Machine learning: a predication model of outcome of SARS-CoV-2 pneumonia

Gang Wu
Department of Radiology, Tongji Hospital of Tongji Medical College of Huazhong University of Science and Technology

Shuchang Zhou
Department of Radiology, Tongji Hospital of Tongji Medical College of Huazhong University of Science and Technology

Yujin Wang
Department of Radiology, Tongji Hospital of Tongji Medical College of Huazhong University of Science and Technology

Xiaoming Li (lilyboston2002@qq.com)
Department of Radiology, Tongji Hospital of Tongji Medical College of Huazhong University of Science and Technology

---

Research Article

Keywords: SARS-CoV-2 pneumonia, laboratory features selected by machine learning, prediction model of outcome

Posted Date: April 15th, 2020

DOI: https://doi.org/10.21203/rs.3.rs-23196/v1

Version of Record: A version of this preprint was published on August 20th, 2020. See the published version at https://doi.org/10.1038/s41598-020-71114-7.

Abstract

The severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) has resulted in thousands of deaths in the world. Information about prediction model of prognosis of SARS-CoV-2 infection is scarce. We used machine learning for processing laboratory findings of 110 patients with SARS-CoV-2 pneumonia (including 51 non-survivors and 59 discharged patients). The maximum relevance minimum redundancy (mRMR) algorithm and the least absolute shrinkage and selection operator (LASSO) logistic regression model were used for selection of laboratory features. Seven laboratory features selected by machine learning were: prothrombin activity, urea, white blood cell, interleukin-2 receptor, indirect bilirubin, myoglobin, and fibrinogen degradation products. The signature constructed using the seven features had 98% [93%, 100%] sensitivity and 91% [84%, 99%] specificity in predicating outcome of SARS-CoV-2 pneumonia. Thus it is feasible to establish an accurate prediction model of outcome of SARS-CoV-2 pneumonia with machine learning.

Introduction

Most human coronavirus infections are mild. However, several betacoronaviruses can cause serious diseases or even death.1,2 The mortality rates of severe acute respiratory syndrome coronavirus (SARS-CoV) and Middle East respiratory syndrome coronavirus (MERS-CoV) were 10% and 37% respectively. SARS-CoV-2 is the pathogen for 2019 novel coronavirus disease (COVID-19),3,4 which has resulted in thousands of deaths in the world since the beginning of 2020.

The diagnosis of SARS-CoV-2 infection must be confirmed by the real-time reverse transcriptase polymerase chain-reaction (RT-PCR) or gene sequencing of specimens of patients.5,6 Chest radiograph and laboratory findings are both important for accessing the severity of the disease.7,8,9 Critical patients should be admitted to Intensive Care Unit (ICU) of infectious disease hospital, while mild patients could be kept and treated at isolation. It is very important to effectively prioritize resources for patients with the highest risk because of the large number of infected people.10

The laboratory findings can help to distinguish critical ill patients with SARS-CoV-2 infection from general patients. ICU patients and non-ICU patients differed significantly in some blood parameters, including: leukocytes, neutrophils, prothrombin time, D-dimer, total bilirubin, lactate dehydrogenase, high sensitivity cardiac troponin I and procalcitonin.11,5, 7 Ruan et al12 retrospectively analyzed laboratory findings of 68 nonsurvivors and 82 discharged patients, and found significant differences in lymphocytes, platelets, albumin, TB, urea nitrogen, creatinine, myoglobin, C-reactive protein and interleukin-6 between the two groups. These laboratory findings seemed useful in predicting outcome of SARS-CoV-2 infection. However, their accuracies need to be validated by other studies, and more laboratory tests need to be verified for their predication value. In addition, an advanced prediction model involving multiple laboratory parameters is urgently required to be applied in a clinical-decision support system to improve the predictive and prognostic accuracy.

As a branch of artificial intelligence, machine learning (ML) is helpful to establish accurate prediction model.13,14,15 However, there are few publications reporting predication of the outcome of SARS-CoV-2 pneumonia using ML methods based on laboratory findings. Thus we retrospectively collected laboratory findings of discharged patients and non-survivors. These data were dealt with a ML method similar to radiomics.16,17 We aim to establish a prediction model of outcome of SARS- CoV-2 pneumonia based on laboratory data.

Methods

All methods were carried out in accordance with relevant guidelines and regulations.

Study design and participants

This study was approved by the Ethics Commission of Hospital (TJ-2020-075). Written informed consent was waived by the Ethics Commission of hospital.

The author's center was the designated hospital for severe and critical SARS-CoV-2 pneumonia. Specialists from the top hospitals in the country gathered here and formed a consensus on the assessment and treatment for patients. Patients underwent repeated RT-PCR tests to confirm the presence of SARS-CoV-2. Laboratory tests for SARS-CoV-2 pneumonia included: blood routine test, serum biochemical test (including glucose, renal and liver function, creatine kinase, lactate dehydrogenase, and electrolytes), coagulation profile, cytokine test, markers of myocardial injury, infection-related makers, and other enzymes. Repeated tests were done every three to six days for monitoring the patient's condition.

Oxygen support (from nasal cannula to invasive mechanical ventilation) was administered to patients according to the severity of hypoxaemia. All patients were administered with empirical antibiotic treatment, and received antiviral therapy. Most of patients improved after regular treatment. Fitness for discharge was based on abatement of fever for at least 10 days, significantly improved respiratory function, and negative RT-PCR for SARS-CoV-2 twice in succession. However, the condition of a few critical patients continued to deteriorate and eventually died.

The presence of SARS-CoV-2 in respiratory specimens was detected by real-time RT-PCR methods. The primers and probe target to envelope gene of CoV were used and the sequences were as follows: forward primer 5'-TCAGAATGCCAATCTCCCCAAC-3'; reverse primer 5'- AAAGGTCCACCCGATACATTGA-3'; and the probe 5'CY5- CTAGTTACACTAGCCATCCTTACTGC-3'BHQ1. Conditions for the amplifications were 50°C for 15 min, 95°C for 3 min, followed by 45 cycles of 95°C for 15 s and 60°C for 30 s.

Data collection

58 fatal cases of SARS-CoV-2 pneumonia (39 male, median age 66 years) were collected by the electronic medical record system. 68 discharged patients with SARS-CoV-2 pneumonia whose age and gender matched the non-survivors were selected (46 male, median age 66 years). The admission date of these patients was from Feb 16, 2020 to Mar 20, 2020. We reviewed all laboratory findings for each patient.

Results of repeated tests were carefully compared to find the greatest deviation from normal value. In general, the greatest number in series of values was recorded. However, for platelets, red blood cell, lymphocytes, hemoglobin, calcium, total protein, albumin, estimated glomerular filtration rate (eGFR), and prothrombin activity (PTA), the minimum was recorded. These recorded laboratory findings were considered as features of a patient. A initial data set of features of 126 patients (non- survivor 58, discharge 68) was thus built.

There were 16 patients who did not have the entire group of laboratory features, thus their data were deleted from the dataset. The remaining data of 110 patients (51 non-survivor, 59 discharge) were analyzed by machine learning. Age and gender were added in the data set for statistical comparison purpose. Training cohort and validation cohort were randomly divided according to 8:2. Thus there were 88 patients in training cohort and 22 patients in validation cohort.

Statistical analysis and modeling for training cohort

First, all the laboratory features were compared between non-survivors and discharged patients using the Mann-Whitney U test for non-normally distributed features or the independent t-test for normally

distributed features. 16,17 Features with $p < 0.05$ were considered significant variables and selected. 16,17 Second, Spearman's correlation coefficient was used to compute the relevance and redundancy of the features. 16,17 Third, we applied the maximum relevance minimum redundancy (mRMR) algorithm to assess the relevance and redundancy of the features. 16,17 The features were ranked according to their mRMR scores. 16,17 Fourth, the top 15 features with high-relevance and low-redundancy were selected for least absolute shrinkage and selection operator (LASSO) logistic regression model. The LASSO logistic regression model with 5-fold cross-validation was adopted for further features selection and construction of signature. 16,17 Some candidate features coefficients were shrunk to zero and the remaining variables with non-zero coefficients were finally selected. 16,17 The features and corresponding coefficients were used for calculating signature for each patient. Mann-Whitney U test and receiver operator characteristic (ROC) analysis were used for comparing signature between two groups. 16,17

Statistical analysis for validation cohort

The model derived from training cohort was used for the validation cohort.16 The signature was calculated for each patient, and compared between non-survivors and discharged patients using a Mann-Whitney U test. ROC analysis was used to determine AUC, sensitivity and specificity.

The statistical analyses were performed using R software (version 3·3·4; https://www.r-project.org). 16,17 The following R packages were used: the "corrplot" package was used to calculate Spearman's correlation coefficient; the "mRMRe" package was used to implement the mRMR algorithm; the "glmnet" was used to perform the LASSO logistic regression model, and the "pROC" package was used to construct the ROC curve. 16,17

Results

Nine features were eliminated in the first step of feature selection because of non-significance. The remaining thirty-eight features were significantly different between two groups ($P < 0.05$), and then mRMR scores were obtained for them. There were seven features having non-zero coefficients after 5- fold cross-validation of LASSO algorithm, and were selected to construct a new signature. Table 1 shows the fifteen features with the highest mRMR scores. Figure 1 shows the correlation matrix

heatmap of the thirty-eight significant features. Figure 2 shows the feature selection process with LASSO algorithm. Figure 3 shows the contribution of the seven features to the new signature. Figure 4 and figure 5 show the signatures of all patients in training and validation cohort respectively.

For training cohort, non-survivors and discharged patients differed significantly in the constructed signature ($P<0·0001$). The AUC was 0·997 [95% CI: 0·99, 1·00]. The sensitivity and specificity in predicating outcome of SARS-CoV-2 pneumonia were 98% [93%, 100%] and 91% [84%, 99%] respectively.

For validation cohort, non-survivors and discharged patients differed significantly in the constructed signature ($P<0.0001$). The AUC was 0·992 [0·97, 0·99]. The sensitivity and specificity in predicating outcome of SARS-CoV-2 pneumonia were 100% [95%, 100%,] and 92% [76%, 100%] respectively.

The seven features included in the predication model were as follows: PTA, urea, white blood cell (WBC), interleukin-2 receptor (IL-2r), indirect bilirubin (IB), myoglobin, and fibrinogen degradation products (FgDP). All features had coefficients of positive number except PTA. PTA and FgDP are from coagulation profile. Urea and IB are from renal and liver function respectively. WBC is from blood routine test. Myoglobin is a marker of myocardial injury. IL-2r is related to immune response. These features and their coefficients were used for the calculation of signature, which could be positive or negative number.

Non-survivors (n=51) and discharged patients (n=59) did not differ in age or gender (median age 67 vs. 66, $P=0·75$; percentage of males, 66% vs. 64%, $P=0·66$). The comparisons of laboratory finding between non-survivors and discharged patients in training cohort (n=88) are shown in Table 2.

Blood routine test

WBC and neutrophils were significantly higher in non-survivor group versus discharge group. Lymphocyte, platelets and red blood cells were significantly lower in non-survivor group versus discharge group. AUC for them were 0·646~0·910.

Electrolyte

Potassium, chlorine and sodium were significantly higher in non-survivor group versus discharge group. Calcium was significantly lower in non-survivor group versus discharge group. AUC for them were 0·634~0·652

Serum biochemical test

Glucose and globulin were significantly higher in non-survivor group versus discharge group. Albumin and total protein were significantly lower in non-survivor group versus discharge group. AUC for them were 0·649~0·736.

Renal function

Urea and creatinine were significantly higher in non-survivor group versus discharge group. The eGFR was significantly lower in non-survivor group versus discharge group. AUC for them were 0·672~0·907.

Liver function

Total bilirubin, direct bilirubin, IB and glutamic oxaloacetic transaminase were significantly higher in non-survivor group versus discharge group. AUC for them were 0·647~0·806.

Coagulation profile

Prothrombin time, activated partial thromboplastin time, D-dimer, international normalized ratio (INR), fibrinogen and FgDP were significantly higher in non-survivor group versus discharge group. PTA was significantly lower in non-survivor group versus discharge group. AUC for them were 0·847~0·886.

Cytokine

IL-2r and IL-6 were significantly higher in non-survivor group versus discharge group. AUC for them were 0·689~0·909.

Infection-related markers and myocardial injury markers

Procalcitonin, high sensitive C-reactive protein, ferritin and N-terminal pro-brain natriuretic peptide (NT-proBNP) were significantly higher in non-survivor group versus discharge group. Myoglobin, MB

isoenzyme of creatine kinase and high sensitive cardiac troponin I were significantly higher in non-survivor group versus discharge group. AUC for them were 0·843~0·915.

Discussion

Non-survivors and discharged patients with SARS-CoV-2 pneumonia differed significantly in thirty- eight laboratory findings. By using machine learning method, we established a predication model involving seven laboratory features. The model was found highly accurate in distinguishing non- survivors from discharged patients. The seven features selected by artificial intelligence also indicated that dysfunction of multiple organs or systems correlated with the prognosis of SARS-CoV-2 pneumonia.

The SARS-CoV-2 spreads and invades through respiratory mucosa, triggers a series of immune responses and induces cytokine storm in vivo, resulting in changes in immune components.18,5 When immune

response is dysregulated, it will result in an excessive inflammation, even cause death.19,7 We found leukocyte and neutrophils count were significantly higher in non-survivors than in survivors.

Excessive neutrophils may contribute to acute lung damage, and are associated with fatality.20 The absolute value of lymphocytes was reduced in SARS-CoV-2 non-survivors, suggesting depletion of lymphocytes caused by strong innate inflammatory immune response. Higher serum levels of pro-inflammatory cytokines (IL-2r and IL-6) and C-reactive protein were found in non-survivors, also indicating excessive immune response. In addition, high leukocyte count in SARS-CoV-2 patients may be also due to secondary bacterial infection.21,5 Elevated procalcitonin was seen in fatal cases, representing more prominent inflammation.22 All these laboratory parameters mentioned above may be associated with prognosis of SARS-CoV-2 pneumonia.

Lung lesions have been considered as the major damage caused by SARS-CoV-2 infection. Severe cases may develop acute respiratory distress syndrome (ARDS) and respiratory failure. However, liver injury has also been reported to occur during the course of the disease,23,24 and is associated with the severity of diseases. Abnormal transaminase levels accompanied by decreased serum albumin and increased serum bilirubin levels were observed in fatal cases. The levels of liver function associated markers were significantly higher in non-survivors compared to survivors. Acute kidney injury could

have been related to direct effects of the virus, hypoxia, or shock.25,26 Blood urea, and creatinine levels continued to increase, until death occurred. Non-survivors had lower eGFR and higher blood urea compared to survivors. Myocardial injury was seen in non-survivors, which was suggested by elevated level of myoglobin, high sensitive cardiac troponin I, or MB isoenzyme of creatine kinase. The pathologic mechanisms of multiple organ dysfunction or failure may be associated with the death of patients with SARS-CoV-2 pneumonia. Some patients with SARS-CoV-2 infection progressed rapidly with sepsis shock, which is well established as one of the most common causes of disseminated intravascular coagulation (DIC).27 Conventional coagulation parameters during course may be also associated with prognosis of SARS-CoV-2 pneumonia. The non-survivors in our cohort revealed significantly longer prothrombin time and APTT compared to survivors. At the late stages of SARS- CoV-2 infection, levels of fibrin-related markers (D-dimer and FgDP) markedly elevated in most cases, suggesting a secondary hyperfibrinolysis condition in these patients.

A number of laboratory features were compared between non-survivors and discharged patients with SARS-CoV-2 pneumonia. The two groups differed significantly in as many as thirty-eight features.

However, none of the futures provided adequate accuracy in predicating the outcome of SARS-CoV-2 pneumonia. Thus, a novel accurate predication model involving multiple features was established in the study. With machine learning methods previously used in radiomics, a predication model combining seven out of the thirty-eight laboratory features was highly accurate in predicating the outcome of SARS-CoV-2 pneumonia, for either training cohort or validation cohort.

The mRMR algorithm was used for assessing significant features to avoid redundancy between features. The features were ranked according to their relevance-redundancy scores. The mRMR score of a feature is defined as the mutual information between the status of the patients and this feature minus the average mutual information of previously selected features and this feature.28,29,17 The top fifteen features with high mRMR scores were selected for the next step of modeling. The least absolute shrinkage and selection operator logistic regression model was used to processing the features selected by mRMR algorithm. LASSO is actually a regression analysis method that improves the mode prediction accuracy and interpretability.30 Some candidate features coefficients were shrunk to zero and the remaining variables with non-zero coefficients were selected. After using LASSO, new signature could be calculated with selected features and their coefficients. The signature used for predication of

outcome can be positive or negative number, corresponding with poor and good prognosis respectively.

Our results showed that the signature provides excellent efficiency for discriminating survivor from

non-survivor. The sensitivity and specificity were both excellent. The AUC of the signature was 10~40% higher than AUC of a single laboratory feature.

As this predication model was established by artificial intelligence, all we did was to match the age and gender of discharged patients and non-survivors before providing laboratory findings to computer. Although the modeling process is a black box to us, the choice of features seems reasonable. PTA can more accurately reflect the coagulation function compared to prothrombin time, and can also reflect the degree of liver injury. Urea is a good index to reflect the degree of renal function damage. WBC can not only reflect immune status, but also be used to evaluate secondary infection. IL-2r is an indicator of inflammation and immune response.20 IB is related to both liver function and possible hemolysis.

Myoglobin reflects the degree of myocardial injury. The increase of FgDP is related to coagulation disorders including DIC. Thus the current model involves multiple important systems closely related to the prognosis. Based on the high accuracy of the prediction model, it seems that we can deduce the following conclusions: liver, kidney, myocardial damage, coagulation disorder and excess immune response all contribute to the outcome of SARS-CoV-2 pneumonia.

One limitation of this model is that it did not cover all laboratory tests. Some important laboratory tests, such as lymphocyte, albumin or creatinine, were not included. Fortunately, there are moderate to high correlations between the unselected and selected features, which is confirmed by our statistical analysis. Furthermore, models involving too many features are not easy for clinicians to use. Another limitation of the model is that it did not involve clinical variables, because we focused on maximizing the predication value of objective laboratory variables.

Our study has some limitations. First, this is a single-center retrospective study with relatively small sample size. There were only 88 patients in training cohort and 22 patients in validation cohort. Multi-center large-sample studies are required to validate our predication model. Second, due to the difference

of instrument among centers, the same patient may have different values for the same laboratory test in different hospitals. Our model based on the laboratory data from the author's center may not be directly used in other centers. However, they could easily establish a predication model using their own data with machine learning method. Third, age and gender were matched for

discharged patients and non-survivors in the current study. It is well established age and gender influence the results of laboratory tests. Because we eliminated the interference of age and gender, the difference of laboratory feature was caused by the disease severity. This study focused on the real predictive value of laboratory tests and aimed to improve prediction accuracy by combining multiple laboratory findings. However, a more complex model combining laboratory features and clinical variables should be constructed in future study. Fourth, it is difficult for general clinicians to understand the method of artificial intelligence. With more and more artificial intelligence used in medical diagnosis, this prediction model will be paid more attention to.

In conclusion, it is feasible to establish a accurate prediction model of outcome of SARS-CoV-2 pneumonia using machine learning method. Injury of liver, kidney and myocardium, coagulation disorder and excess immune response all correlate with the outcome of SARS-CoV-2 pneumonia.

Declarations

Data availability

After publication, the data will be made available to others on reasonable requests to the corresponding author.

Acknowledgments

We thank all patients and their families involved in the study.

Author contributions

GW, SZ and YW collected the epidemiological and clinical data. GW, SZ and YW summarised all data. GW, XL drafted the manuscript. SZ and XL revised the final manuscript.

Competing interests

We declare no competing interests.

Materials & Correspondence

After publication, the data will be made available to others on reasonable requests to the corresponding author.

References

1. Drosten, , et al. Identification of a novel coronavirus in patients with severe acuterespiratory syndrome. *N Engl J Med* **348,** 1967-1976 (2003).
2. Zaki, M. , et al. Isolation of a novel coronavirus from a man with pneumonia in Saudi Arabia. *N Engl J Med* **367,** 1814-1820 (2012).
3. Phelan, L., et, al. The Novel Coronavirus Originating in Wuhan, China: Challenges for Global Health Governance. *JAMA* (2020); published online Jan 30. doi: 10.1001/jama.2020.1097.
4. Li, , et al. Early Transmission Dynamics in Wuhan, China, of Novel Coronavirus-Infected Pneumonia. *N Engl J Med* (2020); published online Jan 29. doi: 10.1056/NEJMoa2001316.
5. Huang, , et al. Clinical features of patients infected with 2019 novel coronavirus in Wuhan, China. *Lancet* 2020; **395**: 497-506 ().
6. Zhu, N., et al. A Novel Coronavirus from Patients with Pneumonia in China, 2019. *N Engl J Med* **382,**727-733 (2020).
7. Wang, , et al. Clinical Characteristics of 138 Hospitalized Patients With 2019 Novel Coronavirus-Infected Pneumonia in Wuhan, China. *JAMA* (2020); published online Feb 7. doi: 10.1001/jama.2020.1585.
8. Bernheim, A., et al. Chest CT Findings in Coronavirus Disease-19 (COVID-19): Relationship to Duration of *Radiology* (2020); published online Feb 20. doi: 10.1148/radiol.2020200463.
9. Fang, , et al. Sensitivity of Chest CT for COVID-19:Comparison to RT-PCR. *Radiology* (2020); published online Feb 19. doi: 10.1148/radiol.2020200432.
10. General Office of the National Health Commission of Diagnosis and Treatment Protocol for 2019-nCoV. 5th ed. Beijing, China: National Health Commission of China; 2020.
11. Yang, , et al. Clinical course and outcomes of critically ill patients with SARS-CoV-2 pneumonia in Wuhan, China: a single-centered, retrospective, observational study. *Lancet Respir Med* (2020); published online Feb 24. doi: 10.1016/S2213-2600(20)30079-5.
12. Ruan, , et al. Clinical predictors of mortality due to COVID-19 based on an analysis of data of 150 patients from Wuhan, China. *Intensive Care Med* (2020); published online Mar 3. doi: 10.1007/s00134-020-05991-x.
13. Shiri, I., et al. Next-Generation Radiogenomics Sequencing for Prediction of EGFR and KRAS Mutation Status in NSCLC Patients Using Multimodal Imaging and Machine Learning *Mol Imaging Biol* (2020); published online Mar 17. doi: 10.1007/s11307-020-01487-8.
14. Matsuzaka, , et al. Prediction Model of Aryl Hydrocarbon Receptor Activation by a Novel QSAR Approach, DeepSnap-Deep Learning. *Molecules* (2020); published online Mar 13. doi: 10.3390/molecules25061317.
15. Katić, K., et al. Machine learning algorithms applied to a prediction of personal overall thermal comfort using skin temperatures and occupants' heating *Appl Ergon* (2020); published online Feb 19. doi: 10.1016/j.apergo.2020.103078.

16. Jiang, M., et al. Nomogram Based on Shear-Wave Elastography Radiomics Can Improve Preoperative Cervical Lymph Node Staging for Papillary Thyroid *Thyroid* (2020); published online Mar 11. doi: 10.1089/thy.2019.0780.
17. Zhang, , et al. T2-Weighted Image-Based Radiomics Signature for Discriminating Between Seminomas and Nonseminoma. *Front Oncol* (2019); published online Nov 28. doi: 10.3389/fonc.2019.01330.
18. Qin, C., et al. Dysregulation of immune response in patients with COVID-19 in Wuhan, *Clin Infect Dis* (2020); published online Mar 12. doi: 10.1093/cid/ciaa248.
19. Mahallawi, , et al. MERS-CoV infection in humans is associated with a pro-inflammatory Th1 and Th17 cytokine profile. *Cytokine* **104,** 8-13 (2018).
20. Channappanavar, , et al. Pathogenic human coronavirus infections: causes and consequences of cytokine storm and immunopathology. *Semin Immunopathol* **39,** 529-539 (2017).
21. Chen, , et al. Epidemiological and clinical characteristics of 99 cases of 2019 novel coronavirus pneumonia in Wuhan, China: a descriptive study. *Lancet* **395,** 507-513 (2020).
22. Guan, W., et al. Clinical characteristics of 2019 novel coronavirus infection in China. *N Engl J Med* (2020); published online Feb 28. doi: 10.1056/NEJMoa2002032.
23. Tang, , et al. Abnormal coagulation parameters are associated with poor prognosis in patients with novel coronavirus pneumonia. *J Thromb Haemost* (2020); published online Feb 19. doi: 10.1111/jth.14768.
24. Xu, , et al. Liver injury during highly pathogenic human coronavirus infections. *Liver Int* (2020); published online Mar 14. doi: 10.1111/liv.14435.
25. Estenssoro, , et al. Pandemic 2009 influenza A in Argentina: a study of 337 patients on mechanical ventilation. *Am J Respir Crit Care Med* **182,** 41-48 (2010).
26. Li, , et al. The Clinical and Chest CT Features Associated with Severe and Critical COVID-19 Pneumonia. *Invest Radiol* (2020); published online Feb 29. doi: 10.1097/RLI.0000000000000672.
27. Abe, , et al. Complement Activation in Human Sepsis is Related to Sepsis-Induced Disseminated Intravascular Coagulation. *Shock* (2020); published online Jan 7. doi: 10.1097/SHK.0000000000001504.
28. Lin, X., et al. A new feature selection method based on symmetrical uncertainty and interaction gain. *Comput Biol Chem* (2019); published online Nov doi: 10.1016/j.compbiolchem.2019.107149.
29. Wang, J., et al. Machine learning-based analysis of MR radiomics can help to improve the diagnostic performance of PI-RADS v2 in clinically relevant prostate *Eur Radiol* **27,** 4082- 4090 (2017).
30. Sauerbrei, , et al. Selection of important variables and determination of functional form for continuous predictors in multivariable model building. *Stat Med* **26,** 5512-5528 (2007).

Tables

Table 1 The fifteen features with higher mRMR scores were selected for the step of LASSO logistic regression. Some candidate features coefficients were shrunk to zero and the remaining variables with non-zero coefficients were selected.

Rank	Features	mRMR score	Coefficient after LASSO
1	PTA	0.353104608838358	-0.414862837231762
2	WBC	0.192111443157334	0.321436744864511
3	urea	0.186791914803107	0.395488285450932
4	IL-2r	0.177391665151341	0.229719312940612
5	IB	0.12492445282599	0.0951242440489642
6	myoglobin	0.111913747881061	0.0526559045119329
7	TB	0.110049432665794	0
8	FgDP	0.107334189436362	0.0243316382092674
9	hs-CRP	0.102543830552162	0
10	Ferritin	0.0952418221441096	0
11	LDH	0.0870072400036755	0
12	D-dimer	0.0860137252146133	0
13	eGFR	0.0820793884917148	0
14	Neutrophils	0.0638615988145559	0
15	sodium	0.0626942491753672	0

mRMR= maximum relevance minimum redundancy; LASSO=least absolute shrinkage and selection operator; PTA=prothrombin activity; WBC=white blood cell; IL-2r=interleukin-2 receptor; IB=indirect bilirubin; TB=total bilirubin; FgDP=fibrinogen degradation products; hs-CRP=hypersensitive C- reactive protein; LDH=lactate dehydrogenase; eGFR=estimated glomerular filtration rate

Table 2 Medians [inter-quartile range] of laboratory findings of patients with SARS-CoV-2 pneumonia were provided in the table. Features were compared between non-survivors and discharged patients using the Mann-Whitney U test for non-normally distributed features or the independent t-test for normally distributed features.

	Non-survivors	Discharged patients	P
Leucocyte (109/L)	11.64 [9.37, 15.61]	5.22 [4.1, 8.79]	<0.0001
Platelet (109/L)	118 [63, 179]	144.5 [113.5, 227.75]	0.004
Erythrocyte (1012/L)	3.48 [2.71, 3.89]	3.76 [3.59, 4.17]	0.001
Neutrophils (109/L)	9.44 [7.36, 12.71]	4.8 [2.45, 7.37]	<0.0001
Lymphocyte (109/L)	0.5 [0.32, 0.74]	0.70 [0.47, 0.93]	<0.0001
Hemoglobin (g/L)	115.5 [91, 127]	120 [112.5, 129]	0.16
Potassium (mmol/L)	4.39 [4.11, 5.19]	4.11 [3.70, 4.76]	0.032
Calcium (mmol/L)	2.01 [1.94, 2.10]	2.06 [2.00, 2.13]	0.024
Chlorine (mmol/L)	99.4 [97.2, 105.3]	98 [95.83, 99.4]	0.041
Sodium (mmol/L)	139.55 [135.2, 145.1]	135.55 [133.4, 137.85]	0.017
Glucose (mmol/L)	9.11 [7.62, 13.66]	7.44 [6.46, 9.48]	0.019
Total protein (g/L)	59.9 [56.8, 65.4]	63.9 [59.8, 67.9]	0.030
Globulin (g/L)	37.9 [33.3, 40.7]	32.2 [29.25, 33.85]	<0.0001
Albumin (g/L)	28 [25.08, 30.95]	32.3 [28.2, 37.2]	0.009
Creatinine (μmol/L)	86 [66, 179.5]	72 [59, 103.5]	0.008
Uric acid (μmol/L)	190.5 [114.5, 309]	188 [148, 362.4]	0.54
Total bilirubin (μmol/L)	17.4 [11.5, 20.4]	11.4 [8.6, 13.4]	<0.0001
Direct bilirubin (μmol/L)	9.75 [6.4, 14.4]	6.8 [5.35, 9.93]	<0.0001
Indirect bilirubin (μmol/L)	9.15 [5.45, 11]	7.6 [5.8, 9.35]	0.043
Urea (mmol/L)	14.55 [10, 20.08]	7.85 [6.98, 10.17]	<0.0001
Estimated glomerular filtration rate (ml/min/1.73m2)	69.3 [41.35, 89.4]	82 [73.4,88.6]	0.008
Glutamic oxaloacetic transaminase (U/L)	43 [24, 104]	37 [21, 57]	0.044
Glutamic-pyruvic transaminase (U/L)	39 [17, 77.25]	38 [24.25, 61.75]	0.70
Myoglobin (ng/mL)	280.6 [152.15, 736.8]	67 [24.45, 129.55]	0.002
High sensitive cardiac troponin I (pg/mL)	202.2 [68.95, 460.18]	30.55 [14.5, 47.6]	<0.0001
MB isoenzyme of creatine kinase (ng/mL)	5.85 [2.63, 11.93]	1.2 [0.7, 1.9]	0.014

Lactate dehydrogenase (U/L)	490 [358.5, 591]	306.5 [281.5, 368.25]	<0.0001
Glutamate dehydrogenase (U/L)	16.6 [9, 44]	13.6 [8.23, 25.13]	0.18
Creatine kinase (U/L)	180 [43, 503]	178 [84, 230.5]	0.24
Prothrombin time (second)	16.5 [15.3, 19.4]	14.3 [13.25, 15.83]	<0.0001
Fibrinogen (g/L)	5.92 [4.82, 6.3]	5.22 [4.59, 5.78]	0.039
Activated partial thromboplastin time (second)	46.4 [42.5, 56]	43.5 [40.65, 46.8]	0.0021
Thrombin time (second)	17.5 [15.7, 20.6]	15.95 [15.13, 17.68]	0.026
D-D dimer (μg/mL)	5.47 [2.73, 12.52]	2.22 [1.82, 2.92]	<0.0001
Prothrombin activity	62% [55%, 75%]	79% [64%, 91%]	<0.0001
International standardized ratio	1.3 [1.22, 1.56]	1.08 [0.99, 1.21]	<0.0001
Fibrinogen degradation products (μg/mL)	29.65 [17.13, 62.65]	6.9 [3.9, 11.5]	<0.0001
Procalcitonin (ng/mL)	0.97 [0.27, 2.58]	0.16 [0.11, 0.21]	<0.0001
N-terminal pro-brain natriuretic peptide (pg/mL)	3375.5 [1491.75, 8102.75]	963 [522, 1483.5]	<0.0001
Ferritin (μg/L)	1064.5 [814.25, 2658.5]	826.8 [616.75, 1481.5]	0.037
Hypersensitive C-reactive protein (mg/L)	142.7 [80.9, 209]	43.1 [22.3, 127.1]	<0.0001
Interleukin-1β (pg/mL)	5.95 [3.65, 10.28]	5 [3.5, 9.5]	0.17
Interleukin-2 receptor (U/mL)	1280.5 [1059.25, 1486.25]	482 [238, 901]	<0.0001
Interleukin-6 (pg/mL)	157.3 [51.62, 227.3]	75.99 [42.96, 148.80]	<0.0001
Interleukin-10 (pg/mL)	11.6 [10.25, 15.6]	9.9 [5.3, 14.4]	0.056

SARS-CoV-2=severe acute respiratory syndrome coronavirus 2.

Figures

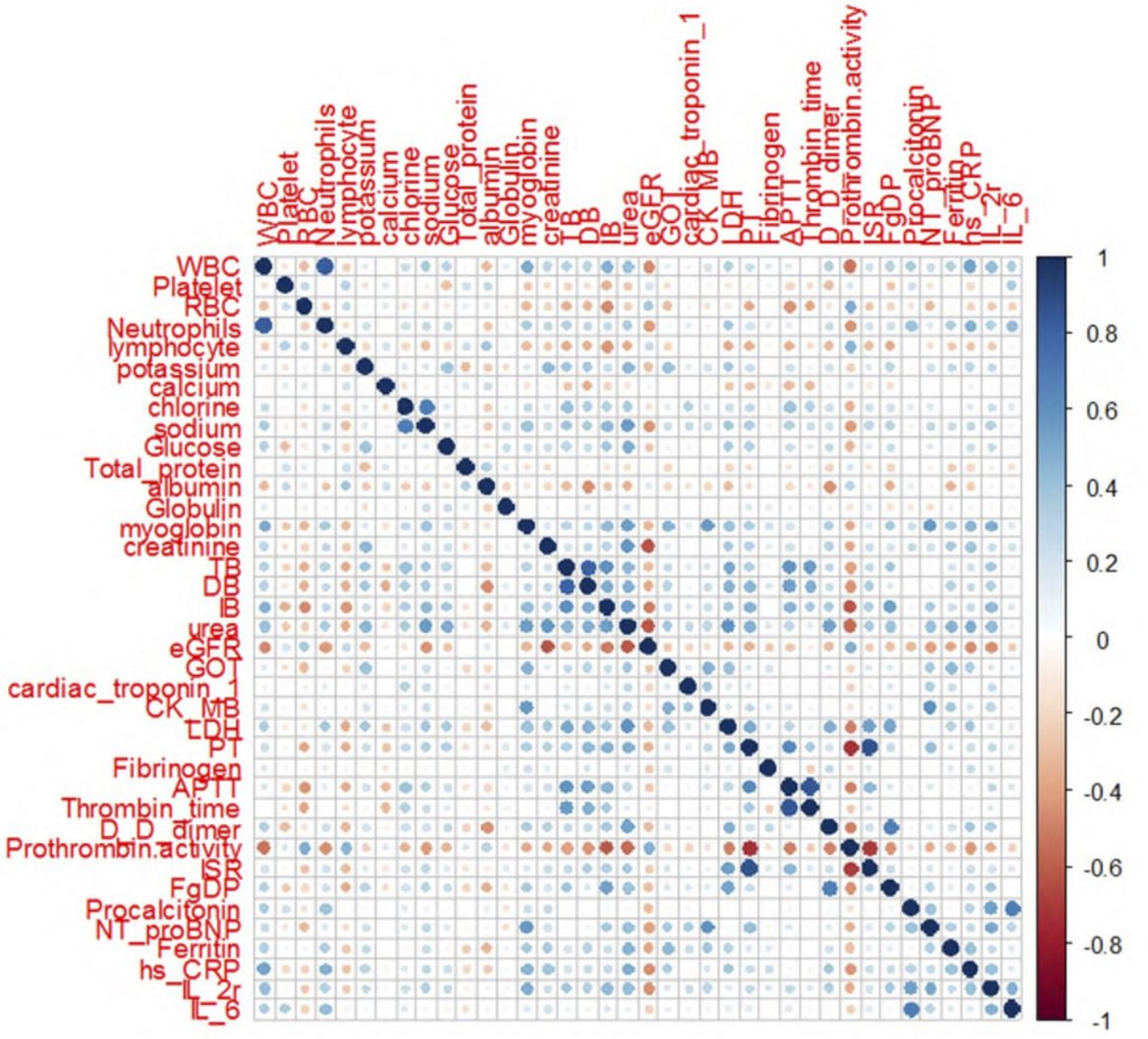

Figure 1

Correlation matrix heatmap of 38 significant features. Spearman's correlation coefficient was used to compute the relevance and redundancy of the features.

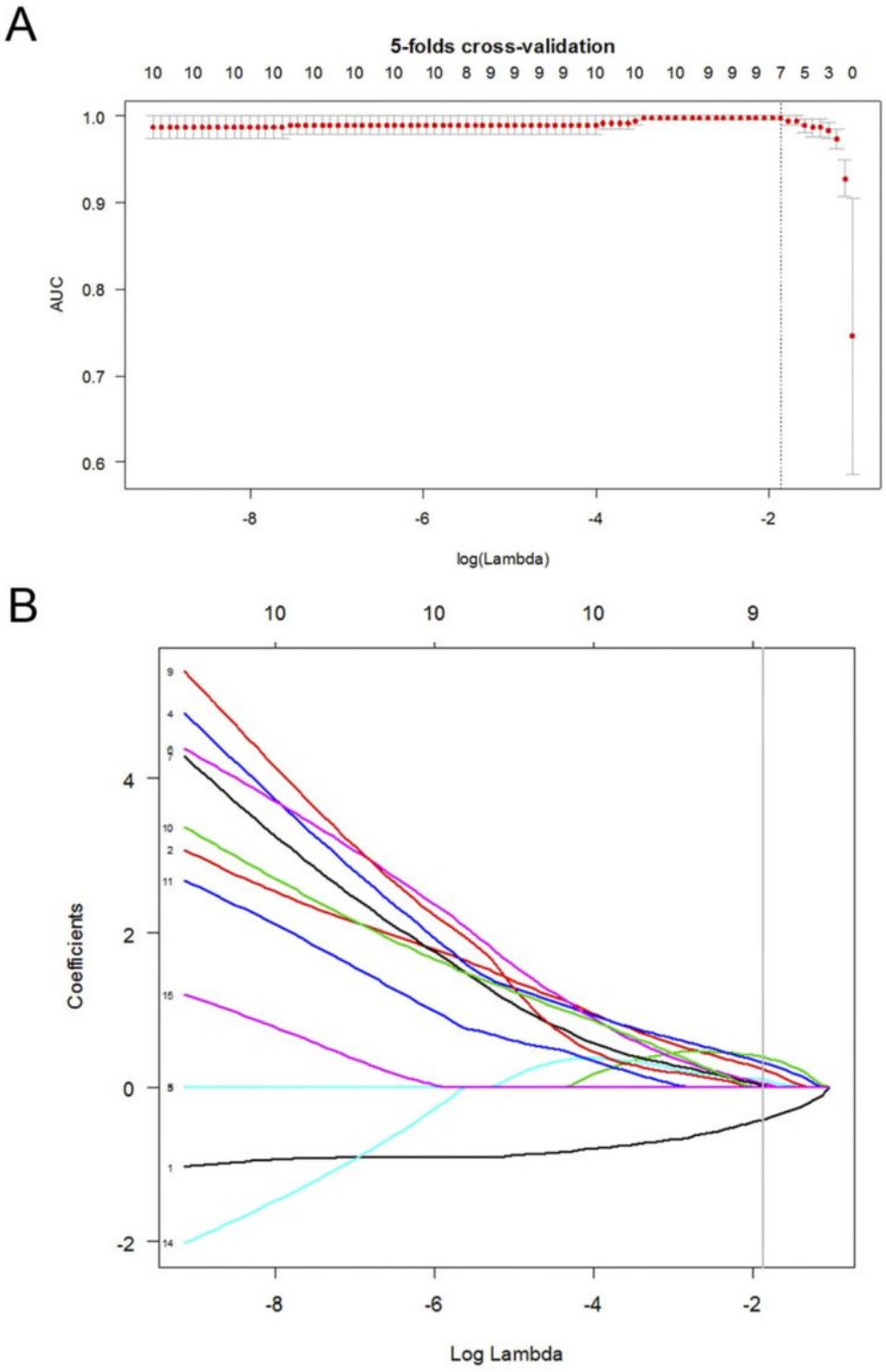

Figure 2

The 5-fold cross-validation (A) of the least absolute shrinkage and selection operator algorithm for feature selection process. A vertical line was drawn at the optimal value. Some candidate features coefficients were shrunk to zero (B) and the remaining seven variables with non-zero coefficients were finally selected to construct the signature.

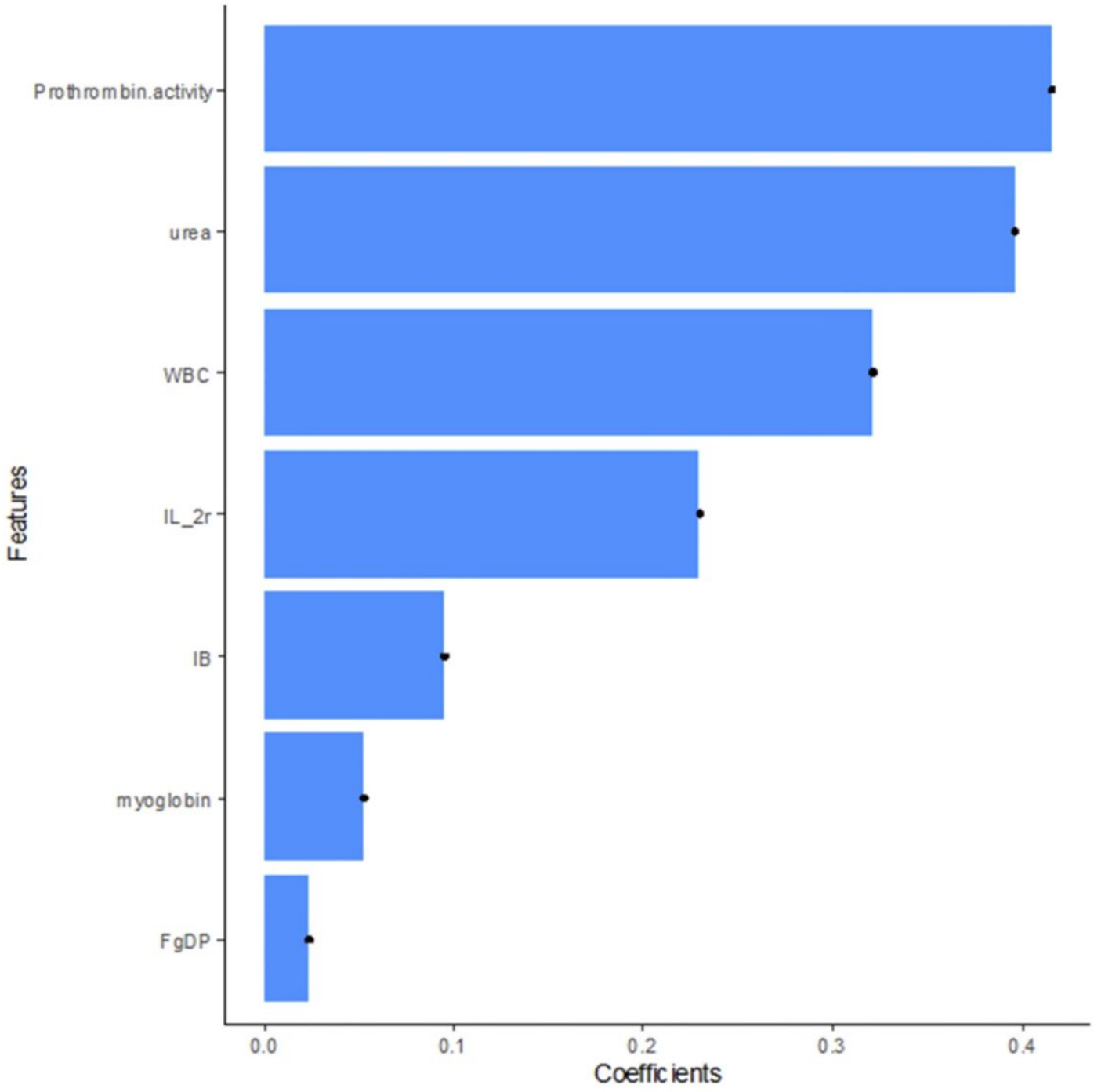

Figure 3

Contribution of the features to the signature. The histogram shows the contribution of the seven features with non-zero coefficients to the signature. The features are plotted on the y-axis, and their coefficients are plotted on the x-axis. The features and corresponding coefficients were used for calculation of signature. All features had coefficients of positive number except prothrombin activity.

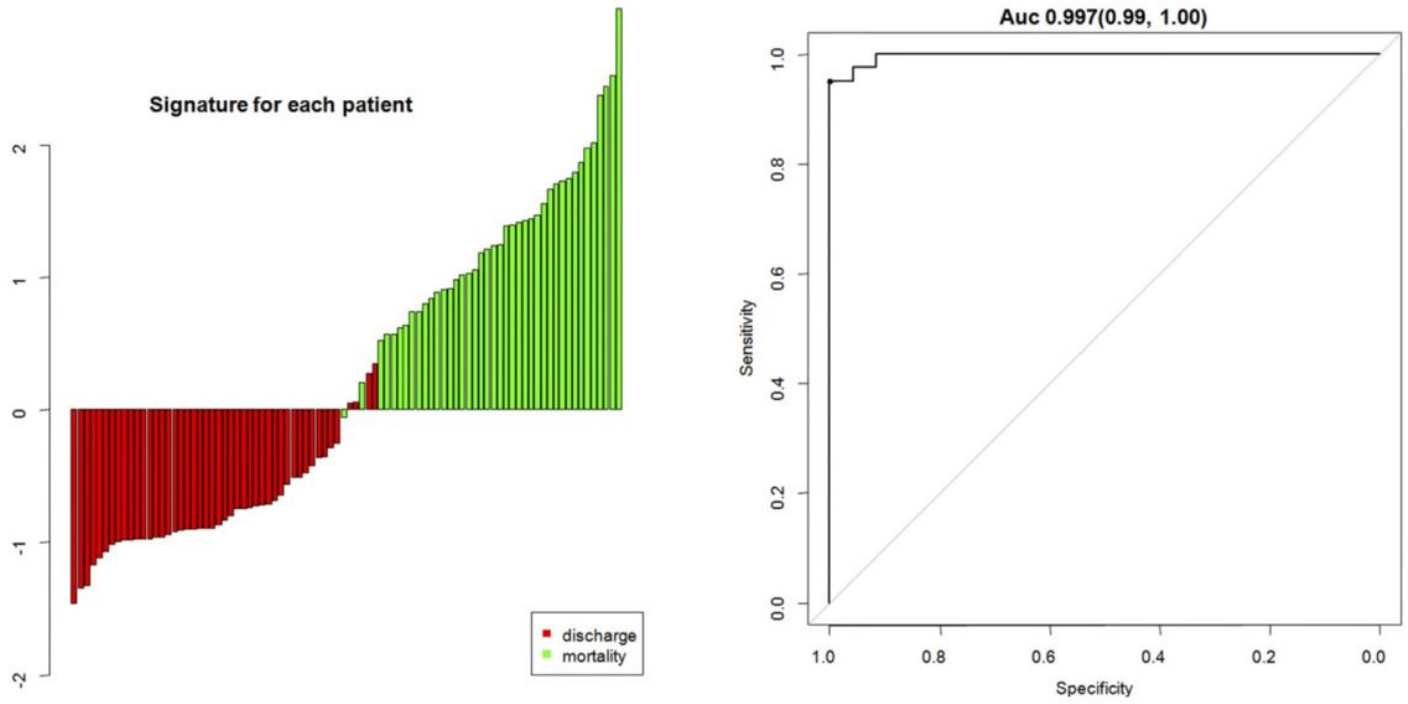

Figure 4

Bar charts of the signature for patients in training cohort. The red bars indicate the signatures of discharged patients, while the light green bars indicate the signatures of non-survivors. The signature AUC was 0.997 in the ROC analysis.

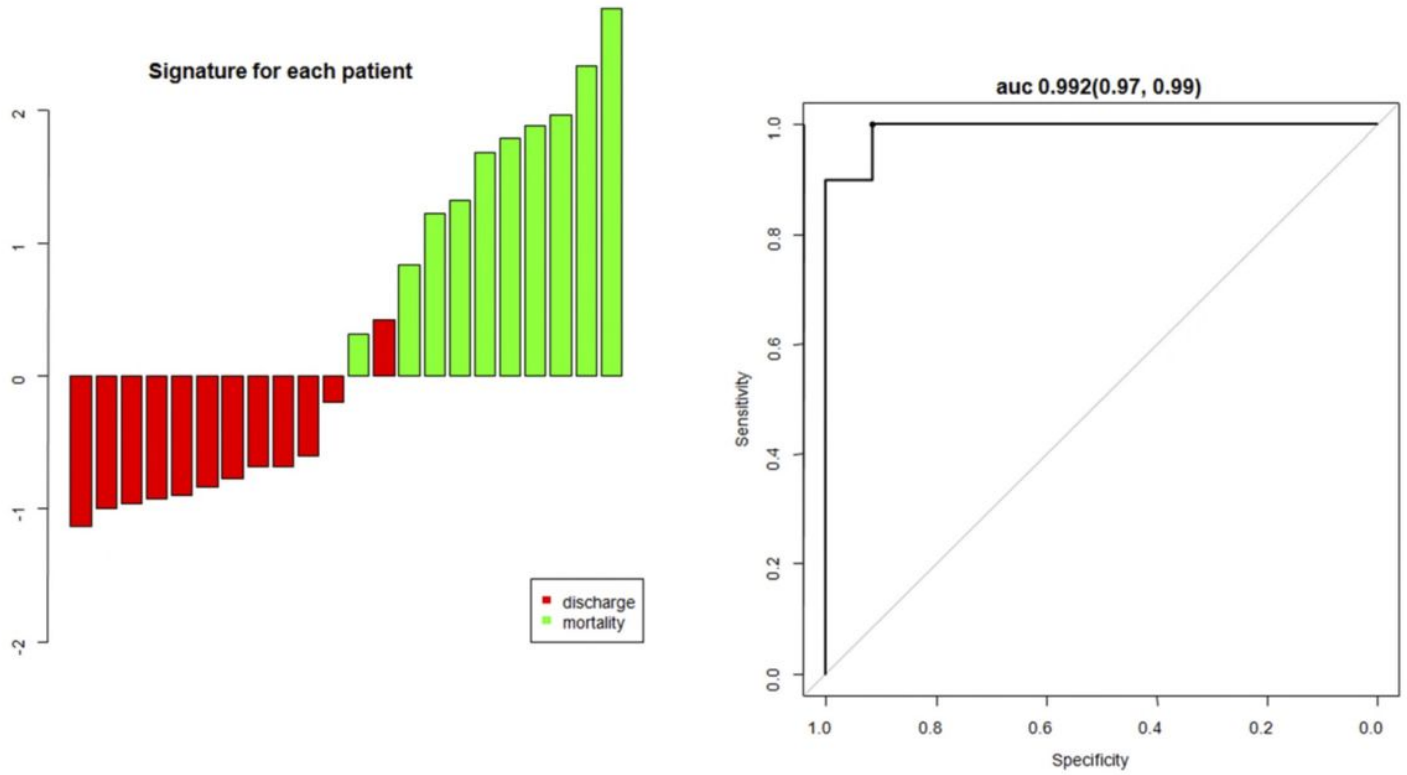

Figure 5

Bar charts of the signature for patients in validation cohort. The model derived from training cohort were used for calculating signatures. The red bars indicate the signatures of discharged patients, while the light green bars indicate the signatures of non-survivors. The signature AUC was 0.992.