Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Dynamic Prognosis Model for Predicting Survival in Severe and Critically Ill COVID-19 Patients Using Machine Learning

Yongyue Wei
Nanjing Medical University
Jieyu He
Nanjing Medical University
Jiao Chen
Nanjing Medical University Sir Run Run Hospital
Ying Zhu
Nanjing Medical University https://orcid.org/0000-0002-8153-4412
Jiajin Chen
Nanjing Medical University
Jingjing Ding
Sir Run Run Hospital, Nanjing Medical Universiy
Hao Wang
Sir Run Run Hospital, Nanjing Medical University
Yahua Hu
the Affiliated Huangshi Central Hospital of Hubei Polytechnic University
Yingzi Huang
Zhongda Hospital, Southeast University
Yue Jiang
Nanjing Medical University
Zoucheng Pan
Nanjing Medical University
Sipeng Shen
Nanjing Medical University
Wei Zhao
Sir Run Run, Nanjing Medical University
Wei Gao
Sir Run Run, Nanjing Medical University
Feng Chen
Nanjing Medical University
Xiang Lu (luxiang66@njmu.edu.cn)
Sir Run Run Hospital, Nanjing Medical University, 109 Longmian Avenue, Nanjing, 211166, China.

Research

Keywords: COVID-19, prognosis, predictive model, random forest, trajectory

Posted Date: September 2nd, 2020

DOI: https://doi.org/10.21203/rs.3.rs-64083/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License. Read Full License

Abstract

Background Novel coronavirus disease (COVID-19) is an emerging, rapidly evolving situation. At present, the prognosis of severe and critically ill patients has become an important focus of attention. We strived to develop a prognostic prediction model for severe and critically ill COVID-19 patients.

Methods

To assess the factors associated with the prognosis of those patients, we retrospectively investigated the clinical, laboratory characteristics of confirmed 112 cases of COVID-19 admitted between 21 January to 6 March 2020 from Huangshi Central Hospital, Huangshi Hospital of Traditional Chinese Medicine, and Daye People's Hospital. We applied machine learning method (survival random forest) to select predictors for 28-day survival and taken into account the dynamic trajectory of laboratory indicators.

Results

Fifteen candidate prognostic features, including 11 baseline measures (including platelet count (PLT), urea, creatine kinase (CK), fibrinogen, creatine kinase isoenzyme activity, aspartate aminotransferase (AST), activation of partial thromboplastin time (APTT), albumin, standard deviation of erythrocyte distribution width (RBC-SD), neutrophils (%) and red blood cell count (RBC)) and 4 trajectory clusters (changes during hospitalization in the white blood cell (WBC), PLT large cell ratio (P-LCR), PLT distribution width (PDW) and AST), combined with covariates achieved 100% (95%CI: 99%-100%) AUC and reached 87% (95%CI: 84%-91%) AUC in an external validation set.

Conclusions

Taking advantage of random forest technique and laboratory dynamic measures, we developed a forest model to predict survival outcome of COVID-19 patients, which achieved 87% AUC in the external validation set. Our online tool will help to facilitate the early recognition of patients with high risk.

Introduction

The outbreak of coronavirus disease 2019 (COVID-19) caused by severe acute respiratory syndrome coronavirus-2 (SARS-CoV-2) is a major worldwide public health concern[1]. By July 11, 2020, this virus has rapidly spread to approximately 215 countries, causing a total of 12,102,328 cases with 551,046 deaths, with no sign of stopping[2].

The typical initial symptoms of COVID-19 usually consist of cough, fever, nasal congestion, expectoration, fatigue, and other signs of upper respiratory tract infections[3]. Approximately 80% of COVID-19 cases are classified as mild and the symptoms usually disappear within two weeks[4]. However, the remaining patients, classified as severe or critical illness, experience clinical deterioration with acute respiratory distress syndrome, septic shock, metabolic acidosis, and coagulation dysfunction, which can progress into multi-organ failure and death[3, 5].

Mortality among overall COVID-19 patients is 4.4%, but mortality of severe and critical illness can be as high as 49%[4]. Due to the absence of specific therapeutics and an effective vaccine, current COVID-19 treatment mainly focuses on symptomatic and respiratory support[6]. Therefore, identifying patients in high-risk groups is vital for patient management and medical resource allocation to decrease the case-fatality rate.

Potential factors predicting poor COVID-19 outcomes include older age, male, organ dysfunction, elevated levels of d-dimer and inflammatory markers, and lymphocytopenia[7-10]. Liang *et al.* developed a risk score incorporating 10 characteristics at admission of COVID-19 patients to predict 'the risk of developing critical illness during hospitalization, achieving an area under the receiver operating characteristic curve (AUC) of 88% in both development and validation cohorts[11]. In addition, Yan and colleagues used machine learning to select three biomarkers from patients' final biospecimens that can predict the mortality of COVID-19 patients >10 days in advance with more than 90% accuracy[12]. Of note, Dong *et al.* developed a nomogram based survival assessment model incorporating three predictors at or immediately after admission which reached approximately 90% AUC[13]. However, the generalization and stability of existing models warrants further validation. In addition, the impact of the dynamic changes of laboratory indices during hospitalization has not been well considered, especially for severe and critically ill patients with high mortality.

Thus, the present study aimed to construct a prognostic prediction model for severe and critically ill COVID-19 patients. Random forest incorporating the demographic and clinical characteristics and laboratory metrics at both admission and during hospitalization was used to develop the model and identify COVID-19 patients with high risks of mortality.

Methods

Study Population, Data Sources and Processing

In the discovery set, all severe and critical-ill COVID-19 patients were recruited in three hospitals in Huangshi City, Hubei Province, China between 21 January and 6 March, 2020, including Huangshi Central Hospital, Huangshi Hospital of Traditional Chinese Medicine, and Daye People's Hospital. Huangshi City is located about 100 km far away from Wuhan City, the center of the outbreak in domestic China. COVID-19 diagnoses were confirmed by real-time reverse-transcription polymerase-chain-reaction (RT-PCR) assay or high-throughput DNA sequencing for nasal and pharyngeal swab specimens[14]. Patients were defined as severe illness if they had any of the following criteria: respiratory rate of at least 30 breaths per min, oxygen saturation of 93% or lower in a resting state, ratio of arterial partial pressure of oxygen and oxygen concentration no greater than 300 mm Hg, or more than 50% lesion progression in lung imaging within 24-48 h. We defined critical COVID-19 illness as a composite of admission to the intensive care unit (ICU), invasive ventilation, or death[15]. All the 112 patients diagnosed as severe or critically ill at admission or during hospitalization in Huangshi City were included in this study.

The ethics committee of the hospitals waived the written informed consent from patients with COVID-19. Detailed demographics and clinical characteristics including initial symptoms, comorbidities and disease severity were recorded at admission. A team of experienced respiratory clinicians reviewed, abstracted and cross-checked the data. Each record was checked independently by 2 clinicians. Laboratory examinations including routine blood tests, lymphocyte subsets, inflammatory or infection-related biomarkers, cardiac, renal, liver and coagulation function tests were obtained at admission and during hospitalization. Baseline laboratory measures with over 40% missing values were excluded from the analysis.

In the validation set, data were obtained from recently published literature[12]. Briefly, the validation population was collected in Tongji Hospital between 10th January and 18th February 2020 and included 375 patients, 197 with general, 27 with severe and 151 with critical COVID-19. Medical records, including epidemiological, demographic, clinical, laboratory and mortality outcome information, were collected. The follow-up time was defined as hospital admission to death or discharge.

Outcome

Death within 28 days after admission to the hospital was the primary outcome of this study. Discharge from the hospital within 28 days or remaining hospitalized after 28 days were considered censored. Time-to-event outcome was defined for the following statistical models.

Statistical analysis

Continuous variables were summarized as mean and standard deviation (SD), and categorized variables were described by frequency (n) and proportion (%). The K-nearest neighbor method (KNN) was used to impute missing laboratory values at baseline using R package *KNN*$_{Imputation}$[16].

Trajectory identification for laboratory measurements

For each laboratory indicator with repeated measures during hospitalization, trajectory analysis was performed to cluster patients based on the dynamic time-series trend of the indicator using R package *traj*[17]. According to trajectory analysis requirements, patients with < 4 observations of the specific indicator were manually classified to the cluster with insufficient data points.

Feature selection

Survival random forest (RF) is a powerful nonparametric and decision-driven machine learning approach to handle high-dimensional data and time-to-event outcome, but false positive or spurious association may occur if confounding factors are not corrected. Ranger, a weighted version of random forest, gives 100% probability for potential confoundings to be selected as candidates for tree construction.

In the discovery set, the importance of all features, including demographics, baseline clinical characteristics, laboratory examination at admission, and laboratory trajectory during hospitalization were evaluated by Ranger[18]. Variable importance scores (VIS) for the features were estimated and ranked in a descending order. In addition, the sliding windows sequential forward feature selection method (SWSFS) was used to identify the top important biomarkers[19]. The SWSFS method adds one variable at a time in the order of variable importance to the Ranger model. The plot of 'out of bagging (OOB)' error was plotted to measure the performance of each model consisting of a specific set of biomarkers. The set of biomarkers having the lowest OOB error was selected as candidate prognostic factors for further analysis.

Further, a Cox proportional hazards model adjusted for age, gender, number of comorbidities was applied to test the association between the candidate factors and overall survival in both the discovery and validation sets. The difference of hazards was illustrated via Kaplan-Meier survival curves.

Prediction forest model construction

For the discovery set, candidate prognostic factors and covariates were put into Ranger to construct prediction forest models, including 1000 decision trees in the forest. This was further validated in the validation set.

Applying the values of features of each patient creates a prediction forest that outputs survival probability by each tree in the forest, thus forming a survival probability distribution for each patient. The median of the probability distribution represents an estimate of the survival probability of each patient.

Assessment of accuracy

Time-dependent ROC analysis was performed with R packages *ROCR*[20] and *pROC*[21] to estimate the area under the ROC curve (AUC) at day 28 since admission to the hospital. The c-index, sensitivity, and specificity idicate the accuracy of the prediction forest model.

To assess the stability of the prediction forest model, the discovery set was randomly divided into training (55 samples) and testing sets (57 samples). The prediction model was trained in the training set, followed by internal evaluation in the testing set and external validation. This was repeated for 1000 times.

Statistical analyses were performed using R version 4.0.1 (The R Foundation of Statistical Computing). The *P* values less than 0.05 were considered statistical significance unless otherwise specified.

Results

Study population characteristics

In the discovery phase, we enrolled 112 severe or critically ill COVID-19 patients from three hospitals in Huangshi City, Hubei Province, China. There were 49 (43.75%) critical illnesses and 31 deaths (27.68%). The mean (SD) age of patients was 61.0 (14.9) years, and 73 (65.2%) patients were male. The symptoms were fever (81.2%), cough (76.8%), chest tightness (65.2%), fatigue (58.0%), shortness of breath (30.4%), phlegm (25.0%), and diarrhea (17.0%) among others, and most patients had two or more (Table 1). There were 66 (58.9%) patients with one or more comorbidities (Table 1). Eighteen (16%) patients had abnormal chest imaging findings at admission. The laboratory measures at admission are presented in Table 2. The imputed data results were generally consistent with the original data (Table S1).

The characteristics of the 375 patients in the validation set were detailed in the original literature[12]. In brief, the mean (SD) age of these patients was 58.83 (16.46), and 224 (59.7%) were male. In the validation set, generally, severe and critically ill patients accounted for 52.5%, 7.2%, and 40.3%, respectively, and the mortality rates in the three groups were 6.09%, 51.85%, 98.01%, respectively (Table S2). Overall, 174 (46.4%) patients in the validation died during hospitalization (Table S2).

Feature selection

In the discovery set, there were 52 laboratory tests with sufficient (≥4) numbers of repeated measures for use in trajectory classification. In total, 3 covariates, 61 laboratory measures at admission, and 52 laboratory trajectory clusters were included in the Ranger model. SWSFS analysis identified the 15 top laboratory features with minimal OOB errors, including 11 laboratory measures at admission: platelet count (PLT), urea, creatine kinase (CK), fibrinogen, creatine kinase isoenzyme activity, aspartate aminotransferase (AST), activation of partial thromboplastin time (APTT), albumin, standard deviation of erythrocyte distribution width (RDW-SD), neutrophils (%) and red blood cell count (RBC), as well as 4 trajectory clusters including the trajectory during hospitalization of white blood cell (WBC), PLT large cell ratio (P-LCR), PLT distribution width (PDW) and AST (Fig. 1).

Cox regression adjusted for common covariates showed that patients at admission with higher neutrophils proportion [Hazard ratio (HR), 3.85; 95% confidence interval (CI), 1.70-8.70; *P*=0.0012], higher Urea (HR, 5.20; 95%CI, 2.15-12.59; *P*=0.0003), higher CK (HR, 4.86; 95%CI, 1.78-13.25; *P*=0.0020) and CK-MB (HR, 3.57; 95%CI, 1.59-8.01; *P*=0.0020), lower PLT (HR, 0.28; 95%CI, 0.11-0.69; *P*=0.0057), higher AST (HR, 2.33; 95%CI, 1.11-4.89; *P*=0.0258), lower albumin (HR, 0.18; 95%CI, 0.04-0.78; *P*=0.0222), higher variation of RDW (HR, 2.63; 95%CI, 1.10-6.30; *P*=0.0297), lower RBC (HR, 0.40; 95%CI, 0.17-0.95; *P*=0.0369) and lower fibrinogen (HR, 0.47; 95%CI, 0.22-0.98; *P*=0.0445) had worse survival outcomes (Fig. 2). In addition, Cox regression for trajectory features showed that persistently higher and more varied WBC, P-LCR, PDW, and AST were associated with increased hazard of death (Fig. 3). After correcting for false discovery rates, all variables except APTT were significant (Table S3).

Prediction forest model construction and assessment

The prediction forest was constructed in the discovery set using all 15 candidate prognostic features combined with covariates. The random forest model achieved 100% (95%CI: 99%-100%) AUC for predicting mortality within 28 days of admission to the hospital with a 3-fold internal cross validation to control for over-fitting. Further, the prediction forest model was validated in the external validation set. The trajectory cluster of each laboratory measure in the validation set was classified by adding one case at a time to the trajectory model trained in the discovery set. In the validation set, association between the baseline indicators and outcome was significant except for APTT, consistent with the results of the discovery set (Figure S1). The AUC in the external validation set reached 87% (95%CI: 84%-91%) (Fig. 4a), which was 14% higher than the AUC of the model using the covariates only (P=6.87×10^{-7}). The optimal cut-off of the survival probability at decision-making determined based on the Youden index was 0.58; the corresponding sensitivity and specificity for predicting 28-day mortality were 0.73 and 0.88, respectively; the specificity was 0.62 according to a fixed sensitivity of 0.90 (Fig. 4b).

In addition, to evaluate the stability of the modeling strategy, the discovery set was randomly split into training set (55 samples) and testing set (57 samples); the prediction forest was developed in the training set, followed by an internal validation in the testing set, and further evaluated in the external validation set. This was repeated 1000 times. The mean AUC was 0.87 (95% CI: 76%-98%) in the testing set, and 0.85 (95% CI: 0.80-0.89) in the validation set (Figure S2).

Comparison with existing prognostic prediction models

Further, we verified the prognostic prediction models reported in published studies in our discovery dataset. The c-index ranged from 0.64 to 0.74, and AUC from 0.66 to 0.82 (Table S4).

Web-based application tool

To facilitate the application of our prediction forest model, we developed an online tool that can be accessed at http://106.15.72.70:3838/COSP. By uploading the values of prognostic factors, the tool will output the distribution of likelihood that a given COVID-19 patient will die at a specific time point (Fig. 5).

Discussion

Using machine learning, we developed and validated a prognostic prediction model incorporating both baselines and dynamic trajectories of laboratory tests that are routinely performed at admission or tenure in the hospital to predict the survival outcome of severe or critically ill COVID-19 patients. Compared to existing models used to predict COVID-19 prognostic outcomes, our model is more accurate and uses more readily available metrics. Yan and colleagues developed a decision tree with three laboratory factors and achieved 95% AUC[12]. However, the hs-CRP test is not typically performed, as the standard CRP is more cost-effectiveness. Hence, we could not validate their tree-model in our discovery dataset, but our prediction forest model achieved considerable accuracy (AUC 87%) in Yan's dataset. Notably, Dong *et al.* built a predictive nomogram by using only three indicators: hypertension, neutrophil-to-lymphocyte ratio and NT-proBNP at admission to hospital, which surprisingly achieved 89.2% of C-index in internal validation set[13]. However, NT-proBNP is usually examined in patients with symptoms of cardiac dysfunction rather than regularly measured at admission, which may limit its application. In addition, a model without external validation should be generalized with caution. Liang *et al.* built a risk score with 10 predictors (COVID-GRAM) to assess the risk of developing critical illness in hospitalized patients with COVID-19, which achieved plausible accuracy (88% AUC) in both development and external validation sets [11]. Of note, risk of developing critical illness is highly correlated with hazard of death. Thus, to test if the COVID-GRAM model can be generalized to predict survival, we validated COVID-GRAM in our discovery set, which obtained 77% AUC to predict the survival of COVID-19 patients. Additional prognostic models proposed by Wu *et al.*[22] and Xie *et al.*[23] obtained 93% and 96% AUC, respectively, but these models resulted in unsatisfactory values of AUC and C-indices of 0.64. Our prediction forest model appears to be superior in accuracy compared to the existing models for predicting COVID-19 patient survival (Table S4).

We identified 11 laboratory measures at admission which appear to associate with the poor COVID-19 outcomes. Among the regularly measured laboratory indicators at admission, the high neutrophil count, low lymphocyte count, high neutrophil-to-lymphocyte ratio, high direct bilirubin, and elevated lactate dehydrogenase have previously been identified as prognostic predictors for an unfavorable outcome [12, 13, 22, 23]. However, in our discovery set, these predictors were not ranked at the top of VIS list for all candidate factors possibly because all patients in our study had progressed to severe or critical disease and the previous predictors may be less prognostic for severity. Notably, in our study, PLT had the highest variable importance score, and 2 of 4 trajectory predictors were platelet related: PDW and P-LCR. PDW and P-LCR are novel markers, but associations of platelet count with the risk or outcome of critically ill patients are well known[24-26]. Since human lung is the site of platelet biogenesis, the abnormal trajectory of PDW and P-LCR during COVID-19 illness may suggest the lung dysfunction and injury and add value to models predicting COVID-19 prognosis[27]. Also, platelets count and functional abnormality increase the risk of bleeding, secondary to clotting disease, such as heparin-induced thrombocytopenia (HIT) or disseminated intravascular coagulation (DIC), and thus increased the risk of death[28].

There are several strengths of this study. First, random forests have better modeling efficiency than most of the regular methods and can avoid overfitting, resist noise interference to a certain extent, are not sensitive to the distribution of variables, can handle non-linear relationships, and detect the interaction between features in the training process to improve predictive ability. Second, by integrating multiple classification trees, a model can output a series of predicted probabilities forming survival likelihood and accounting for uncertainty. Third, our model includes dynamic changes in laboratory measures during hospitalization, which are more relevant to the progression of the illness than baseline predictors. Despite these strengths, the sample size is relatively small in the discovery set which may limit the machine learning technique to build a more accurate model, although our model retained considerable predictive accuracy in the external validation dataset. Additionally, some patients in this study had non negligible missing values for laboratory tests. We thus used the KNN method to impute the missing values by “borrowing” the information from the correlated variables. Also, our model may be difficult to generalize as it was created using data from just severely and critically ill patients. The model was verified in an external validation set of patients with general, severe, or critical COVID-19, but requires more external validation in general and severe populations to ensure stability. Finally, the prediction model of this study was trained and validated using Chinese population. Hence, caution is warranted when extending these findings to other populations.

Conclusions

In conclusion, taking advantage of random forest technique and laboratory dynamic measures, we developed a highly accurate model to predict COVID-19 patient survival, and we have developed our prognostic model into a user-friendly web-based application tool, which outputs the distribution of a patient’s survival likelihood to capture predictive uncertainties. Our online tool will help to facilitate the early recognition of patients with high risk.

List Of Abbreviations

COVID-19, Coronavirus disease 2019; HR: Hazard ratio; SARS-CoV-2, Severe acute respiratory syndrome coronavirus 2; receiver operating characteristic curve, ROC; area under the ROC, AUC; random forest, RF; the sliding windows sequential forward feature selection method, SWSFS; Variable importance scores, VIS; out of bagging, OOB.

Declarations

Acknowledgements

Not applicable.

Author Contributions

Conception and design of the study: Y.W., J.H. W.G., W.Z, F.C., X.L. Drafting of the manuscript: Y.W., Y.Z., J.C., J.C., J.H., X.L. Statistical analysis and interpretation: Y.W., Y.Z., J.H., J.C. Sample collection: J.C., J.D., H.W., W.G., W.Z, X.L. Data entry: Y.J., J.H., Z.P., Y.W., J.C. Study supervision: F.C., X.L., W.G., W.Z. Manuscript revision and approval of the final manuscript: all authors.

Funding

This work was supported by the Special Program of the National Natural Science Foundation of China (81770440 and 81970218 to XL, 81970217 to WG, 82041024 to FC, 81973142 to YW), the Jiangsu Province Health Development Project with Science and Education (QNRC201685 to WG), and a grant from the Six Talent Peaks Project of Jiangsu Province (WSN-175 to WG).

Availability of data and materials

The data in discovery set that support the findings of this study are available from Qingyuan Zhan, but restrictions apply to the availability of these data, which were used under license for the current study, so these data are not publicly available. The data in validation set is available at https://doi.org/10.1038/s42256-020-0180-7.

Ethics approval and consent to participate

The study was approved by the ethics committee of Huangshi Central Hospital, Huangshi Hospital of Traditional Chinese Medicine, and Daye People’s Hospital.

Competing interests

The authors declare no conflicts of interests.

Author details

[1]Department of Epidemiology and Biostatistics, School of Public Health, Center for Global Health, Nanjing Medical University, Nanjing 211166, China.

[2]Department of Geriatrics, Sir Run Run Hospital, Nanjing Medical University, 109 Longmian Avenue, Nanjing, 211166, China.

[3]Department of Critical Care Medicine, Sir Run Run Hospital, Nanjing Medical University, Nanjing, China.

[4]Department of Emergency, Sir Run Run Hospital, Nanjing Medical University, Nanjing, China.

[5]Department of Critical Care Medicine, the Affiliated Huangshi Central Hospital of Hubei Polytechnic University, Huangshi, China

[6]Department of Critical Care Medicine, Zhongda Hospital, Southeast University, Nanjing, China.

Consent for publication

Not applicable.

References

1. Zhu N, Zhang D, Wang W, Li X, Yang B, Song J, Zhao X, Huang B, Shi W, Lu R *et al*: **A Novel Coronavirus from Patients with Pneumonia in China, 2019**. *The New England journal of medicine* 2020, **382**(8):727-733.
2. WHO: **Coronavirus disease 2019 (COVID-19) Situation Report.Updated April 10, 2020**. 2020.
3. Huang C, Wang Y, Li X, Ren L, Zhao J, Hu Y, Zhang L, Fan G, Xu J, Gu X *et al*: **Clinical features of patients infected with 2019 novel coronavirus in Wuhan, China**. *Lancet* 2020, **395**(10223):497-506.
4. Wu Z, McGoogan JM: **Characteristics of and Important Lessons From the Coronavirus Disease 2019 (COVID-19) Outbreak in China: Summary of a Report of 72314 Cases From the Chinese Center for Disease Control and Prevention**. *Jama* 2020.
5. Guan WJ, Ni ZY, Hu Y, Liang WH, Ou CQ, He JX, Liu L, Shan H, Lei CL, Hui DSC *et al*: **Clinical Characteristics of Coronavirus Disease 2019 in China**. *The New England journal of medicine* 2020.
6. Yang X, Yu Y, Xu J, Shu H, Xia J, Liu H, Wu Y, Zhang L, Yu Z, Fang M *et al*: **Clinical course and outcomes of critically ill patients with SARS-CoV-2 pneumonia in Wuhan, China: a single-centered, retrospective, observational study**. *The Lancet Respiratory medicine* 2020, **8**(5):475-481.
7. Zhou Y, Zhang Z, Tian J, Xiong S: **Risk factors associated with disease progression in a cohort of patients infected with the 2019 novel coronavirus**. *Annals of palliative medicine* 2020, **9**(2):428-436.
8. Zhou F, Yu T, Du R, Fan G, Liu Y, Liu Z, Xiang J, Wang Y, Song B, Gu X *et al*: **Clinical course and risk factors for mortality of adult inpatients with COVID-19 in Wuhan, China: a retrospective cohort study**. *Lancet* 2020, **395**(10229):1054-1062.
9. Pan F, Yang L, Li Y, Liang B, Li L, Ye T, Liu D, Gui S, Hu Y, Zheng C: **Factors associated with death outcome in patients with severe coronavirus disease-19 (COVID-19): a case-control study**. *International journal of medical sciences* 2020, **17**(9):1281-1292.
10. Wu C, Chen X, Cai Y, Xia J, Zhou X, Xu S, Huang H, Zhang L, Du C, Zhang Y *et al*: **Risk Factors Associated With Acute Respiratory Distress Syndrome and Death in Patients With Coronavirus Disease 2019 Pneumonia in Wuhan, China**. *JAMA internal medicine* 2020.
11. Liang W, Liang H, Ou L, Chen B, Chen A, Li C, Li Y, Guan W, Sang L, Lu J *et al*: **Development and Validation of a Clinical Risk Score to Predict the Occurrence of Critical Illness in Hospitalized Patients With COVID-19.***JAMA Intern Med* 2020(2168-6114 (Electronic)).
12. Yan L, Zhang H-T, Goncalves J, Xiao Y, Wang M, Guo Y, Sun C, Tang X, Jing L, Zhang M *et al*: **An interpretable mortality prediction model for COVID-19 patients**. *Nature Machine Intelligence* 2020, **2**(5):283-288.
13. Dong YM, Sun J, Li YX, Chen Q, Liu QQ, Sun Z, Pang R, Chen F, Xu BY, Manyande A *et al*: **Development and Validation of a Nomogram for Assessing Survival in Patients with COVID-19 Pneumonia. LID - ciaa963 [pii] LID - 10.1093/cid/ciaa963 [doi]**. *Clin Infect Dis* 2020(1537-6591 (Electronic)).
14. **Clinical management of COVID-19** [https://www.who.int/publications/i/item/clinical-management-of-covid-19]
15. **Diagnosis and treatment protocol for novel coronavirus pneumonia** [http://www.nhc.gov.cn/yzygj/s7653p/202003/46c9294a7dfe4cef80dc7f5912eb1989/files/ce3e6945832a438eaae415350a8ce964.pdf]
16. Torgo L: **Data Mining with R: Learning with Case Studies, Second Edition**: Chapman & Hall/CRC; 2017.

17. Leffondré K, Abrahamowicz M, Regeasse A, Hawker GA, Badley EM, Mccusker J, Belzile E: **Statistical measures were proposed for identifying longitudinal patterns of change in quantitative health indicators**. *Journal of Clinical Epidemiology* 2004, **57**(10):0-1062.
18. Wright MN, Ziegler A: **ranger: A Fast Implementation of Random Forests for High Dimensional Data in C++ and R**. *Journal of Statistical Software* 2017, **077**(1).
19. Jiang R, Tang W, Wu X, Fu W: **A random forest approach to the detection of epistatic interactions in case-control studies**. *BMC Bioinformatics* 2009, **10 Suppl 1**(Suppl 1):S65.
20. Sing T, Sander O, Beerenwinkel N, Lengauer T: **ROCR: visualizing classifier performance in R**. *Bioinformatics (Oxford, England)* 2005, **21**(20):3940-3941.
21. Robin X, Turck N, Hainard A, Tiberti N, Lisacek F, Sanchez J-C, Müller M: **pROC: an open-source package for R and S+ to analyze and compare ROC curves**. *BMC Bioinformatics* 2011, **12**(1):77.
22. Wu S, Du Z, Shen S, Zhang B, Yang H, Li X, Cui W, Chen F, Huang J: **Identification and validation of a novel clinical signature to predict the prognosis in confirmed COVID-19 patients.** . *Clin Infect Dis* 2020(1537-6591 (Electronic)).
23. Xie J, Shi D, Bao M, Hu X, Wu W, Sheng J, Xu K, Wang Q, Wu J, Wang K *et al*: **A Predictive Nomogram for Predicting Improved Clinical Outcome Probability in Patients with COVID-19 in Zhejiang Province, China**. *Engineering* 2020.
24. Akca S, Haji-Michael P Fau - de Mendonça A, de Mendonça A Fau - Suter P, Suter P Fau - Levi M, Levi M Fau - Vincent JL, Vincent JL: **Time course of platelet counts in critically ill patients**. *Crit Care Med* 2002, **30**(0090-3493 (Print)).
25. Wei Y, Tejera P, Wang Z, Zhang R, Chen F, Su L, Lin X, Bajwa EK, Thompson BT, Christiani DC: **A Missense Genetic Variant in LRRC16A/CARMIL1 Improves Acute Respiratory Distress Syndrome Survival by Attenuating Platelet Count Decline**. 2017, **195**(1535-4970 (Electronic)):1353-1361.
26. Rice TW, Wheeler AP: **Coagulopathy in critically ill patients: part 1: platelet disorders**. (1931-3543 (Electronic)).
27. Lefrançais E, Ortiz-Muñoz G, Caudrillier A, Mallavia B, Liu F, Sayah DM, Thornton EE, Headley MB, David T, Coughlin SR *et al*: **The lung is a site of platelet biogenesis and a reservoir for haematopoietic progenitors**. *Nature* 2017, **544**(1476-4687 (Electronic)).
28. Michiels JJ, Berneman Z Fau - Van Bockstaele D, Van Bockstaele D Fau - van der Planken M, van der Planken M Fau - De Raeve H, De Raeve H Fau - Schroyens W, Schroyens W: **Clinical and laboratory features, pathobiology of platelet-mediated thrombosis and bleeding complications, and the molecular etiology of essential thrombocythemia and polycythemia vera: therapeutic implications**. *Semin Thromb Hemost* 2006, **32**(0094-6176 (Print)).

Tables

Table 1 Demographic and clinical characteristics at hospitalization of severe or critically ill COVID-19 patients.

Characteristics	$N_{Missing}$ (%)	Total (n=112)	Survived (n=81)	Died (n=31)
Age, years, mean(SD)		61.0(14.9)	57.1(13.8)	71.0(13.0)
Male, n(%)		73(65.2)	54(66.7)	19(61.3)
Vital signs, mean(SD)				
Temperature, °C	2(1.8)	37.3(0.8)	37.3(0.8)	37.2(0.8)
Heart rate, beats/min	29(25.9)	89.4(17.7)	87.1(16.7)	94.4(19.0)
Respiratory rate, breaths/min	5(4.5)	24.8(5.6)	25.1(5.9)	24.1(4.9)
Blood pressure, mm Hg				
Diastolic	5(4.5)	73.2(13.7)	73.3(14.7)	72.8(11.0)
Systolic	5(4.5)	124.9(17.3)	124.0(18.0)	127.0(15.7)
Symptoms, n(%)				
Fever		91(81.2)	67(82.7)	24(77.4)
Cough		86(76.8)	62(76.5)	24(77.4)
Chest tightness		73(65.2)	56(69.1)	17(54.8)
Fatigue		65(58.0)	54(66.7)	11(35.5)
Shortness of breath		34(30.4)	21(25.9)	13(41.9)
Phlegm		28(25.0)	20(24.7)	8(25.8)
Dyspnea		25(22.3)	14(17.3)	11(35.5)
Diarrhoea		19(17.0)	15(18.5)	4(12.9)
Headache		9(8.0)	7(8.6)	2(6.5)
Myalgia		6(5.4)	5(6.2)	1(3.2)
Sore throat		5(4.5)	4(4.9)	1(3.2)
Nausea and vomiting		5(4.5)	2(2.5)	3(9.7)
Imaging abnormality[a]		18(16.1)	13(16.0)	5(16.1)
No. of Symptoms, n(%)				
0		2(1.8)		2(6.5)
1		4(3.6)	4(4.9)	
2		15(13.4)	10(12.3)	5(16.1)
3		20(17.9)	15(18.5)	5(16.1)
4		30(26.8)	23(28.4)	7(22.6)
5		23(20.5)	16(19.8)	7(22.6)
6		12(10.7)	8(9.9)	4(12.9)
≥7		6(5.4)	5(6.2)	1(3.2)
Comorbidities, n(%)				
Hypertension		40(35.7)	26(32.1)	14(45.2)
Respiratory failure		27(24.1)	16(19.8)	11(35.5)
Cardiovascular disease		17(15.2)	10(12.3)	7(22.6)
Diabetes		21(18.8)	15(18.5)	6(19.4)

Characteristics	$N_{Missing}$ (%)	Total (n=112)	Survived (n=81)	Died (n=31)
Acute lung injury		14(12.5)	9(11.1)	5(16.1)
COPD[b]		5(4.5)	2(2.5)	3(9.7)
Bacterial pneumonia		3(2.7)	2(2.5)	1(3.2)
Hepatic injury		3(2.7)	3(3.7)	
Septic shock		3(2.7)	2(2.5)	1(3.2)
cerebral infarction		2(1.8)	1(1.2)	1(3.2)
Acute kidney injury		1(0.9)	1(1.2)	
cerebral hemorrhage		1(0.9)	1(1.2)	
Sepsis		1(0.9)	1(1.2)	
N **of Comorbidities, n(%)**				
0		46(41.1)	36(44.4)	10(32.3)
1		26(23.2)	20(24.7)	6(19.4)
2		19(17.0)	13(16.0)	6(19.4)
3		13(11.6)	7(8.6)	6(19.4)
4		4(3.6)	2(2.5)	2(6.5)
5		1(0.9)	1(1.2)	0(0)
≥6		3(2.7)	2(2.5)	1(3.2)
Worst severity in hospital				
Severe		63	63	0
Critical illness, n(%)		49	18	31

Abbreviation: SD, standard error.

[a] Including chest radiography and Computed tomography (CT).

[b] Chronic obstructive pulmonary disease.

Table 2 Laboratory findings of severe or critically ill COVID-19 patients.

Variable	$N_{Missing}$ (%)	Total (n=112)	Survived (n=81)	Died (n=31)
Blood Routine				
Platelet-related				
Platelet count (PLT), 10^9/L	24(21.4)	183.0(81.9)	195.9(82.6)	148.8(70.7)
Mean platelet volume (MPV), fL	25(22.3)	11.2(1.0)	11.1(1.1)	11.5(0.8)
Platelet distribution width (PDW), fL	24(21.4)	14.2(2.3)	14.2(2.4)	14.3(2.2)
Platelet large cell ratio (PLCR), %	29(25.9)	35.9(7.4)	35.5(7.7)	36.9(6.7)
Thrombocytocrit, %	24(21.4)	0.2(0.1)	0.2(0.1)	0.2(0.1)
Red blood cell related				
Red blood cell count (RBC), 10^12/L	23(20.5)	4.3(0.6)	4.4(0.6)	4.0(0.5)
RDW-CV, %	26(23.2)	13.3(2.0)	13.3(2.2)	13.1(0.8)
RDW-SD, fL	28(25.0)	42.3(4.5)	41.4(4.7)	44.6(2.9)
Hematocrit (HCT), %	26(23.2)	39.3(5.9)	39.6(6.1)	38.4(5.1)
Mean corpuscular volume (MCV), fL	23(20.5)	127.1(19.2)	128.2(20.6)	124.1(14.4)
Haemoglobin (Hb), g/L	23(20.5)	127.1(19.2)	128.2(20.6)	124.1(14.4)
Mean corpuscular hemoglobin (MCH), pg	23(20.5)	33.2(32.5)	34.1(37.8)	30.7(2.8)
MCH concentration (MCHC), g/L	23(20.5)	325.2(17.3)	326.0(15.2)	322.8(22.4)
White blood cell related				
White blood cell count (WBC), 10^9/L	18(16.1)	7.3(5.2)	6.2(3.5)	10.1(7.4)
Lymphocyte count (LYM), 10^9/L	20(17.9)	0.8(0.4)	0.8(0.4)	0.6(0.3)
Lymphocyte proportion (LYM%), %	19(17.0)	14.3(8.5)	15.9(8.5)	9.6(6.7)
Neutrophil count (NEU), 10^9/L	20(17.9)	5.7(4.1)	4.9(3.3)	7.9(5.4)
Neutrophil proportion (NEU%), %	18(16.1)	78.1(12.0)	75.4(11.8)	85.1(9.4)
Neutrophils / lymphocyte ratio (NLR)	21(18.8)	9.7(10.2)	8.5(10.2)	13.3(9.4)
Monocyte count, 10^9/L	22(19.6)	0.5(0.3)	0.5(0.3)	0.5(0.4)
Monocyte proportion, %	19(17.0)	14.3(8.5)	15.9(8.5)	9.6(6.7)
Blood Biochemistry				
Liver function				
Alanine transaminase, U/L	29(25.9)	37.7(32.8)	37.9(36.4)	37.2(19.7)
Aspartate transaminase, U/L	28(25.0)	42.1(27.2)	39.3(23.5)	50.0(35.2)
Gamma-glutamyltransferase, U/L	34(30.4)	59.7(63.4)	56.5(63.4)	70.3(63.8)
Alkaline phosphatase, U/L	33(29.5)	76.9(41.1)	73.6(41.9)	87.9(36.9)
Lactate dehydrogenase, U/L	42(37.5)	385.0(189.1)	349.1(167.8)	496.9(212.5)
Total protein, g/L	29(25.9)	63.6(7.0)	64.1(7.0)	62.0(6.8)
Albumin, g/L	30(26.8)	36.7(4.6)	37.3(4.6)	34.2(3.8)

Variable	$N_{Missing}$ (%)	Total (n=112)	Survived (n=81)	Died (n=31)
Globulin, g/L	32(28.6)	26.8(5.0)	26.4(5.1)	28.1(4.4)
Albumin–globulin ratio	29(25.9)	1.4(0.3)	1.5(0.3)	1.2(0.2)
Indirect bilirubin, umol/L	31(27.7)	4.3(3.2)	4.2(3.2)	4.7(3.1)
Direct bilirubin, umol/L	31(27.7)	6.8(3.6)	6.5(3.1)	8.0(4.9)
Total bilirubin, umol/L	31(27.7)	11.1(5.1)	10.7(4.9)	12.6(5.5)
Total bile acid, umol/L	35(31.2)	7.1(18.9)	7.8(21.3)	4.7(3.9)
Renal function				
Uric acid, μmol/L	32(28.6)	266.4(147.5)	265.9(144.5)	267.9(159.3)
Cystatin C, mg/L	36(32.1)	1.0(0.3)	1.0(0.3)	1.2(0.3)
Creatinine, mmol/L	26(23.2)	71.6(34.8)	70.5(37.2)	74.8(27.3)
Glucose, mmol/L	26(23.2)	8.0(5.0)	7.9(5.7)	8.3(3.0)
Urea, mmol/L	23(20.5)	5.8(3.3)	5.0(2.8)	7.8(3.5)
eGFR, ml/min	29(25.9)	91.4(24.2)	95.8(24.5)	79.9(19.5)
Creatine kinase, U/L	41(36.6)	234.5(336.4)	183.0(243.3)	398.2(511.5)
CK-MB, IU/L	36(32.1)	6.6(14.4)	4.9(14.2)	11.6(14.3)
CO2-CP, mmol/L	30(26.8)	24.4(4.0)	24.9(4.1)	23.1(3.6)
Electrolysis				
K, mmol/L	24(21.4)	4.1(0.6)	4.1(0.6)	4.1(0.7)
Ca, mmol/L	23(20.5)	2.1(0.1)	2.1(0.1)	2.0(0.1)
Cl, mmol/L	23(20.5)	102.8(12.3)	103.3(4.8)	101.7(22.2)
Na, mmol/L	24(21.4)	137.0(15.9)	135.9(18.2)	139.7(6.1)
Blood coagulation				
Prothrombin time, S	29(25.9)	12.1(1.4)	12.0(1.4)	12.5(1.3)
Thrombin time, S	29(25.9)	13.0(2.5)	12.7(2.4)	13.9(2.5)
Activated partial thromboplastin time, S	29(25.9)	38.0(8.3)	39.1(8.3)	34.6(7.3)
Fibrinogen, g/L	30(26.8)	4.8(1.4)	5.0(1.3)	4.3(1.7)
International normalized ratio (INR)	30(26.8)	1.0(0.1)	1.0(0.1)	1.0(0.1)
D-dimer, ug/ml	35(31.2)	0.8(1.2)	0.6(1.1)	1.6(1.4)

Data are shown as the mean (SD).

Figures

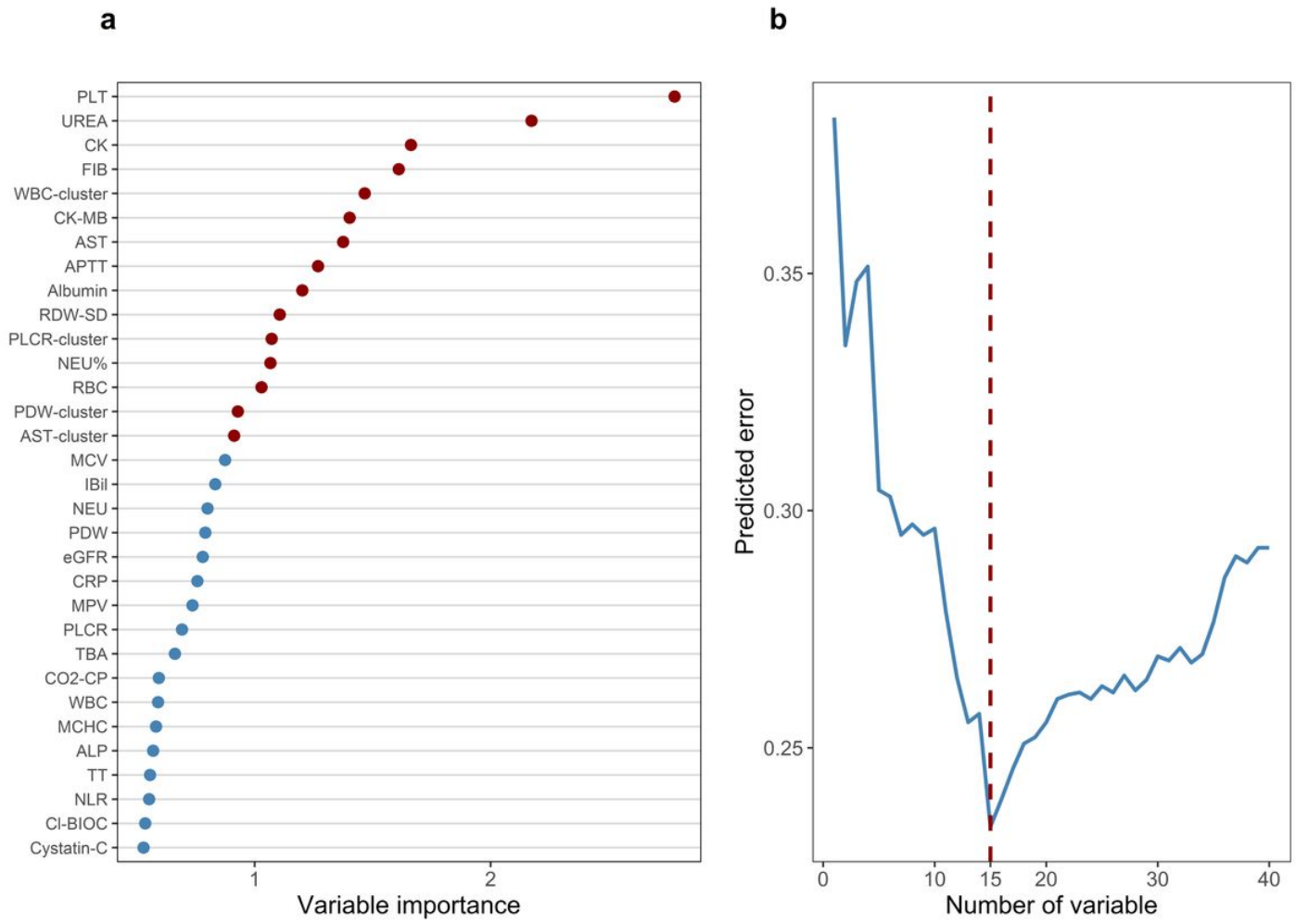

Figure 1

Random forest screening for prognostic factors for severe/critically ill COVID-19 patients in the discovery dataset. a variable importance score (VIS) plot for top indicators, where covariates were given 100% weight; b out-of-bag (OOB) error as the number of variables in the random forest (RF) model.

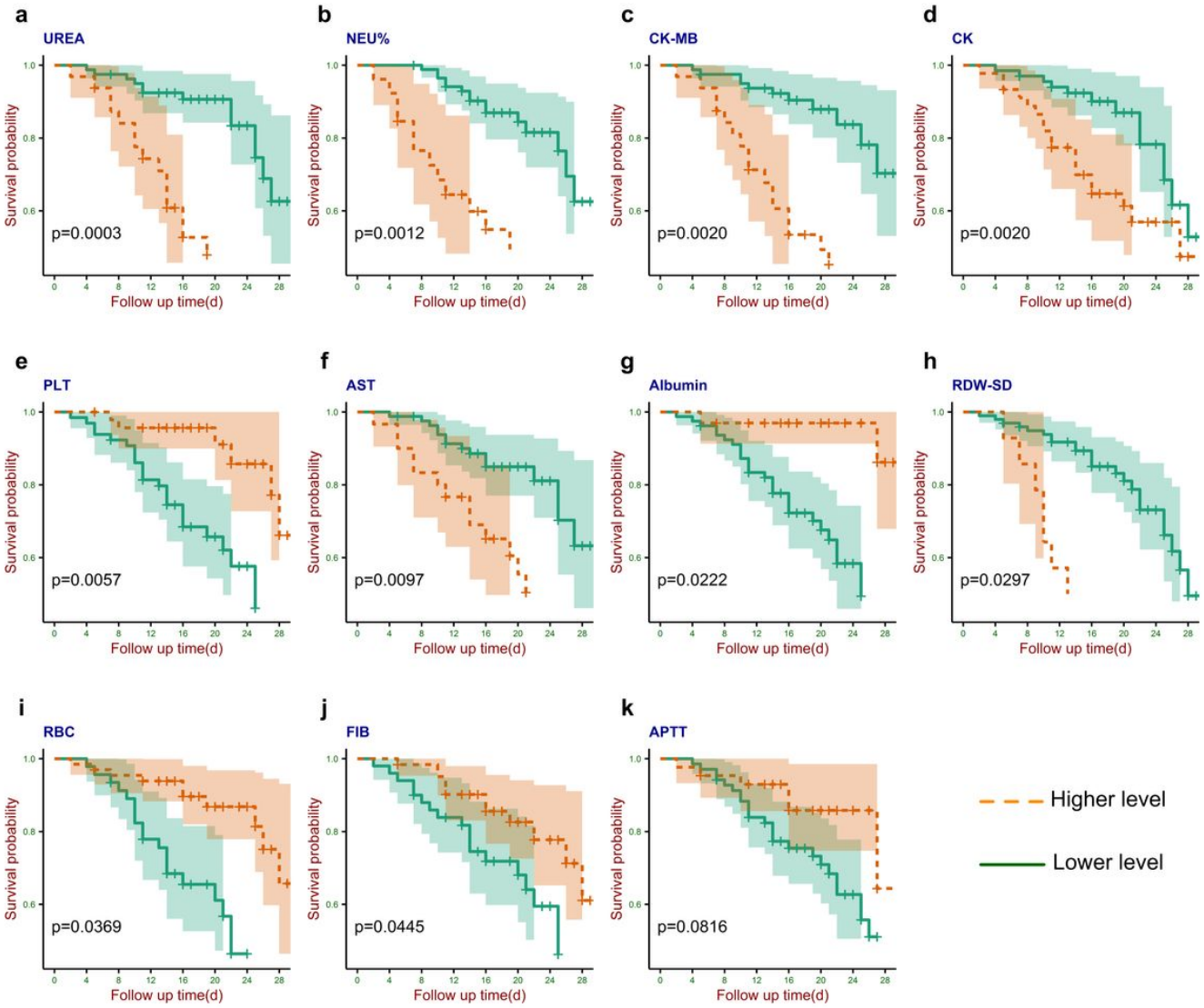

Figure 2

Baseline laboratory factors that associated with prognosis of severe or critically ill COVID-19 patients. a urea; b neutrophils (%); c creatine kinase isoenzyme activity (CK-MB); d creatine kinase (CK); e platelet count (PLT); f aspartate aminotransferase (AST); g albumin; h standard deviation of erythrocyte distribution width (RDW-SD); i red blood cell count (RBC); j fibrinogen; k activated partial thromboplastin time (APTT); red dashed lines indicate higher-level group and green solid lines for lower-level group; the optimal cutoff points were obtained by the maximum of discrimination of the survival curve.

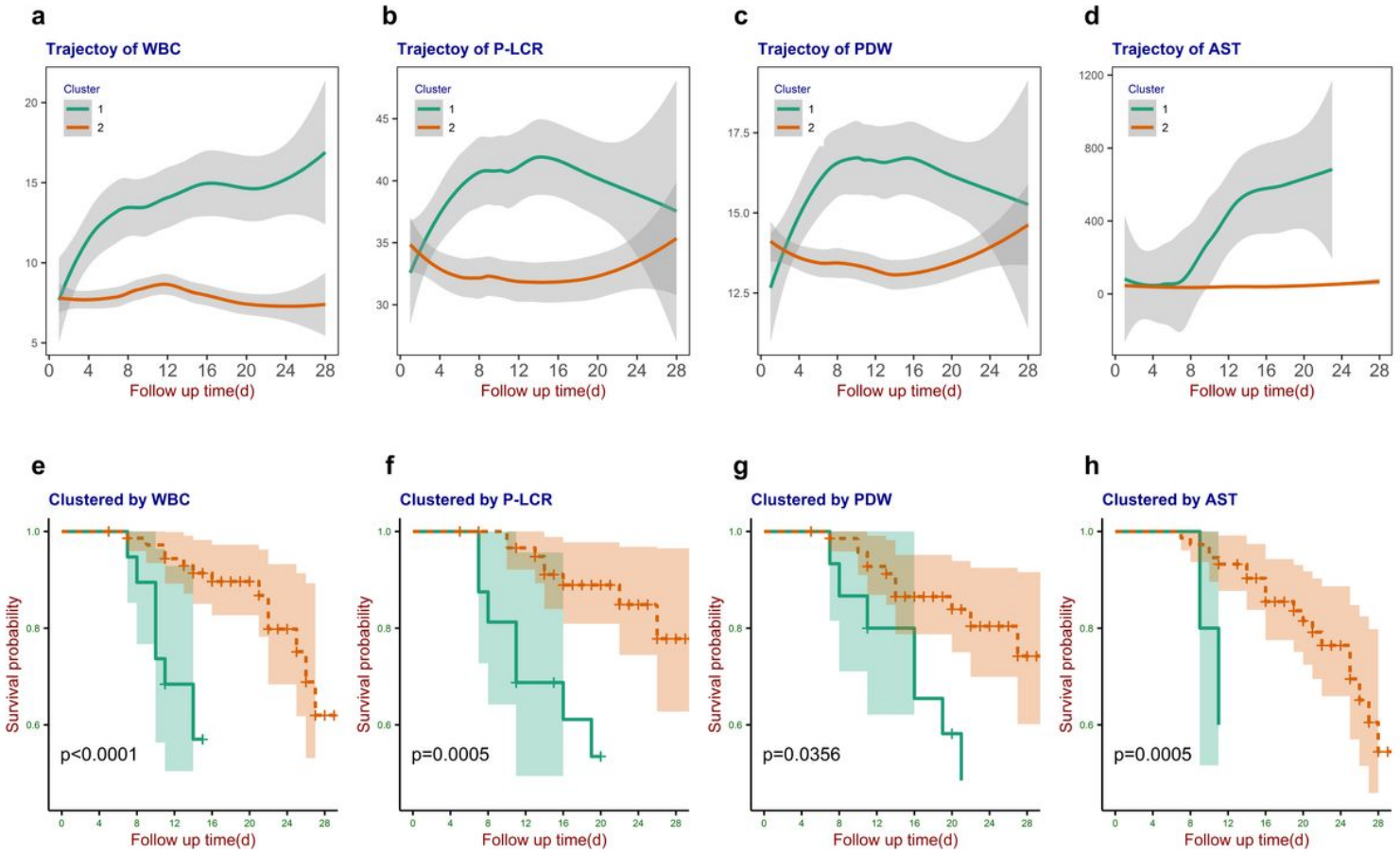

Figure 3

Trajectory of laboratory factors during hospitalization associated with prognosis of severe/critically ill COVID-19 patients. a trajectory of white blood cells (WBC) in patients during hospitalization; b trajectory of platelet large cell ratio (P-LCR) in patients during hospitalization; c trajectory of PLT distribution width (PDW) in patients during hospitalization; d trajectory of aspartate aminotransferase (AST) in patients during hospitalization; e association between WBC and prognosis of patients; f association between P-LCR and prognosis of patients; g association between PDW and prognosis of patients; h association between AST and prognosis of patients; red lines indicate lower-level and slighter-variation and green higher-level and larger-variation in a, b, c and d; red dashed lines represent groups with lower-level and slighter-variation and green solid lines represent groups with higher-level and greater-variation in e, f, g and h.

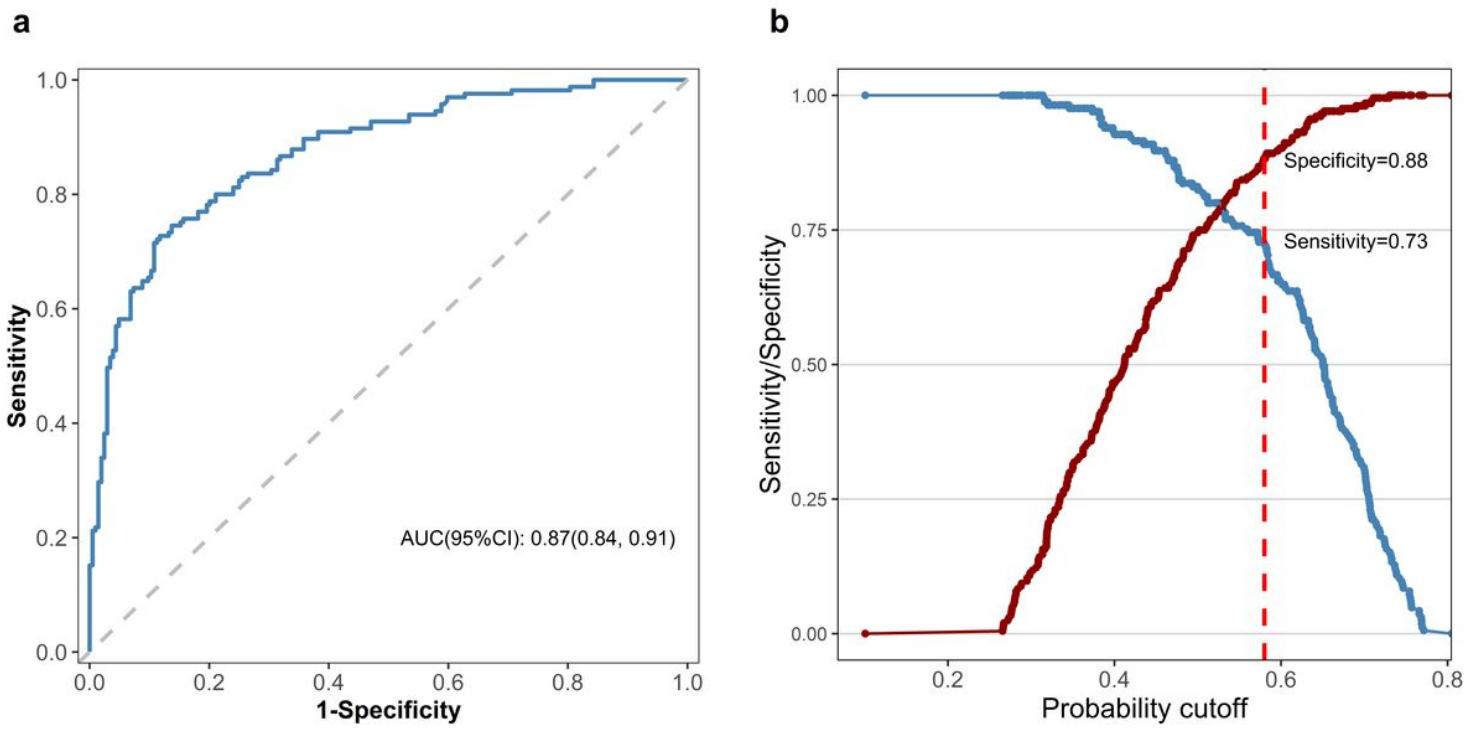

Figure 4

ROC curve and sensitivity/specificity curve of the optimal cutoff in validation dataset. a ROC curve and AUC of the RF model in predicting 28-day survival in the validation cohort; b an illustration of the sensitivity and specificity levels retrieved from the ROC-curve analysis, in which sensitivity (blue) and specificity (red) were plotted separately against the potential cutoff probability and the cutoff was specified with red dashed line where the Youden index was maximal.

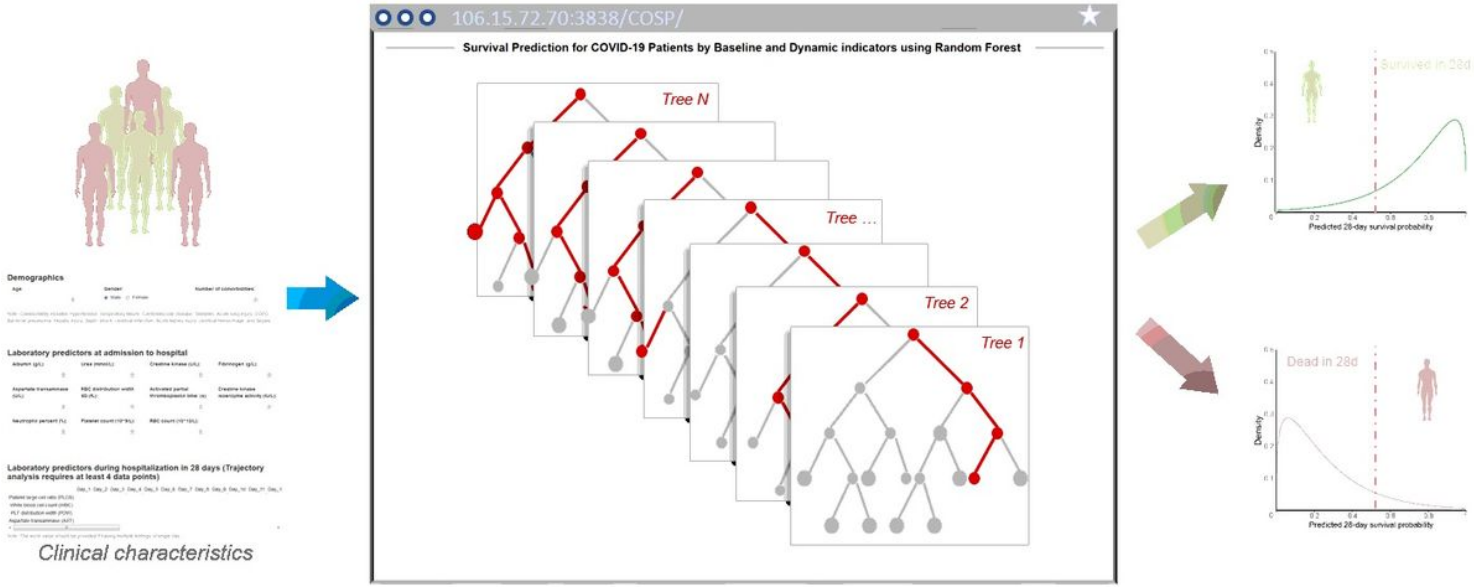

Figure 5

Schematic diagram of prediction model. The left panel is the input of a patient's clinical characteristics information, the middle panel is the random forest prediction model, and the right panel is the predicted survival probability distribution of a patient.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- Additionalfile1.docx