Construction and Validation of a Machine Learning-based Nomogram: A TOOL to Predict the Risk of Getting Severe Corona Virus Disease 2019 (COVID-19)

Xinyi Zheng
Huashan Hospital
Zhixian Yao
Shanghai General Hospital
Zhong Zheng
Shanghai General Hospital
Ke Wu (doctorwuke@sjtu.edu.cn)
Shanghai Jiao Tong University
Junhua Zheng
Shanghai General Hospital

Research

Keywords: COVID-19, Machine Learning, Nomogram, Severe COVID-19 prediction

Posted Date: June 18th, 2020

DOI: https://doi.org/10.21203/rs.3.rs-35149/v1

Abstract

Background

Identifying patients who may develop severe coronavirus disease 2019 (COVID-19) will facilitate personalized treatment and optimize the distribution of medical resources.

Methods

In this study, 590 COVID-19 patients during hospitalization were enrolled (Training set: n = 285; Internal validation set: n = 127; Prospective set: n = 178). After filtered by 2 machine learning methods in the training set, 5 out of 31 clinical features were selected into model building to predict the risk of developing severe COVID-19 disease. Multivariate logistic regression was applied to build the prediction nomogram and validated in 2 different sets. Receiver operating characteristic (ROC) analysis and decision curve analysis (DCA) were used to evaluate its performance.

Results

From 31 potential predictors in the training set, 5 independent predictive factors were identified and included in the risk score: C-reactive protein (CRP), Lactate dehydrogenase (LDH), Age, Charlson/Deyo comorbidity score (CDCS) and Erythrocyte sedimentation rate (ESR). Subsequently, we generated the nomogram based on the above features for predicting severe COVID-19. In the training cohort, the Area under curves (AUCs) were 0.822 (95% CI 0.765–0.875) and the internal validation cohort was 0.762 (95% CI 0.768–0.844). Further, we validated it in a prospective cohort with the AUCs of 0.705 (95% CI 0.627–0.778). The internally bootstrapped calibration curve showed favorable consistency between prediction by nomogram and actual situation. And DCA analysis also conferred high clinical net benefit.

Conclusion

In this study, our predicting model based on 5 clinical characteristics of COVID-19 patients will enable clinicians to predict the potential risk of developing critical illness and thus optimize medical management.

Background

Severe Corona Virus Disease 2019 (COVID-19) outbreaks worldwide during early December 2019. As of May 5th 2020, the number of cumulative cases has surpassed 3500,000 with over 240,000 deaths worldwide. So far, the global epidemic situation is still very serious. Coronavirus has specific immune response and immune escape characteristics, and then causes severe pathogenic mechanisms through inflammation, which leaded to severe pneumonia, pulmonary edema, ARDS, or multiple organ failure and

even died[1]. Therefore, rapid and accurate prediction of COVID-19 pneumonia trends can provide effective treatment.

As research progresses, more and more information about COVID-19 pneumonia has been revealed. Lauer et al. reported that under conservative assumptions, the estimated median incubation period of COVID-19 is 5.1 days, and 101 of every 10,000 cases would develop symptoms after 14-day active monitoring or isolation [2]. As reported by Huang et al, patients with COVID-19 present primarily with fever, dry cough, and myalgia or fatigue. Although prognosis of most patients are thought to be favorable, older people and those with weakened immunity may have worse outcomes, even dead[3]. One week after onset of the disease is a critical period, patients with severe illness may develop dyspnea and hypoxemia within, which may quickly progress to acute respiratory distress syndrome(ARDS) or end-organ failure[4].

Therefore, to identify high risk patients whose disease may likely progress is of great importance both in delivering personalized medical care and optimizing medical resource distribution on the macro level. Gong et al, provided a nomogram to help clinicians to early identify patients who will exacerbate to severe COVID-19 but they didn't take clinical factors like underlying comorbidities in consideration which was a universally acknowledged risk factor[5]. Dong et al, used CALL score model to estimate the progressive risk of COVID-19 patients but the sample size was limited, which may cause the volatility of the result, for example the hazard ratio of LDH > 500 is 9.8 (2.8–33.8)[6].

As a systemic disease, it is necessary to take multiple indicators into account. Comorbidities, age, biochemical Indicators: lactate dehydrogenase (LDH), C-reactive protein (CRP) and blood urea nitrogen (BUN) et al and blood indicators are all potential influencing factors. In this study, we filtered out 5 effective indexes among 31 items and established an effective 5-feature based nomogram by machine learning methods. More importantly, to ensure the prediction accuracy, we then further verify this system in a prospective cohort. This meddle could help clinicians to predict the progression of COVID-19 and provide better centralized management.

Method

Study Population

We retrospectively collected 412 patients from January 1st to February 6th 2020 in Jinyintan Hospital of Wuhan City who were centrally treated and diagnosed with Common or Severe type of COVID-19. For extra validation, 178 patients were prospective recruited from February 6th 2020 to March 10th. This study was approved by the Ethics Review Committee of Wuhan Jinyintan Hospital and Shanghai General Hospital.

Diagnostic Criteria

According to the "New Coronavirus Pneumonia Diagnosis and Treatment Program (Trial Version 6)" promulgated by the General Office of the National Health Commission [7] Clinical classification: (1) Light type: mild clinical symptoms, no pneumonia manifestations in imaging; (2) Common type: fever, respiratory tract and other symptoms, pneumonia manifestations can be seen in imaging; (3) Severe Type: meet any of the following: ① Respiratory distress, RR ≥ 30 times / min; ②In resting state, it means oxygen saturation ≤ 93%; ③Arterial blood oxygen partial pressure (PaO2) / oxygen concentration (FiO2) ≤ 300 mmHg; (4) Critical Type: Those meet with one of the following conditions: ① have respiratory failure and need mechanical ventilation; ② have shock; ③ combined with other organ failure which require ICU monitoring treatment. More details are provided in Supplement material.

Data Collection

We collect the hospitalization history of all the subjects and analyze their clinical data, including gender, age, date of onset, time of first diagnosis, time of diagnosis, time of admission, time of discharge, occupation, history of exposure, underlying disease, first symptoms, body signs, laboratory tests, imaging data and treatment status, etc. Two radiologists were assigned to read the chest radiographs and CT of the selected patients, recorded the type of lung lesions and the distribution characteristics of the lung lobes at the time of onset with reached consensus. The New Coronavirus Infection Pneumonia Diagnosis and Treatment Program (Trial Version 6) conducts epidemiological investigations, including whether there is a history of travel or residence in Hubei province and its surrounding areas, whether it has contact with diagnosed patients, and whether there are clustered diseases.

Study Design and Data Processing

Since most of the Light type of COVID-19 victims do not need medical support or hospitalization and Critical type patients were limited in the hospital, we only analyzed the Common and Severe type which occupied most of the medical system and equipment.

For the research design, we incorporated three sections to identify and validate clinical signature-based nomogram to predict whether a Common COVID-19 patient will progress to the Severe type. Initially, we collected 439 patients and we filtered 2 common type, 6 critical type and 24 with incomplete medical information. The subsequent 412 patients were divided in to the training set (n = 285) and internal validation set (n = 127) by random seed with a 7:3 ratio.

For continuous variables, the Maximally Selected Rank Statistics (MSRS) was used to generate the optimal cutoff value and all variables are transformed into dichotomous data[8]. After primary filtration, a Least Absolute Shrinkage and Selector Operation (LASSO) algorithm[9] with penalty parameter tuning conducted by 10-fold cross-validation, was built to select candidate features. Simultaneously, another algorithm, Support Vector Machine-Recursive Feature Elimination (SVM-RFE)[10] was also used for feature selection. Finally, we intersect clinical features from the LASSO and the SVM-RFE algorithms, and then use the multivariate logistic regression (LG) and random forest (RM) to test the predictive power of

the model. Superior to random forest performance, multivariate logistic regression was used to build the predictive nomogram in 285 patients and internally validated in 127 patients. In the independent validation phase, candidate features were validated in a prospective cohort (n = 178) (Fig. 1).

Statistical Analysis

The R packages of *glmnet* and *e1071* were applied to operate LASSO and SVM-RFE algorithm, respectively. The performance of the nomogram for the validation data was measured by the concordance index and was explored graphically by calibration plots, using the *rms* package of R software. For clinical use, the predictive performance of the nomogram was measured receiver operating characteristic (ROC) analysis with area under the curve (AUC) values. Decision curve analysis (DCA) were performed to plot net benefit (NB) and assess the utility of models for decision making[11]. Hosmer-Lemeshow goodness of fit test is used to validate the model fitness[12]. Statistical analyses were performed in the R (version 3.6.1) language environment and P-value < 0.05 (two-sided) is considered to be significant.

Results

Patient Characteristics

Taken together, 590 confirmed cases with COVID-19 were recruited from February 1st through March 10th. 226 (38.5%) of them developed severe diseases during hospitalization and the median hospitalization time is 12 days. The study flowchart is shown in Fig. 1. We used MSRS to seek the cutoff value in the training set. The optimal cutoff value and the distribution of patients' characteristics in three cohorts are presented in Table 1.

Table 1
Baseline Characteristics of COVID-19 patients

	Training set (%) N = 285	Internal validation(%) N = 127	Prospective validation(%) N = 178	*P* value
Type				
Common	184 (64.6)	81 (63.8)	98 (55.1)	0.104
Severe	101 (35.4)	46 (36.2)	80 (44.9)	
Gender				
Male	152 (53.3)	64 (50.4)	92 (51.7)	0.847
Femal	133 (46.7)	63 (49.6)	86 (48.3)	
Age <=48	136 (47.7)	70 (55.1)	78 (43.8)	0.148
>48	149 (52.3)	57 (44.9)	100 (56.2)	
CDCS				
>=1	82 (28.8)	33 (26.0)	53 (29.8)	0.761
0	203 (71.2)	94 (74.0)	125 (70.2)	
WBC(10^9/L)				
<=9.16	250 (87.7)	108 (85.0)	153 (86.0)	0.727
>9.16	35 (12.3)	19 (15.0)	25 (14.0)	
Hb(g/L)				
<=133	189 (66.3)	88 (69.3)	122 (68.5)	0.798
>133	96 (33.7)	39 (30.7)	56 (31.5)	
PLT(10^9/L)				
<=155	67 (23.5)	26 (20.5)	49 (27.5)	0.348
>155	218 (76.5)	101 (79.5)	129 (72.5)	

Abbreviations: CDCS, Charlson/Deyo comorbidity score; WBC, white blood cell; Hb,hemoglobin; PLT, platelet ; ESR, erythrocyte sedimentation rate; CRP, C-reactive protein; TBIL, total bilirubin; DBIL, direct bilirubin; ALT, alanine transaminase; AST, glutamic oxalacetic transaminase; ALB, albumin; GLB, globulin; BUN, blood urea nitrogen; Cr, creatinine; GLU, glucose; CK, reatine kinase; CK-MB, reatine kinase isoenzyme-MB; LDH, lactate dehydrogenase; AP, amyloid protein; Mb, myohemoglobin; Tn, troponin; PCT, procalcitonin; IL-6, interlukin-6;AMY, amylase; LPS, lipase; FER, ferritin. P value was calculated by Fisher exact test.

	Training set (%) N = 285	Internal validation(%) N = 127	Prospective validation(%) N = 178	*P* value
Lymphcyte(10^9/L)				
<=0.8	81 (28.4)	32 (25.2)	46 (25.8)	0.733
>0.8	204 (71.6)	95 (74.8)	132 (74.2)	
ESR(mm/h)				
<=44	99 (34.7)	46 (36.2)	69 (38.8)	0.681
>44	186 (65.3)	81 (63.8)	109 (61.2)	
CRP(mg/L)				
<=48.3	196 (68.8)	94 (74.0)	123 (69.1)	0.536
>48.3	89 (31.2)	33 (26.0)	55 (30.9)	
TBIL(umol/L)				
<=8.2	47 (16.5)	23 (18.1)	35 (19.7)	0.682
>8.2	238 (83.5)	104 (81.9)	143 (80.3)	
DBIL(umol/L)				
<=2.9	91 (31.9)	29 (22.8)	43 (24.2)	0.075
>2.9	194 (68.1)	98 (77.2)	135 (75.8)	
ALT(u/L)				
<=14	37 (13.0)	25 (19.7)	24 (13.5)	0.181
>14	248 (87.0)	102 (80.3)	154 (86.5)	
AST(u/L)				
<=21	58 (20.4)	22 (17.3)	23 (12.9)	0.123
>21	227 (79.6)	105 (82.7)	155 (87.1)	
ALB(g/L)				

Abbreviations: CDCS, Charlson/Deyo comorbidity score; WBC, white blood cell; Hb,hemoglobin; PLT, platelet ; ESR, erythrocyte sedimentation rate; CRP, C-reactive protein; TBIL, total bilirubin; DBIL, direct bilirubin; ALT, alanine transaminase; AST, glutamic oxalacetic transaminase; ALB, albumin; GLB, globulin; BUN, blood urea nitrogen; Cr, creatinine; GLU, glucose; CK, reatine kinase; CK-MB, reatine kinase isoenzyme-MB; LDH, lactate dehydrogenase; AP, amyloid protein; Mb, myohemoglobin; Tn, troponin; PCT, procalcitonin; IL-6, interlukin-6;AMY, amylase; LPS, lipase; FER, ferritin. P value was calculated by Fisher exact test.

	Training set (%) N = 285	Internal validation(%) N = 127	Prospective validation(%) N = 178	*P* value
<=31.7	122 (42.8)	55 (43.3)	81 (45.5)	0.845
>31.7	163 (57.2)	72 (56.7)	97 (54.5)	
GLB(g/L)				
<=28.9	70 (24.6)	33 (26.0)	56 (31.5)	0.256
>28.9	215 (75.4)	94 (74.0)	122 (68.5)	
BUN(mmol/L)				
<=4.48	159 (55.8)	63 (49.6)	88 (49.4)	0.312
>4.48	126 (44.2)	64 (50.4)	90 (50.6)	
Cr(umol/L)				
<=71.5	169 (59.3)	79 (62.2)	103 (57.9)	0.745
> 71.5	116 (40.7)	48 (37.8)	75 (42.1)	
GLU(mmol/L)				
<=6.9	208 (73.0)	88 (69.3)	124 (69.7)	0.647
>6.9	77 (27.0)	39 (30.7)	54 (30.3)	
CK(u/L)				
<=106	214 (75.1)	100 (78.7)	125 (70.2)	0.228
>106	71 (24.9)	27 (21.3)	53 (29.8)	
CK-MB(u/L)				
<=11.1	100 (35.1)	43 (33.9)	54 (30.3)	0.569
>11.1	185 (64.9)	84 (66.1)	124 (69.7)	
LDH(u/L)				
<=291	174 (61.1)	85 (66.9)	106 (59.6)	0.394

Abbreviations: CDCS, Charlson/Deyo comorbidity score; WBC, white blood cell; Hb,hemoglobin; PLT, platelet ; ESR, erythrocyte sedimentation rate; CRP, C-reactive protein; TBIL, total bilirubin; DBIL, direct bilirubin; ALT, alanine transaminase; AST, glutamic oxalacetic transaminase; ALB, albumin; GLB, globulin; BUN, blood urea nitrogen; Cr, creatinine; GLU, glucose; CK, reatine kinase; CK-MB, reatine kinase isoenzyme-MB; LDH, lactate dehydrogenase; AP, amyloid protein; Mb, myohemoglobin; Tn, troponin; PCT, procalcitonin; IL-6, interlukin-6;AMY, amylase; LPS, lipase; FER, ferritin. P value was calculated by Fisher exact test.

	Training set (%) N = 285	Internal validation(%) N = 127	Prospective validation(%) N = 178	*P* value
>291	111 (38.9)	42 (33.1)	72 (40.4)	
K+(mmol/L)				
<=4.2	154 (54.0)	70 (55.1)	94 (52.8)	0.922
>4.2	131 (46.0)	57 (44.9)	84 (47.2)	
Ca+(mmol/L)				
<=1.99	120 (42.1)	50 (39.4)	79 (44.4)	0.682
>1.99	165 (57.9)	77 (60.6)	99 (55.6)	
AP(mg/L)				
<=127.5	123 (43.2)	61 (48.0)	70 (39.3)	0.318
>127.5	162 (56.8)	66 (52.0)	108 (60.7)	
Mb(ng/ml)				
<=92.4	253 (88.8)	118 (92.9)	153 (86.0)	0.164
>92.4	32 (11.2)	9 (7.1)	25 (14.0)	
Tn(pg/ml)				
<=0.9	72 (25.3)	38 (29.9)	46 (25.8)	0.598
>0.9	213 (74.7)	89 (70.1)	132 (74.2)	
PCT(ng/ml)				
<=0.069	222 (77.9)	102 (80.3)	145 (81.5)	0.631
>0.069	63 (22.1)	25 (19.7)	33 (18.5)	
D-dimer(ug/ml)				
<=3.03	260 (91.2)	115 (90.6)	156 (87.6)	0.445
>3.03	25 (8.8)	12 (9.4)	22 (12.4)	

Abbreviations: CDCS, Charlson/Deyo comorbidity score; WBC, white blood cell; Hb,hemoglobin; PLT, platelet ; ESR, erythrocyte sedimentation rate; CRP, C-reactive protein; TBIL, total bilirubin; DBIL, direct bilirubin; ALT, alanine transaminase; AST, glutamic oxalacetic transaminase; ALB, albumin; GLB, globulin; BUN, blood urea nitrogen; Cr, creatinine; GLU, glucose; CK, reatine kinase; CK-MB, reatine kinase isoenzyme-MB; LDH, lactate dehydrogenase; AP, amyloid protein; Mb, myohemoglobin; Tn, troponin; PCT, procalcitonin; IL-6, interlukin-6;AMY, amylase; LPS, lipase; FER, ferritin. P value was calculated by Fisher exact test.

	Training set (%) N = 285	Internal validation(%) N = 127	Prospective validation(%) N = 178	*P* value
IL-6(pg/ml)				
<=8.17	183 (64.2)	91 (71.7)	126 (70.8)	0.195
>8.17	102 (35.8)	36 (28.3)	52 (29.2)	
AMY(U/L)				
<=61.2	136 (47.7)	58 (45.7)	82 (46.1)	0.905
>61.2	149 (52.3)	69 (54.3)	96 (53.9)	
LPS(U/L)				
<=29	92 (32.3)	31 (24.4)	50 (28.1)	0.245
>29	193 (67.7)	96 (75.6)	128 (71.9)	
FER(ng/ml)				
<=273.54	82 (28.8)	35 (27.6)	33 (18.5)	0.04
>273.54	203 (71.2)	92 (72.4)	145 (81.5)	

Abbreviations: CDCS, Charlson/Deyo comorbidity score; WBC, white blood cell; Hb,hemoglobin; PLT, platelet ; ESR, erythrocyte sedimentation rate; CRP, C-reactive protein; TBIL, total bilirubin; DBIL, direct bilirubin; ALT, alanine transaminase; AST, glutamic oxalacetic transaminase; ALB, albumin; GLB, globulin; BUN, blood urea nitrogen; Cr, creatinine; GLU, glucose; CK, reatine kinase; CK-MB, reatine kinase isoenzyme-MB; LDH, lactate dehydrogenase; AP, amyloid protein; Mb, myohemoglobin; Tn, troponin; PCT, procalcitonin; IL-6, interlukin-6;AMY, amylase; LPS, lipase; FER, ferritin. P value was calculated by Fisher exact test.

Selection Of Candidate Clinical Features

Using the primary filter criteria mentioned in the Methods section, we used 10 fold cross validated LASSO method and 5 fold cross validated SVM-RFE (Fig. 2A-C), two different machine learning algorithms and derived five clinical features, which are CRP, LDH, Age, CDCS and ESR (Fig. 2D). Next, we used LG and RM to build the predictive model based on the 5 selected features (Fig. 3). And the LG model demonstrated AUCs among all the data sets than the RM model, and thus were chosen for subsequent analysis. Subsequent multivariate logistic analysis revealed that ESR > 44 mm/h (HR 1.97, 95%CI 1.03–3.89, P = 0.04), CDCS > = 1 (HR 2.31, 95%CI 1.23–4.35, P = 0.009), age > 48 years (HR 2.27, 95%CI 1.24–4.19, P = 0.008), LDH > 291 U/L (HR 3.08, 95% CI 1.7–5.61, P < 0.001) and CRP > 48.3 mg/L (HR 4.1, 95%CI 2.23–7.62, P < 0.001) were critical risk factors closely related to the progression of COVID-19 (Table 2).

Table 2
Univariate and Multivariate Logistic Regression Analysis of Progression of COVID-19 patients in Training Cohort

	Univariate logistic analysis		Multivariate logistic analysis	
	HR (95% CI)	P value	HR (95% CI)	P value
ESR(mm/h)				
<=44	1	—	1	—
>44	3.05(1.75–5.49)	< 0.001	1.97 (1.03–3.89)	0.04
CDCS				
0	1	—	1	—
>=1	2.97 (1.75–5.07)	< 0.001	2.31(1.23–4.35)	0.009
Age <=48	1	—	1	—
>48	3.7 (2.21–6.33)	< 0.001	2.27(1.24–4.19)	0.008
LDH(u/L)				
<=291	1	—	1	—
>291	4.7 (2.81–7.97)	< 0.001	3.08 (1.7–5.61)	< 0.001
CRP(mg/L)				
<=48.3	1	—	1	—
>48.3	7.21(4.17–12.73)	< 0.001	4.1 (2.23–7.62)	< 0.001

Abbreviations: ESR, erythrocyte sedimentation rate; CDCS, Charlson/Deyo comorbidity score; LDH, lactate dehydrogenase;CRP, C-reactive protein; HR, Hazard Ratio; CI, confidence interval.

Building A Predictive Signature

To develop a clinically applicable tool that could predict the probability of whether a COVID-19 patient can develop severe disease, we constructed a nomogram to build a predictive model, taking clinical covariates into consideration (Fig. 4A). The predictors included CRP, LDH, Age, CDCS and ESR and the risk score of each covariates produced by the LG model are listed (Fig. 4B). To take one patient for example (blue track in Fig. 4A), basing on the selected features, the total points adds up to 197 and thus the corresponding probabilities of progressing to severe COVID-19 is 0.495.

Validating And Assessing The Model

To substantiate the stability of the nomogram, validation analyses were performed in an internal validation cohort (n = 127) and a prospective validation cohort (n = 178). ROC analysis revealed that the nomogram displayed similar AUC values of 0.822 (95% CI 0.765–0.875, P < 0.001), 0.762 (95% CI 0.768–0.844, P < 0.001) and 0.705 (95% CI 0.627–0.778, P < 0.001) for the estimation of possible severe COVID-19 cases (Fig. 3A-C).

The bootstrapped calibration plot for the probability showed consistency between prediction by nomogram and actual observation (Fig. 5A-C) and favorable discriminative power (Fig. 5D-F). And Hosmer-Lemeshow test P value are 0.501 (training set), 0.239 (internal validation set) and 0.453 (prospective validation set) which indicate the good fitness of the model. And the DCA plots show the favorable net benefit for clinical use (Fig. 6A-C). To take 2 patients with distinguished risk scores for example, the low risk one added up to 74 points with the probability of 0.12 to progress and the CT scan showed no worsen pneumonia after 10 days hospitalization. Another one added up to 361 points with the probability of 0.89 went through severe lung lesion in 10 days.

Discussion

Since the COVID-19 outbroke in Hubei, even with the optimal control in China, the cumulative confirmed cases in the globe had overpassed three million and the threat of coronavirus is still out there. In the patients who suffered COVID-19, CRP was increased in 86.22% of them, and ESR in 90.22%[13]. During the disease course, longitudinal evaluation of lymphocyte count dynamics and inflammatory indices, including LDH, CRP and IL-6 may help to identify cases with dismal prognosis and prompt intervention in order to improve outcomes[14]. Zhou et al. showed that increasing odds of in-hospital death associated with older age (odds ratio 1.10, 95% CI 1.03–1.17, per year increase; p = 0.0043), which could be the potential risk factor[15]. Another research demonstrated that elder age, underlying hypertension, high cytokine levels (IL-2R, IL-6, IL-10, and TNF-a), and high LDH level were significantly associated with severe COVID-19 on admission[16]. Until now, more and more independent risk factors have been determined and a number of systemic score systems have been built to analyze the status of disease progression and prevent severe outcomes. Dong et al. [6]reported a novel scoring model, named as CALL, established for disease condition prediction, which included comorbidity, age, lymphocyte, and LDH, with the AUC reaching 0.91 (95% CI 0.86–0.94). However, only one verification of this model has been made, the efficacy of it is doubted, and the number of patients is insufficient. Another model to early predict severe type of COVID-19 showed older age, higher LDH, CRP, RDW, DBIL, BUN, and lower ALB on admission correlated with higher odds of severe COVID-19, with the AUC reached 0.912 (95% CI 0.846–0.978) in the training set, and 0.853 (95% CI 0.790–0.916) in the validation set[5]. Nonetheless, the small sample size could be the deficiency of this model.

For the sake of controlling the major health incident and bettering the medical resource allocation, we extracted the clinical data of 590 cases from The Wuhan Jinyintan Hospital and the prediction model was been established, which included ESR, CDCS, Age, LDH, and CRP. Notably, CDCS, the Charlson/Deyo comorbidity score, is a method of categorizing comorbidities of patients based on the International

Classification of Diseases (ICD) diagnosis codes found in administrative data, such as hospital abstracts data. Each comorbidity category has an associated weight (from 1 to 6), based on the adjusted risk of mortality or resource use, and the sum of all the weights results in a single comorbidity score for a patient. A score of zero indicates that no comorbidities were found. The higher the score, the more likely the predicted outcome will result in mortality or higher resource use. This scoring systems have been reported to be associated with overall survival in various types of cancer and the death rate of other morbidities, such as ischemic stroke, acute cholecystitis, acute hip fracture, and so forth[17–21]. Due to the efficacy of CDCS, we ground breakingly utilized this scoring system in our prediction model with meeting the rule of The TRIPOD Statement [22], which could also interpret the high mortality of COVID-19 with multiple comorbidities. Hereby, the five significant indices were overlapped by the LASSO and SVM analysis, which are machine learning used for classification and regression analysis in order to enhance the prediction accuracy and interpretability of the statistical model it produces. Then, the ROC, DCA, and Calibration were performed for performance assessment and the triple verification were applied. The predictive nomogram indicated that the possibility of the progression from common type to severe type could reach 50%, when the total points meet 197. Thereafter, the AUC of internal training set, testing set, and external testing set reached 0.822, 0.762, and 0.705, respectively.

However, there are still some limitation should be majorized in the future investigation. The AUC values are lower than 0.9 and more cases should be recruited to optimize our prediction model for more precise forecasting. And the data of patients were derived from Wuhan, Hubei Province, which means the situation outside Hubei Province could be distinct and multi-center analysis is urgently needed.

Conclusion

Through the filtering by LASSO and SVM-RFE, two machine learning method, 5 independent predictive feature for severe COVID-19 were selected which are: CRP, LDH, Age, CDCS and ESR. Based on this, we build a predicting tool which can early predict severe COVID-19 and aid medical decisions for COVID-19 patients.

Declarations

Ethics approval and consent to participate

This study was approved by the Ethics Review Committee of Wuhan Jinyintan Hospital and Shanghai General Hospital.

Consent for publication

All the authors have agreed for the publication.

Availability of data and materials

Not applicable

Competing interests

The authors have no conflicts of interest to declare.

Funding

The reported work was supported in part by research grants from the Natural Science Foundation of China (No. 81972393, 81772705, 31570775).

Authors' contributions

JHZ conceived the initial concept and designed the study, KW, ZZ and ZXY participated to design the study and in the data extraction. XYZ and ZXY analyzed the data and wrote the manuscript. All authors read and approved the final manuscript.

Acknowledgments

We appreciate the support from Youth Science and Technology Innovation Studio of Shanghai Jiao Tong University School of Medicine.

References

1. Wu C, Chen X, Cai Y, et al. Risk Factors Associated With Acute Respiratory Distress Syndrome and Death in Patients With Coronavirus Disease 2019 Pneumonia in Wuhan, China. JAMA Intern Med **2020**;
2. Lauer SA, Grantz KH, Bi Q, et al. The Incubation Period of Coronavirus Disease 2019 (COVID-19) From Publicly Reported Confirmed Cases: Estimation and Application. Ann Intern Med **2020**; 172:577–582.
3. Huang C, Wang Y, Li X, et al. Clinical features of patients infected with 2019 novel coronavirus in Wuhan, China. Lancet **2020**; 395:497–506.
4. Chen N, Zhou M, Dong X, et al. Epidemiological and clinical characteristics of 99 cases of 2019 novel coronavirus pneumonia in Wuhan, China: a descriptive study. Lancet **2020**; 395:507–513.
5. Gong J, Ou J, Qiu X, et al. A Tool to Early Predict Severe Corona Virus Disease 2019 (COVID-19) : A Multicenter Study using the Risk Nomogram in Wuhan and Guangdong, China. Clin Infect Dis **2020**;
6. Ji D, Zhang D, Xu J, et al. Prediction for Progression Risk in Patients with COVID-19 Pneumonia: the CALL Score. Clin Infect Dis **2020**;

7. Zu ZY, Jiang MD, Xu PP, et al. Coronavirus Disease 2019 (COVID-19): A Perspective from China. Radiology **2020**; :200490. Available at: https://pubs.rsna.org/doi/10.1148/radiol.2020200490.
8. Wright MN, Dankowski T, Ziegler A. Unbiased split variable selection for random survival forests using maximally selected rank statistics. Stat Med **2017**; 36:1272–1284.
9. Tibshirani R. Regression Shrinkage and Selection via the Lasso. Journal of the Royal Statistical Society Series B (Methodological) **1996**; 58:267–288. Available at: https://www.jstor.org/stable/2346178.
10. Huang M-L, Hung Y-H, Lee WM, Li RK, Jiang B-R. SVM-RFE Based Feature Selection and Taguchi Parameters Optimization for Multiclass SVM Classifier. Hindawi, 2014: e795624. Available at: https://www.hindawi.com/journals/tswj/2014/795624/.
11. Van Calster B, Wynants L, Verbeek JFM, et al. Reporting and Interpreting Decision Curve Analysis: A Guide for Investigators. Eur Urol **2018**; 74:796–804.
12. Paul P, Pennell ML, Lemeshow S. Standardizing the power of the Hosmer-Lemeshow goodness of fit test in large data sets. Stat Med **2013**; 32:67–80.
13. Li R, Tian J, Yang F, et al. Clinical characteristics of 225 patients with COVID-19 in a tertiary Hospital near Wuhan, China. J Clin Virol **2020**; 127:104363.
14. Terpos E, Ntanasis-Stathopoulos I, Elalamy I, et al. Hematological findings and complications of COVID-19. Am J Hematol **2020**;
15. Zhou F, Yu T, Du R, et al. Clinical course and risk factors for mortality of adult inpatients with COVID-19 in Wuhan, China: a retrospective cohort study. Lancet **2020**; 395:1054–1062.
16. Li X, Xu S, Yu M, et al. Risk factors for severity and mortality in adult COVID-19 inpatients in Wuhan. J Allergy Clin Immunol **2020**;
17. Suzuki H, Hanai N, Nishikawa D, et al. The Charlson comorbidity index is a prognostic factor in sinonasal tract squamous cell carcinoma. Jpn J Clin Oncol **2016**; 46:646–651.
18. Yao Z, Zheng Z, Ke W, et al. Prognostic nomogram for bladder cancer with brain metastases: a National Cancer Database analysis. J Transl Med **2019**; 17:411.
19. Hall RE, Porter J, Quan H, Reeves MJ. Developing an adapted Charlson comorbidity index for ischemic stroke outcome studies. BMC Health Serv Res **2019**; 19:930.
20. Bonaventura A, Leale I, Carbone F, et al. Pre-surgery age-adjusted Charlson Comorbidity Index is associated with worse outcomes in acute cholecystitis. Dig Liver Dis **2019**; 51:858–863.
21. Quach LH, Jayamaha S, Whitehouse SL, Crawford R, Pulle CR, Bell JJ. Comparison of the Charlson Comorbidity Index with the ASA score for predicting 12-month mortality in acute hip fracture. Injury **2020**; 51:1004–1010.
22. Collins GS, Reitsma JB, Altman DG, Moons KGM. Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD): the TRIPOD statement. BMJ **2015**; 350:g7594.

Figures

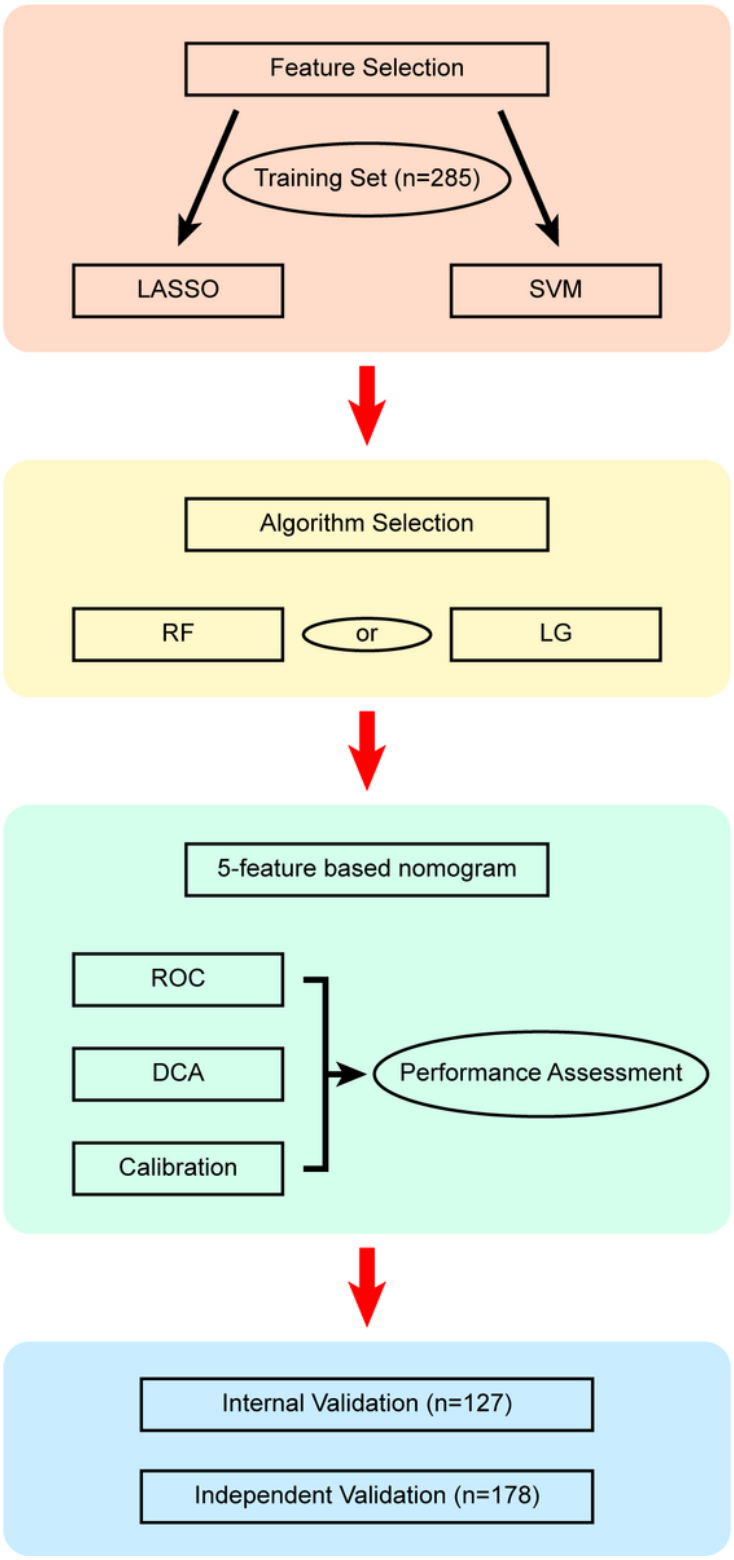


Figure 1

Study flowchart. LASSO, Least Absolute Shrinkage and Selector Operation; SVM-RFE, Support Vector Machine-Recursive Feature Elimination; LG, logistic regression; RM, random forest.

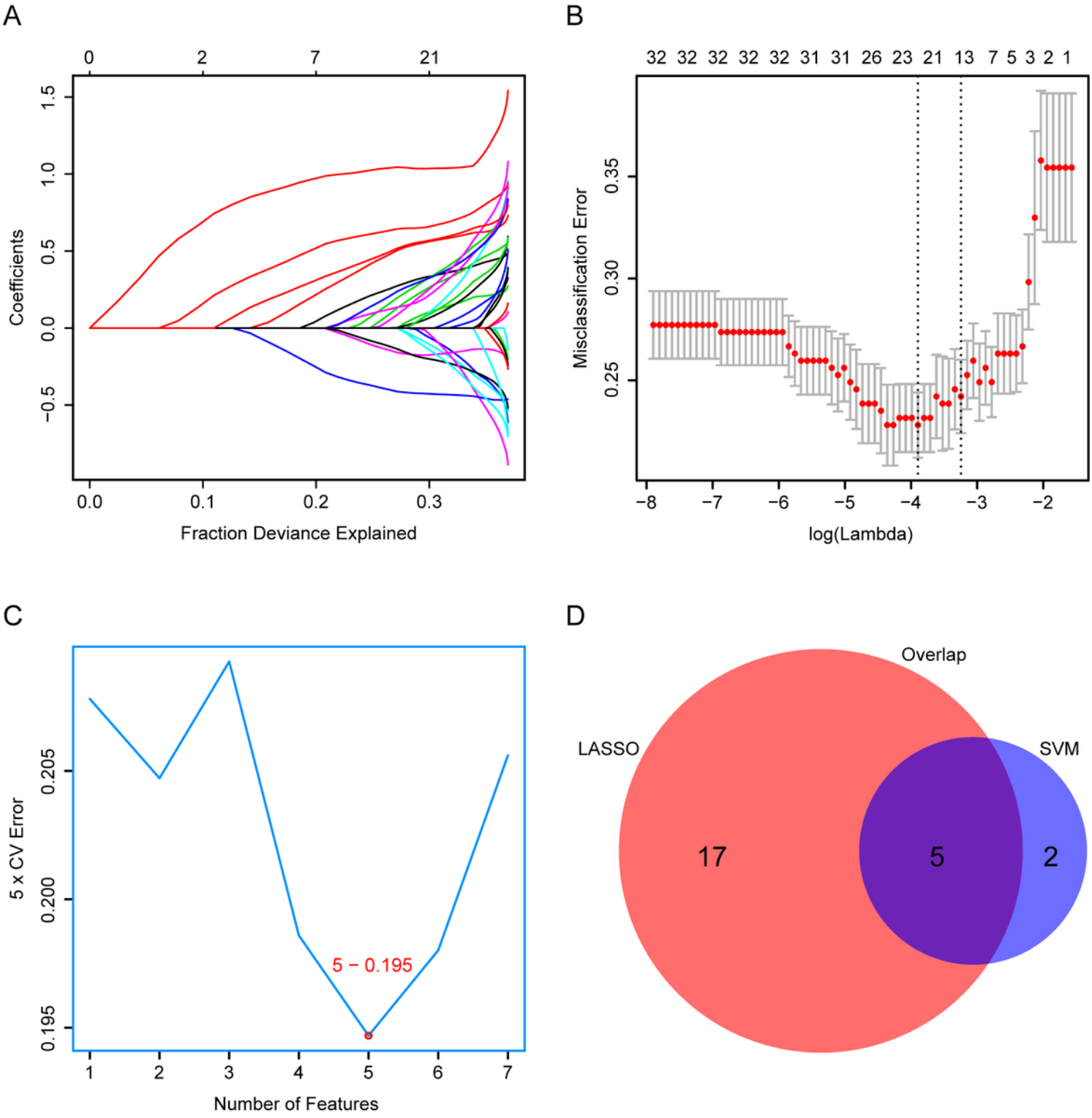


Figure 2

Two algorithms were used for feature selection. (A) 10 fold cross-validated error (first vertical line equals the minimum error (lambda = 0.033) and the second one shows the cross-validated error within 1 standard error of the minimum) in LASSO. (B) Profiles of coefficients with penalization were plotted against the log (lambda) sequence. (C) 5 fold cross validated SVM-RFE algorithms in the training cohort. (D) Intersection of important features.

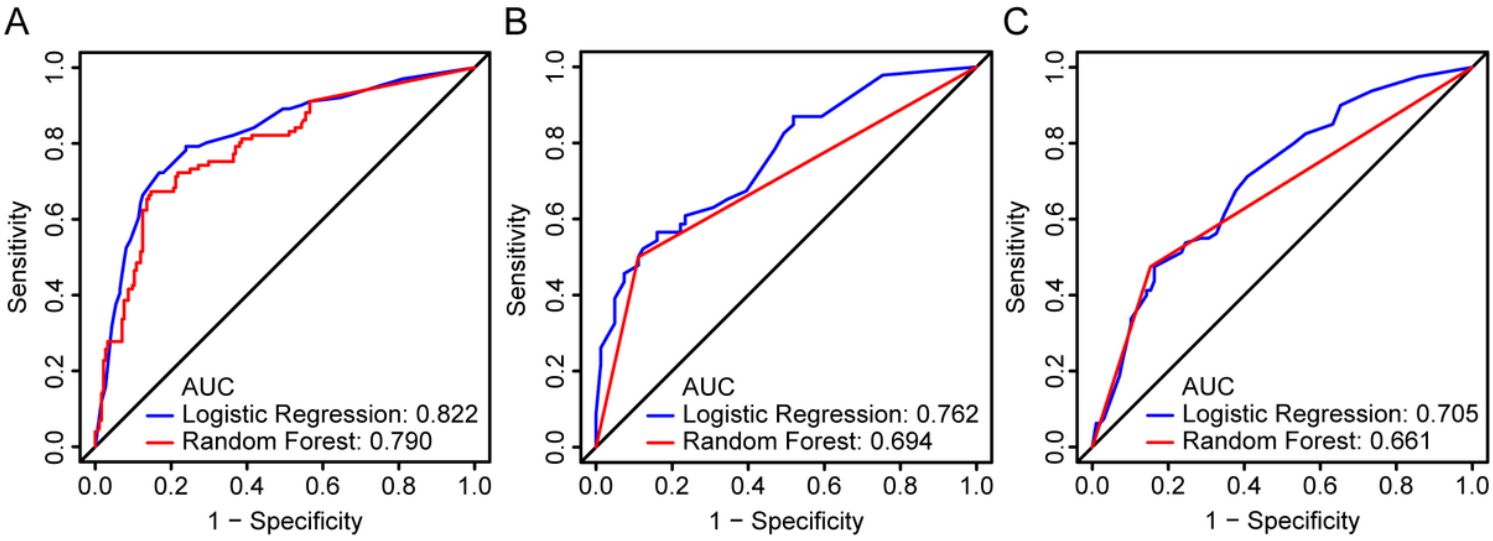
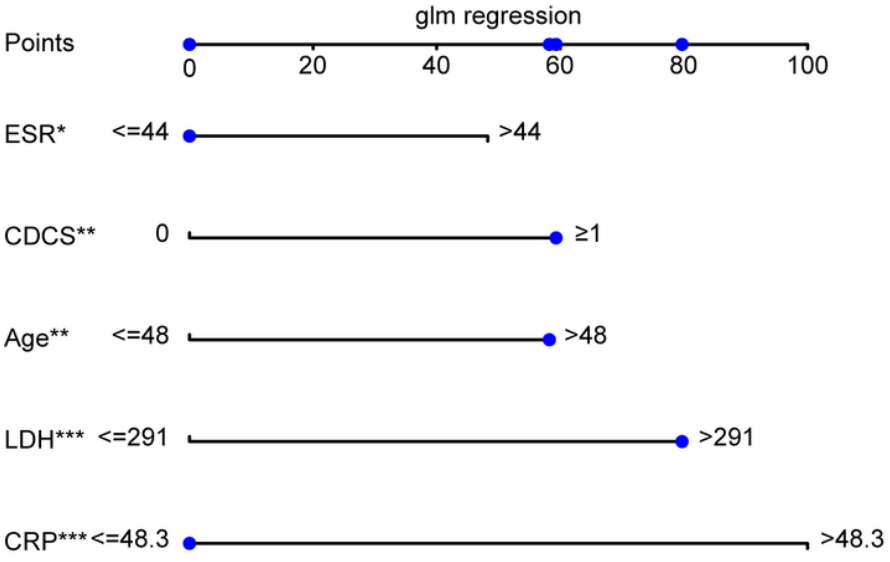


Figure 3

ROC analysis to assess the performance of LG and RM in predicting the severe type COVID-19 in the three cohorts. (A) training cohort, (B) internal validation cohort (C) external validation cohort.

A



B

Feature	Cut-off	Points
ESR	≤44	0
	>44	48
CDCS	0	0
	≥1	59
Age	≤48	0
	>48	58
LDH	≤291	0
	>291	80
CRP	≤48.3	0
	>48.3	100

C

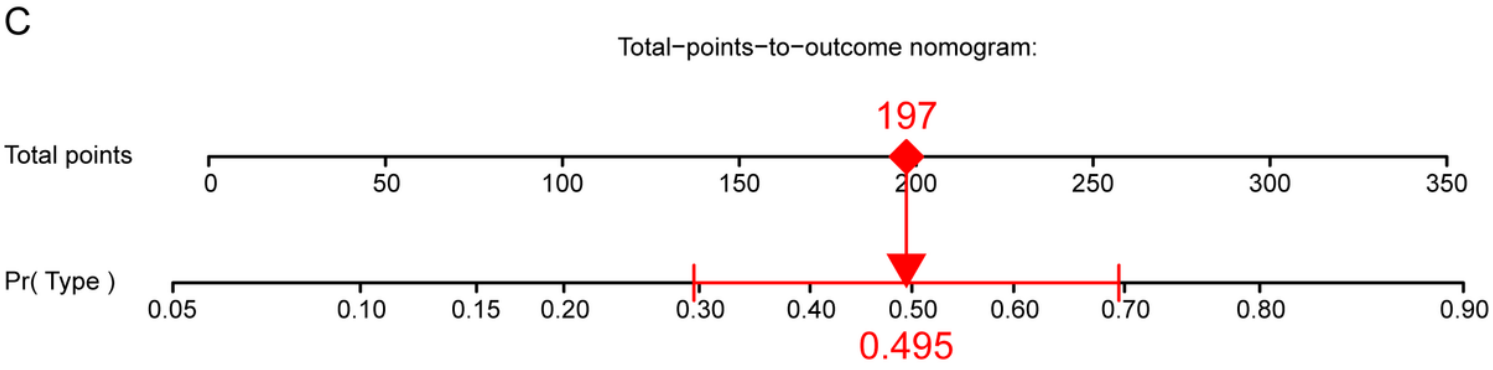


Figure 4

Predictive model and feature risk scores. (A) Nomogram to predict the severe type COVID-19. (B) Risk scores for each feature. (*P < 0.05, **P < 0.01, ***P < 0.001)

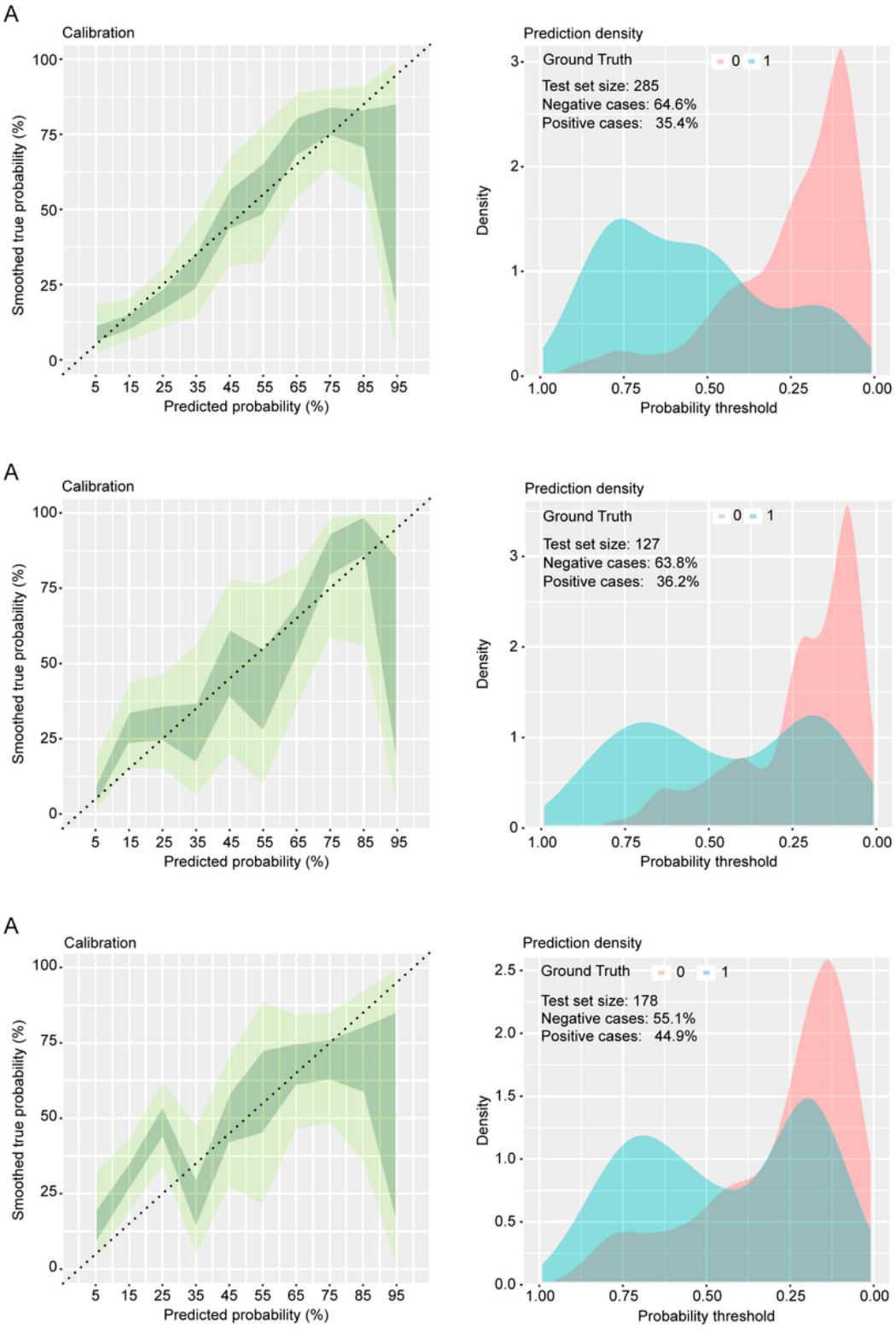


Figure 5

Calibration curve and prediction density plot for nomogram model. (A, D) training cohort (B, E) internal validation cohort (C, F) prospective validation cohort. The dotted line represents the ideal nomogram, and

the deep green track represents the observed nomogram with the diluted green track representing confidence interval.

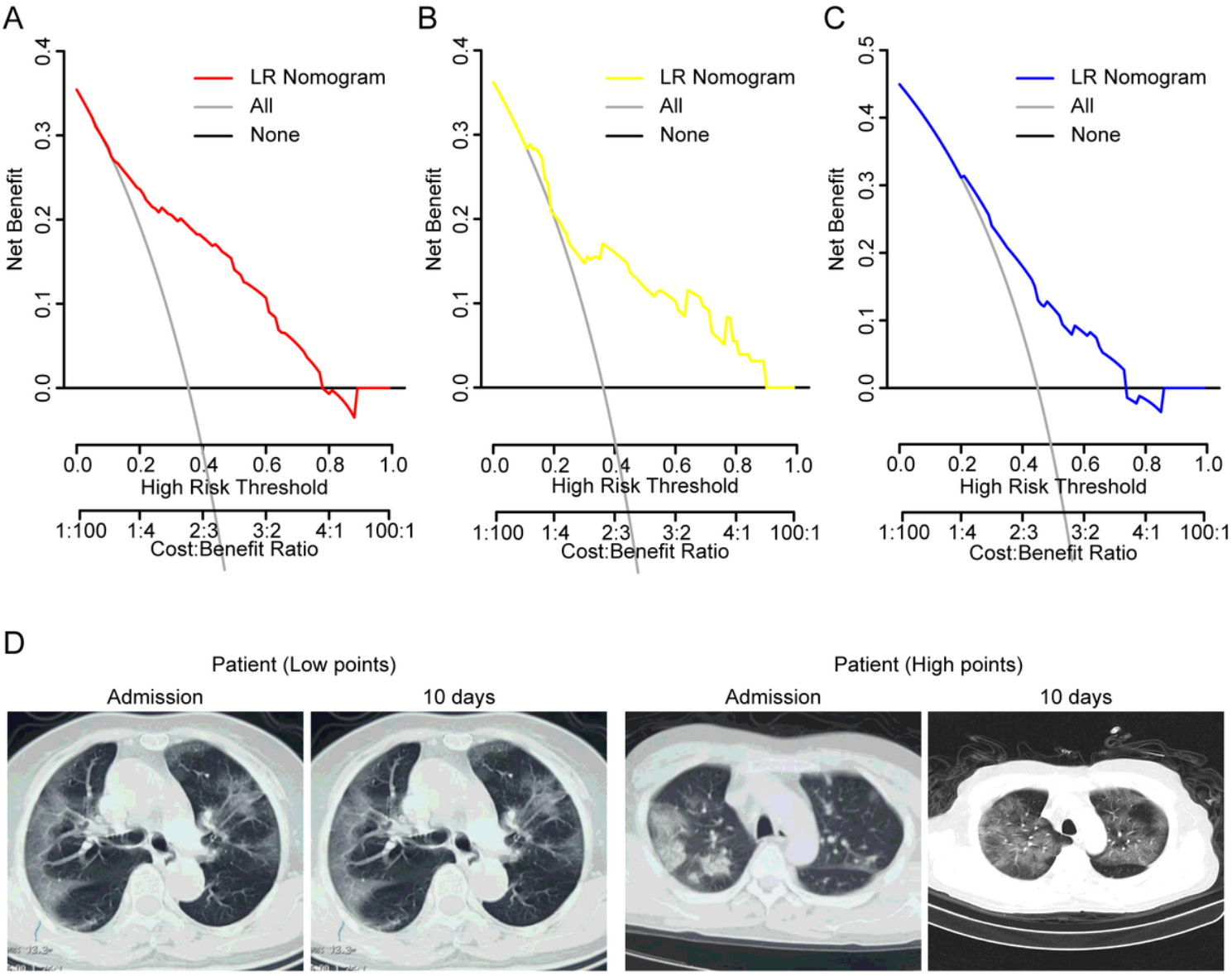


Figure 6

Decision curve analysis for nomogram. (A) training cohort (B) internal validation cohort (C) prospective validation cohort. (D) Two example CT scans of victims with distinguished risk points.

Supplementary Files

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

- Supplemtentmaterial.docx