A descriptive study of machine learning algorithms for predicting COVID-19 patients outcome

Jie Wang [1], **Heping Yu** [2], **Qingquan Hua** [1], **Shuili Jing** [1], **Zhifen Liu** [3], **Xiang Peng** [4], **Cheng'an Cao** [Corresp. 4]

[1] Department of Otolaryngology-Head and Neck Surgery, Renmin Hospital of WUHAN University, Wuhan, Hubei, China

[2] Department of Nail and breast surgery, Wuhan Forth Hospital, Wuhan, Hubei, China

[3] Department of Nephrology, Wuhan Forth Hospital, Wuhan, Hubei, China

[4] Department of Neurosurgery, Wuhan Forth Hospital, Wuhan, Hubei, China

Corresponding Author: Cheng'an Cao
Email address: 2017202040101@whu.edu.cn

Background: The outbreak of coronavirus disease 2019 (COVID-19) occurred in Wuhan, has become a global public health threat. It is necessary to find the optimal predictors for the clinical outcomes of COVID-19 patients.

Methods: This is a retrospective cohort analysis including 126 patients diagnosed with COVID-19 from Wuhan Fourth Hospital, hospitalized for treatment during Feb. 1th to Mar. 15th, 2020. Among them, 7 patients were excluded because there was no clinical outcome. Clinical characteristics were analyzed between the alive and died patients via a random forest algorithm. A random forest classification model was contributed to find the optimal diagnostic predictors for patients' clinical outcomes between two groups, the area under the ROC curve (AUC) of train data (100%) and test data (93.3%) showed the high accuracy of a classification model. Partial dependence correlation was used to evaluate the relationship between COVID-19 survival and predictors.

Results: Of 119 patients, 103 of them were discharged and 16 died in hospital. Random forest (RF) algorithm found two optimal clinical characteristic predictors of COVID-19 patient outcome, which were LDH and Myo, partial correlation showed negative correlations between the survival and these two variables. Moreover, a substantial increase was found in the risk of in-hospital mortality for the increase of Myo (OR=7.54 95%CI, 3.42 to 16.63) and LDH (OR=4.90, 95%CI, 2.13 to 11.25).

Conclusion: In summary, we applied an integrated machine learning approach to find that LDH higher than 500U/L, and Myo higher than 80ng/ml were considered as optimal risk predictors for the prognosis of COVID-19 patients.

A descriptive study of machine learning algorithms for predicting COVID-19 Patients outcome

Jie Wang, MD[1]; Heping Yu, MD[2]; Qingquan Hua, MD[1] ;Shuili Jing[1];Zhifen Liu, MD[3]; Xiang Peng, MD[4]; Cheng'an Cao* MD[4]

[1] Department of Otolaryngology-Head and Neck Surgery, Renmin Hospital of Wuhan University, Wuhan, China

[2]Nail and breast surgery Department, Wuhan Fourth Hospital, Puai Hospital, Tongji Medical College, Huazhong University of Science and Technology, Wuhan, 430033, China.

[3]Department of Nephrology, Wuhan Fourth Hospital, Puai Hospital, Tongji Medical College, Huazhong University of Science and Technology, Wuhan, China.

[4]Neurosurgery Department, Wuhan Fourth Hospital, Puai Hospital, Tongji Medical College, Huazhong University of Science and Technology, Wuhan, 430033, China.

***Corresponding author:**

Dr. Cao, Email: cca24@163.com. Address: No.473 Hanzheng Street, Qiaokou District, Wuhan, Hubei, China

Abstract:

Background: The outbreak of coronavirus disease 2019 (COVID-19) occurred in Wuhan, has become a global public health threat. It is necessary to find the optimal predictors for the clinical outcomes of COVID-19 patients.

Methods: This is a retrospective cohort analysis including 126 patients diagnosed with COVID-19 from Wuhan Fourth Hospital, hospitalized for treatment during Feb. 1th to Mar. 15th, 2020. Among them, 7 patients were excluded because there was no clinical outcome. Clinical characteristics were analyzed between the alive and died patients via a random forest algorithm. A random forest classification model was contributed to find the optimal diagnostic predictors for patients' clinical outcomes between two groups, the area under the ROC curve (AUC) of train data (100%) and test data (93.3%) showed the high accuracy of a classification model. Partial dependence correlation was used to evaluate the relationship between COVID-19 survival and predictors.

Results: Of 119 patients, 103 of them were discharged and 16 died in hospital. Random forest (RF) algorithm found two optimal clinical characteristic predictors of COVID-19 patient outcome, which were LDH and Myo, partial correlation showed negative correlations between the survival and these two variables. Moreover, a substantial increase was found in the risk of in-hospital mortality for the increase of Myo（OR=7.54　95%CI, 3.42 to 16.63）and LDH （OR=4.90, 95%CI , 2.13 to 11.25).

Conclusion: In summary, we applied an integrated machine learning approach to find that LDH higher than 500U/L, and Myo higher than 80ng/ml were considered as optimal risk predictors for the prognosis of COVID-19 patients.

Key words：COVID-19, predictors, machine learning, patient outcome

Introduction

In December2019, an acute respiratory syndrome coronavirus pneumonia occurred in Wuhan, Hubei Province, China(Phelan et al. 2020), and attracted an intense amount of attention worldwide. WHO named it 2019-nCoV by identifying it from a patient's pharyngeal swab sample(Jan 11,2020.; COVID & Team 2020). The scientific community and infection control agencies were facing enormous challenges in controlling the increasing intensity of epidemics. However, the disease is developing rapidly around the world. By April 14, 2020, the COVID-19 has affected 210 countries, with over 1929000 confirmed cases and 119754 deaths and the epidemic situation in Wuhan is the substantial focus of Chinese attention(Dhungana 2020). Fever and bilateral infiltration of chest imaging appear to be the most common manifestation of the clinical features of pneumonia, followed by coughing and dyspnea(Wang et al. 2020). Study shows severe COVID-19 patients can develop into severe pneumonia, ARDS, and multiple organ failure leading to death, while non-severe COVID-19 patients can present as a general symptom of respiratory infection. (Chen et al. 2020; Huang et al. 2020)

Nowadays, machine learning has been widely used in the field of medical diagnosis, such as medical imaging, drug mining, diagnosis prediction. An RF contains a large number of classification accuracy obtained by using the set of trees, and each tree in the set is grown according to random parameters(Biau 2012; Matusiewicz et al. 1993). It can analyze complex interactions between clinical characteristics, improving the performance of risk prediction. Among the COVID-19 patients, although the individual condition varied between patients, the clinical characteristics of optimal diagnostic predictors for patients' clinical outcome were worth exploring.

Methods

Study population

This is a retrospective cohort analysis including 126 patients aged from 27 to 87, from Wuhan Fourth Hospital, they were all diagnosed as COVID-19 according to the World Health Organization interim guidance. Among them, 7 patients were excluded for losing the outcome. These patients were hospitalized for treatment during Feb. 1th to Mar. 15th, 2020. This research has passed the approval of the Ethics Committee of Wuhan Fourth Hospital（KY 2020-032-01）and informed consent of the study participants was waived by the Ethics Committee of the hospital for appeared highly transmissible disease.

77 **Data collection**

78 Clinical characteristics including medical history, exposure history, clinical symptoms,
79 demographic information , and laboratory findings were obtained from Wuhan Fourth Hospital
80 electronic medical record system. Three independent researchers collected and judged all the
81 information. The access was granted by the director of the hospital.

82

83 **Statistical analysis**

84 Quartiles and medians were used to compare the differences between descriptive data,χ^2 test,
85 or the Fisher exact test were employed to test the categorical data. The normal distribution of
86 laboratory results was analyzed by two independent sample t-tests, meanwhile, non-parametric
87 the Mann-Whitney-Wilcoxon test was used to detect the data that does not satisfy the normal
88 distribution, all the data above were processed by spss26. RF algorithm, combining several
89 random decision trees and aggregates their predictions by averaging(Biau & Scornet 2016), has
90 achieved great success in empirical research and its mechanism of action was being actively
91 studied(Heitner et al. 2010).

92 All data were processed with Rstudio (R 3.6.3), RF model was constructed by randomForest
93 (https://cran.r-project. org/web/packages/randomForest/), and rpart package (https://cran.r-
94 project.org/web/packages/rpart/index.html),and the validation cohort was processed with caret
95 package (http://CRAN.R- project.org/ package=caret). CART method was used to calculate the

96 decision tree, the final result was obtained by voting results of a combined prediction. Using the
97 gini index as the split criterion. The greater change in the Gini of the nodes before and after the
98 split, the more important the variables.

99 Moreover, to give a graphical depiction of the marginal effect of a variable on the
100 classification during the calculation process, a partial correlation was employed to analyze the
101 relationships between clinical data and patient prognosis. The function being plotted was defined
102 as:

103 $$\tilde{f}(x) = \frac{1}{n}\sum_{i=1}^{n} f(x, x_{iC}),$$

104

105

106 x was the variable for chosen clinical characteristics, and xiC was the other variables in the
107 clinical information. The summand was the predicted logits (log of a fraction of votes) for
108 classification:

109 $$f(x) = \log p_k(x) - \frac{1}{K}\sum_{j=1}^{K} \log p_j(x),$$

110

111

112 where K was the number of classes, and pj was the proportion of votes for class j. Pearce
113 correlation was used to calculate the correlation with the important variables of predictors of

prognosis in patients with COVID-19 to avoid over-fitting of the model caused by excessive correlation. Graphpad 8.0 was used to analyze the level of two variables in survival and non-survival patients.

Results

Clinical demographics of COVID-19 patients on admission

This study contained 126 patients who were hospitalized in Wuhan Fourth Hospital with COVID-19 (**Table1**). 78 of the patients (61.9%) were younger than 65 years old, the median age of patients was 60 years (IQR 53-69.5). These patients with COVID-19 were generally accompanied by fever(116 [92.0%] patients), 39 (34.8%) of patients had the highest temperature above 39 ° C, the median temperature was 38.6 °C (IQR 37.4°C-40°C) (Table 1). The infection of the COVID-19 was basically gender-neutral, the proportion of male and female patients barely the same. Among them, 7 of the 126 patients (5.6%) have visited the South China Seafood Market in Wuhan. Most of these patients on admission have cough (n=95 75.4%), followed by fatigue (n=74 58.7 %), dyspnea (n=70 55.6%), myalgia (n=41 32.5%), and diarrhea14 (n= 14 11.1%). In addition, many patients also suffer from other comorbidities, including hypertension (n=44 34.9%), diabetes(n=21 16.7%), cardiovascular and macrovascular disease(n=15 11.9%) ,chronic lung disease (n=13 10.3%) , gastric disease(n=7 5.6%), tumor(n=6 4.8%), chronic kidney disease (n=3 2.4%) , endocrine system diseases (n=2 1.6%). In the process of

133 treatment,83 of patients (65.9%) used Nasal cannula, 35(27.8%) of them used NMV, 5(4.0%) of
134 them used IMV. In terms of clinical severity, 61 of these patients (50.0%) were in a moderate
135 state, 38 (31.1%) were in a severe state, and 23 (18.9%) were in a critically ill state. Judging
136 from the current treatment results, 103 (86.6%) patients have been alive, while another 16
137 (13.4%) patients have died.

138

139 **Laboratory findings of COVID-19p atients on admission**

140 The laboratory results of 126 patients with Corona Virus Disease 2019 (COVID-19) were
141 shown in **Table 2**. More than 80% of patients had lymphopenia decrement, especially the
142 reduction of CD4 + and CD8 + T lymphocytes (91.3%), and about half of patients had a decrease
143 in Th / Ts ratio. C-reactive protein (CRP) increased in 85.6% of patients, and rarely procalcitonin
144 (PCT) was elevated. The coagulation function of some patients was affected, with prothrombin
145 time (PT) prolonged in about 1/2 patients and fibrinogen (FIB) increased in 2/3 patients. D-
146 Dimer was increased by76.2% of patients. Cardiac dysfunction may be present in some patients
147 because 60% of patients had elevated B-type natriuretic peptide (BNP) and 25% of patients had
148 increased creatine kinase MB (CK-MB),. A small proportion of patients have elevated aspartate
149 aminotransferase (AST) and alanine aminotransferase (ALT), and about 1/3 have elevated
150 triglyceride (TG). In addition, patients with elevated lactate dehydrogenase (LDH) accounted for
151 76.2% of the totality.

152

153 **Clinical Characteristics comparison between alive and died patients**

154 **Table 3** shows that the patients in the died group were older than those in the alive group
155 ($p<0.001$) and the majority are male (75%). The proportion of dyspnea was remarkably increased
156 in the died group ($p=0.018$), while the rest of the clinical symptoms on admission, such as
157 fatigue, myalgia and diarrhea, were not obvious. Compared with the alive group, PCO2($p=0.023$),
158 PO2($p<0.001$), SO2($p=0.029$) and admission oxygenation($p<0.001$) were significantly reduced
159 in the arterial blood gas analysis of the patients in the died group at the early stage of admission.
160 Laboratory analysis revealed that the died group had a higher proportion of neutrophils ($p=0.042$)
161 and a lower proportion of lymphocytes ($p=0.047$) than the alive group. Additionally, NLR
162 notably increased ($p=0.005$) and LMR significantly decreased ($p=0.005$) in the died group.
163 Moreover, T lymphocytes in the died group were remarkable lower than those in the alvie group
164 ($p<0.001$), both CD4+($p=0.006$) and CD8+($p<0.001$), and the Th/Ts ratio increased in the died
165 group ($p=0.002$). Compared with the alive group, the inflammation-related indices, CRP
166 ($p=0.080$) and PCT ($p=0.009$), were significantly higher in the died group. There was no obvious
167 difference in coagulation function indices between the died group and the alive group, except
168 that D-Dimer($p=0.003$) increased significantly in the died group. In addition, there were some
169 elevated biochemical indices in the died group which represented the condition of cardiac
170 dysfunction such as Myo($p<0.001$), CK($p=0.019$), CK-MB ($p=0.024$), and LDH ($p<0.001$).

171

Construction of a classification model to predict the important factors for clinical outcome

After the screening of significant clinical characteristics ($p<0.05$) that were associated with COVID-19 patients from Table3, the RF classification procedures were employed on these screened factors for the identification of important clinical characteristics to predict prognosis of COVID-19 patients. RF has become a very popular tool for analyzing high-dimensional data(Statnikov et al. 2008). We used a bagging algorithm to collect 500 random samples from the clinical performance and laboratory data of all COVID-19 patients, each of them was calculated by a decision tree, all of the results vote for the final decision in RF signature(Albert et al. 2008). Building integrations from basic learners, such as trees, can greatly improve predictive performance. We divided the data into a training set and a test set at a ratio of 1: 4 （23：96）. These training sets and test sets were independent of each other. Good confirmation of the performance and reliability of each model. Through the use of CART for multiple calculations and the accuracy of step-by-step testing, variables that significantly affect the prognosis of COVID-19 are found. As shown in Figure 1, the larger the Gini coefficient, the more important the information content of the independent variables. LDH and Myo were considered as the two optimal diagnostic clinical characteristics for COVID-19 patients' prognosis (**Figure 1A**). The accuracy of these variables screened by RF was shown in (**Figure 1B**)，the accuracy of Myo ranked the first，followed by CD45 and LDH.

Identification of accuracy of prediction signature

192 The heat map shows the correlation between the variables in the form of a matrix, where
193 each element in the matrix is the Pearson correlation coefficient between the variables, and the
194 range [−1, 1] is used to evaluate the relevant significance between two continuous variables.
195 When the correlation coefficient is greater than 0.6, the correlation is strong, indicating that the
196 factors are not relatively independent but is affected by more complex interactions, the test of
197 Pearson correlation coefficient avoided over-fitting of models caused by excessive correlation.

198 It can be found that the increase of all total T Lymphocyte in COVID-19 owning to CD8,
199 CD45 and CD4 T cell increment, and the increase of CD45 can improve the production of CD8
200 and CD4 in patients' body. When faced with viruses, the growth of NLR was due to the increase
201 in the number of neutrophils and the decrease in the number of lymphocytes. Except for these
202 associations, these clinical characteristics did not have strong correlations, a special process was
203 not necessary (**Figure 2A**). A ROC curve represented the diagnostic capability of RF
204 classification calculations. The area under the ROC curve (AUC) is the accuracy of the model. It
205 can be seen that the accuracy of the training group is 100% and that of the test group is also
206 93.3%. The accuracy of both is very high (**Figure 2B**). Secondly, out-of-bag (OOB) samples
207 representing the generalization ability of RF to calculate the proportion of
208 misclassification(Genuer 2010; Ishwaran et al. 2010). A voting process progressed when each
209 independent decision tree in the forest calculated the corresponding classification result, and the
210 OOB error rate gradually decreases and stabilizes as the forest size increases (**Figure 2C**).

211

Relationship of characteristics and survival of COVID-19 patients

To identify the difference between LDH and Myo levels of alive patients and died patients, we compared the mortality rates of patients with different levels of LDH, and Myo(**Figure 3A**). The mortality rate increased significantly($P<0.05$) when Myo higher than 80ng/ml or LDH higher than 500U/L，there is a substantial increase in the risk of in-hospital mortality for the increase of Myo（OR=7.54　95%CI，3.42 to 16.63）and LDH（OR=4.90，　95%CI , 2.13 to 11.25). The changes of LDH and Myo in survival and died groups were compared and analyzed(**Figure 3B**) , the median and IRQ of these two variables of the died group were higher than that of alive patients($p<0.001$). The partial dependence plot showed the impact of various clinical symptoms and laboratory results on survival when controlling for marginal effects in the process of RF classification. LDH and Myo were analyzed by partial dependence plot to study their impact on survival rate, as the figure shows (**Figure 3C**): There is a clear negative correlation between the survival and LDH or Myo, their increase was a precursor to the poor prognosis of COVID-19 patients. Their respective ROCs of predicting COVID-19 patients' prognosis was processed with spss26.0：Myo:0.857，LDH:0.807 (**Figure 3D**). They all have high prediction accuracy of COVID-19 patient prognosis, while their accuracy was lower than that of the RF classification model.

Discussion：

The spread of COVID-19 in Wuhan was highly contagious and had a high critical illness rate. In this case statistics, about 50% of severe cases and 13.4% mortality rate. Most of these patients have cough (75.4%), fatigue (58.7 %), dyspnea (55.6%) and myalgia (n=41 32.5%), fever（92.0%）was the most common symptoms. 65% of patients had at least comorbidities, high pressure up to 34.9%，moreover nasal was the most common oxygen therapy approach. When it comes to immunity systems, more than 80% patients' lymphopenia decreased, CD4 + and CD8 + T lymphocytes account for 91.3%, and about half of the patients had a decrease in Th / Ts ratio, in the same time, inflammatory factors such as C-reactive protein (CRP) increased in 85.6% of patients. After comprehensive treatment such as antiviral treatment, dialectical treatment of traditional Chinese medicine, and symptomatic support, most of the disease gradually improved and the prognosis is better, but there were still patients who died due to abnormal physiological changes, so looking for optimum indicators that affect the prognosis was meaningful.

In table 3, we observed that some factors significantly associated with a mortality rate of COVID-19 patients, older men were susceptibility factor, dyspnea, neutrophil, lymphocyte counts, NLR, LMR, total all T lymphocyte counts, CD4, CD8, CD45, T-cell counts, Th/Ts, Myo, CK, PCT, LDH, CK-MB, D-Dimer, PCO_2, PO_2, SO_2, admission oxygenation all significantly changed between survival and died groups($P<0.05$). CD45 was closely related to CD8 and CD4 in the Pearce correlation in figure 5, research showed it played an important role in the activation of immune cells (Hermiston et al. 2003; Rheinländer et al. 2018). As we studied, the level of

CD45 was significantly higher (p=0.011) in non-survivors (635.82±43.43x10^6/ml) than survivors (346.70±57.66x10^6/ml), showing as the increase of CD45, patient's own immunity was strengthened and prognosis of patients becomes better.

They were chosen to build an RF classification model to analyze the optimal predictor of COVID-19 patients. Finally, LDH higher than 500U/L, and Myo higher than 80ng/ml were found that associated with their increased odds of dying.

Myoglobin (Myo) (Premru et al. 2013)is a myocardial marker that has important significance in the clinical detection of patients with severe pneumonia, patients with severe pneumonia are often accompanied by different degrees of myocardial injury, so they are more prone to heart failure and other complications. Clinical reports indicate that when myocardial cells are damaged, it diffuses into the blood faster than CK，cTn1(Ohman et al. 1990).In this study, we found that 75% of dead patients whose Myo>80ng/mL were accompanied with high pressure, which showed a possibility of high pressure accelerated COVID-19 patients' heart damage. After evaluating the different level of Myo in death and survival patients, we found that a high level of Myo which beyond 80ng/mL leads to high mortality rate (61.5%) of COVID-19, and the risk of patients' mortality rate elevated significantly (p=0.013). Partial correlation analyzed that as the increase of Myo, the survival was less，in summary, the damage of myocardial cells made the prognosis of pneumonia worse, Myo was a sign for the patients' myocardial cells condition.

Study showed when tissue damage occurs, LDH will be released outside the cell, causing

271 increased blood circulation LDH (Reis et al. 1988). When lung tissue damages, LDH is
272 positive(Pan et al. 1991). Most of the COVID-19 patients have severely reduced lung ventilation,
273 leading to hypoxia and carbon dioxide retention(Matusiewicz et al. 1993). In this study, 96
274 (76.2%) patients' LDH values were higher than the normal reference range, and the average level
275 of LDH of non-survivors was higher than that of survivors（$p<0.001^{***}$）. The level beyond
276 500U/L of LDH lead to high mortality rate risk，moreover，partial correlation showed a
277 negative correlation between the survival and LDH，in COVID-19 patients, microcirculation
278 disorders caused by infection and insufficient tissue perfusion lead to lung tissue damage and
279 accumulate LDH, therefore the increase of LDH was a risk factor for death.

280 **Conclusion**

281 Outbreak of COVID-19 occurred in Wuhan, has become a global public health threat. The
282 clinical characteristics of confirmed COVID-19 cases suffered from many abnormal laboratory
283 findings, according to a machine learning approach, LDH higher than 500U/L, and Myo higher
284 than 80ng/ml were considered as optimal risk predictors for patients outcome.

285

286 Limitations

287 This study has several limitations. First, due to the inclusion and exclusion of a large number
288 of patients, it is inevitable to omit some relatively important factors for the disease, such as
289 smoking, history of allergies, etc. Second, we only studied some patients who were relatively

severe during the epidemic which may lead to statistical bias because of limited medical resources. Third, a small part of data was lost in the information on the COVID-19 patients list. In the process of RF classification modeling, the count variable is filled in with the median, and the categorical variable is filled in with the mode, which perhaps leads to tiny bias.

Abbreviations

RF: Random forest ; RUC: receiver operating characteristic ; AUC: Area under ROC curve; IQR: Interquartile range ; ARDS: acute respiratory distress syndrome ; Lym: Lymphocyte; Myo: Myoglobin ; NMV：Noninvasive mechanical ventilation；IMV：Invasive mechanical ventilation；CRP：C-reactive protein；PCT：procalcitonin ; FIB: fibrinogen; BNP: B-type natriuretic peptide; CK-MB: ceatine kinase-MB ; AST: aminotransferase; ALT: alanine aminotransferase ;TG: triglyceride; LDH: Lactate dehydrogenase; LMR: Lymphocyte to monocyte ratio ;NLR: Neutrophil lymphocyte ratio; Mon: Monocyte ;OOB: out-of-bag; Neu: Neutrophil.

Authors' contributions

J. Wang,, Q. Q.Hua, C.A. Cao conceived and devised study , J. Wang, Q. Q. Hua, H.P. Yu. analyzed and interpreted the data. J. Wang, S. L. Jing and H.P. Yu. draft the manuscript. C.A. Cao, X. Peng, Zhifen. Liu revised the important intellectual content of the manuscript. S. L. Jing,

X. Peng, H.P. Yu gave Administrative, technical, or material support. C.A. Cao,Q. Q. Hua supervised the manuscript. Dr. Cao had full access to all of the data in the study and took responsibility for the integrity of the data and the accuracy of the data analysis. All authors read and approved the final manuscript.

Ethics approval and consent to participate

Ethics approval was obtained from the Ethics Committee of Wuhan Fourth Hospital (KY 2020-032-01), and informed consent of the study participants was waived by Ethics Committee of hospital for appeared highly transmissible disease.

Funding

The authors received no funding for this work.

Consent for publication

Not applicable.

Availability of data and material

All data generated or analyzed during this study are included in this article and supplementary

materials.

Conflict of interest

The authors have conflict of interest to disclose.

Acknowledgments

We thank the patients and their family members for participating in our study.

References

COVID-19 Coronavirus – Update https://virusncov.comaccessed April 14, 2020).

Jan 11,2020. WHO Clinical management of severe acute respiratory infection when Novel coronavirus (nCoV) infection is suspected: interim guidance.

Albert J, Aliu E, Anderhub H, Antoranz P, Armada A, Asensio M, Baixeras C, Barrio J, Bartko H, Bastieri DJNI, Methods in Physics Research Section A: Accelerators S, Detectors, and Equipment A. 2008. Implementation of the random forest method for the imaging atmospheric Cherenkov telescope MAGIC. 588:424-432.

Biau G, and Scornet E. 2016. A random forest guided tour. *TEST* 25:197-227. 10.1007/s11749-016-0481-7

Biau GJJoMLR. 2012. Analysis of a random forests model. 13:1063-1095.

Chen N, Zhou M, Dong X, Qu J, Gong F, Han Y, Qiu Y, Wang J, Liu Y, Wei Y, Xia J, Yu T, Zhang X, and Zhang L. 2020. Epidemiological and clinical characteristics of 99 cases of 2019 novel coronavirus pneumonia in Wuhan, China: a descriptive study. *Lancet* 395:507-513. 10.1016/S0140-6736(20)30211-7

COVID C, and Team RJMMMWR. 2020. Severe outcomes among patients with coronavirus disease 2019 (COVID-19)—United States, February 12–March 16, 2020. 69:343-346.

348 Dhungana HN. 2020. Comments on "Preliminary estimation of the basic reproduction number of novel Coronavirus
349 (2019-nCoV) in China, from 2019 to 2020: A data-driven Analysis in the early phase of the outbreak". *Int J*
350 *Infect Dis*. 10.1016/j.ijid.2020.02.024

351 Genuer RJapa. 2010. Risk bounds for purely uniformly random forests.

352 Heitner SB, Hollenberg SM, and Colilla SA. 2010. Heat maps, random forests, and nearest neighbors: a peek into
353 the new molecular diagnostic world. *Crit Care Med* 38:296-298. 10.1097/CCM.0b013e3181c545ed

354 Hermiston ML, Xu Z, and Weiss AJAroi. 2003. CD45: a critical regulator of signaling thresholds in immune cells.
355 21:107-137.

356 Huang C, Wang Y, Li X, Ren L, Zhao J, Hu Y, Zhang L, Fan G, Xu J, Gu X, Cheng Z, Yu T, Xia J, Wei Y, Wu W,
357 Xie X, Yin W, Li H, Liu M, Xiao Y, Gao H, Guo L, Xie J, Wang G, Jiang R, Gao Z, Jin Q, Wang J, and
358 Cao B. 2020. Clinical features of patients infected with 2019 novel coronavirus in Wuhan, China. *Lancet*
359 395:497-506. 10.1016/S0140-6736(20)30183-5

360 Ishwaran H, Kogalur UBJS, and letters p. 2010. Consistency of random survival forests. 80:1056-1064.

361 Matusiewicz S, Williamson I, Sime P, Brown P, Wenham P, Crompton G, and Greening AJERJ. 1993. Plasma
362 lactate dehydrogenase: a marker of disease activity in cryptogenic fibrosing alveolitis and extrinsic allergic
363 alveolitis? 6:1282-1286.

364 Ohman EM, Casey C, Bengtson JR, Pryor D, Tormey W, and Horgan JH. 1990. Early detection of acute myocardial
365 infarction: additional diagnostic information from serum concentrations of myoglobin in patients without
366 ST elevation. *Br Heart J* 63:335-338. 10.1136/hrt.63.6.335

367 Pan L, Beverley P, Isaacson PJC, and Immunology E. 1991. Lactate dehydrogenase (LDH) isoenzymes and
368 proliferative activity of lymphoid cells—an immunocytochemical study. 86:240-245.

369 Phelan AL, Katz R, and Gostin LO. 2020. The Novel Coronavirus Originating in Wuhan, China: Challenges for
370 Global Health Governance. *Jama*. 10.1001/jama.2020.1097

371 Premru V, Kovac J, and Ponikvar R. 2013. Use of myoglobin as a marker and predictor in myoglobinuric acute
372 kidney injury. *Ther Apher Dial* 17:391-395. 10.1111/1744-9987.12084

373 Reis GJ, Kaufman HW, Horowitz GL, and Pasternak RCJTAjoc. 1988. Usefulness of lactate dehydrogenase and
374 lactate dehydrogenase isoenzymes for diagnosis of acute myocardial infarction. 61:754-758.

375 Rheinländer A, Schraven B, and Bommhardt UJIl. 2018. CD45 in human physiology and clinical medicine.
376 196:22-32.

377 Statnikov A, Wang L, and Aliferis CF. 2008. A comprehensive comparison of random forests and support vector
378 machines for microarray-based cancer classification. *BMC Bioinformatics* 9:319. 10.1186/1471-2105-9-319

379 Wang D, Hu B, Hu C, Zhu F, Liu X, Zhang J, Wang B, Xiang H, Cheng Z, Xiong Y, Zhao Y, Li Y, Wang X, and
380 Peng Z. 2020. Clinical Characteristics of 138 Hospitalized Patients With 2019 Novel Coronavirus-Infected
381 Pneumonia in Wuhan, China. *JAMA*. 10.1001/jama.2020.1585

382

383

384

385 **Figure legend**

386

387 Table 1. Demographic Characteristics of Patients With COVID-19.

388

389 Table 2. Initial Laboratory Indices of Patients With COVID-19.

390 Table3. Clinical characteristics between the alive and died groups.

391 Figure 1. Identification of optimal diagnostic clinical characteristics for the prognosis of
392 COVID-19 patients. (A) Ranking of clinical characteristics according to Gini. (B) Ranking of
393 clinical characteristics according to standardized drop in prediction accuracy.

394 Figure 2. The accuracy of RF classification models.(A)Heat map visualization shows Pearce
395 correlation coefficient of clinical characteristics

396 (B) ROC curve shows the accuracy of training data and test data in RF classification models. (C)
397 Tendency chart of the relationship between OOB error rate and the number of decision trees.

Figure 3. The different levels of Myo and LDH in death and survival groups.

(A) The table shows the mortality rate increased significantly as the level of Myo and LDH elevated.（B）The scatter plot shows the different levels of Myo /LDH in death and survival groups. (C）The tendency chart shows the partial dependence correlation of Myo /LDH and survival. (D) ROC curve shows Myo and LDH accuracy of predicting the COVID-19 patients' outcome.

Table 1(on next page)

Demographic Characteristics of Patients With COVID-19

Demographic Characteristics of Patients With COVID-19

1 **Table 1. Demographic Characteristics of Patients With COVID-19**

Variable	Number of patients (%)
No. of patients	126
Age, median (IQR), y	60(53 -69.5)
≥65	48(38.1)
<65	78(61.9)
Highest patient temperature, median (IQR),℃	38.6(38- 39)
≥39（high fever）	39(34.8)
<39	73 (65.2)
Gender	
Male	65(51.6)
Female	61 (48.4)
Contact history of epidiemic area	7(5.6)
Initial common symptoms	
Fever	112 (88.9)
Cough	95(75.4)
Productive cough	21 (16.7)
Hemoptysis	6(4.8)
Dyspnea	70 (55.6)
Fatigue	74 (58.7)
Myalgia	41(32.5)
Diarrhea	14 (11.1)
Comorbidities	
Hypertension	44(34.9)
Diabetes	21(16.7)
Cardriovascular and Macrovascular disease	15(11.9)
Liver and gall disease	5(4.0)
Nervous system disease	6(4.8)
Chronic lung disease	13(10.3)
Chronic kidney disease	3(2.4)
Endocrine system disease	2 (1.6)

Immunological disease	1 (0.8)
Hyperlipidemia	3(2.4)
Gastric disease	7(5.6)
Tumor	6(4.8)
Highest level of oxygen therapy	
Nasal cannula	83(65.9)
NMV	35(27.8)
IMV	5(4.0)
IMV with ECMO	0
Severity of clinical condnition	
Moderate	61(50.0)
Severe	38(31.1)
Critical	23(18.9)
Clinical outcomes	
Cure Death	103(86.6)
Death	16 (13.4)

Abbreviations: IQR, interquartile range; NMV, noninvasive mechanical ventilation (including high flow supply and face mask); IMV, invasive mechanical ventilation; ECMO, extracorporeal membrane oxygenation.

Table 2(on next page)

Initial Laboratory Indices of Patients With COVID-19.

Initial Laboratory Indices of Patients With COVID-19.

1 **Table2：Initial Laboratory Indices of Patients With COVID-19**

2

Laboratory Indices	Reference values	Patient amount	median (IQR)	Patient value of deviation
Hematology				
White blood cells, ×109/mL	3.5-9.5	126	6.14(3.96-8.29)	26 (20.6) [a]
Neutrophils, ×109/mL	1.8-6.3	126	4.51(2.77-7.34)	41(32.5) [a]
Lymphocytes, ×109/mL	1.1-3.2	126	0.73(0.53-1.01)	102(81.0)[b]
Monocytes, ×109/m L	0.1-0.6	126	0.29(0.20-0.43)	6(4.8)[a]
NLR	NA	126	5.99(3.07-12.59)	
LMR	NA	126	2.39(1.65-3.68)	
CD4+ Tlym, ×106/mL	450-1440	116	142.16(78.50-271.84)	1 08(93.1)[b]
CD8+ Tlym, ×106/mL	320-1250	116	109.84(61.35-154.52)	1 08(93.1)[b]
Th/Ts	1.5-2.9	116	1.52(0.96-2.08)	55 (51.9)[b]
CD45, ×106/mL	NA	116	481.92(338.36-724.95)	
Biochemical analysis				
AST, U/L	15-40	1 26	27.5(19-44)	48(38.1) [a]
ALT, U/L	9-50	126	25(15-43.5)	22(17.5)[b]
TG, mmol/L	0.45-1.69	126	1.49(1.16-1.89)	41(32.5) [a]
Creatine, μM	57-111	126	66(54-81.25)	7(5.6) [a]
TnI, μg/L	0-0.6	91	0.03(0.03-0.03)	1(1.1) [a]
Myo, ng/mL	0-80	111	27.2(18.1-38.05）	13(11.7)[a]
CK, U/L	0-171	119	63.2(35.25-138.05)	18(15.1)

CK-MB, ng/m L	0-2.37	92	1.1(1-2.33)	23(25.0) [a]
BNP, ng/mL	0-100	88	196.5(42.25-754.25)	53(60.2) [a]
CEA , μg/L	0-5	57	2.08(1.51-5.53)	15(26.3) [a]
LDH, U/L	120-150	126	306.50(241-389)	123(97.6) [a]
Infection indices				
CRP, mg/L	0-5	126	40.31(21.27-86.56)	166 (85.6)[a]
PCT, ng/mL	0-0.5	121	0.04(0.04-0.08)	5(4.1) [a]
Coagulation function				
PT, s	9-13	126	13.6(11.3-41.2)	61(48.4) [a]
PTT, s	20-40	126	35.5(22.4-69.9)	22(17.4)[a]
TT, s	14-21	126	16.3(12.8-72.3)	3(2.4)[a]
INR	0.8-1.25	126	1.10(0.86-4.33)	7(5.6)[a]
FIB, g/L	2-4	126	4.79(1.01-37.9)	83(65.9)[a]
D-Dimer, mg/L	0-0.2	126	1.96(0.03-60.14)	96(76.2)[a]

4 Abbreviations: IQR, interquartile range; NLR, neutrophil lymphocyte ratio; LMR, lymphocyte monocyte ratio.

5 [a]Above reference； [b]Below reference

Table 3(on next page)

Clinical characteristics between the alive and died groups.

Clinical characteristics between the alive and died groups.

1 **Table3. Clinical characteristics between the alive and died groups**

Variables	No. of patients	Alive（103）	Died(n=16)	statistics	p-value
Demographics					
Male	61	49(52.8)	12(8.2)	4.170d	0.041
Female	58	54(50.2)	4(7.8)		
Age	119	58.65±1.21	71.81±1.85	-5.948a	0.000
Highest temperature	106	38.54±0.06	38.73±0.16	-1.137a	0.258
Dyspnea					
Yes	68	54(58.9)	14(9.1)	5.598d	0.018
No	51	49(44.1)	2(6.9)		
Fatigue					
Yes	72	61(62.3)	11(9.7)	0.526c	0.468
No	47	42(40.7)	5(6.3)		
Hematology					
WBC, $\times 10^9$/mL	119	6.49±0.37	8.22±1.22	-1.366a	0.190
Neu, $\times 10^9$/mL	119	5.21±0.34	7.24±1.18	-2.060a	0.042
Lym, $\times 10^9$/mL	119	0.87±0.06	0.58±0.07	2.006a	0.047
Mon, $\times 10^9$/mL	118	0.32±0.02	0.35±0.04	-0.686a	0.494
NLR	119	8.46±0.83	16.16±3.15	-2.785b	0.005
LMR	119	3.29±0.22	1.64±0.27	2.880a	0.005
Total Tlym, $\times 10^6$/mL	109	369.89±27.62	168.71±27.53	-3.677b	0.000
CD4+ Tlym,$\times 10^6$/mL	109	202.67±15.38	115.62±22.70	-2.741b	0.006
CD8+ Tlym,$\times 10^6$/mL	109	150.17±12.37	51.35±7.92	3.164a	0.000
Th/Ts	109	1.57±0.10	2.45±0.32	-3.222a	0.002
CD45,$\times 10^6$/mL	109	635.82±43.43	346.70±57.66	2.595a	0.011
Biochemical analysis	119	32.3±1.8	41.9±5.7	-1.016a	0.312
AST, U/L					
ALT, U/L	119	33.6±3.0	42.1±9.1	-1.266b	0.205
TG, mmol/L	119	1.64±0.07	1.57±0.13	0.379a	0.705

Creatine, μM	119	69.76±2.91	82.25±0.88	-1.611a	0.110
Myo, ng/mL	111	31.80±3.19	109.4±23.93	-10.77b	0.000
CK, U/L	112	100.33±14.21	152.73±30.36	-2.354b	0.019
CK-MB, ng/mL	85	2.37±0.44	2.89±0.58	-2.250b	0.024
Infection indices					
LDH, U/L	119	312.95±12.54	481.94±43.23	-3.981b	0.000
CRP, mg/L	119	49.49±3.91	67.37±10.38	-1.753a	0.080
Coagulation function					
PCT, ng/mL	114	0.07±0.10	0.20±0.24	-2.610b	0.009
APTT, s	119	35.06±0.62	36.80±0.06	-1.042a	0.300
TT, s	118	15.99±0.27	15.72±0.43	-0.830b	0.407
PT, s	119	13.58±0.31	14.19±0.66	-1.068b	0.286
INR	119	1.09±0.03	1.13±0.05	-1.169b	0.242
FIB, g/L	119	4.82±0.85	4.64±0.40	0.196a	0.845
D-Dimer, mg/L	115	1.25±0.29	3.19±1.27	-3.003b	0.003
Blood gas analysis					
PH	119	7.43±0.01	7.41±0.04	-1.667b	0.095
PCO2, mmHg	118	38.75±0.47	33.80±1.91	2.514a	0.023
PO2, mmHg	119	81.98±3.07	55.06±3.49	5.790a	0.000
SO2, %	119	93.88±0.47	83.63±4.83	-3.582b	0.029
oxygenation, mmHg	119	282.8±13.9	124.3±10.6	9.072a	0.000

Abbreviations: NLR, neutrophil lymphocyte ratio; LMR, lymphocyte monocyte ratio;

[a]t-test，[b]Mann-Whitney U test，[c]χ2 test，[d]Continuity Correction

Figure 1

Identification of optimal diagnostic clinical characteristics for prognosis of COVID-19 patients.

(A) Ranking of clinical characteristics according to gini. (B) Ranking of clinical characteristics according to standardized drop in prediction accuracy.

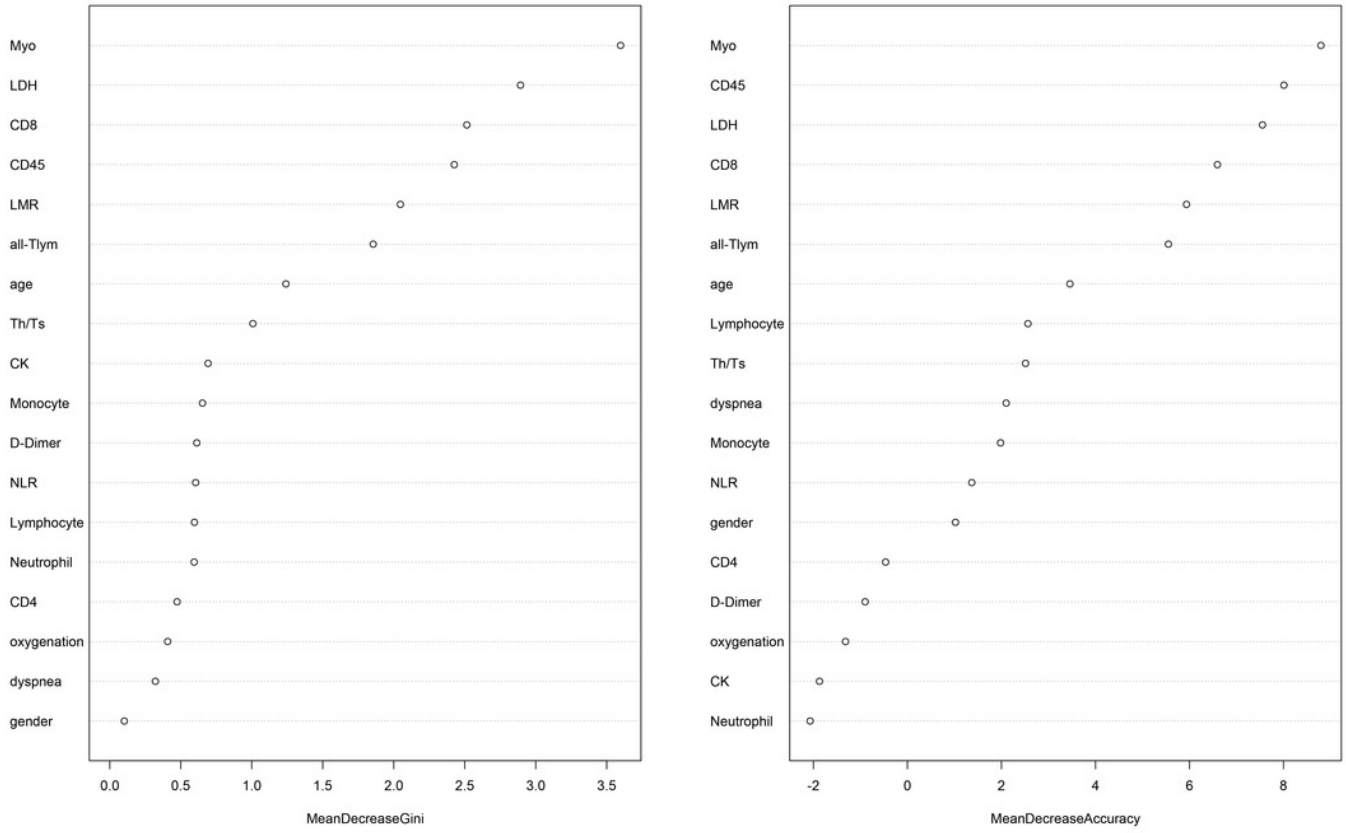


Figure 2

The accuracy of RF classification model.

(A)Heat map visualization shows Pearce correlation coefficient of clinical characteristics. (B)ROC curve shows the accuracy of training data and test data in RF classification models.(C) Tendency chart of relationship between OOB error rate and number of decision trees.

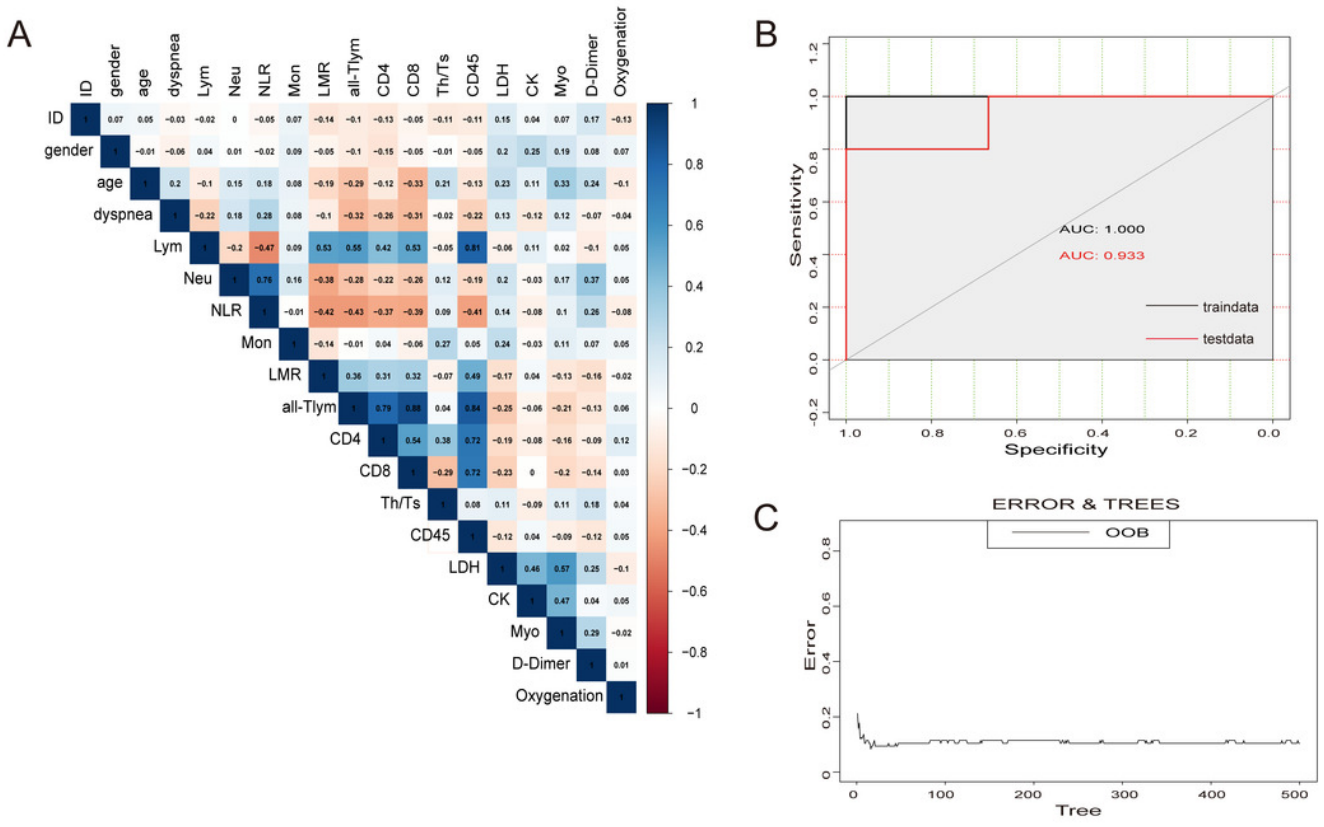
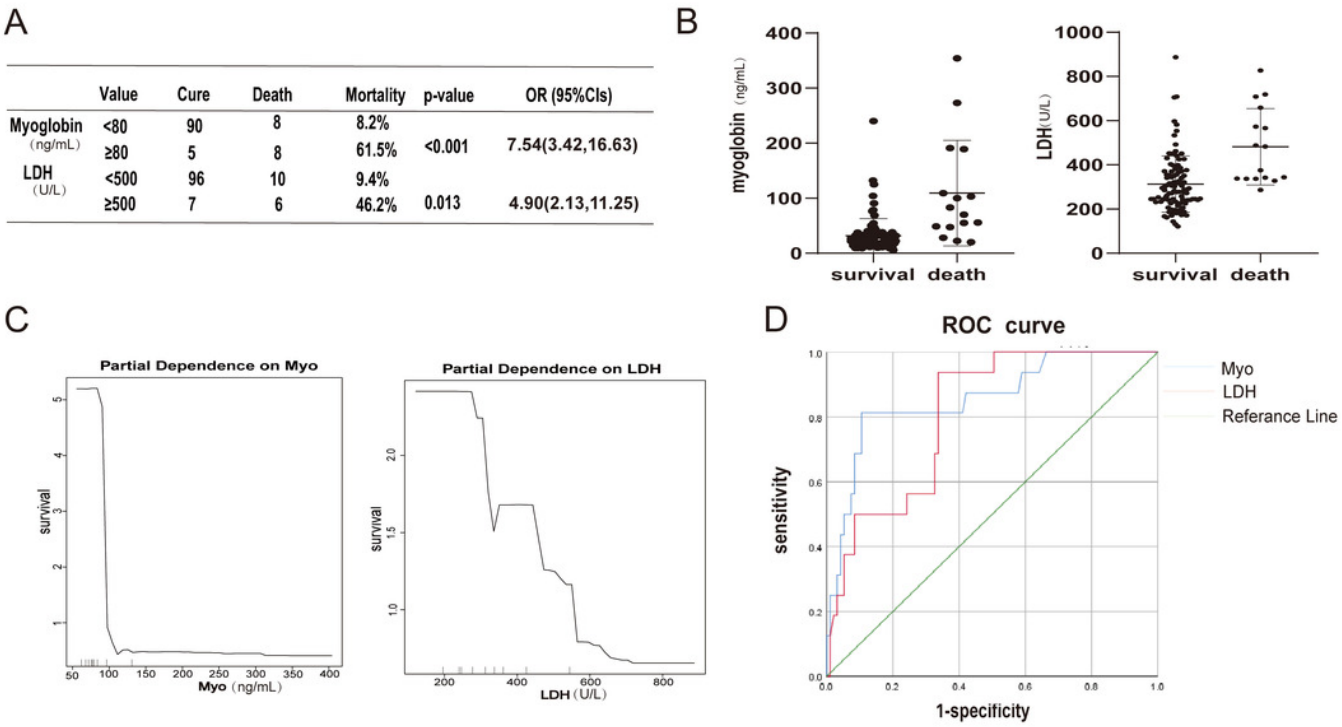


Figure 3

The different level of Myo and LDH in death and survival groups.

(A)The table shows the mortality rate increased significantly as the level of Myo and LDH elevated.(B)The scatter plot shows the different levels of Myo /LDH in death and survival groups. (C)The tendency chart shows the partial dependence correlation of Myo /LDH and survival. (D) ROC curve shows Myo and LDH accuracy of predicting the COVID-19 patients' outcome.



	Value	Cure	Death	Mortality	p-value	OR (95%CIs)
Myoglobin (ng/mL)	<80	90	8	8.2%		
	≥80	5	8	61.5%	<0.001	7.54(3.42,16.63)
LDH (U/L)	<500	96	10	9.4%		
	≥500	7	6	46.2%	0.013	4.90(2.13,11.25)