tool for Predicting Mortality Risk of COVID-19 Patients Using Machine Learning

Muhammad E. H. Chowdhury¹ · Tawsifur Rahman² · Amith Khandakar¹ · Somaya Al-Madeed³ · Susu M. Zughaier⁴ · Suhail A. R. Doi⁵ · Hanadi Hassen³ · Mohammad T. Islam⁶

Received: 16 July 2020 / Accepted: 28 December 2020
© The Author(s) 2021

Abstract

COVID-19 pandemic has created an extreme pressure on the global healthcare services. Fast, reliable, and early clinical assessment of the severity of the disease can help in allocating and prioritizing resources to reduce mortality. In order to study the important blood biomarkers for predicting disease mortality, a retrospective study was conducted on a dataset made public by Yan et al. in [1] of 375 COVID-19 positive patients admitted to Tongji Hospital (China) from January 10 to February 18, 2020. Demographic and clinical characteristics and patient outcomes were investigated using machine learning tools to identify key biomarkers to predict the mortality of individual patient. A nomogram was developed for predicting the mortality risk among COVID-19 patients. Lactate dehydrogenase, neutrophils (%), lymphocyte (%), high-sensitivity C-reactive protein, and age (LNLCA)—acquired at hospital admission—were identified as key predictors of death by multi-tree XGBoost model. The area under curve (AUC) of the nomogram for the derivation and validation cohort were 0.961 and 0.991, respectively. An integrated score (LNLCA) was calculated with the corresponding death probability. COVID-19 patients were divided into three subgroups: low-, moderate-, and high-risk groups using LNLCA cutoff values of 10.4 and 12.65 with the death probability less than 5%, 5–50%, and above 50%, respectively. The prognostic model, nomogram, and LNLCA score can help in early detection of high mortality risk of COVID-19 patients, which will help doctors to improve the management of patient stratification.

Keywords Machine learning · Prognostic model · Early warning tool · Predicting mortality risk · COVID-19

This article belongs to the Topical Collection: *Data-Driven Artificial Intelligence approaches to Combat COVID-19*
Guest Editors: Mufti Mahmud, M. Shamim Kaiser, Nilanjan Dey, Newton Howard, Aziz Sheikh

✉ Muhammad E. H. Chowdhury
mchowdhury@qu.edu.qa

- ¹ Department of Electrical Engineering, Qatar University, 2713 Doha, Qatar
- ² Department of Biomedical Physics & Technology, University of Dhaka, 1000 Dhaka, Bangladesh
- ³ Department of Computer Science and Engineering, Qatar University, 2713 Doha, Qatar
- ⁴ Department of Basic Medical Sciences, College of Medicine, Qatar University, 2713 Doha, Qatar
- ⁵ Department of Population Medicine, College of Medicine, Qatar University, 2713 Doha, Qatar
- ⁶ Department of Electrical, Electronics and Systems Engineering, Universiti Kebangsaan Malaysia, 43600 Bangi, Selangor, Malaysia

Introduction

The novel coronavirus disease (COVID-19) spread rapidly throughout the world from Wuhan (Hubei, China) since December 2019 [2–6]. Since the outbreak, the number of reported cases has surpassed 12 million with more than 550 thousand deaths worldwide as of 12 July 2020 [7]. The COVID-19 disease is caused by the severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2), which is a member of the coronavirus family. On 11 March 2020, COVID-19 was declared as a pandemic by the World Health Organization (WHO) [8]. Due to the pandemic, hospital capacity is being exceeded in many places and face issues in terms of limited medical staff, personal protective equipment, life-support equipment, and others [9, 10]. Symptoms of COVID-19 are non-specific, and infected individuals may develop fever (83–99%), cough (59–82%), loss of appetite (40–84%), fatigue (44–70%), shortness of breath (31–40%), coughing up sputum (28–33%), or muscle

aches (11–35%) [11]. The disease can further progress into a severe pneumonia, acute respiratory distress syndrome (ARDS), myocardial injury, sepsis, septic shock, and even death [12]. Though most COVID-19 patients have a mild illness, there are some patients who show rapid deterioration (particularly within 7–14 days) from the onset of symptoms into severe COVID-19 with or without ARDS [13, 14]. Current epidemiological data suggest that the mortality rate of patients with severe COVID-19 is higher than that of patients with non-severe COVID-19 [15, 16]. It has been reported that 26.1–32.0% of patient infected with COVID-19 are prone to progressing critical illness [17]. Recent studies have confirmed a high fatality rate of 61.5% for patients in critical cases, which increase with age and other medical comorbidities [17].

A large cohort study from 2449 patients has reported that during this pandemic, healthcare system can be overwhelmed by hospitalization (20–31%) and intensive care unit (ICU) admission rates (4.9–11.5%) [18]. This can be avoided by prioritizing hospital treatment for patients at high risk of deterioration and death, and treating low-risk patients in ambulatory environments, or by home-based self-quarantine. An effective tool is required to predict the disease trajectory to allocate resources efficiently and also improve the patient's condition. Understanding the great potential of this approach, it is important to identify key patient variables that can help to predict the course of the disease at diagnosis. In other words, early identification of patients at high risk for progression to severe COVID-19 will help in efficient utilization of healthcare resources via patient prioritization to reduce the mortality rate.

Several researches indicate that biomarkers can help to classify COVID-19 patients with elevated risk of serious disease and mortality by providing crucial information regarding the patients' health status. Al Youha et al. [19] proposed a prognostic model called the Kuwait Progression Indicator (KPI) score for predicting progression of severity in COVID-19. The KPI model was based on quantifiable laboratory readings unlike other self-reported symptoms and other subjective parameter-based scoring systems. The KPI score categorizes patients to low risk if the score goes below -7 and high risk if the score goes above 16 ; however, the progression risk in the intermediate group (for patients scores within -6 to 15) deemed by the authors as uncertain. This intermediate category however exists with many prognostic systems. Weng et al. [20] reported an early prediction score called ANDC to predict mortality risk for COVID-19 patients using 301 adult patients' data. Least absolute shrinkage and selection operator (LASSO) regression has identified age, neutrophil-to-lymphocyte ratio (NLR), D-dimer, and C-reactive protein (CRP) recorded during admission as mortality predictors for COVID-19 patients [20]. They have developed a nomogram

demonstrating good performance and also derived an integrated score, ANDC, with its corresponding death probability. They have also developed cutoff ANDC values to classify COVID-19 patients into three groups: low-, moderate-, and high-risk groups. The death probabilities were 5%, 5–50%, and more than 50% in the low-, moderate-, and high-risk group, respectively. Using a cohort of 444 patients, Xie et al. [21] proposed a prognostic model using lactate dehydrogenase, lymphocyte count, age, and SpO₂ as key predictors of COVID-19-related death. The model showed good discrimination for internal and external validation with C-statistics of 0.89 and 0.98, respectively. Even though the model shows promising performance for internal calibration, however, external validation showed over- and under-prediction for low-risk and high-risk patients, respectively.

Yan et al. [1] reported a machine learning approach to select three biomarkers (lactic dehydrogenase (LDH), lymphocyte, and high-sensitivity C-reactive protein (hs-CRP)) and used them to predict individual patient's mortality in on average 10 days ahead with more than 90% accuracy. In particular, high levels of LDH alone have been found to play a crucial role in identifying the vast majority of cases, which require immediate medical attention. However, there is no scoring system reported in this work, which can help the clinicians to identify the patients under risk quantitatively. Moreover, the claim of 10 days earlier prediction is an average matrix while some patients were admitted 32 days earlier than their actual outcome and some patient died in the same day of admission, and therefore, this average matrix will not reflect the performance of the model in individual level.

Another clinical study on 82 COVID-19 patients showed that respiratory, cardiac, hemorrhage, hepatic, and renal injury had caused the death of 100%, 89%, 80.5%, 78.0%, and 31.7% patients, respectively. Most of the patients had increased CRP (100%) and D-dimer (97.1%) [22]. The value of D-dimer as a prognostic factor was also shown to significantly increase odds of death if the amount is greater than $1 \mu\text{g mL}^{-1}$ upon admission [23, 24].

Although several predictive prognostic models are proposed for the early detection of individuals at high risk of COVID-19 mortality, a major gap remains in the design of state-of-the-art interpretable machine learning-based algorithms and high-performance quantitative scoring system to classify the most selective predictive biomarkers of patient death. Identifying and prioritizing those at severe risks is important for both resource planning and treatment therapy. Moreover, the high-risk patients should be possible to continuously monitored using a reliable scoring tool during their hospital stay-time. Likewise, reducing patient admission with very low risk of complications that can be handled safely by self-quarantine will help to minimize the load on healthcare facilities.

The aim of the study is to provide a simple, easy-to-use and reliable scoring system for the prognosis of risk severity of individuals suffering from COVID-19 to stratify them into appropriate risk group and provide them necessary health support accordingly. Therefore, using the state-of-the-art machine learning algorithm, an early warning system for mortality risk-prediction using a nomogram-based scoring system was developed. We have identified the discriminatory biomarkers that contributed the most in the classification of death and survival of COVID-19 patients among 76 biomarkers available in the dataset. The top ranked features and the best performing classification model was used to develop a multivariable logistic regression-based nomogram and validated for the prognosis of death and survival. It was also investigated that how useful this model in monitoring the death risk of the patients longitudinally.

Methodology

Human Subjects and Study Design

Blood samples collected between 10 January and 18 February 2020 from 375 patients in Wuhan, China, were retrospectively analyzed to identify reliable and relevant markers of mortality risk. Medical records were collected using standard case report forms, which included information on epidemiological, demographic, clinical, laboratory, and mortality outcomes. Yan et al. [1] has published the dataset along with the article and the original study was approved by the Tongji Hospital Ethics Committee.

Patients' exclusion criteria for the study were: age (< 18 years), pregnant, breastfeeding, and missing data (> 20%). Out of 375 patients, 187 (49.9%) had fever while cough, fatigue, dyspnea, chest distress, and muscular soreness were present in 52 (13.9%), 14 (3.7%), 8 (2.1%), 7 (1.9%), and 2 (0.5%) patients, respectively.

Statistical Analysis

Stata/MP 13.0 software was used for conducting the statistical analysis. Gender variation was described using number and percentage. Continuous variables, age, and other biomarkers were reported with the number of missing data, median, mean, and quartiles (Q1, Q3) for each biomarker in death, and survival groups. Wilcoxon tests were conducted for all continuous variables while the chi-squared test was conducted for univariate analysis such as gender. Statistically significant difference was defined as a P value < 0.05. There were 76 biomarkers present in the original dataset; however, 14 biomarkers using two-different algorithms were identified as promising and are summarized in Table 1. These 14 biomarkers selected included lactate dehydrogenase

(LDH), neutrophils (%), lymphocyte (%), high-sensitivity C-reactive protein (hs-CRP), serum sodium, eosinophil (%), serum chloride, monocyte (%), international normalized ratio (INR), activated partial thromboplastin time (APTT), high-sensitivity cardiac troponin I, brain natriuretic peptide precursor (NT-proBNP), albumin, and mean corpuscular hemoglobin concentration (MCHC).

Of the 375 patients, 174 (46.4%) died, while 201 (53.6%) patients recovered from COVID-19 and were discharged from hospital. Figure 1 summarizes the outcome of patients based on their initial conditions: general (197), severe (27), and critical (151). The minimal, maximal, and median follow-up times (from hospital admission to death or discharge) for all 375 patients are 0 days, 35 days, and 12 days, respectively.

Table 1 summarizes the demographic characteristics, clinical characteristics, and clinical outcomes of the subjects in the death and survival groups. There were 142 (37.9%) patients, who were Wuhan residents, 2 (0.5%) had contact with confirmed or suspected patients, 24 (6.4%) were from familial cluster, 7 (1.9%) were health workers, 2 (0.5%) had contact with Huanan Seafood Market, and 198 (52.5%) had no contact history.

Two hundred twenty-four (59.7%) patients were male while 151 (40.3%) were female and the mean age of the patients was 58.83 years with a standard deviation of 16.46 years. Even though 76 demographic, laboratory, and clinical characteristics were available in the dataset, 14 biomarkers and two demographic variables were identified using feature ranking. Using two different feature ranking techniques, two different top-10 features were identified as most contributing features (Fig. 2). Some features are found common to both the techniques resulting in 15 different features contributing most for early prediction of death.

The detailed descriptions of 16 characteristics are listed in Table 1. It was found that gender, age, LDH, neutrophils (%), lymphocyte (%), hs-CRP, eosinophil (%), monocyte (%), INR, high-sensitivity cardiac troponin I, NT-proBNP, and albumin had statistically significant differences between the groups ($P < 0.05$), whereas serum sodium, serum chloride, APTT, and MCHC variables were not significantly different ($P > 0.05$) among the two groups. Out of these 16 characteristics, 12 characteristics were observed statistically significant. Therefore, it was important to check the most useful variables for the early prediction of death.

Feature Ranking

Even though multiple blood sample data of the patients were available, only the data from the first sample were used as inputs for model training and validation to identify the key predictors of the disease severity. The model also helps in distinguishing patients that require immediate medical

Table 1 Statistical analysis of the characteristic of the subjects' data

Item	Survived	Death	Total	Method	Statistic	P value
1 Gender				Chi-square test	$X^2 = 21.70$	< 0.00001
• Male (%)	98 (49%)	126 (72%)	224 (60%)			
• Female (%)	103 (51%)	48 (28%)	151 (40%)			
2 Age				Rank-sum test	$Z = -11$	< 0.0001
• N (missing)	201 (0)	174 (0)	375 (0)			
• Mean \pm SD	50.2 \pm 15	68.8 \pm 11.8	58.8 \pm 16.5			
• Median	51	69	62			
• Q1, Q3	37, 62	62.2, 77	46, 70			
• Min, max	18, 88	19, 95	18, 95			
3 Lactate dehydrogenase				Rank-sum test	$Z = -13.18$	< 0.0001
• N (missing)	193 (8)	163 (11)	356 (19)			
• Mean \pm SD	271 \pm 102	642 \pm 341	441 \pm 305			
• Median	250	567	336			
• Q1, Q3	203, 312	428, 762	239, 564			
• Min, max	119, 799	188, 1867	119, 1867			
4 Neutrophils (%)				Rank-sum test	$Z = -12.88$	< 0.0001
• N (missing)	194 (7)	162 (12)	356 (19)			
• Mean \pm SD	65.7 \pm 13.8	87 \pm 9.86	75.4 \pm 16.1			
• Median	66.2	89.5	77.5			
• Q1, Q3	56.5, 75.4	83.2, 93.7	64.3, 89.2			
• Min, max	1.7, 95.1	18.2, 98.7	1.7, 98.7			
5 Lymphocyte (%)				Rank-sum test	$Z = 11.97$	< 0.0001
• N (missing)	194 (7)	162 (12)	356 (19)			
• Mean \pm SD	24.8 \pm 11.4	7.6 \pm 6.22	17 \pm 12.7			
• Median	23.8	5.8	14.4			
• Q1, Q3	16.6, 33.5	3.3, 10.1	6.1, 25.2			
• Min, max	4.1, 60	0, 44.3	0, 60			
6 High-sensitivity C-reactive protein				Rank-sum test	$Z = -11.93$	< 0.0001
• N (missing)	194 (7)	159 (15)	353 (22)			
• Mean \pm SD	36 \pm 44	127 \pm 75.5	77 \pm 75.4			
• Median	19	114	53			
• Q1, Q3	4, 50	62, 179	12, 118			
• Min, max	0, 237	4, 320	0, 320			
7 Serum sodium				Rank-sum test	$Z = -1.57$	0.12
• N (missing)	193 (8)	161 (13)	354 (21)			
• Mean \pm SD	138.9 \pm 3.38	139.9 \pm 8.37	139.3 \pm 6.18			
• Median	139.2	138.9	139			
• Q1, Q3	136.6, 141	135.8, 143	136.3, 142			
• Min, max	125, 146.4	115.4, 179	115.4, 179			
8 Eosinophil (%)				Rank-sum test	$Z = 6.63$	< 0.0001
• N (missing)	194 (7)	162 (12)	356 (19)			
• Mean \pm SD	0.7 \pm 0.941	0.11 \pm 0.38	0.44 \pm 0.79			
• Median	0.3	0.00	0.00			
• Q1, Q3	0, 1.1	0.0, 0.0	0.00, 0.53			
• Min, max	0, 6.40	0, 3.70	0.00, 6.40			
9 Serum chloride				Rank-sum test	$Z = -0.65$	0.52
• N (missing)	193 (8)	161 (13)	354 (21)			
• Mean \pm SD	100.8 \pm 3.8	101.5 \pm 8.56	101.1 \pm 6.42			
• Median	101.3	100.6	101.1			
• Q1, Q3	98.8, 103.3	97.1, 105.5	97.9, 103.9			
• Min, max	85.6, 109.1	71.5, 140	71.5, 140			
10 Monocyte (%)				Rank-sum test	$Z = 8.42$	< 0.0001
• N (missing)	194 (7)	152 (12)	356 (19)			
• Mean \pm SD	8.4 \pm 3.15	5.1 \pm 4.31	6.9 \pm 4.08			
• Median	8.2	4	6.8			
• Q1, Q3	6.6, 10.1	2.4, 6.3	3.8, 9.2			
• Min, max	0.7, 15.8	0.3, 35.2	0.3, 35.2			

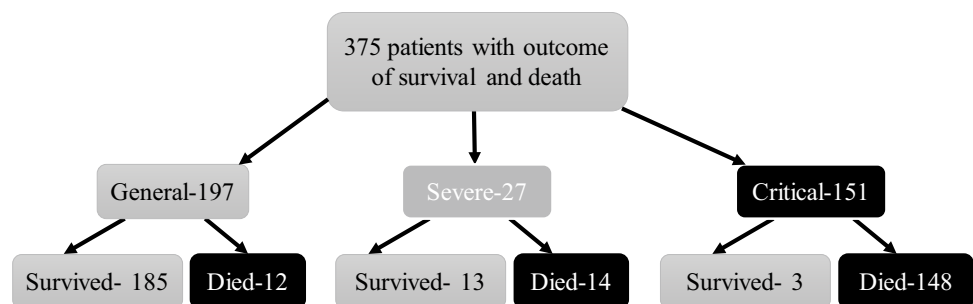
Table 1 (continued)

Item	Survived	Death	Total	Method	Statistic	P value
11 International standard ratio				Rank-sum test	Z = - 9.4	< 0.0001
• N (missing)	189 (12)	163 (11)	352 (23)			
• Mean ± SD	1.055 ± 0.086	1.37(1.01)	1.2 ± 0.709			
• Median	1.040	1.22	1.1			
• Q1, Q3	1, 1.1	1.1, 1.37	1, 1.2			
• Min, max	0.84, 1.33	0.88, 13.48	0.8, 13.5			
12 Activation of partial thromboplastin time				Rank-sum test	Z = - 1.2	0.23
• N (missing)	165 (36)	133 (41)	298 (77)			
• Mean ± SD	40.1 ± 5.7	41.9 ± 11.4	41 ± 8.7			
• Median	39.9	39.4	40			
• Q1, Q3	35.9, 43.5	35, 45.4	36, 44			
• Min, max	22, 56.9	25.3,137	22, 137			
13 Hypersensitive cardiac troponin I				Rank-sum test	Z = - 5.82	< 0.0001
• N (missing)	141 (60)	146 (28)	287 (88)			
• Mean ± SD	12 ± 53.3	1391 ± 5748	714 ± 414			
• Median	3	41	11			
• Q1, Q3	2, 7	15, 271	3, 50			
• Min, max	2, 617	2, 50,000	2, 50,000			
14 Brain natriuretic peptide precursor (NT-proBNP)				Rank-sum test	Z = - 3.87	< 0.0001
• N (missing)	128 (73)	139 (35)	267 (108)			
• Mean ± SD	1039 ± 6620	2806 ± 5906	1959 ± 6308			
• Median	65	827	271			
• Q1, Q3	23, 178	362, 2402	68, 935			
• Min, max	5, 70,000	24, 45,850	5,70,000			
15 Albumin				Rank-sum test	Z = 10.64	< 0.0001
• N (missing)	193 (8)	163 (11)	356 (19)			
• Mean ± SD	37.1 ± 4.53	30.3 ± 4.22	34 ± 5.57			
• Median	37.4	30.1	34.2			
• Q1, Q3	34.2, 40.2	27.6, 33	29.9, 38.3			
• Min, max	22.6, 48.6	18.5, 40.9	18.5, 48.6			
16 Mean corpuscular hemoglobin concentration				Rank-sum test	Z = - 2.27	0.023
• N (missing)	194 (7)	162 (12)	356 (19)			
• Mean ± SD	343 ± 13.9	346 ± 18.7	345 ± 16.3			
• Median	344	346	345			
• Q1, Q3	335, 351	337,354	336, 352			
• Min, max	306,416	299,488	299, 488			
17 Outcome (%)	201(54%)	174(46%)	375			

assistance. Research using clinically captured data often suffers from missing data challenge leading to either bias introduction or negative impact on analytical outcomes. Simple approach to handle this challenge is deleting the

respective rows of data from further analysis. It is stated in [25] that this simple approach of deleting rows with missing data can sometime lead to loss of valuable information that would have been beneficial in the analysis and also can lead

Fig. 1 Patients’ outcome tree with the initial condition of the patients in admission



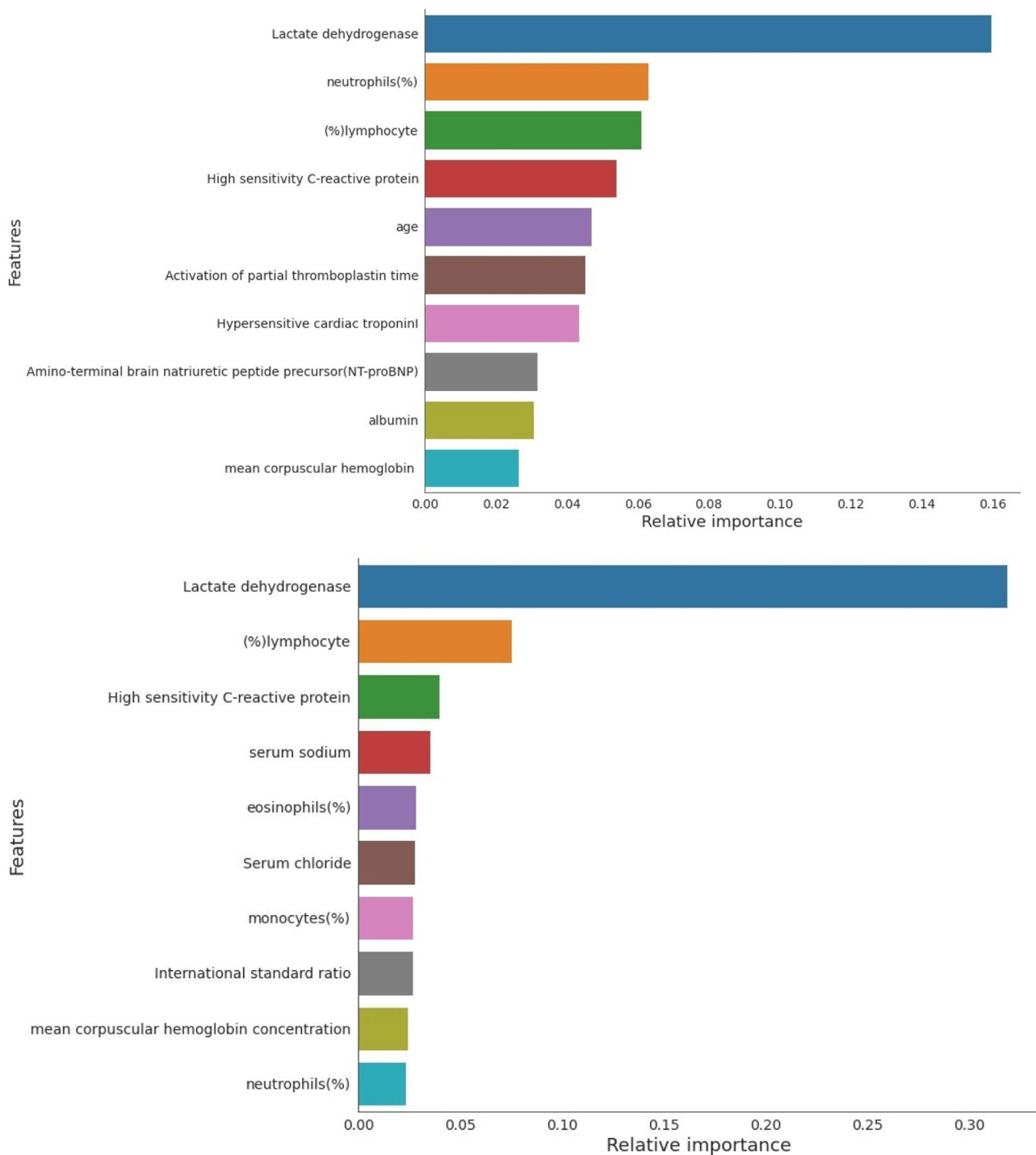


Fig. 2 Comparison of the top-ranked 10 features identified using Multi-Tree XGBoost algorithm from data imputed using MICE (top) and (- 1) (bottom)

to biased estimates. If the dataset is large, deleting the rows will have less impact on the missing variables; however, in this study, there are only 375 patients' data available, and therefore, discarding rows of missing data will significantly affect the model performance.

There are many popular data imputation techniques to deal with the missing data. Multiple imputation using chained equations (MICE) data imputation technique is the most popular technique for clinical data imputation [26]. MICE technique uses multiple regression models to predict

the missing data depending on the other available variables in the dataset. The technique also takes into account the datatype of the missing variables while imputing. Binary variables are predicted using logistic regression while continuous variables were predicted using predictive mean matching [25]. The imputed data is then normalized using “Z score” [27, 28]. Yan et al. [1] have released the database of 76 biomarkers from 375 infected patients of China and developed a prognostic model for the mortality risk of COVID-19 patients. In their study, missing data were padded by “- 1” as also commonly used by the researchers and normalized by Z score. In this study, the two different data imputation techniques are compared: imputation by - 1 and imputation using MICE technique. In both the cases, normalization was carried out using Z score technique.

Each of the 76 parameters of the dataset, after data imputation, was assessed to take decisions and identify the top-14 biomarkers in addition to age and gender, to obtain the top-ranked biomarkers as mortality predictors. Two different sets (top 10 features) were identified using Multi-Tree Extreme Gradient Boost (XGBoost) technique [29], according to their importance from the best imputation approach discussed above. The importance of each individual feature in XGBoost is from its accumulated use in each decision step in trees. The approach is extremely useful when dealing with clinical parameters [1, 30]. Initially, default settings of XGBoost was used, i.e., maximum depth = 4, learning rate = 0.2, tree estimators = 150, regularization parameter $\alpha = 1$, and “subsample” and “colsample bytree” both set to 0.9 to avoid overfitting for cases with many features and limited sample size [1, 31].

Development and Validation of the Logistic Regression Model in Classifying the Outcome

Logistic regression is a model used commonly in medical statistics and is a statistical learning technique categorized in supervised machine learning (ML) methods dedicated to classification tasks [32]. It is really useful when we wish to estimate the probability of a bipartite outcome such as survival or death of a patient [33]. The logistic function is a sigmoid function and shrinks real value continuous inputs into a probability. They also make the independent values more resistant to deviations from normality and thus more consistent coefficients [32]. Therefore, this study uses a supervised logistic regression classifier [34] as the predictor model.

Receiver operating characteristic (ROC) curves for the test data were used to calculate area under the curve (AUC) for the predictor variables separately and also in combination. The AUC values helped in order to evaluate the performance of different top ranked features in classifying death and survival cases. After getting the top ranked features, the logistic regression classifier, discussed in the previous paragraph, was

evaluated for different combinations of features as input to the model. The different combination models were validated using fivefold cross-validation (80% data were used for training and validation while remaining 20% data were used for testing and this is repeated 5 times). The performance of different models were evaluated using several performances metrics including sensitivity, specificity, positive likelihood ratio (PLR) and negative likelihood ratio (NLR) using testing dataset. Per-class values were computed over the overall confusion matrix that accumulates all test (unseen) fold results of the fivefold cross-validation.

$$Sensitivity_{class_i} = \frac{TP_{class_i}}{TP_{class_i} + FN_{class_i}} \quad (1)$$

$$Specificity_{class_i} = \frac{TN_{class_i}}{TN_{class_i} + FP_{class_i}} \quad (2)$$

$$PLR_{class_i} = \frac{Sensitivity_{class_i}}{1 - Specificity_{class_i}} \quad (3)$$

$$NLR_{class_i} = \frac{1 - Sensitivity_{class_i}}{Specificity_{class_i}} \quad (4)$$

where $class_i$ = survival and death.

In the above Eqs. , true positive (TP), true negative (TN), false positive (FP), and false negative (FN) were used to denote the number of died patients identified as died, the number of survivors identified as survivors, the number of survivors incorrectly identified as died patients, and the number of died patients incorrectly identified as survivors, respectively.

Development and Validation of Logistic Regression-Based Nomogram in the Outcome Prediction

A diagnosis nomogram was constructed using Stata/MP software version 13.0 and Nomolog (developed by Alexander Zlotnik) [35], based on multivariate logistic regression analysis. Logistic (logit) regression estimates the parameter in the form of a binary regression. Logistic regression works with probability, odds and regression. In the binary logistic model, there is an outcome/indicator variable which has two possible values. The outcome variable is a dependent variable which is typically labeled as “0” and “1” where “0” represent survival and “1” represents death in this case. The odds are the ratio of the probability (P) of an event happening to the probability of not happening, as shown in Eq. 5. Although the probability can vary between 0 and 1, the odds can vary between 0 and ∞ . In logistic regression, the logarithm of odds

is a linear combination of one or more independent variables (“predictors”) which can be a binary variable (e.g., gender) and continuous variable (e.g., age). The log-odds can be termed as linear prediction (LP), as shown in Eq. 6, and can be related to the probability of a particular outcome. Equations are used to create relationship between death probability and the key predictors using logistic regression.

$$\text{odds} = \frac{P}{1 - P} \quad (5)$$

$$LP = \ln(\text{odds}) = \ln\left(\frac{P}{1 - P}\right) = b_0 + b_1x_1 + b_2x_2 + \dots + b_nx_n \quad (6)$$

$$\frac{P}{1 - P} = e^{b_0 + b_1x_1 + b_2x_2 + \dots + b_nx_n} = e^{LP} \quad (7)$$

$$P = \frac{e^{LP}}{1 + e^{LP}} = \frac{1}{1 + e^{-LP}} \quad (8)$$

The top-ranked features (independent variables) showing best AUC was used for creating the logistic regression-based nomogram. The entire dataset was divided into training (70%) and validation (30%) sets. Calibration curves for internal (with development set) and external (with validation set) validation were plotted to compare predicted and actual death probability of patients with COVID-19. Decision curve analysis (DCA) was carried out to identify the threshold values in which nomograms were clinically useful, using Stata software.

Development and Validation of Early Warning Score

The parameters were drawn as a numerated horizontal axis scale and the values for the patient are put on the numerated scale. A vertical line was drawn down from the different parameter numerated arranged scales downward to a score axis. All five scores on the score axis were added to make a total score, and this was linked to a death probability. It can be noted that according to the nomogram, higher score corresponds to a higher death probability. The model was designed using the initial blood sample of the patients. However, it can be applied to the biomarkers collected in later during the hospital stay period of the patients to predict death probability longitudinally using the LNLCA score.

Results

This section will discuss (a) the performance comparison of the two imputation techniques, (b) validation of the nomogram in predicting death using the best data imputation technique and the five top ranked features, and finally (c) evaluates the prognostic model.

Performance Comparison of Two Imputation Techniques

To identify the best data imputation technique and to determine the most contributing independent variables associated with death, data imputed with two different techniques were investigated with top-1, top-2, and up to top-10 features in each case of imputed data. It is clear from the Fig. 3 that top-ranked 5 features produced highest AUC of 0.97 for data imputed using MICE algorithm while top-ranked 3 features produced highest AUC of 0.95 for the data imputed using -1 (Fig. 3). Table 2 shows the overall accuracies and weighted average performance for other matrices for different models using top 1 to 10 features for fivefold cross-validation using the logistic regression classifier along with the confusion matrices for each case.

Top-ranked 5 features using MICE data imputation showed better performance than the top-ranked 4 features for the data imputed by -1 . Therefore, in the rest of the study, 5 top-ranked MICE imputed independent variables—lactate dehydrogenase, neutrophils (%), lymphocyte (%), high-sensitivity C-reactive protein, and age (in short LNLCA)—were used for nomogram creation and scoring technique development and validation.

Evaluation of Nomogram in Predicting Death

A multivariate logistic regression-based nomogram for predicting early COVID-19 mortality was built using top-ranked five biomarkers that were found important both statistically and using ML-based classifier (as shown in Tables 1 and 2 and Fig. 3). The relationship between linear prediction of death and these biomarkers was evaluated using multivariable logistic regression with bootstrapping (1000 repetition) which is reported in Table 3. Regression coefficient, z value, standard error and its statistical significance along with 95% confidence interval are shown in Table 3. z Value is the ratio of regression coefficient and its standard error. Typically z value indicates the strong and weak contributors in logistic regression. The higher z values (either positive/negative) represent a strong contributor while values close to 0 represent a weak contributor. Therefore, out of 5 variables, neutrophil (%) is not a very strong predictor while age and lactate dehydrogenase are strong contributors. A null hypothesis of particular regression coefficient can determine the P value to relate the significance of a particular X variable in relationship to the Y variable. The X variables for which P is less than 0.05 have significant relationship to Y variables. This also reflects that the neutrophils (%) are weakly related to Y variable. However, the logistic regression classifier shows that 5 variables outperform than 4 variables. Therefore, no variable was discarded out of these 5 variables in developing the nomogram.

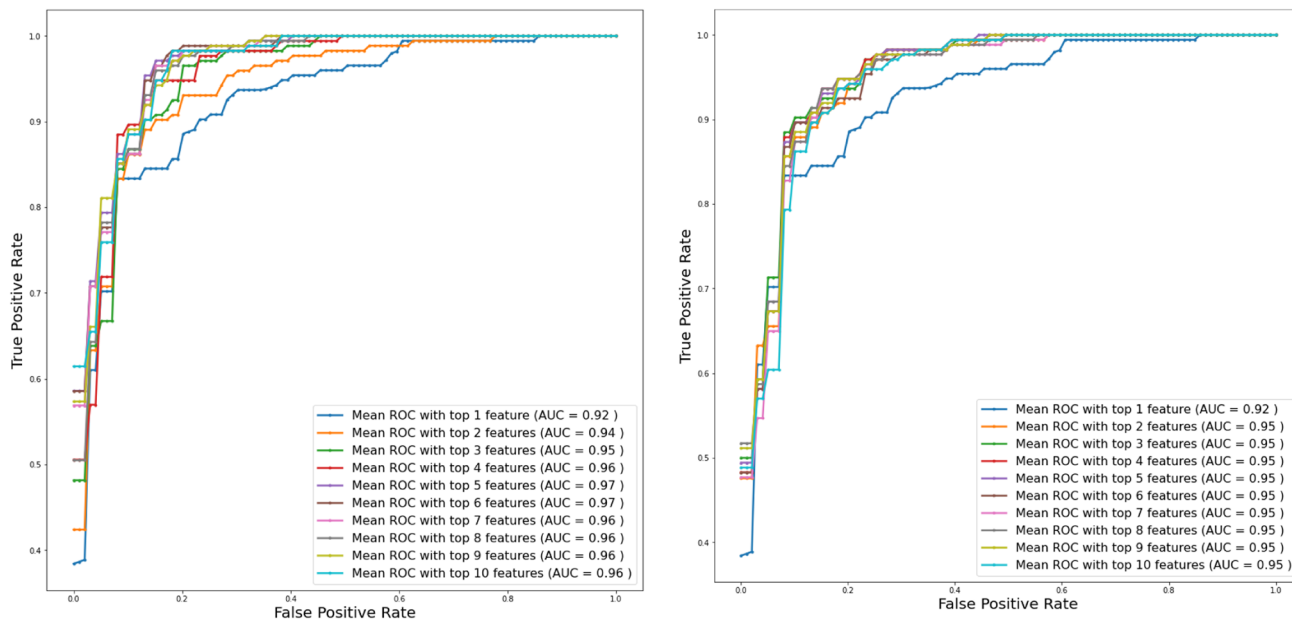


Fig. 3 Comparison of the receive operating characteristic (ROC) plots for top-ranked 1 up to 10 features using the data imputation using MICE (left) and (- 1) (right) while feature selection and classification techniques were same

According to Fig. 4, the calibration plot graphed closely toward the diagonal line both for internal and external validation which were indicative of the reliable model. It is evident from Fig. 5 that the net benefit of every single predictor model is positive until threshold of 0.85. This indicates that all of them contributed to the prediction of outcomes. Interestingly, the full model demonstrated the best performance which also confirmed the need to combine five predictors in the model.

As shown in Fig. 6, the nomogram is comprised of 8 rows while rows 1–5 are representing independent variables. For each variable, an assigned score was obtained by drawing a downward vertical line from the value on the variable axis to the “points” axis using COVID-19 patient data. The points of the five variables correspond to score (row 6), and the scores were added up to the total score, as shown in row 8. Then, a line could be drawn from the “Total Score” axis to the “Prob” axis (row 7) to determine the death probability of COVID-19 patients. However, it is useful to derive the mathematical equations explaining the total score, linear prediction, and death probability based on which the LNLCA score is calculated, using the similar equations derived earlier in Eqs. 5–8:

$$\begin{aligned}
 \text{Linear prediction} = & - 3.662636 + 0.0735038 \times \text{age (years)} \\
 & + 0.0110451 \times \text{hsCRP} \left(\frac{\text{mg}}{\text{L}} \right) \\
 & - 0.1624422 \times \text{lymphocyte}(\%) \\
 & - 0.0327053 \times \text{neutrophils}(\%) \\
 & + 0.0070514 \times \text{lactate dehydrogenase} \left(\frac{\text{u}}{\text{L}} \right) \quad (9)
 \end{aligned}$$

$$\text{Death probability} = 1 / (1 + \exp(-\text{linear prediction})) \quad (10)$$

The corresponding probability of death for a given LNLCA score was determined from the model and is listed in Table 4. In particular, LNLCA score cutoff values of 10.4 and 12.65 correspond to 5% and 50% of death probability (based on [20]); thus, these values can be used to stratify COVID-19 patients into three groups: low-, moderate-, and high-risk groups. The death probability were less than 5%, between 5 and 50%, and more than 50% for low-risk group (LNLCA < 10.4), moderate-risk group (10.4 ≤ LNLCA ≤ 12.65), and high-risk group (LNLCA > 12.65), respectively.

Performance Evaluation of the Model

Figure 7 shows an example nomogram-based scoring system for a COVID-19 patient with the variable values at admission. Individual score for each predictor was calculated and added to produce total score, and death probability was calculated to 80%. This can be done as early as 9 days before the death of the patient.

Furthermore, we have categorized the patients from training and testing subgroups into three subgroups (low, moderate, and high risk) by associating actual outcome with the predicted outcome using the LNLCA score. For training set (Table 5), the proportions of death were 0% (0/83) for low-risk group, 22.6% (12/53) for moderate-risk group, and 88.1% (111/126) for high-risk group while for test set (Table 6), the proportions of death

Table 2 Comparison of the average performance matrix and confusion matrix from five-fold cross-validation for top1 to 10 features using data imputation using (-1) (A) and mice (B)

	Weighted average (95% confidence interval)				Confusion matrix			
	Sensitivity	Specificity	PLR	NLR	Death		Survived	
					TP	FN	FP	TN
A (imputation using - 1)								
Top 1 feature	87 ± 3.92	87.4 ± 3.01	7.4 ± 4.1	0.15 ± 0.1	142	32	14	187
Top 2 features	88.04 ± 3.13	88 ± 3.5	8.1 ± 5.1	0.14 ± 0.08	148	26	17	184
Top 3 features	90 ± 3.8	88.9 ± 3.78	9.3 ± 6.9	0.12 ± 0.09	155	19	19	182
Top 4 features	90.5 ± 3.92	90.7 ± 3.72	11.8 ± 10.1	0.10 ± 0.09	157	17	18	183
Top 5 features	90.1 ± 3.6	90.03 ± 3.5	10.5 ± 7.9	0.11 ± 0.086	155	19	18	183
Top 6 features	90.08 ± 2.7	90 ± 2.4	9.63 ± 5.1	0.11 ± 0.06	154	20	19	182
Top 7 features	89.8 ± 2.3	90.16 ± 3.4	10.5 ± 7.5	0.12 ± 0.05	156	18	21	180
Top 8 features	89.3 ± 3.6	89.1 ± 3	8.96 ± 5.5	0.12 ± 0.08	155	19	21	180
Top 9 features	89.6 ± 3.2	88.9 ± 3.5	9.06 ± 6.2	0.11 ± 0.07	153	21	20	181
Top 10 features	89.01 ± 3.3	89.01 ± 4	9.46 ± 7.3	0.13 ± 0.083	154	20	21	180
B (imputation using MICE)								
Top 1 feature	88.2 ± 7.4	87.6 ± 3.5	7.91 ± 5.6	0.13 ± 0.17	143	31	13	188
Top 2 features	87.7 ± 4.4	87.01 ± 3.5	7.37 ± 4.6	0.14 ± 0.11	145	29	17	184
Top 3 features	87.1 ± 3.5	87 ± 4.1	7.53 ± 5.2	0.15 ± 0.09	148	26	22	179
Top 4 features	89.2 ± 2.8	89 ± 3.2	8.93 ± 5.6	0.12 ± 0.07	155	19	22	179
Top 5 features	92 ± 2.6	92 ± 3	13.52 ± 10.6	0.09 ± 0.06	160	14	16	185
Top 6 features	92.3 ± 2.45	92 ± 4.1	15.86 ± 16.5	0.085 ± 0.06	162	12	17	184
Top 7 features	90.2 ± 5	90.6 ± 3.5	11.37 ± 9.3	0.11 ± 0.12	158	16	22	179
Top 8 features	89.9 ± 4.8	90.2 ± 3.8	11.02 ± 9.3	0.11 ± 0.11	158	16	23	178
Top 9 features	89.2 ± 2.8	89.03 ± 3.2	8.97 ± 5.6	0.12 ± 0.07	155	19	22	179
Top 10 features	88 ± 3.4	89.6 ± 3.7	9.82 ± 7.5	0.14 ± 0.08	156	18	23	178

were 0% (0/41) for low-risk group, 22.7% (5/22) for moderate-risk group, and 94% (3/50) for high-risk group. It was found that the true death rates were significantly different ($P < 0.001$) among the three subgroups. Therefore, this nomogram-based scoring technique can be used to early predict patients' outcome to categorize them into low-, moderate-, and high-risk groups as shown in Table 4 and prioritize the moderate- and high-risk group patients.

There were 52 patients in the test set who had an outcome of death after different duration of hospital stay. Some patients were hospitalized in very late stages while some other patients were admitted in the early stages. The minimum, maximum,

mean (\pm standard deviation), and median of hospital admission to death for the test data set were 3.68, 760.92, 249.2 ± 227.55 , and 172.79 h, respectively. Most patients out of the 375 patients of the cohort had multiple blood samples taken throughout their hospital stay. LNLCA model-based prediction score was calculated on the admission and also calculated for the next available samples and identified when the model is predicting the patient in high-risk group in the earliest possible time after admission. Figure 8 shows the difference in hours between hospital admissions to the event of death and also shows when the model can predict the potential outcome with 100% accuracy. It was evident from Fig. 8 that the model can predict the outcome

Table 3 The logistic regression analysis to construct the nomogram for death prediction

Outcome	Coef	Std. err	z	$P > z $	[95% conf. interval]	
Lactate dehydrogenase	0.0070514	0.0017099	4.12	0.000	0.0037001	0.0104027
Neutrophils	- 0.0327053	0.0568836	- 0.57	0.565	- 0.1441951	0.0787845
Lymphocyte	- 0.1624422	0.0806231	- 2.01	0.044	- 0.3204607	- 0.0044238
High-sensitivity CRP	0.0110451	0.0043462	2.54	0.011	0.0025267	0.0195635
Age	0.0735038	0.0185211	3.97	0.000	0.0372032	0.1098045
_cons	- 3.662636	5.65169	- 0.65	0.517	- 14.73975	7.414473

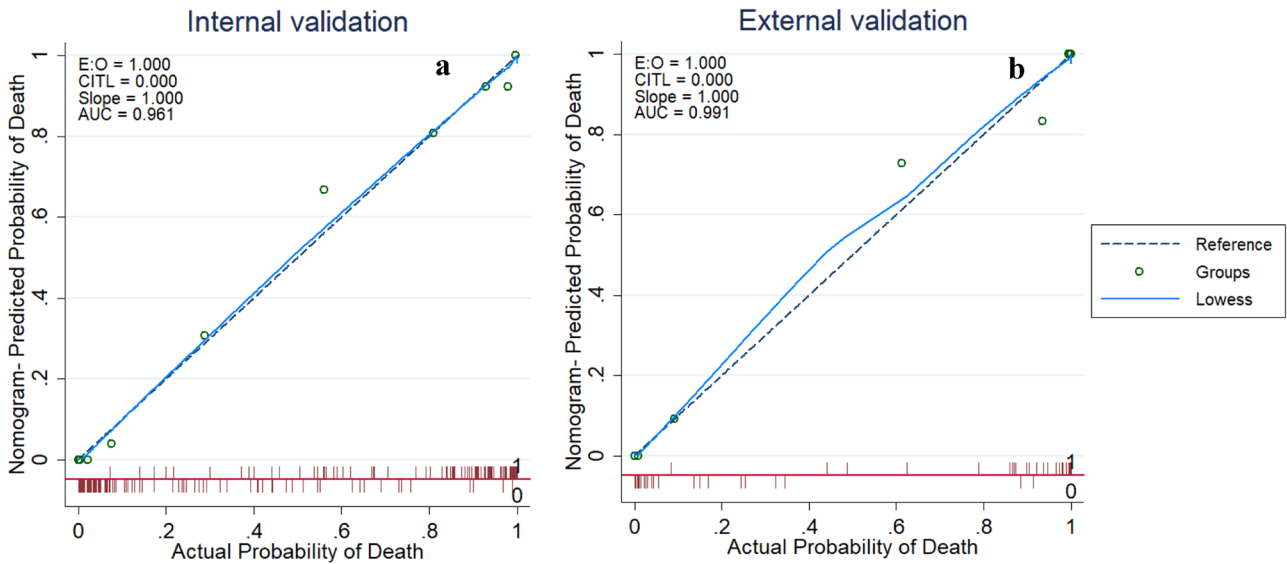


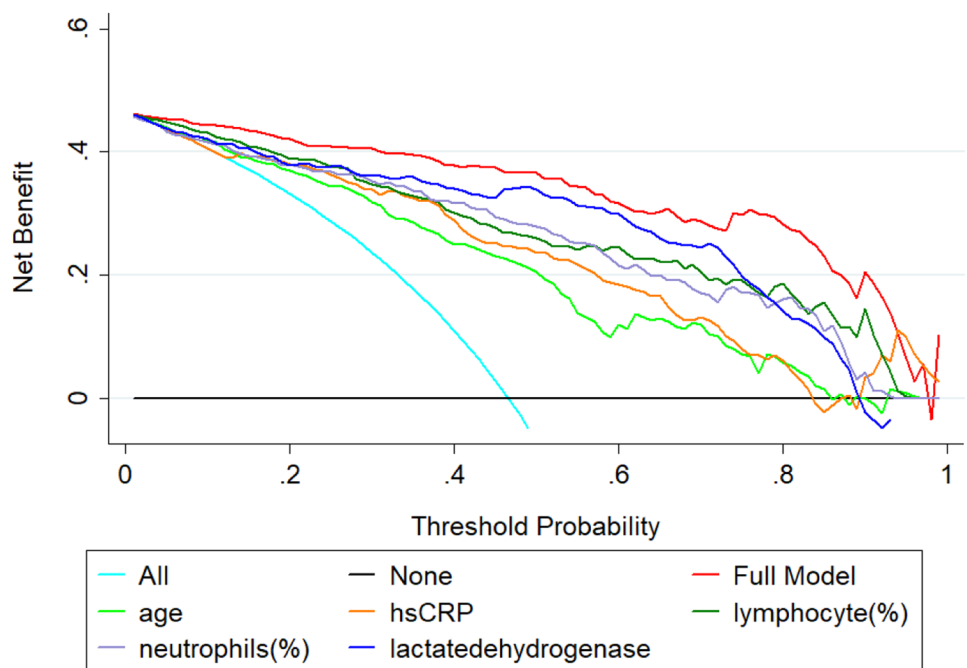
Fig. 4 Calibration plot comparing predicted and actual death probability of patients with COVID-19. **a** Internal validation. **b** External validation

of 52 patients within several hours after admission for most of the patients. The minimum, maximum, mean (\pm standard deviation), and median of model’s high-risk prediction to death for the test data set were 3.68, 756.11, 239.85 ± 228.56 , and 156.36 h, respectively. The model can even predict 31.5 days in advance for a patient about the outcome with a probability of 97%. This early prediction suggests that, where a patient’s condition deteriorates, the clinical route is able to give an early warning to clinicians several days in advance.

Discussion

Yan et al. [1] have ranked three biomarkers (lactic dehydrogenase (LDH), lymphocyte and high-sensitivity C-reactive protein (hs-CRP)) and used them to predict individual patients’ mortality using machine learning model. However, there is no scoring system reported in this work, which can help the clinicians to identify the patients under risk quantitatively. Current study investigated the

Fig. 5 Decision curves analysis comparing different models to predict the death probability of patients with COVID-19. The net benefit balances the mortality risk and potential harm from unnecessary over-intervention for patients with COVID-19



LNLCA - an early warning tool for predicting mortality risk of COVID-19 patients

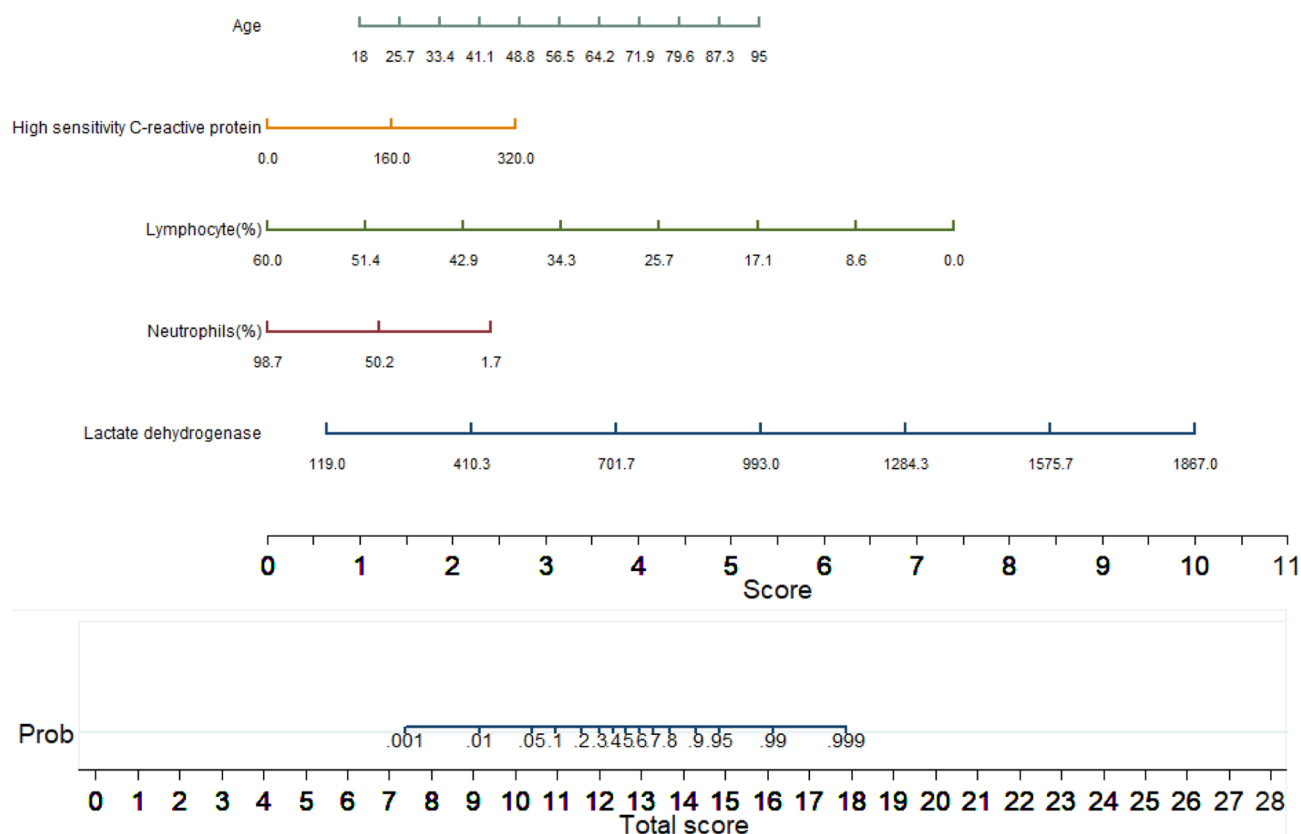


Fig. 6 Multivariate logistic regression-based nomogram to predict the probability of death. Nomogram for prediction of death was created using the following five predictors: lactate dehydrogenase, neutrophils (%), lymphocytes (%), high-sensitivity C-reactive protein, and age

relationship between the disease severity and the different clinical biomarkers and developed a scoring system to

Table 4 LNLCA score from nomogram and corresponding death probability of COVID-19 patients

Patient group	LNLCA score	Death probability
Low	7.45	0.001
	9.2	0.01
	10.4	0.05
	10.95	0.1
	11.6	0.2
Moderate	11.99	0.3
	12.4	0.4
	12.65	0.5
	12.95	0.6
	13.3	0.7
	13.7	0.8
High	14.3	0.9
	14.8	0.95
	16.2	0.99
	17.85	0.999

stratify the severity of the patients so that clinicians can identify the patients under risk easily.

Ten predictors were identified by Multi-Tree XGBoost algorithm as death probability predictors based on the data acquired at hospital admission time. Two different prediction models were compared while missing data were imputed using – 1 and using MICE algorithm. Ten different classification models were trained, validated, and tested for top 1 to 10 features using two different techniques. It was observed from the AUC and performance matrices that the MICE-based technique outperforms other approach with an AUC = 0.97 achieved for 5 top-ranked features. Then, a logistic regression-based nomogram was developed using these five variables. An integrated score (LNLCA) with corresponding likelihood of death was obtained for the early stratification of COVID-19 patients based on the severity prediction. This can help to effectively the use the healthcare facilities without overloading their capability.

Age was identified as a key predictor of mortality in previous studies on coronavirus family such as SARS [36], Middle East respiratory syndrome (MERS) [37], and COVID-19 [38]. This study has also concluded similar findings, and this

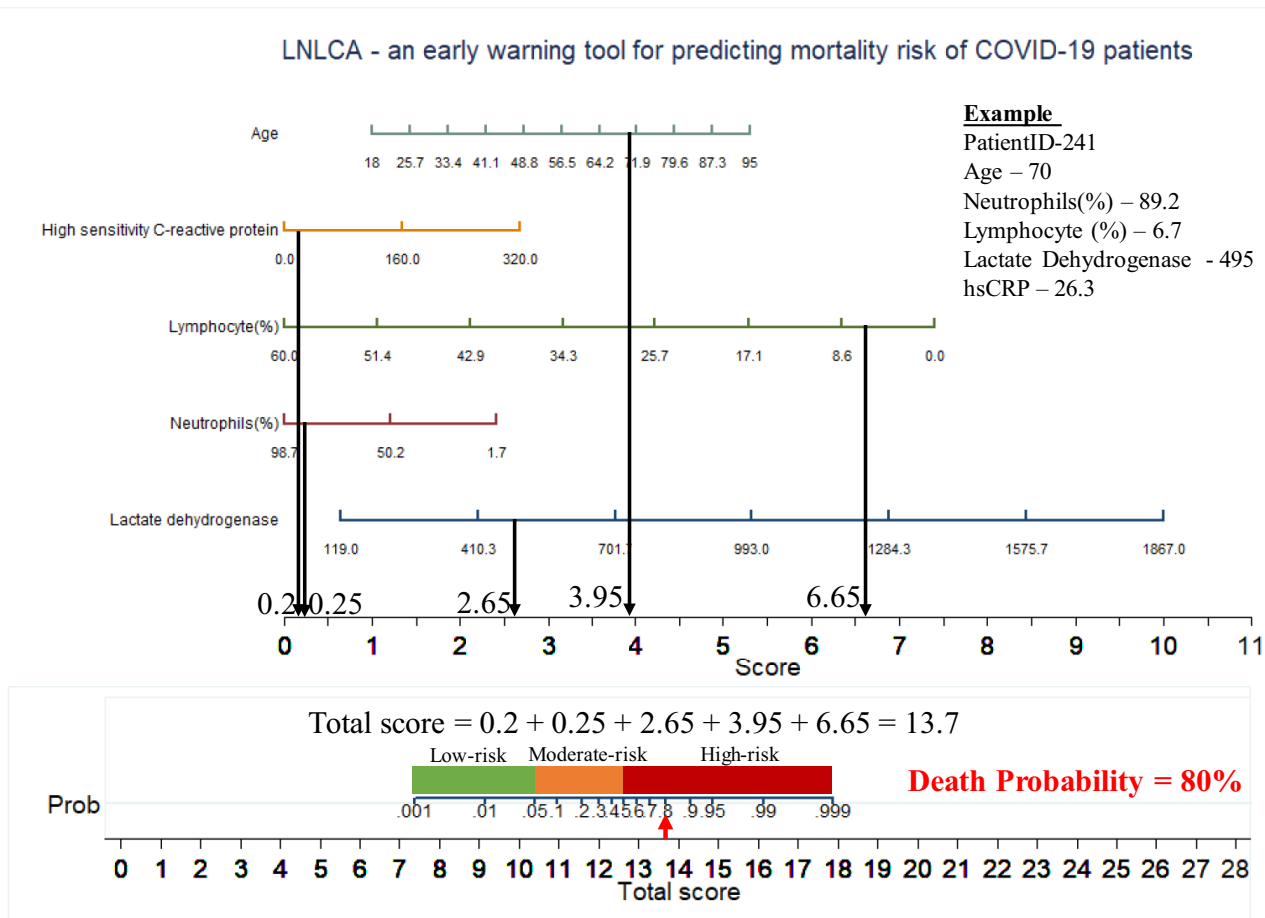


Fig. 7 An example nomogram-based score to predict the probability of death of a COVID-19 patient from test set (9 days before the actual outcome)

is because with the older age the immunosenescence and/or multiple medical conditions tend to make patients more prone to critical COVID-19 illness [20].

Yan et al. [17] showed that in patients with severe pulmonary interstitial disease, there is a significant increase of

LDH and can be associated with indications for lung injury or idiopathic pulmonary fibrosis [39]. Consistent results from the previous research were also found in this study, in which critically ill patients with COVID-19 had elevated levels of LDH, suggesting an increase in activity and severity of

Table 5 Association between different risk groups and actual outcome in the training cohort using Fisher exact probability test

Risk category	Outcome		Overall
	Alive	Death	
Low-risk	83 (100.0%)	0 (0%)	83 (100.0%)
Moderate-risk	41 (77.36%)	12 (22.64%)	53 (100.0%)
High-risk	15 (11.9%)	111 (88.1%)	126 (100.0%)
Overall	139 (53%)	123 (47%)	262 (100.0%)

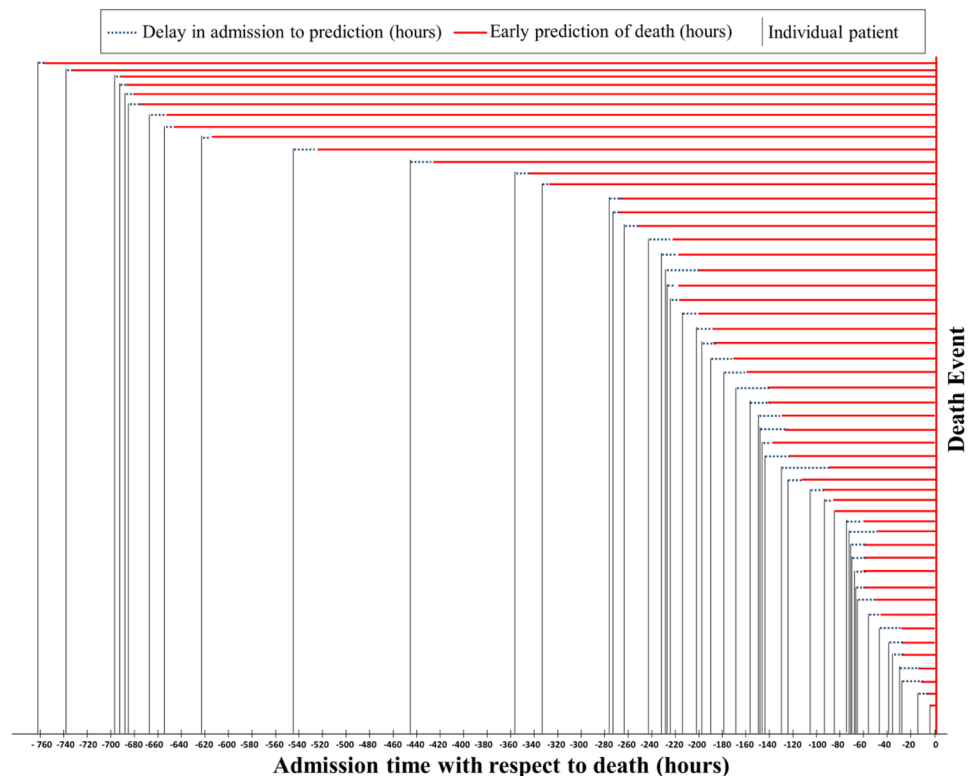
P value among three group is less than 0.001
P value of Low-risk group vs Moderate-risk group is less than 0.001
P value of Low-risk group vs High-risk group is less than 0.001
P value of Moderate-risk group vs High-risk group is less than 0.001

Table 6 Association between different risk groups and actual outcome in the testing cohort using Fisher exact probability test

Risk category	Outcome		Overall
	Alive	Death	
Low-risk	41 (100%)	0 (0%)	41 (100.0%)
Moderate-risk	17 (77.27%)	5 (22.73%)	22 (100.0%)
High-risk	3 (6%)	47 (94%)	50 (100.0%)
Overall	61 (54%)	52 (46%)	113 (100.0%)

P value among three group is less than 0.001
P value of Low-risk group vs Moderate-risk group is 0.0037
P value of Low-risk group vs High-risk group is less than 0.001
P value of Moderate-risk group vs High-risk group is less than 0.001

Fig. 8 Estimation of the prediction of the patients' outcome for 52 test patients with death outcome. The model was trained on the data present at admission, and multiple samples from a patient were used to predict the patient to be in high-risk group in the earliest time after admission. Note: "0" denotes the death outcome event for each patient, and vertical lines represent the time of admission with respect to death. Solid red line starts from the earliest prediction time point of death prediction, and the dotted line represents the delay between admission and death prediction by the model using the LNLCA model



lung injury. LDH is an intracellular enzyme that leaks from damaged cells due to infection and viral replication leading to elevated levels in circulation.

Recently, Liu et al. [40] proposed that increased neutrophil-to-lymphocyte ratio (NLR) can aid in the early prediction of the severity of COVID-19 illness. Both neutrophils and lymphocytes are critical components of the immune system and play very important role in host defense and clearing infections. Lymphopenia, medical condition due to lower number of lymphocytes in the blood, is a typical feature in COVID-19 patients and may be a key factor in disease severity and mortality [41]. In this study, we have used neutrophil and lymphocyte percentage and similar to the previous studies have found that lower percentage of these two quantities were associated with severe COVID-19 patients. According to previous research, patients with community-acquired pneumonia have significant immune system activation and/or immune dysfunction leading to changes in these quantities [41]. In addition, on the event of immunosuppression and apoptosis of lymphocytes caused by specific anti-inflammatory cytokines, bone marrow circulates neutrophils [42], resulting in an increased NLR. However, in contrast to other models, it was observed in this study, both the parameters were small for high-risk patients.

Lu et al. [43] stated that CRP tested upon admission may assist in predicting confirmed or suspected short-term mortality associated with COVID-19. CRP is an acute phase protein formed by hepatocytes caused by leukocyte-derived

cytokines induced by infection, inflammation, or tissue damage [44–46]. Similar findings were found in this study where increased CRP rates were measured at admission for the high mortality risk COVID-19 patients. This indicated that these patients developed a serious lung inflammation or possibly a secondary bacterial infection, and clinical antibiotic treatment might be appropriate for those patients [1].

Non-survivors in our study had low lymphocyte and neutrophil percentages, higher age, hs-CRP, and LDH than those of survivors. In addition to the dysregulation of the coagulation system and immune system, it can be seen that COVID-19 severity was significantly linked to the inflammatory response to the infection. This could lead to other worse medical consequences like ARDS, septic shock, and coagulopathy. Therefore, this kind of prognostic model will aid in the development of a rational and personalized therapeutic plan for the patients with critical illness.

Weng et al. [20] recently suggested that age, NLR, D-dimer, and CRP were individual key predictors correlated with death probability. These key predictors were used to create a nomogram for death prediction due to COVID-19. In our research, the five key predictors recorded at admission were chosen by the XGBoost feature selection to create a nomogram-based prognostic model that exhibits excellent calibration and discrimination in predicting death probability of COVID-19 patients. It was also validated by an unseen validation cohort. Moreover, it was verified with multiple blood sample data collected from the patients

during their hospital stay and the model holds valid for those cases as well. The AUC values for development and validation cohort showed a strong distinction of 0.961 and 0.991, respectively, using the proposed nomogram, which is, to the best of our knowledge, outperforms any other nomogram-based models for COVID-19 mortality prediction. In addition, this nomogram-derived LNLCA score offered a simple, easy-to-understand, and interpretable early detection tool for stratifying the high-risk COVID-19 patients at admission and thereby assist their clinical management. COVID-19 patients were categorized into three risk groups with varying risk of death using LNLCA score measured and calculated at admission. Low-risk group cases could be isolated and treated in an isolation center while the moderate-risk patients could be treated isolation ward in a specialized hospital. On the other hand, patients in high-risk group could be under close monitoring and should be moved to critical medical services or ICU for urgent treatment if required.

This study has scope for further improvement, which will be carried out in the future work. Firstly, the study motivates the possibility of research on COVID-19 clinical data helping in early mortality prediction but the proposed machine learning method is purely data-driven and may vary if starting from different datasets. The model can be further improved with the help of a larger dataset. Secondly, the modelling principle adopted here is to have a minimal number of features for accurate predictions to avoid overfitting, which can be revised with several other models to identify any other sets of best features on a multi-center and multi-country data to produce a generalized model.

Conclusion

In summary, based on multiple risk factors (lactate dehydrogenase, neutrophils (%), lymphocytes (%), high-sensitivity C-reactive protein, and age), our developed nomogram can predict the prognosis of patients with COVID-19 with good discrimination and calibration. The model can predict the patient's outcome far ahead of the day of primary clinical outcome with very high accuracy. Therefore, the application of LNLCA would help clinicians make an efficient and optimized patient stratification management plan without overloading the healthcare resources and also reduce the death with improved and planned response. A Web-based application is created using the model which can be used by the clinicians to take benefit of the model (<https://www.openasapp.net/portal#!/client/app/2309220f-105e-4e7d-9fe0-51cbef5bb18c>). The authors also plan to further improve the performance of the model with the help of larger dataset with multicenter and multicountry data.

Funding Open access funding provided by the Qatar National Library. This publication was made possible by Qatar University Emergency Response Grant (QUERG-CENG-2020-1) from the Qatar University. The statements made herein are solely the responsibility of the authors.

Compliance with Ethical Standards

Conflict of Interest The authors declare that they have no conflict of interest.

Ethical Approval This article uses the clinical data which was made publicly available by Yan et al. [1]. Therefore, the authors of this study were not involved with human participants or animals. However, the original retrospective study carried out by Yan et al. [1] was approved by the Tongji Hospital Ethics Committee.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

References

1. Yan L, Zhang HT, Goncalves J, Xiao Y, Wang M, Guo Y, et al. "An interpretable mortality prediction model for COVID-19 patients. *Nat Mach Intell*. 2020: 1–6.
2. Huang C, Wang Y, Li X, Ren L, Zhao J, Hu Y, et al. Clinical features of patients infected with 2019 novel coronavirus in Wuhan, China. *The lancet*. 2020;395(10223):497–506.
3. Guan W-J, Ni ZY, Hu Y, Liang WH, Ou CQ, He JX, et al. Clinical characteristics of coronavirus disease 2019 in China. *N Engl J Med*. 2020;382(18):1708.
4. Wang D, Hu B, Hu C, Zhu F, Liu X, Zhang J, et al. Clinical characteristics of 138 hospitalized patients with 2019 novel coronavirus–infected pneumonia in Wuhan, China. *JAMA*. 2020;323(11):1061–9.
5. Wu Z, McGoogan JM. Characteristics of and important lessons from the coronavirus disease 2019 (COVID-19) outbreak in China: summary of a report of 72 314 cases from the Chinese Center for Disease Control and Prevention. *JAMA*. 2020;323(13):1239–42.
6. W. H. Organization. Coronavirus disease 2019 (COVID-19): situation report-107 [Online]. Available: https://www.who.int/docs/default-source/coronaviruse/situation-reports/20200506covid-19-sitrep-107.pdf?sfvrsn=159c3dc_2
7. COVID-19 CORONAVIRUS PANDEMIC [Online]. Available: <https://www.worldometers.info/coronavirus/>
8. W. H. Organization. Coronavirus disease 2019 (COVID-19) Situation Report – 68 [Online]. Available: <https://www.who.int/docs/default-source/coronaviruse/situation-reports/20200328-sitrep-68-covid-19.pdf>
9. Grasselli G, Pesenti A, Cecconi M. Critical care utilization for the COVID-19 outbreak in Lombardy, Italy: early experience and forecast during an emergency response. *JAMA*. 2020;323(16):1545–6.

10. Moghadas SM, Shoukat A, Fitzpatrick MC, Wells CR, Sah P, Pandey A, et al. Projecting hospital utilization during the COVID-19 outbreaks in the United States. *Proc Natl Acad Sci*. 2020;117(16):9122–6.
11. D. o. V. D. National Center for Immunization and Respiratory Diseases (NCIRD). Interim Clinical Guidance for Management of Patients with Confirmed Coronavirus Disease (COVID-19) [Online]. Available: <https://www.cdc.gov/coronavirus/2019-ncov/hcp/clinical-guidance-management-patients.html>
12. Mizumoto K, Kagaya K, Zarebski A, Chowell G. Estimating the asymptomatic proportion of coronavirus disease 2019 (COVID-19) cases on board the Diamond Princess cruise ship, Yokohama, Japan, 2020. *Eurosurveillance*. 2020;25(10):2000180.
13. F. Zhou, T. Yu, R. Du, G. Fan, Y. Liu, Z. Liu, et al., "Clinical course and risk factors for mortality of adult inpatients with COVID-19 in Wuhan, China: a retrospective cohort study," *The lancet*, 2020)
14. Goh KJ, Kalimuddin S, Chan KS. Rapid Progression to Acute Respiratory Distress Syndrome: Review of Current Understanding of Critical Illness from Coronavirus Disease 2019 (COVID-19) Infection. *Ann Acad Med Singapore*. 2020;49:108–18.
15. Yang J, Zheng Y, Gou X, Pu K, Chen Z, Guo Q, et al. Prevalence of comorbidities in the novel Wuhan coronavirus (COVID-19) infection: a systematic review and meta-analysis. *Int J Infect Dis*. 2020.
16. W.-j. Guan, W.-h. Liang, Y. Zhao, H.-r. Liang, Z.-s. Chen, Y.-m. Li, et al., "Comorbidity and its impact on 1590 patients with Covid-19 in China: A Nationwide Analysis," *European Respiratory Journal*, vol. 55 (5), 2020)
17. Yan L, Zhang HT, Goncalves J, Xiao Y, Wang M, Guo Y, et al. A machine learning-based model for survival prediction in patients with severe COVID-19 infection. *MedRxiv*. 2020.
18. Covid C, Team R. Severe outcomes among patients with coronavirus disease 2019 (COVID-19)—United States, February 12–March 16, 2020. *MMWR Morb Mortal Wkly Rep*. 2020;69(12):343–6.
19. Al Youha S, Doi SA, Jamal MH, Almazeedi S, Al Haddad S, AlSeidan M, et al. Validation of the Kuwait Progression Indicator Score for predicting progression of severity in COVID19. *medRxiv*. 2020.
20. Weng Z, Chen Q, Li S, Li H, Zhang Q, Lu S, et al. ANDC: an early warning score to predict mortality risk for patients with Coronavirus Disease 2019. 2020.
21. Xie J, Hungerford D, Chen H, Abrams ST, Li S, Wang G, et al. Development and external validation of a prognostic multivariable model on admission for hospitalized patients with COVID-19. 2020.
22. B. Zhang, X. Zhou, Y. Qiu, F. Feng, J. Feng, Y. Jia, et al., "Clinical characteristics of 82 death cases with COVID-19," *MedRxiv*, 2020)
23. McRae MP, Simmons GW, Christodoulides NJ, Lu Z, Kang SK, Fenyo D, et al. Clinical decision support tool and rapid point-of-care platform for determining disease severity in patients with COVID-19. *Lab Chip*. 2020.
24. Zhang L, Yan X, Fan Q, Liu H, Liu X, Liu Z, et al. D-dimer levels on admission to predict in-hospital mortality in patients with Covid-19. *J Thromb Haemost*. 2020;18(6):1324–9.
25. Hegde H, Shimpi N, Panny A, Glurich I, Christie P, Acharya A. MICE vs PPCA: Missing data imputation in healthcare. *Inform Med Unlocked*. 2019;17:100275.
26. Buuren SV, Groothuis-Oudshoorn K. Mice: Multivariate imputation by chained equations in R. *J Stat Soft*. 2010:1–68.
27. Jain A, Nandakumar K, Ross A. Score normalization in multimodal biometric systems. *Pattern Recogn*. 2005;38(12):2270–85.
28. Patro S, Sahu KK. Normalization: a preprocessing stage. *arXiv preprint arXiv : 1503.06462*.
29. Chen T, He T, Benesty M, Khotilovich V, Tang Y. Xgboost: extreme gradient boosting. R package version 0.4–2. 2015:1–4.
30. Torlay L, Perrone-Bertolotti M, Thomas E, Baciú M. Machine learning–XGBoost analysis of language networks to classify patients with epilepsy. *Brain Informatics*. 2017;4(3):159–69.
31. Li W, Yin Y, Quan X, Zhang H. Gene Expression Value Prediction Based on XGBoost Algorithm. *Front Genet*. 2019;10:1077.
32. Subasi C. LOGISTIC REGRESSION CLASSIFIER [Online]. Available: <https://towardsdatascience.com/logistic-regression-classifier-8583e0c3cf9>
33. Anderson RP, Jin R, Grunkemeier GL. Understanding logistic regression analysis in clinical reports: an introduction. *Annals Thoracic Surg*. 2003;75(3):753–7.
34. Le Cessie S, Van Houwelingen JC. Ridge estimators in logistic regression. *J R Stat Soc: Ser C (Appl Stat)*. 1992;41(1):191–201.
35. Zlotnik A, Abaira V. A general-purpose nomogram generator for predictive logistic regression models. *Stata J*. 2015;15(2):537–46.
36. J. C. Chan, E. L. Tsui, V. C. Wong, H. A. S. C. Group. Prognostication in severe acute respiratory syndrome: A retrospective time-course analysis of 1312 laboratory-confirmed patients in Hong Kong. *Respirology*. 2007;12(4):531–42.
37. Assiri A, Al-Tawfiq JA, Al-Rabeeh AA, Al-Rabiah FA, Al-Hajjar S, Al-Barrak A, et al. Epidemiological, demographic, and clinical characteristics of 47 cases of Middle East respiratory syndrome coronavirus disease from Saudi Arabia: a descriptive study. *Lancet Infect Dis*. 2013;13(9):752–61.
38. R. Chen, W. Liang, M. Jiang, W. Guan, C. Zhan, T. Wang, et al (2020) "Risk factors of fatal outcome in hospitalized subjects with coronavirus disease 2019 from a nationwide analysis in China," *Chest*,
39. Kishaba T, Tamaki H, Shimaoka Y, Fukuyama H, Yamashiro S. Staging of acute exacerbation in patients with idiopathic pulmonary fibrosis. *Lung*. 2014;192(1):141–9.
40. Liu J, Liu Y, Xiang P, Pu L, Xiong H, Li C, et al. Neutrophil-to-lymphocyte ratio predicts severe illness patients with 2019 novel coronavirus in the early stage. *MedRxiv*. 2020.
41. Huang I, Pranata R. Lymphopenia in severe coronavirus disease-2019 (COVID-19): systematic review and meta-analysis. *J Intensive Care*. 2020;8(1):1–10.
42. Adamzik M, Broll J, Steinmann J, Westendorf AM, Rehfeld I, Kreissig C, et al. An increased alveolar CD4+ CD25+ Foxp3+ T-regulatory cell ratio in acute respiratory distress syndrome is associated with increased 30-day mortality. *Intensive Care Med*. 2013;39(10):1743–51.
43. Lu J, Hu S, Fan R, Liu Z, Yin X, Wang Q, et al. ACP risk grade: a simple mortality index for patients with confirmed or suspected severe acute respiratory syndrome coronavirus 2 disease (COVID-19) during the early stage of outbreak in Wuhan, China. 2020.
44. Ko J-H, Park GE, Lee JY, Lee JY, Cho SY, Ha YE, et al. Predictive factors for pneumonia development and progression to respiratory failure in MERS-CoV infected patients. *J Infect*. 2016;73(5):468–75.
45. Wang J, Wu X, Tian Y, Li X, Zhao X, Zhang M. Dynamic changes and diagnostic and prognostic significance of serum PCT, hs-CRP and s-100 protein in central nervous system infection. *Exp Ther Med*. 2018;16(6):5156–60.
46. Yildiz B, Poyraz H, Cetin N, Kural N, Colak O. High sensitive C-reactive protein: a new marker for urinary tract infection, VUR and renal scar. *Eur Rev Med Pharmacol Sci*. 2013;17(19):2598–604.