A descriptive study of random forest algorithm 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], **Yongwen Luo** [Corresp. 5]

[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

[5] Department of Urology, Zhongnan Hospital of Wuhan University, Wuhan, Hubei, China

Corresponding Authors: Cheng'an Cao, Yongwen Luo
Email address: 2017202040101@whu.edu.cn, luoywen@whu.edu.cn

Background The outbreak of coronavirus disease 2019 (COVID-19) that 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 was 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 they have no clinical outcome. Clinical characteristics were analyzed between discharged patients and those who died via a random forest algorithm. An oversampling method was employed for a serious imbalance of patients outcome, and a random forest classification model was used to find the optimal diagnostic predictors for the patients' clinical outcomes between two groups, the areas under the ROC curve (AUC) of the training data (100%) and test data (100%) showed the high accuracy of the 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. The random forest (RF) algorithm identified two optimal diagnostic clinical characteristic predictors of COVID-19 patient outcome after an oversampling method, which were LDH and Myo, and their 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 a random forest algorithm to find that LDH higher than 500U/L, and Myo higher than 80ng/ml were considered as potential risk factors to help doctors find patients with poor COVID-19 prognoses at an early stage and provide scientific data for mortality reduction.

A descriptive study of random forest algorithm 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] ; Yongwen Luo *MD[5]

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

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

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

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

[5]Department of Urology, Zhongnan Hospital of Wuhan University, Wuhan, 430073, China.

*Corresponding author:

Cheng'an Cao, Email: cca24@163.com.

Address: No.473 Hanzheng Street, Qiaokou District, Wuhan, Hubei, China

Yongwen Luo, Email: luoywen@whu.edu.cn;

Abstract:

Background: The outbreak of coronavirus disease 2019 (COVID-19) that 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 was 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 they have no clinical outcome. Clinical characteristics were analyzed between discharged patients and those who died via a random forest algorithm. An oversampling method was employed for a serious imbalance of patients outcome, and a random forest classification model was used to find the optimal diagnostic predictors for the patients' clinical outcomes between two groups, the areas under the ROC curve (AUC) of the training data (100%) and test data (100%) showed the high accuracy of the 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. The random forest (RF) algorithm identified two optimal diagnostic clinical characteristic predictors of COVID-19 patient outcome after an oversampling method, which were LDH and Myo, and their 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 a random forest algorithm to find that LDH higher than 500U/L, and Myo higher than 80ng/ml were considered as potential risk factors to help doctors find patients with poor COVID-19 prognoses at an early stage and provide scientific data for mortality reduction.

Key words：**COVID-19, predictors, random forest algorithm, patient outcome**

Introduction

In December2019, an outbreak of acute respiratory syndrome coronavirus pneumonia occurred in Wuhan, Hubei Province, China (Phelan et al. 2020), and attracted an intense amount of attention worldwide. The WHO named the virus 2019-nCoV by identifying it from a patient's pharyngeal swab sample. The scientific community and infection control agencies are facing enormous challenges in controlling the increasing intensity of epidemics. However, the disease is developing rapidly around the world. By April 14, 2020, COVID-19 had affected 210 countries, with over 1929000 confirmed cases and 119754 deaths, and the epidemic situation in Wuhan come under substantial focus in China (Dhungana 2020). As of April 2020, the mortality rate of COVID-19 was relatively lower than those of Middle East respiratory syndrome and SARS, the death toll in Hubei Province accounts for about 97% of the country according to WHO. Severe COVID-19 patients can develop severe pneumonia, ARDS, and multiple organ failure leading to death, while nonsevere COVID-19 patients can present with general symptoms of respiratory infection (Chen et al. 2020; Huang et al. 2020a). Studies shows that high-dose corticosteroid, the elder, diabetes, cardiovascular disease, chronic respiratory diseases, and cancer are all associated

with increased risk of COVID-19 death rate in the world (Guan et al. 2020; Ji et al. 2020; Wu & McGoogan 2020; Zhou et al. 2020, Huang et al. 2020b; Wang et al. 2020b), so it was important to explore the risk factors to predict patients‘outcome in Hubei province.

Currently, machine learning is widely used in the field of medical diagnosis, an RF is one of machine learning that can analyze complex interactions between clinical characteristics, resulting in good risk prediction and it can yield a large number of classification accuracies obtained by using a set of decision trees(Touw et al. 2013). Although individual conditions vary between different COVID-19 patients, clinical characteristics that could serve as optimal diagnostic predictors for patients' clinical outcome are worth exploring. In this study, we aim to describe the clinical characteristics of patients during hospitalization and apply an RF model to find risk factors of in-hospital mortality for COVID-19 patients to provide scientific data for mortality reduction.

Methods

Study population

This is a retrospective cohort analysis including 126 patients aged 27 to 87 years, from Wuhan Fourth Hospital, who were diagnosed with COVID-19 according to the World Health Organization interim guidance. Among them, 7 patients were excluded for losing the outcomes. These patients were hospitalized for treatment from Feb. 1st to Mar. 15th, 2020. This study

obtained the approval of the Ethics Committee of Wuhan Fourth Hospital（KY 2020-032-01), and informed consent from the study participants was waived by the Ethics Committee of the hospital due to the apparently high transmissibility of the disease.

Among patients, the criteria of discharge were as follows: the highest temperature has recovered to normal for more than three days, the chest CT imaging showed significant inflammation absorption, the respiratory symptoms have substantially improved. Two consecutive throat swabs nucleic acid tests were negative and the time interval was at least one day. Finally, after the evaluation and unanimous decision by the expert team, a patient can be discharged.

Data collection

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

Statistical analysis

Descriptive data were compared with quartiles and medians, and the $\chi 2$ test, or Fisher's exact test was employed to analyze the categorical data. Kolmogorov-Smirnov test was used to analyze the normality of discharged patients (n=103), and Shapiro-Wilk test was employed to analyze the normality of died patients (n=16). After that, normally distributed laboratory results

were employed to analyze by independent samples paired t tests, while the nonparametric Mann-Whitney-Wilcoxon test was used for data that did not satisfy a normal distribution. All the data above were processed by IBM SPSS Version 26.0.

A distribution-sensitive oversampling procedure for imbalanced data was employed to analyze the clinical characteristics between COVID-19 discharged patients and those who died , and the final ratio of these two groups were 1:1 (103：103). Then a random forest classification model and a five-fold cross-validation were used to the process to persuasively find out the optimal diagnostic predictors for the clinical outcomes. The mean value of the AUC in the 5 different testing sets of 1 mtry was the highest with 0.87, so the 1 mtry of random forest classification model was chosen, and a bagging algorithm was used to randomly collect the clinical characteristic samples of all patients with COVID-19 for a total of 500 times, with all of the results voted upon the final decision in the RF prediction. Moreover, the data was divided into a training set and a test set with a ratio of about 4: 1 (166：40). Graphpad 8.0 was performed to analyze the level of two variables in discharged and died groups.

All data were processed with Rstudio (R 3.6.3), the oversampling algorithm was processed with caret package (http://CRAN.R- project.org/ package=caret)，the RF model was constructed by randomForest package (https://cran. r-project. org/web/packages/randomForest/),and the rpart package (https://cran.r-project. org/web/packages/rpart/index.html). The CART method was used to calculate the decision tree, and the final result was obtained by voting for the results of the combined predictions, using the Gini index as the split criterion. The greater the change in the

Gini of the nodes before and after the split, the more important the variables are.

Moreover, to give a graphical depiction of the marginal effect of a variable on the classification during the calculation process, a partial correlation was employed to analyze the relationships between clinical data and patient prognosis (Greenwell 2017). The function being plotted was defined as:

$$\tilde{f}_{(x)} = \frac{1}{n}\sum_{n=1}^{n} f(x, x_{ic})$$

where x is the variable corresponding to the chosen clinical characteristic, and x_{iC} represents the other variables in the clinical information. The summand was the predicted logits (log of a fraction of votes) for classification:

$$f(x) = \log_{Pk}(x) - \frac{1}{k}\sum_{j=1}^{k} \log_{Pj}(x)$$

where K is the number of classes, and P_j is the proportion of votes for class j. Pearson correlation was used to calculate the correlations among the important variables that could predict the prognosis in patients with COVID-19 to avoid overfitting the model from excessive correlation. Graphpad 8.0 was used to analyze the levels of two variables in survival and non-survival patients.

Results

Clinical demographics of COVID-19 patients on admission

To study the demographic characteristics of patients with COVID-19, 126 patients who were hospitalized at Wuhan Fourth Hospital were analyzed (**Table1**). Seventy-eight patients (61.9%) were younger than 65 years old, and the median age of the patients was 60 years (IQR 53-69.5). The COVID-19 patients were generally accompanied by fever (116 [92.0%] patients), 39 (34.8%) of whom had peak temperatures above 39 °C, and the median temperature was 38.6 °C (IQR 37.4°C-40°C) (Table 1). COVID-19 infection was basically gender-neutral, that is, the proportions of male and female patients was almost identical. Seven of the 126 patients (5.6%) had visited the South China Seafood Market in Wuhan. Most of these patients on admission had cough (n=95，75.4%), followed by fatigue (n=74, 58.7%), dyspnea (n=70, 55.6%), myalgia (n=41, 32.5%), and diarrhea(n= 14, 11.1%). In addition, many patients also suffered 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%) and endocrine system diseases (n=2, 1.6%). In the process of treatment, 83 patients (65.9%) used nasal cannula, 35 (27.8%) used NMV, and 5 (4.0%) used IMV. In terms of clinical severity, 61 of these patients (50.0%) were in a moderate state, 38 (31.1%) were in a severe state, and 23 (18.9%) were in a critically ill state. Judging from the current treatment results, 103 (86.6%) patients were cured, while another 16 (13.4%) patients died.

Laboratory findings of COVID-19 patients on admission

To analyze the physiological and biochemical changes of patients with COVID-19, the

laboratory results of the 126 patients with Corona virus Disease 2019 (COVID-19) were collected in **Table 2**. More than 80% of patients had lymphopenia reduction, especially CD4 + and CD8 + T lymphocytes (91.3%), and approximately half of the patients had a decrease in the Th/Ts ratio. C-reactive protein (CRP) was elevated in 85.6% of patients, and procalcitonin (PCT) showed a slight increase. The coagulation function of some patients was affected, with prothrombin time (PT) prolonged in approximately 1/2 of the patients and fibrinogen (FIB) increased in 2/3 of the patients. D-dimer was increased in 76.2% of patients. Cardiac dysfunction may have been present in some patients because 60% of patients had elevated B-type natriuretic peptide (BNP) and 25% of patients had increased creatine kinase MB (CK-MB). A small proportion of patients had elevated aspartate aminotransferase (AST) and alanine aminotransferase (ALT) levels, and about 1/3 have elevated triglyceride (TG). In addition, patients with elevated lactate dehydrogenase (LDH) accounted for 76.2% of the total.

Comparison of clinical characteristics between discharged and deceased patients

The different clinical variables were produced after comparison between discharged patients and deceased patients with univariate analysis. **Table 3** shows that the patients in the deceased group were older than those in the discharged group ($p<0.001$), and the majority were male (75%). The proportion of dyspnea was remarkably increased in the deceased group ($p=0.018$), while the rest of the clinical symptoms on admission, such as fatigue, myalgia and diarrhea, showed no obvious difference. Compared with the discharged group, PCO2($p=0.023$),

PO2($p<0.001$), SO2($p=0.029$) and admission oxygenation ($p<0.001$) were significantly reduced in the arterial blood gas analysis of the patients in the deceased group on admission. Laboratory analysis revealed that the deceased group had a higher proportion of neutrophils ($p=0.042$) and a lower proportion of lymphocytes ($p=0.047$) than the discharged group. Additionally, the NLR was notably increased ($p=0.005$) and the LMR was significantly decreased ($p=0.005$) in the deceased group. Moreover, T lymphocytes in the deceased group were remarkably lower than those in the discharged group ($p<0.001$), and CD4+ ($p=0.006$), CD8+($p<0.001$), and the Th/Ts ratio were increased in the deceased group ($p=0.002$). Compared with the discharged group, the inflammation-related indices, CRP ($p=0.080$) and PCT ($p=0.009$), were significantly higher in the deceased group. There was no obvious difference in coagulation function indices between the deceased group and the discharged group, except that D-dimer($p=0.003$) was increased significantly in the deceased group. In addition, there were some elevated biochemical indices in the deceased group which represented the condition of cardiac dysfunction such as Myo($p<0.001$), CK($p=0.019$), CK-MB ($p=0.024$), and LDH ($p<0.001$).

Construction of a classification model to predict important factors for clinical outcomes

After screening significant clinical characteristics ($p<0.05$) that were associated with COVID-19 from Table 3, an oversampling procedure for imbalanced data was employed to analyze the clinical characteristics between COVID-19 discharged patients and those who died，the final ratio between two groups reached 1：1（103：103）， the RF classification procedures were employed on these screened factors for the identification of important clinical

194 characteristics to predict the prognosis of COVID-19 patients. RF has become a very popular
195 tool for analyzing high-dimensional data (Statnikov et al. 2008). Through the use of CART for
196 multiple calculations and the accuracy of step-by-step testing, variables that significantly
197 affected the prognosis of COVID-19 were found. As shown in Figure 1, the larger the Gini
198 coefficient, the more important the information content of the independent variables was. The
199 importance ranking of the variables was Myo，CD8, LDH, age，CK，D-dimer，all T
200 lymphocyte，CD4，CD45，Th/Ts, LMR, lymphocyte, NLR, neutrophil, monocyte,
201 oxygenation, dyspnea, gender. LDH and Myo were considered the two optimal diagnostic risk
202 factors for COVID-19 patient prognosis (**Figure 1A**). The accuracy of these variables screened
203 by RF is shown in (**Figure 1B**)，the accuracy of Myo ranked the first，followed by CD45 and
204 LDH.

205 **Identification of the accuracy of the prediction signature**

206 To evaluate the accuracy of RF classification model, a Pearson correlation coefficient test
207 were processed, and a heat map shows the correlations between variables in the form of a matrix,
208 where each element in the matrix is the Pearson correlation coefficient between the variables,
209 and the range [−1, 1] is used to evaluate the relevant significance between two continuous
210 variables. When the correlation coefficient is greater than 0.6, the correlation is strong, indicating
211 that the factors are not relatively independent but are affected by more complex interactions. The
212 Pearson correlation coefficient test avoids model overfitting of from excessive correlation.

213 The increase in total T lymphocytes in COVID-19 was the result of increases in CD8+,

CD45+ and CD4+ T cells, and increases in CD45 can improve the production of CD8 and CD4 in patients' bodies. Furthermore, when faced with the virus, the NLR increases as a result of the increase in the number of neutrophils and the decrease in the number of lymphocytes. Except for these associations, the clinical characteristics did not have strong correlations, and a special process was not necessary (**Figure 2A**). An ROC curve was used to represent the diagnostic capability of the RF classification calculations. The area under the ROC curve (AUC) represents the accuracy of the model. It can be seen that the accuracy of both the training group is 100% and that of the test group is also 100% (**Figure 2B**) and the threshold of training data ROC was 0.445. Moreover, out-of-bag (OOB) samples represent the generalization ability of RF to calculate the proportion of misclassification (Ishwaran & Kogalur 2010). The OOB error gradually decreased and stabilizesd as the forest size increased, finally it reduced to lower than 0.05 (**Figure 2C**), meanwhile the death and survival errors were gradually reduced to the level same as OOB.

Relationship between clinical characteristics and survival in COVID-19 patients

To identify the difference between LDH and Myo levels of discharged patients and deceased 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 was higher than 80ng/ml or when LDH was higher than 500U/L. In these cases, there is a substantial increase in the risk of in-hospital mortality with 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 discharged

234 alive and deceased groups were compared and analyzed (**Figure 3B**). The median and IRQ of
235 these two variables of the deceased group were higher than that of discharged patients($p<0.001$).
236 The partial dependence plot showed the impact of various clinical symptoms and laboratory
237 results on survival when controlling for marginal effects in the process of RF classification. LDH
238 and Myo were analyzed by partial dependence plot to study their impact on survival rate. As the
239 figure shows (**Figure 3C**): there is a clear negative correlation between the survival and LDH or
240 Myo, their increase was a precursor to the poor prognosis of COVID-19 patients. Their
241 respective ROC of predicting COVID-19 patients' prognosis was processed with IBM SPSS
242 Version 26.0, the AUC of Myo was 0.857 and the cutoff was 46.95ng/mL, the AUC of LDH was
243 0.807, and the cutoff was 327.5U/L (**Figure 3D**). They all have high accuracy of COVID-19
244 patient prognosis prediction, while their accuracy was lower than that of the RF classification
245 model.

246

247 **Discussion:**

248 The spread of COVID-19 in Wuhan is highly contagious and has a high critical illness rate.
249 In this case, approximately 50% of the cases were severe, and the disease had a 13.4% mortality
250 rate. Most of the patients with COVID-19 had cough (75.4%), fatigue (58.7 %), dyspnea
251 (55.6%). Fever (92%) was the most common symptoms. 65% of the patients with COVID-19
252 had comorbidities like hypertension (34.9%). Nasal endotracheal intubation seems to be the most
253 common oxygen therapy in treating COVID-19. More than 80% patients showed lymphopenia,

CD4 + and CD8 + were decreased account for 91.3% patients, 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) was increased in 85.6% patients. After comprehensive treatment, such as antiviral treatment, dialectical treatment with traditional Chinese medicine, and symptomatic support, most of the infection patients gradually improved and the prognosis was better, but there were still patients who died. Therefore, it is important to seek optimum indicators that can affect the prognosis.

In Table 3, we observed that several factors were associated with the mortality of COVID-19 patients, such as older men were susceptibility factor, dyspnea, neutrophil count, lymphocyte count, NLR, LMR, total T lymphocyte count, CD4, CD8, CD45, Th/Ts, Myo, CK, PCT, LDH, CK-MB, D-dimer with p-values below 0.05. As for the immunity system, studies have found that the body cannot produce sufficient protective immunity to eliminate SARS-CoV, however, the immune response to this pathogen may actually exacerbate their disease(Chen et al. 2005). CD45 was closely related to CD8 and CD4 according to Pearson correlation. Research shows that it plays an important role in the activation of immune cells (Hermiston et al. 2003; Rheinländer et al. 2018). As we studied as the increase of CD45, patient's own immunity was strengthened and prognosis of patients becomes better.

Due to a serious imbalance of patients' outcome, an oversampling algorithm was used to eliminate the error caused by minority samples from died group, and some samples were randomly selected from died group for copying to increase the proportion of the non-survivor

group, to reach the same sample size as the discharged group. The RF algorithm, combining with several random decision trees and aggregating their predictions by averaging, has achieved great success in empirical research，and its mechanism has been actively studied(Heitner et al. 2010). All of the data chosen by univarite analysis were selected for building an RF classification model to analyze the optimal predictor of COVID-19 patients. Results showed that LDH higher than 500U/L, and Myo higher than 80ng/ml were associated with the high risk 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, in this study, we found that 75% of the non-survivors whose Myo higher than 80ng/mL were accompanied with hypertension, suggesting that hypertension may accelerate heart damage in patients with COVID-19. After evaluating the different level of myo in non-survivors and survivors, we found that a high level of Myo which beyond 80ng/mL can results in high mortality rate (61.5%) of COVID-19, and the risk of the mortality rate elevated significantly (p=0.013). Partial correlation analyzed that as the increase of Myo, the survival became less. Patients with severe pneumonia are often accompanied by different degrees of myocardial injury, so they are more prone to heart failure and other complications. Study shows that the elevated concentration in venous blood of Myo was found to predict the severity of the COVID-19 (Han et al. 2020) and patients in the deceased group are susceptible to multiple organ failure, especially heart failure and respiratory failure(Du et al. 2020). This study identified the importance of Myo in COVID-19 disease，and doctors need careful observation of patients' heart condition，and control their hypertension

with early intervention to prevent the mortality rate of COVID-19.

In this study, 96 (76.2%) patients' LDH values were higher than the normal reference range, and the average level of LDH of the nonsurvivors was higher than that the survivors ($p<0.001$***). The level beyond 500U/L of LDH lead to high mortality rate risk. Study showed that when lung tissue damage occurs, LDH is released outside the cell, causing an increase in its blood circulation (Yang et al. 2020). Most COVID-19 patients have severely reduced lung ventilation, leading to hypoxia and carbon dioxide retention (Huang et al. 2020a). Moreover, in influenza, LDH is considered to be a marker that reflects the degree of virus damage to the tissue and the severity of the disease, especially the damage to myocardial cells(Mamas et al. 2008; Warren-Gash et al. 2009). In COVID-19 patients, microcirculation disorders caused by the infection and insufficient tissue perfusion lead to lung tissue damage and LDH accumulated, and in severe patients, its abnormal elevation often represents rapid disease progression, and an acute respiratory failure (Wang et al. 2020a), therefore an increase in LDH was a risk factor for death.

Moreover, we are in support of the suggestion that the measurement of cardiac and lung damage, patients' comorbidities should get enough attention for a COVID-19 treatment.

Conclusion

The outbreak of COVID-19 that occurred in Wuhan, has become a global public health threat. According to a random forest algorithm, LDH higher than 500U/L, and Myo higher than

312 80ng/ml were considered as optimal risk predictors to help doctors find patients with poor
313 COVID-19 prognoses at an early stage and provide scientific data for mortality reduction.

314 **Limitations**

315 This study has several limitations. First, due to the inclusion and exclusion of a large
316 number of patients, it is inevitable that some relatively important factors for the disease were
317 omitted, such as smoking, and a history of allergies, etc. Second, we only studied some patients
318 who were relatively severe during the epidemic, which may have led to a statistical bias because
319 of limited medical resources. Third, a small portion of the data were lost from the COVID-19
320 patient list. In the process of RF classification modeling, the count variable uses median, and the
321 categorical variable uses the mode, which perhaps led to a small bias.

322 **Abbreviations**

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

Authors' contributions

J. Wang,, Q. Q.Hua, and C.A. Cao conceived and devised the study , J. Wang, Q. Q. Hua and H.P. Yu. analyzed and interpreted the data. J. Wang, S. L. Jing and H.P. Yu. drafted the manuscript. C.A. Cao, X. Peng and Zhifen. Liu revised the important intellectual content of the manuscript. S. L. Jing, X. Peng and H.P. Yu provided administrative, technical, or material support. C.A. Cao and Q. Q. Hua supervised the manuscript. Y.W. Luo instructed us to revise 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 the Ethics Committee of the hospital due to the apparently high transmissibility of the 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. We appreciate all the efforts of our colleagues in treating the patients.

References

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

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

Chen H, Hou J, Jiang X, Ma S, Meng M, Wang B, Zhang M, Zhang M, Tang X, Zhang F, Wan

T, Li N, Yu Y, Hu H, Yang R, He W, Wang X, and Cao X. 2005. Response of memory CD8+ T cells to severe acute respiratory syndrome (SARS) coronavirus in recovered SARS patients and healthy individuals. *J Immunol* 175:591-598. 10.4049/jimmunol.175.1.591

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

Dhungana HN. 2020. Comments on "Preliminary estimation of the basic reproduction number of novel Coronavirus (2019-nCoV) in China, from 2019 to 2020: A data-driven Analysis in the early phase of the outbreak". *Int J Infect Dis* 94:72-73. 10.1016/j.ijid.2020.02.024

Du RH, Liang LR, Yang CQ, Wang W, Cao TZ, Li M, Guo GY, Du J, Zheng CL, Zhu Q, Hu M, Li XY, Peng P, and Shi HZ. 2020. Predictors of mortality for patients with COVID-19 pneumonia caused by SARS-CoV-2: a prospective cohort study. *Eur Respir J* 55. 10.1183/13993003.00524-2020

Greenwell BM. 2017. pdp: An R package for constructing partial dependence plots. *The R Journal* 9:421-436.

Guan W-j, Ni Z-y, Hu Y, Liang W-h, Ou C-q, He J-x, Liu L, Shan H, Lei C-l, and Hui DS. 2020. Clinical characteristics of 2019 novel coronavirus infection in China. *MedRxiv*.

Han H, Xie L, Liu R, Yang J, Liu F, Wu K, Chen L, Hou W, Feng Y, and Zhu C. 2020. Analysis of heart injury laboratory parameters in 273 COVID-19 patients in one hospital in Wuhan, China. *J Med Virol*. 10.1002/jmv.25809

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

Hermiston ML, Xu Z, and Weiss A. 2003. CD45: a critical regulator of signaling thresholds in immune cells. *Annu Rev Immunol* 21:107-137. 10.1146/annurev.immunol.21.120601.140946

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, 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 Cao B. 2020a. Clinical features of patients

infected with 2019 novel coronavirus in Wuhan, China. *Lancet* 395:497-506. 10.1016/S0140-6736(20)30183-5

Huang R, Zhu L, Xue L, Liu L, Yan X, Wang J, Zhang B, Xu T, Ji F, and Zhao Y. 2020b. Clinical findings of patients with coronavirus disease 2019 in Jiangsu province, China: A retrospective, multi-center study. *PLOS Neglected Tropical Diseases* 14:e0008280.

Ishwaran H, and Kogalur UB. 2010. Consistency of Random Survival Forests. *Stat Probab Lett* 80:1056-1064. 10.1016/j.spl.2010.02.020

Ji H-L, Zhao R, Matalon S, and Matthay MA. 2020. Elevated plasmin (ogen) as a common risk factor for COVID-19 susceptibility. *Physiological reviews* 100:1065-1075.

Mamas MA, Fraser D, and Neyses L. 2008. Cardiovascular manifestations associated with influenza virus infection. *Int J Cardiol* 130:304-309. 10.1016/j.ijcard.2008.04.044

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

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

Rheinländer A, Schraven B, and Bommhardt U. 2018. CD45 in human physiology and clinical medicine. *Immunol Lett* 196:22-32. 10.1016/j.imlet.2018.01.009

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

Touw WG, Bayjanov JR, Overmars L, Backus L, Boekhorst J, Wels M, and van Hijum SA. 2013. Data mining in the Life Sciences with Random Forest: a walk in the park or lost in the jungle? *Brief Bioinform* 14:315-326. 10.1093/bib/bbs034

Wang B, Li R, Lu Z, and Huang Y. 2020a. Does comorbidity increase the risk of patients with COVID-19: evidence from meta-analysis. *Aging (Albany NY)* 12:6049-6057. 10.18632/aging.103000

Wang D, Hu B, Hu C, Zhu F, Liu X, Zhang J, Wang B, Xiang H, Cheng Z, and Xiong Y. 2020b. Clinical characteristics of 138 hospitalized patients with 2019 novel coronavirus–infected pneumonia in Wuhan, China. *Jama* 323:1061-1069.

Warren-Gash C, Smeeth L, and Hayward AC. 2009. Influenza as a trigger for acute myocardial infarction or death from cardiovascular disease: a systematic review. *Lancet Infect Dis* 9:601-610. 10.1016/s1473-3099(09)70233-6

Wu Z, and McGoogan JM. 2020. 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* 323:1239-1242.

Yang P, Wang P, Song Y, Zhang A, Yuan G, and Cui Y. 2020. A retrospective study on the epidemiological characteristics and establishment of early warning system of severe COVID-19 patients. *J Med Virol*. 10.1002/jmv.26022

Zhou F, Yu T, Du R, Fan G, Liu Y, Liu Z, Xiang J, Wang Y, Song B, and Gu X. 2020. Clinical course and risk factors for mortality of adult inpatients with COVID-19 in Wuhan, China: a retrospective cohort study. *The lancet*.

Figure legend

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

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

Table3. Clinical characteristics between the discharged and deceased groups.

Figure 1. Identification of optimal diagnostic clinical characteristics for the 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 models.

452 (A)Heat map visualization shows Pearson correlation coefficient of clinical characteristics

453 (B) ROC curve shows the accuracy of training data and test data in RF classification models. (C)
454 Tendency chart of the relationship between OOB, death, survival error rate and the number of
455 decision trees.

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

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

462

463

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	
Discharged alive	103(86.6)
Died	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

3

Laboratory Indices	Reference values	Number of all patients	Median (IQR)	Number of patient with value
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) [a]

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]
APTT, 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 discharged and deceased groups.

Clinical characteristics between the discharged and deceased groups.

1 **Table3. Clinical characteristics between the discharged and deceased groups**

Variables	No. of patients	Discharged	Deceased (n=16)	statistics	p-value
Demographics					
Male	61	49(52.8)	12(8.2)	4.170 [d]	0.041
Female	58	54(50.2)	4(7.8)		
Age	119	58.65±1.21	71.81±1.85	-5.948 [a]	0.000
Peak temperature	106	38.54±0.06	38.73±0.16	1.645 [b]	0.100
Dyspnea					
Yes	68	54(58.9)	14(9.1)	5.598 [d]	0.018
No	51	49(44.1)	2(6.9)		
Fatigue					
Yes	72	61(62.3)	11(9.7)	0.526 [c]	0.468
No	47	42(40.7)	5(6.3)		
Hematology					
WBC, $\times10^9$/mL	119	6.49±0.37	8.22±1.22	1.079 [b]	0.281
Neu, $\times10^9$/mL	119	5.21±0.34	7.24±1.18	2.060 [a]	0.042
Lym, $\times10^9$/mL	119	0.87±0.06	0.58±0.07	-2.551 [b]	0.11
Mon, $\times10^9$/mL	118	0.32±0.02	0.35±0.04	1.091 [a]	0.275
NLR	119	8.46±0.83	16.16±3.15	-2.785 [b]	0.005
LMR	119	3.29±0.22	1.64±0.27	2.880 [a]	0.005
Total Tlym, $\times10^6$/mL	109	369.89±27.62	168.71±27.53	-3.677 [b]	0.000
CD4+ Tlym,$\times10^6$/mL	109	202.67±15.38	115.62±22.70	-2.741 [b]	0.006
CD8+ Tlym,$\times10^6$/mL	109	150.17±12.37	51.35±7.92	4.468 [b]	0.000
Th/Ts	109	1.57±0.10	2.45±0.32	2.951 [b]	0.003
CD45,$\times10^6$/mL	109	635.82±43.43	346.70±57.66	3.070 [b]	0.002
Biochemical analysis					
AST, U/L	119	32.3±1.8	41.9±5.7	-1.831 [b]	0.067
ALT, U/L	119	33.6±3.0	42.1±9.1	-1.266 [b]	0.205
TG, mmol/L	119	1.64±0.07	1.57±0.13	0.055 [b]	0.957

Creatine, μM	119	69.76±2.91	82.25±0.88	-1.611 [a]	0.110
Myo, ng/mL	111	31.80±3.19	109.4±23.93	-10.77 [b]	0.000
CK, U/L	112	100.33±14.21	152.73±30.36	-2.354 [b]	0.019
CK-MB, ng/mL	85	2.37±0.44	2.89±0.58	-2.250 [b]	0.024
Infection indices					
LDH, U/L	119	312.95±12.54	481.94±43.23	-3.981 [b]	0.000
CRP, mg/L	119	49.49±3.91	67.37±10.38	-1.753 [a]	0.080
Coagulation function					
PCT, ng/mL	114	0.07±0.10	0.20±0.24	-2.610 [b]	0.009
APTT, s	119	35.06±0.62	36.80±0.06	1.176 [b]	0.239
TT, s	118	15.99±0.27	15.72±0.43	-0.830 [b]	0.407
PT, s	119	13.58±0.31	14.19±0.66	-1.068 [b]	0.286
INR	119	1.09±0.03	1.13±0.05	-1.169 [b]	0.242
FIB, g/L	119	4.82±0.85	4.64±0.40	0.651 [b]	0.515
D-Dimer, mg/L	115	1.25±0.29	3.19±1.27	-3.003 [b]	0.003
Blood gas analysis					
PH	119	7.43±0.01	7.41±0.04	0.507 [a]	0.619
PCO2, mmHg	118	38.75±0.47	33.80±1.91	2.514 [a]	0.023
PO2, mmHg	119	81.98±3.07	55.06±3.49	-3.837 [b]	0.000
SO2, %	119	93.88±0.47	83.63±4.83	-3.582 [b]	0.029
oxygenation, mmHg	119	282.8±13.9	124.3±10.6	9.072 [a]	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 the 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.

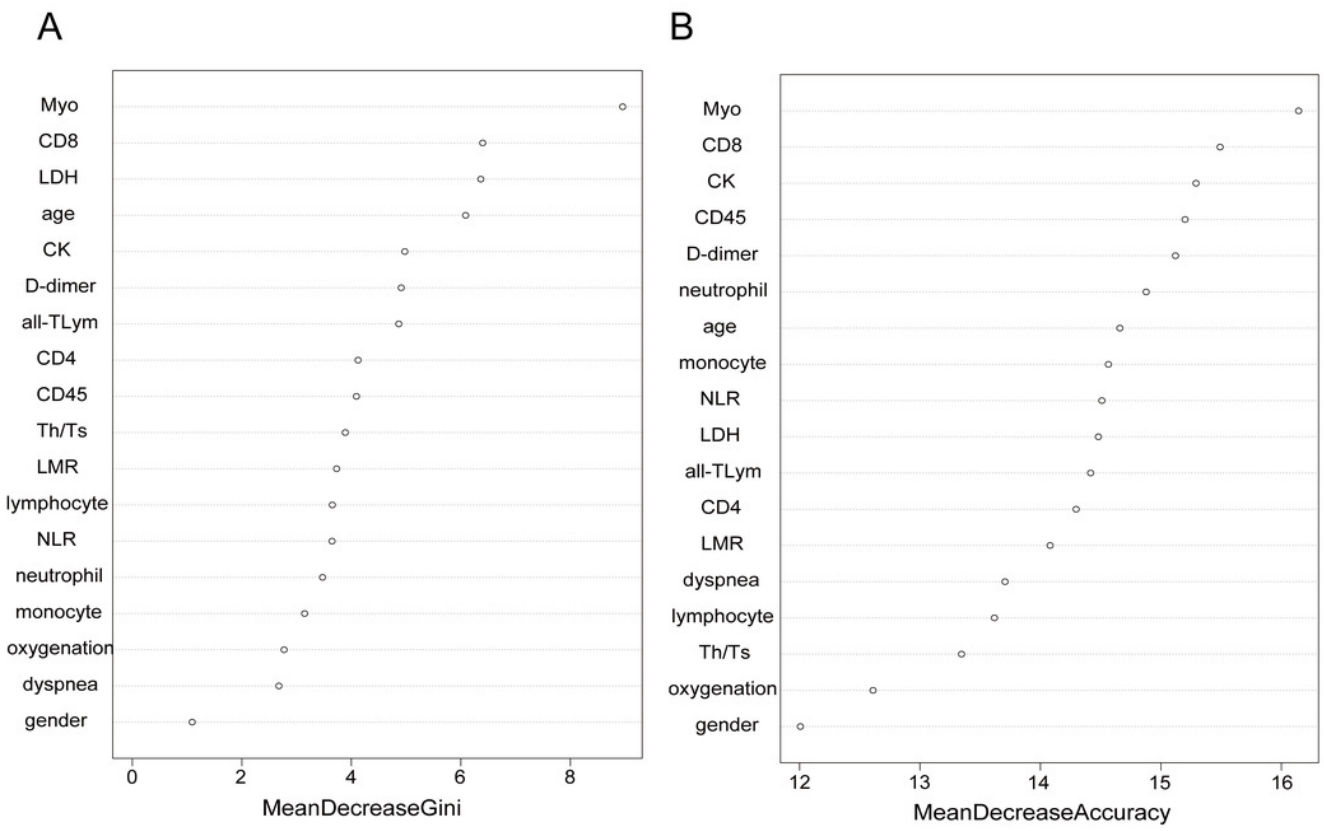


Figure 2

The accuracy of RF classification models.

(A)Heat map visualization shows Pearson 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 the relationship between OOB, death, survival error rate and the number of decision trees.

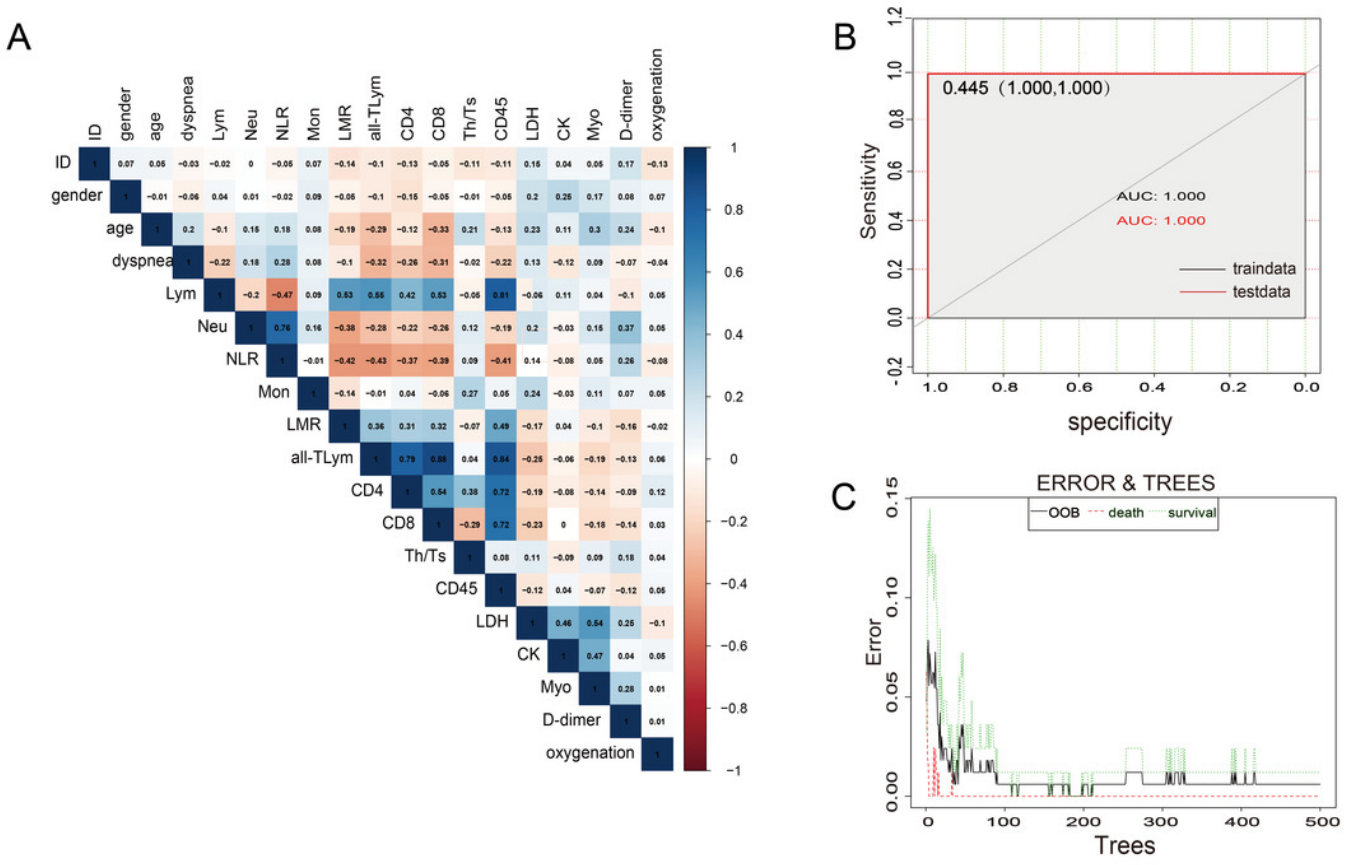
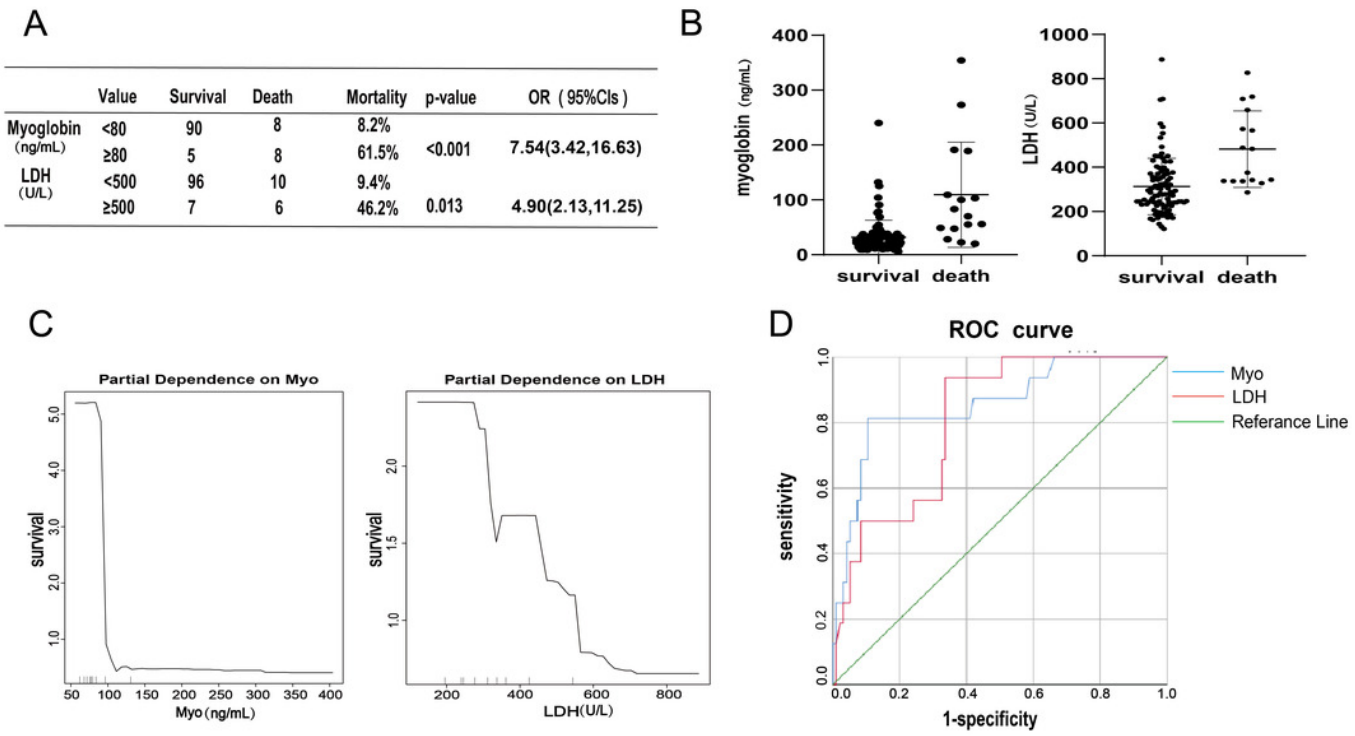


Figure 3

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.



	Value	Survival	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)