COMPUTATIONAL ANDSTRUCTURAL BIOTECHNOLOGY JOURNAL

journal homepage: www.elsevier.com/locate/csbj

3044 Cases reveal important prognosis signatures of COVID-19 patients

Shijie Qin [a,b,1], Weiwei Li [a,1], Xuejia Shi [b,1], Yanjun Wu [c,h,1], Canbiao Wang [b], Jiawei Shen [a], Rongrong Pang [a,d], Bangshun He [a,e], Jun Zhao [a], Qinghua Qiao [f,h], Tao Luo [a], Yanju Guo [a,b], Yang Yang [a], Ying Han [a], Qiuyue Wu [a], Jian Wu [a], Wei Dai [a], Libo Zhang [a,d], Liming Chen [b], Chunyan Xue [a], Ping Jin [b,*], Zhenhua Gan [g,h,*], Fei Ma [b,*], Xinyi Xia [a,h,*]

[a] *COVID-19 Research Center, Institute of Laboratory Medicine, Jinling Hospital, Nanjing University School of Medicine, Nanjing Clinical College of Southern Medical University, Nanjing, Jiangsu 210002, China*
[b] *Laboratory for Comparative Genomics and Bioinformatics, College of Life Science, Nanjing Normal University, Nanjing 210046, China*
[c] *Department of Information, Jinling Hospital, Nanjing University School of Medicine, Nanjing 210002, China*
[d] *Department of Laboratory Medicine, Nanjing Red Cross Blood Center, Nanjing 210003, Jiangsu, China*
[e] *General Clinical Research Center, Nanjing First Hospital, Nanjing Medical University, Nanjing, China*
[f] *Medical and Technical Support Department, Pingdingshan Medical District, the 989th Hospital, Pingdingshan, Henan 467000, China*
[g] *Department of Medical Administration, Jinling Hospital, Nanjing University School of Medicine, Nanjing 210002, China*
[h] *Joint Expert Group for COVID-19, Department of Laboratory Medicine & Blood Transfusion, Wuhan Huoshenshan Hospital, Wuhan, Hubei 430100, China*

ARTICLE INFO

Article history:
Received 1 August 2020
Received in revised form 21 January 2021
Accepted 26 January 2021
Available online 9 February 2021

Keywords:
COVID-19
SARS-CoV-2
Clinical characteristics
Prognostic factors
China

ABSTRACT

Critical patients and intensive care unit (ICU) patients are the main population of COVID-19 deaths. Therefore, establishing a reliable method is necessary for COVID-19 patients to distinguish patients who may have critical symptoms from other patients. In this retrospective study, we firstly evaluated the effects of 54 laboratory indicators on critical illness and death in 3044 COVID-19 patients from the Huoshenshan hospital in Wuhan, China. Secondly, we identify the eight most important prognostic indicators (neutrophil percentage, procalcitonin, neutrophil absolute value, C-reactive protein, albumin, interleukin-6, lymphocyte absolute value and myoglobin) by using the random forest algorithm, and find that dynamic changes of the eight prognostic indicators present significantly distinct within differently clinical severities. Thirdly, our study reveals that a model containing age and these eight prognostic indicators can accurately predict which patients may develop serious illness or death. Fourthly, our results demonstrate that different genders have different critical illness rates compared with different ages, in particular the mortality is more likely to be attributed to some key genes (e.g. ACE2, TMPRSS2 and FURIN) by combining the analysis of public lung single cells and bulk transcriptome data. Taken together, we urge that the prognostic model and first-hand clinical trial data generated in this study have important clinical practical significance for predicting and exploring the disease progression of COVID-19 patients

1. Introduction

The new coronavirus disease (COVID-19) is caused by severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2). According to statistics up to July 28th, the spread of COVID-19 has infected more than 16 million people worldwide and caused more than 650,000 deaths. The number of confirmed cases and deaths in some regions may be even higher than data available due to multiple factors, such as detection methods, medical resource lists, political and cultural differences [1]. COVID-19 has resulted in considerable morbidity and mortality. Thus early rapid diagnosis, surveillance, risk assessments, and medical resource managements are essential in the prevention and control of epidemics before protective vaccines are applied clinically. Generally, the COVID-19 patients with the highest mortality rate are the critically ill and ICU patients, but they only account for a small proportion of the

[*] Corresponding authors at: COVID-19 Research Center, Institute of Laboratory Medicine, Jinling Hospital, Nanjing University School of Medicine, Nanjing Clinical College of Southern Medical University, Nanjing, Jiangsu 210002, China (X. Xia).

E-mail addresses: jinping8312@163.com (P. Jin), janeandbilly@126.com (Z. Gan), mafei01@tsinghua.org.cn (F. Ma), xinyixia@nju.edu.cn (X. Xia).

[1] Contributed equally to this article.

https://doi.org/10.1016/j.csbj.2021.01.042

hospitalized mild and severe patients [2]. Therefore, establishing reliable methods are crucial to distinguish high-risk patients from others. Although combined nucleic acid detection, antibody detection and computed tomography (CT) imaging can effectively diagnose COVID-19, the severity and prognosis of patients cannot be predicted [3,4].

Previous studies have reported multiple organ dysfunctions and prognostic markers including neutrophils and interleukin-6 [5]. Myoglobin and C-reactive protein are related to myocardial injuries [6]. Alkaline phosphatase is related to liver damage whereas increasing D-Dimer causes new impaired blood coagulation [7,8]. These findings reveal the potential of laboratory indexes serving as indicators of COVID-19 severity. Our previous work also indicated that the level of tumor biomarkers is associated with the severity of patients and could predict clinical outcomes [9]. However, most of the published data was based on a relatively small sample size, which may reduce the statistical reliability. Therefore, in order to establish a risk stratification model for finding key laboratory indicators, which predict the disease progression and clinical outcomes of COVID-19 patients, we here further conduct the study based on a large sample size.

2. Methods

2.1. Study design and participants

3059 COVID-19 patients, who were hospitalized between February 4th, 2020 and April 13th, 2020 in Huoshenshan hospital Wuhan, China, were eligible for inclusion in the study. 3044 patients were further included in the final research cohort after excluding these patients with incomplete medical records (e.g., transfer to any other hospital). The study was approved by the Research Ethics Commission of Huoshenshan Hospital, and written informed consent was obtained from each patient.

2.2. Data collection

According to the world health organization (WHO)/International Severe Acute Respiratory and Emerging Infection Consortium case record form for severe acute respiratory infections, baselines of participants, epidemiological and clinical manifestations, laboratory findings and outcome data were extracted from electronic medical records. Major basic information (i.e., age, sex, the highest historical classification, preliminary diagnosis, discharge diagnosis and discharge conditions) were collected except for patients' personal information (e.g., name and identification) and comorbidities were also included in clinical symptoms.

2.3. Definition

All COVID-19 patients were implemented in accordance with the New Coronavirus Pneumonia Diagnosis and Treatment Plan (7th edition) issued by the National Health Commission of China. The diagnosis of COVID-19 was confirmed by the positive real-time PCR results of throat swab performed in the laboratory. The chest CT images were also used to assist in the diagnosis and assessment of the condition of patients. SARS-CoV-2 nucleic acid was detected by reverse transcription and real-time PCR assays using a commercial detection kit (Changsha Sansure Biotech). Routine hematological test was analyzed by a Mindray Automatic Blood Cell Analyzer (BC-5390CRP). The different treatment outcomes and clinical severity classification during hospitalization were defined as follows. "Cured" refers to patients who can be discharged from hospital and they must meet all of the following conditions. 1. The body temperature returns to normal for more than 3 days. 2. Respiratory symptoms improved significantly. 3. Lung imaging showed significant improvement in acute exudative lesions. 4. Negative nucleic acid test of respiratory tract specimens such as sputum and nasopharyngeal swab twice in a row (sampling time interval of at least 24 h). "Improved" means that the overall symptoms of the patient are significantly improved after treatment, but still do not meet the criteria for discharge. "Severity classification" represents the worst state (mild, severe, critical) of patients during the entire hospitalization. "Mild" symptoms are described as follows: 1. Clinical symptoms are mild and no imaging findings of pneumonia are found. "Severe" symptoms should meet one of the following: 1. Shortness of breath, respiration rate (RR) $\geq$ 30 times/min. 2. In the resting state, the oxygen saturation is $\leq$ 93%. 3. Arterial partial pressure of oxygen (PAOZ)/oxygen concentration (FIO2 $\leq$ 300 mmHg) (1 mmHg = 0.133 kPa). "Critical" symptoms should meet one of the following: 1. Respiratory failure and the need for mechanical ventilation. 2. Shock. 3. Complicated with other organ failure, ICU monitoring and treatment are required.

2.4. Single cell sequencing data analysis

The single-cell data of 8 normal tissues came from the gene expression omnibus (GEO) database (GSE122960) [10]. The bulk transcriptome data of normal tissues adjacent to lung cancer were derived from The Cancer Genome Atlas (TCGA), and the standardized fragments per kilobase per million mapped reads (FPKM) expression data of these samples was obtained from the UCSC Xena database (https://xenabrowser.net/hub/). The Seurat3.0 R package was applied for quality control, filtering, standardization and subsequent analysis [11]. The inclusion criteria for cell quality control included 200–5000 genes detected in a single cell (nFeature_RNA), and <5% mitochondrial gene expression. The logNormalize function was used to normalize and normalize the expression matrix. The clustering performance of the cells was performed using the top 2000 most variable genes with a resolution of 0.5. The cell scatter gram method was obtained from t-distributed stochastic neighbor embedding (t-SNE) [12].

2.5. Statistical analysis

Median or average value indicates continuous variables, and the n (%) stands for categorical variables. Two-tailed wilcoxon rank sum test was applied to compare the differences of continuous variables of two groups. When there were three groups (mild, severe and critical), they were compared with each other. Chi-square test was used to compare the frequency of different groups, and the fisher exact test was applied instead when the theoretical prediction value of the chi-square test is <5. In order to avoid non-convergence in modeling, the extreme value of each indicator was processed by the block method in which data greater than 99 quantiles was replaced with 99 quantiles, while data <1 quantile was substituted with 1 quantile. The impact of various laboratory indicators on clinical critical illness and death was explored by logistic regression. These most important variables, which may give rise to the severity and mortality of COVID-19, were assessed by random forest machine learning algorithm. Logistic regression was applied to model age, gender and 8 important prognostic indicators. Receiver operating characteristic (ROC) curve was used to evaluate the quality of the model in the training set and verification set. All statistical analyses were performed using R software (version 3.5.3), and *p*-value under 0.05 were considered statistically significant.

3. Result

3.1. The demographic and clinical characteristics of 3044 COVID-19 patients

In this study, we intended to use the following process to find the laboratory indicators which could be served as prognostic factors for developing critical illness and death (Fig. 1A). At the beginning, 3044 patients with complete clinical information were screened from 3059 COVID-19 patients (Fig. 1A). Secondly, statistics were made on the demographics, hospitalization, baseline characteristics, underlying diseases and complications of 3044 COVID-19 patients, whilst these laboratory indicators from patients with different severity and death outcomes were further compared to determine differences among them. Age and gender were then introduced as covariates to screen for significant laboratory indicators, and these screened laboratory indicators can affect critical illness and death by category. Next these 29 significant laboratory indicators extracted from the above results were used to undergo random forest algorithm screening, and a prognostic model containing 8 prognostic indicators and age was constructed and verified by performing ROC. The dynamic changes of 8 prognostic indicators were also evaluated. Finally, public single-cell and bulk transcriptome data were jointly analyzed to explore the underlying molecular mechanisms of different COVID-19 types with different ages and genders.

As shown in Fig. 1, these enrolled patient's ages were from 10 to 100 years old, and the most inpatients concentrated (n = 932, 30.62%) between 61 and 70 years old (Fig. 1B). Another major age population for hospitalization was from 40 to 60 years old or from 70 to 80 years older, respectively (Fig. 1B). Our study indicated that the elderly constituted the main population among infected patients, which agrees with the previous report [13]. Our results demonstrated that the number of cured and improved patients could respectively reached 2930 (96.25%) and 48 (1.41%) in the clinical setting, indicating that most patients had good treatment outcomes, but the number of deaths remained 66 (2.17%) (Fig. 1C). Most patients could be attributable to mild and severe level while the rest minority could progress to critical level (5.2%) according to the classification (Fig. 1D). Similar to the proportion of critical ones, the proportion of ICU patients accounted for 4.2% and there was a great overlap between ICU patients and critical patients (Fig. 1E). In view of gender, the proportion of males was only 1.58% higher than that of females (Fig. 1F), which seems to imply that the infection rate is no significant gender difference.

3.2. Clinical baseline characteristics of COVID-19 patients with different outcomes

In this study, patients were assigned into three severity groups according to their highest severity classification, including 1467 mild, 1418 severe and 159 critical patients with median ages of 56.0 years (IQR: 45–65), 63.0 years (IQR: 53.0–71.0), and 68.0 years (IQR: 61.5–76.5), respectively (Table 1). Except for 40 patients with unknown survival outcome information, the median age of 2038 survivors was 60.0 (IQR: 49.0–68.0), while the median age of 66 deaths was 71.5 years (IQR: 65.5–78.0), indicating the elderly patients have worse clinical status and a higher mortality rate ($p < 0.001$, Table 1). In mild and moderate patients, the ratio of men to women was equal, but there was a significant increase number of male developing critical severity ($p = 0.002$, Table 1), and the mortality rate of males was also markedly higher than that of females ($p = 0.013$, Table 1). For ICU treatment, critically ill patients were significantly higher than mild and severe patients ($p < 0.001$), accounting for 64.78% (Table 1). For the clinical outcomes, there were 4 severe and 61 critical cases among these deceased patients ($p < 0.001$, Table 1). In addition, most deceased patients were critical patients and underwent ICU treatment. The median length of hospital stay of 3044 COVID-19 patients was 13.0 days (IQR: 8.0–19.0), and patients with higher disease severity had a significant longer hospital stay (severe: 14.0 days (IQR: 8.0–22.0); critical: 19.0 days (IQR: 11.0–32.0) (Table 1).

In our work, three most common comorbidities of 3044 COVID-19 patients were hypertension, diabetes and coronary atherosclerosis, which was similar to the reports from the United States and other countries [14]. Interestingly, we found that people with underlying diseases, such as hypertension, diabetes and coronary atherosclerosis, tumors, chronic obstructive pulmonary disease, and abnormal renal function, were more likely to develop severe and critical illness (Table 1, $p < 0.05$). Remarkably, four most common comorbidities related with clinical death were hypertension ($p < 0.001$), diabetes ($p = 0.001$), coronary heart disease ($p = 0.001$) and chronic obstructive pulmonary disease ($p < 0.001$). The result of logistic regression adjusted by age and gender revealed that hypertension (OR = 1.483, $p = 0.014$), diabetes (OR = 1.557, $p = 0.016$) and tumor (OR = 2.315, $p = 0.022$) were main risk factors (Table 1 and Table S1). Besides, respiratory failure, acute respiratory distress syndrome and thrombocytopenia also were the most common comorbidities among critical and dead patients, which could turn out to be potential lethal factors (Table 1 and Table S1). Especially, we found that 1369 (45.57%) infected patients had no comorbidity ($p < 0.001$, Table 1 and Table S1).

3.3. Laboratory test results of patients with different clinical prognosis of COVID-19

After sorting and summarizing the laboratory examination indicators of COVID-19 patients, 54 indicators were screened for subsequent analysis. Here, we urged that not all 3044 patients have undergone all laboratory tests, and the specific number of people tested for each indicator will be shown in the results below. Moreover, most patients underwent massive tests of multiple indicators, but we here only used the earliest test values for subsequent analysis. These 54 indicators were roughly divided into 8 categories, including blood routine examination, electrolytes, liver function, urine tests, kidney function, heart function, blood coagulation indicators and others. Red blood cells and several white blood cells, including monocytes, lymphocytes, basophils, eosinophils and neutrophils, were involved in the blood routine examination. Electrolytes included sodium, potassium, chlorine, phosphorus, serum magnesium and calcium. The indicators related to liver function mainly contained alanine aminotransferase, aspartate aminotransferase, total protein, albumin, total bilirubin, direct bilirubin, total bile acid, indirect bilirubin, globulin, alkaline phosphatase, γ-glutamyl transpeptidase, etc. Urine examination included cystatin C, pH, white cells, red cells, etc. The related torenal function indicators contained urea nitrogen, creatinine, total carbon dioxide, uric acid and so on. The indicators related to cardiac function included creatine kinase, lactate dehydrogenase, α-hydroxybutyrate dehydrogenase, creatine kinase isoenzyme, myoglobin, hypersensitive troponin I, B-type natriuretic peptide, etc. Coagulation indexes included fibrinogen, activated partial thromboplastin time, prothrombin time, thrombin time, international standardized ratio, DD dimer, etc. Other indicators included C-reactive protein, hypersensitive C-reactive protein, interleukin-6, procalcitonin, and blood glucose.

Multiple analysis revealed that the levels of many indicators were significantly different within diverse severity circumstances. For example, the median of neutrophil percentage and neutrophil

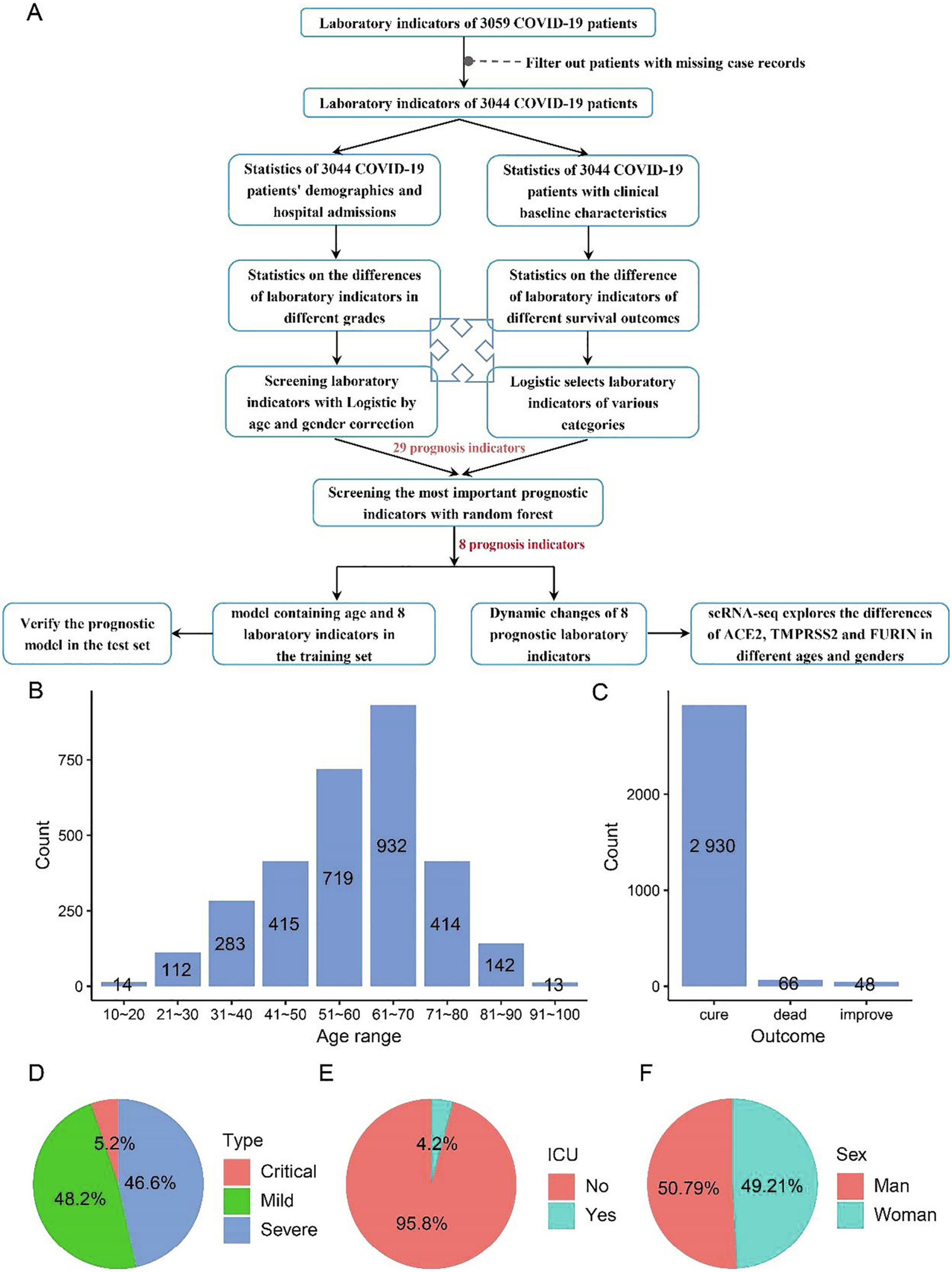


Fig. 1. The research flow and overall distribution of 3044 COVID-19 patients. A: The roadmap of research. B: Age distribution of COVID-19 patients. C: Treatment outcome chart of COVID-19 patients. D: Ratio chart of the highest historical classification (mild, severe, critical) of COVID-19 patients. E: Scale diagram of COVID-19 patients staying in the ICU. F: Sex ratio chart of COVID-19 patients.

Table 1
Baseline characteristics of 3044 COVID-19 patients.

Total. NO. (Highest classification)	Total (n = 3044)	Mild (n = 1467)	Severe (n = 1418)	Critical (n = 159)	*p-value*
Demographics and characteristics					
Age, median [IQR]	60.0 (49.0–68.0)	56.0 (45.0–65.0)	63.0 (53.0–71.0)	68.0 (61.5–76.5)	p < 0.001
Sex					
Female	1498 (49.21%)	719 (49.01%)	722 (50.91%)	57 (35.85%)	0.002
Male	1546 (50.78%)	748 (50.99%)	696 (49.08%)	102 (64.15%)	
Stay in ICU	127 (4.17%)	1 (0.07%)	23 (1.62%)	103 (64.78%)	p < 0.001
State of Death	66 (2.17%)	0 (0.00%)	4 (0.35%)	61 (38.36%)	p < 0.001
Basic comorbidity					
Hypertension	935 (30.72%)	366 (24.95%)	492 (34.7%)	77 (48.43%)	p < 0.001
Diabetes	435 (14.29%)	169 (11.52%)	224 (15.8%)	42 (26.42%)	p < 0.001
Coronary atherosclerosis	165 (5.42%)	50 (3.41%)	98 (6.91%)	17 (10.69%)	p < 0.001
Tumor	61 (2.00%)	15 (1.02%)	38 (2.68%)	8 (5.03%)	p < 0.001
Chronic obstructive pulmonary disease	30 (0.99%)	5 (0.34%)	21 (1.48%)	4 (2.52%)	0.001
Hyperlipidemia	23 (0.76%)	11 (0.75%)	12 (0.85%)	0 (0%)	0.505
Abnormal liver function	59 (1.94%)	27 (1.84%)	27 (1.9%)	5 (3.14%)	0.522
Gastritis	36 (1.18%)	17 (1.16%)	17 (1.20%)	2 (1.26%)	0.991
Cirrhosis	16 (0.53%)	4 (0.27%)	11 (0.71%)	1 (0.63%)	0.245
Hepatitis	55 (1.81%)	30 (2.04%)	24 (1.69%)	1 (0.63%)	0.403
Nephritis	7 (0.23%)	3 (0.2%)	3 (0.21%)	1 (0.63%)	0.558
Benign prostatic hyperplasia	32 (1.05%)	6 (0.41%)	22 (1.55%)	4 (2.52%)	0.002
Prostatitis	4 (0.13%)	1 (0.07%)	3 (0.21%)	0 (0%)	0.509
Asthma	8 (0.26%)	3 (0.2%)	4 (0.28%)	1 (0.63%)	0.599
Complications					
Respiratory failure	52 (1.71%)	0 (0%)	16 (1.13%)	46 (28.93%)	p < 0.001
Acute respiratory distress syndrome	24 (0.79%)	0 (0%)	2 (0.14%)	22 (13.84%)	p < 0.001
Abnormal kidney function	26 (0.85%)	6 (0.41%)	17 (1.20%)	3 (1.89%)	0.024
Heart failure	14 (0.46%)	2 (0.13%)	6 (0.42%)	6 (3.77%)	p < 0.001
Venous thrombosis	10 (0.33%)	3 (0.20%)	5 (0.35%)	2 (1.25%)	0.079
Thrombocytopenia	20 (0.66%)	5 (0.34%)	11 (0.78%)	4 (2.51%)	0.010
Days in hospital, median [IQR]	13.0 (8–19)	12.0 (8–17)	14.0 (8–22)	19.0 (11–32)	p < 0.001
No additional diseases	1386 (45.53%)	751 (51.19%)	604 (42.60%)	31 (19.50%)	p < 0.001
Effects on clinical outcome adjusting for age and gender					
Basic comorbidity	**OR**	**$\log_2$OR**	**95% CI lower**	**95% CI upper**	***p-value***
Hypertension	1.483	0.394	1.081	2.029	0.014
Diabetes	1.557	0.443	1.076	2.216	0.016
Coronary atherosclerosis	1.174	0.161	0.671	1.947	0.553
Cancer	2.315	0.839	1.068	4.570	0.022
Chronic obstructive pulmonary disease	0.822	−0.196	0.233	2.241	0.728
Hyperlipidemia	0	−13.76	NA	NA	0.978
Abnormal liver function	1.573	0.453	0.530	3.758	0.355
Abnormal renal function	1.313	0.272	0.303	3.950	0.667
Gastritis	0.778	−0.252	0.124	2.648	0.735
Cirrhosis	1.918	0.651	0.288	7.507	0.411
Hepatitis	0.651	−0.430	0.105	2.173	0.559
Nephritis	2.426	0.886	0.123	15.895	0.430
Benign prostatic hyperplasia	0.919	−0.084	0.264	2.461	0.880
Prostatitis	0	−11.778	NA	NA	0.978
Asthma	2.605	0.957	0.132	16.834	0.393

Note: IQR: The 25% and 75% quantiles. ICU: intensive care unit. OR: odds ratio. $\log_2$OR: log2 (odds ratio). 95% CI lower: The lower of 95% confidence interval. 95% CI upper: The upper of 95% confidence interval.

count increased with the disease grade (60.2vs. 63.35vs. 82.4, 3.3vs. 3.57vs. 6.27, p < 0.001, Table S2), but the lymphocyte percentage and lymphocyte count gradually decreased (29vs. 25.4vs. 10.3, 1.58vs. 1.43vs. 0.82, p < 0.001, Table S2). Various electrolytes such as sodium, potassium and calcium had large disturbances among patients with different grades (Table S2). Remarkably, aspartate aminotransferase and alkaline phosphatase related to liver function gradually increased with disease level upgraded (p < 0.001, Table S2), and total protein and albumin decreased gradually (p < 0.001, Table 2), whereas alanine aminotransferase and indirect bilirubin had no significant changes (p = 0.670, p = 0.340, Table S2). Urine test-related cystatin C increased as the progression of infection (0.9vs. 0.96vs. 1.08, p < 0.001, Table S2). Creatinine in renal function also increased with the worsen situation of disease (0.9vs. 0.96vs. 1.08, p < 0.001, Table S2). Furthermore, the amount of many other indicators were more highly increased in more severe situations. For example, myoglobin and B-type natriuretic peptide were related to cardiac function, while fibrinogen and increase in D-Dimer indicates new blood coagulation. Notably, C-reactive protein, interleukin-6, procalcitonin and blood glucose showed a significant increase in severe and critical patients (p < 0.001, Table S2). Similar changes in all above indicators were also observed in both ICU and non-ICU groups (Table S3).

By comparing laboratory indicators in patients with different survival outcomes and severity classification, we found that their trends were not identical. Some usual prognostic indicators such as neutrophils, interleukin-6, D-Dimer, and C-reactive protein increased markedly, while lymphocytes, eosinophils, total protein, and albumin abnormally decreased in both critical and dead patients (p < 0.001, Table S2 and Table S4). In addition, sodium, chloride, fibrinogen, globulin and other indicators had significant differences between different grades (p < 0.001, Table S2) but not the clinical outcomes (p > 0.05, Table S4). Taken together, we suggested that the disturbance of these indicators may be related to the disease progression but not survival rate because they are not obvious lethal factor.

Table 2
The impact of different laboratory test indicators on clinical death and critical illness outcomes.

Laboratory testing index	OR	log_2OR	95% CI lower	95% CI upper	*p-value*	Total
Neutrophil percentage	1.136	0.127	1.118	1.155	$p < 0.001$	2976
Neutrophil absolute value	1.466	0.382	1.387	1.553	$p < 0.001$	2976
Basophil percentage	0.014	−4.260	0.006	0.035	$p < 0.001$	2976
Absolute value of basophil	0.000	−48.688	0.000	0.000	$p < 0.001$	2976
Eosinophil percentage	0.539	−0.618	0.462	0.622	$p < 0.001$	2976
Eosinophil absolute value	0.003	−5.689	0.000	0.023	$p < 0.001$	2976
Monocyte percentage	0.657	−0.420	0.610	0.705	$p < 0.001$	2976
Monocyte absolute value	0.937	−0.065	0.415	2.003	0.872	2976
Lymphocyte percentage	0.854	−0.157	0.836	0.873	$p < 0.001$	2976
Lymphocyte absolute value	0.150	−1.895	0.103	0.217	$p < 0.001$	2976
Blood leukocytes	1.349	0.299	1.282	1.422	$p < 0.001$	2976
Red blood cells	0.803	−0.212	0.600	1.079	0.1424	2976
Potassium	1.113	0.107	0.833	1.474	0.4623	2851
sodium	0.966	−0.034	0.925	1.009	0.126	2851
chlorine	0.903	−0.102	0.867	0.942	$p < 0.001$	2851
calcium	0.002	−6.116	0.001	0.008	$p < 0.001$	2850
phosphorus	0.193	−1.644	0.091	0.406	$p < 0.001$	2390
Serum magnesium	18.143	2.898	3.165	105.618	0.0012	2389
Alanine aminotransferase	1.004	0.004	1.001	1.008	0.0184	2898
Aspartate aminotransferase	1.007	0.007	1.003	1.012	0.0072	2907
Total protein	0.918	−0.085	0.894	0.942	$p < 0.001$	2901
albumin	0.800	−0.223	0.769	0.832	$p < 0.001$	2901
Total bilirubin	1.038	0.037	1.020	1.057	$p < 0.001$	2900
Direct bilirubin	1.080	0.077	1.046	1.124	$p < 0.001$	2900
Total bile acid	0.971	−0.029	0.939	0.998	0.064	2899
Indirect bilirubin	1.050	0.049	1.010	1.094	0.017	2378
globulin	1.014	0.014	0.976	1.051	0.475	2380
Alkaline phosphatase	1.008	0.008	1.005	1.011	$p < 0.001$	2899
γ-glutamyl transpeptidase	1.005	0.005	1.003	1.007	$p < 0.001$	2899
Cystatin C	1.837	0.608	1.431	2.390	$p < 0.001$	2894
PH	1.002	0.002	0.750	1.332	0.99	2313
Urine red blood cells	1.001	0.001	1.001	1.002	0.004	2377
Urine leukocyte	1.000	0.000	1.000	1.001	0.184	2378
Urea nitrogen	1.250	0.223	1.187	1.319	$p < 0.001$	2903
Creatinine	1.003	0.003	1.001	1.005	0.0023	2903
Uric acid	0.996	−0.004	0.994	0.998	$p < 0.001$	2898
Total carbon dioxide	0.962	−0.039	0.912	1.015	0.1551	2897
Creatine kinase	1.003	0.003	1.001	1.004	$p < 0.001$	2847
Lactate dehydrogenase	1.010	0.010	1.009	1.012	$p < 0.001$	2850
alpha-hydroxybutyrate dehydrogenase	1.012	0.012	1.010	1.013	$p < 0.001$	2850
Creatine kinase isoenzyme	1.006	0.006	1.000	1.017	0.0821	2846
Myoglobin	1.008	0.008	1.005	1.012	$p < 0.001$	1270
Hypersensitive troponin I	1.525	0.422	1.168	2.353	0.012	1276
B-type natriuretic peptide	1.001	0.001	1.001	1.002	$p < 0.001$	1638
Fibrinogen	1.064	0.062	0.895	1.222	0.385	2528
Activated partial thromboplastin time	1.046	0.045	1.021	1.078	0.001	2529
Prothrombin time	1.328	0.284	1.224	1.445	$p < 0.001$	2529
Thrombin time	1.186	0.170	1.100	1.291	$p < 0.001$	2529
International standardized ratio	27.539	3.316	10.362	75.970	$p < 0.001$	2529
DD dimer	1.265	0.235	1.200	1.339	$p < 0.001$	2510
C-reactive protein	1.023	0.023	1.020	1.027	$p < 0.001$	2926
Hypersensitive C-reactive protein	1.227	0.204	1.179	1.278	$p < 0.001$	2923
Interleukin-6	1.012	0.012	1.008	1.017	$p < 0.001$	1472
Procalcitonin	1.881	0.632	1.368	2.899	0.002	2018
Glucose	1.252	0.225	1.184	1.326	$p < 0.001$	2900

Note: OR: odds ratio. log_2OR: log2 (odds ratio). 95% CI lower: The lower of 95% confidence interval.
95% CI upper: The upper of 95% confidence interval.

3.4. Logistic regression analyzes the impact of laboratory indicators on the critical and death

We further assessed the contribution of the 54 laboratory indicators above to the clinical critical illness and survival. Critical patients accounted for 92.4% (61/66) in deaths and 81.1% (103/123) in ICU patients, indicating a great overlap between ICU, dead and critical patients. Thus we here defined a composite endpoint event. Specifically, the composite endpoint event represents the died, critical and ICU patients as event occurrences, while the rest as no event occurrences. Then, we set age and gender as covariate corrections, and used multivariate logistic regression to calculate the impact of various indicators, which turned out that multiple laboratory indicators affected the prognosis and disease progression of patients. For example, neutrophils percentage (OR = 1.136, $p < 0.001$), neutrophil count (OR = 1.466, $p < 0.001$), white blood cells amount (OR = 1.349, $p < 0.001$), cystatin C level (OR = 1.837, $p < 0.001$), D-Dimer level (OR = 1.265, $p < 0.001$), interleukin-6 level (OR = 1.012, $p < 0.001$), C-reactive protein level (OR = 1.023, $p < 0.001$), blood glucose level (OR = 1.252, $p < 0.001$) and other multiple detection indicators are risk indicators of clinical critical illness and death (Table 2). Conversely, lymphocyte percentage (OR = 0.854, $p < 0.001$), lymphocyte count (OR = 0.150, $p < 0.001$), eosinophil percentage (OR = 0.539, $p < 0.001$), albumin level (OR = 0.800) are all protective factors against clinical critical illness and death (Table 2). Some indicators such as monocyte count (OR = 0.937, p = 0.872), red blood cells amount (OR = 0.803, p = 0.142), pH (OR = 1.002, p = 0.990) and creatine kinase isoenzyme (OR = 1.006, p = 0.082) are not the main factors correlating with clinical critical illness and death (Table 2). In short, we found that 44 indicators (p < 0.05) might affect the patient's disease process and survival outcomes (Table 2).

3.5. Effects of different categories of laboratory indicators on critical illness and death

In order to further explore the main indicators in each category, a multi-factor stepwise regression analysis was carried out on each category, including blood routine test, electrolytes, urine tests, and function assessments of kidney, liver, heart and blood coagulation. In terms of cell ratio, we found that white blood cells amount (OR = 1.063, p = 0.026) and neutrophil percentage (OR = 1.130, p < 0.001) were the main risk factors (Table 3) that can promote disease progression. On the absolute level, lymphocytes (OR = 0.215, p < 0.001) and neutrophils (OR = 1.415, p < 0.001) were the main prognostic factors which can inhibit virus infection and increase inflammation, respectively (Table 3). At the electrolyte level, although differences were significant in potassium, sodium, and magnesium (p < 0.001), the regression coefficient of the overall model was not significant (p = 0.090, Table 3). In renal function assessments, cystatin C (OR = 3.782, p < 0.001) and urine red blood cells (OR = 1.001, p = 0.006) were the main risk factors while urea nitrogen (OR = 1.526, p < 0.001), creatinine (OR = 0.994, p < 0.001) and uric acid (OR = 0.991, p < 0.001) played more important roles in urine examination (Table 3). Among the liver function indexes, aspartate aminotransferase, albumin, alkaline phosphatase, etc. were the main prognostic factors. Meanwhile, we found that alanine aminotransferase (OR = 0.990, p = 0.01), globulin (OR = 1.042, p = 0.041) and other indicators were significantly independent of age and gender (Table 3 vs. Table 2). These findings proposed that these indicators are greatly affected by age and gender or they are not sufficiently robust as prognostic indicators. Among cardiac-related indicators, lactate dehydrogenase (OR = 1.013, p < 0.001), myoglobin (OR = 1.012, p < 0.001) and creatine kinase (OR = 0.995, p < 0.001) were the main prognostic factor (Table 3) while prothrombin time (OR = 1.187, p < 0.001), fibrinogen (OR = 1.277, p < 0.0141), thrombin time (OR = 1.099, p < 0.017) and D-Dimer (OR = 1.255, p < 0.001) were the main prognostic factor in coagulation parameters (Table 5). The remaining index items, including C-reactive protein (OR = 1.019, p < 0.001), interleukin-6 (OR = 1.014, p < 0.001), procalcitonin (OR = 2.362, p < 0.001) and blood glucose (OR = 1.227, p < 0.001) had nothing to do with clinical critical illness and death (Table 3). Finally, we found that 34 laboratory indicators could serve as independent prognostic signatures (Table 3).

3.6. Random forest screening of prognostic indicators causing critical illness and death

Monitoring such large amounts of laboratory indicators is a heavy burden for clinical doctors in anti-virus therapy. Therefore, 29 significant prognostic indicators obtained from 491 patients were selected and further tested in random forest machine learning algorithms at the same time. Interestingly, they could clearly distinguish the event group (Critical or ICU or Dead) and non-event group according to the principal component results (Fig. 2A). Moreover, 5 times 10-fold cross-validation was used to screen the best number of variables included in the model, and eight turns out to be the most suitable for its the smallest error (Fig. 2B). Combined with the importance of indicators given by the random forest algorithm (Fig. 2C), eight indicators were selected as the final prognostic indicators including neutrophil percentage, procalcitonin, neutrophil absolute value, C-reactive protein, albumin, interleukin-6, lymphocyte absolute value and myoglobin due to their significant differences in different disease grades, survival outcomes and ICU grouping (Fig. S1).

More importantly, these 8 prognostic indicators at different times in the event and non-event patients showed stable and significant differences. In particular, neutrophil percentage, procalcitonin, neutrophil absolute value, C-reactive protein, myoglobin and interleukin-6 in patients with compound endpoint events were always higher than the non-event group (Fig. 2D). On the opposite, these protective factors, such as lymphocyte count and albumin obtained from patients with a composite endpoint event, were always lower than those without a composite endpoint event. Hence, these 8 laboratory testing indicators indeed be treated as the prognostic factor of patients, because they were significantly different in both the critical and mild groups from onset to a long time before the end event (Fig. 2D).

3.7. Establishing a COVID-19 patient prediction model

In order to assist doctors in defining patients who are more likely to be critically ill or even die, we here combined age and eight prognostic indicators presented above to establish a clinically available prognostic model. Prior to that procedure, patients were divided into normal and abnormal groups according to eight prognostic indicators and the cumulative event rate was counted between the two groups. The cumulative event rates in the abnormal risk factor group were significantly higher than the non-abnormal group (p < 0.001, Fig. 3A ~ F). Similarly, the cumulative event rate in the abnormal protection factor group was significantly higher than the healthy group (p < 0.001, Fig. 3G and p < 0.001, Fig. 3H).

In order to further analyze the clinical reliability of the model in the 491 samples training set, logistic regression was applied to construct a joint model which contains age and the selected eight prognostic indicators. Receiver operating characteristic curve (ROC) indicated that the model had a good AUC value of 0.878 (95% CI: 0.829–0.927) (Fig. 3I). The model regression equation for calculating the probability of event occurrence was as follows: $y = 1 / (1 + e^{-z})$, z = -4.038 + 0.051 * (neutrophil percentage) + 0.763 * (procalcitonin) + 0.003 * (mymyoglobin) + 0.128 * (neutrophil absolute) + 0.005 * (C reactive protein) + 0.003 * (interleukin 6) – 0.148 * (lymphocyte absolute) – 0.089 * (albumin) + 0.015 * (age). This model was further verified in another independent cohort containing 170 patients and its AUC value reached 0.897 (95% CI: 0.787–1.000) (Fig. 3J). These results clearly demonstrated that this model has great robustness. Finally, a nomogram containing all 611 patients was drawn to facilitate clinical use and explain the relationship between model variabilities, which may not only query the risk scores of patients' various model indicators conveniently, but also predict the risk of disease progression and death in patients according to the sum of scores.

3.8. Analyses of lung single cell and full transcriptome data of different ages and genders

During the analysis, we noticed that age and gender are always important factors leading to critical illness and death compared with various testing indicators. Therefore, we compared three key genes ACE2, TMPRSS2 and FURIN, which were related to virus infection at both single cell and whole tissue levels under different ages and genders [15–17]. As shown in Fig. 4, 13 cell populations were identified from 8 normal lungs (Fig. 4A ~ B, Table S5). ACE2 was mainly expressed in alveolar epithelial type 2 cells (AT2), basal cells and tuft cells (Fig. 4C). Based on our analysis, the expression of ACE2 was detected in limited cells. In AT2 subpopulation, <1% of

Table 3
Independent prognostic factors for laboratory inspection indicators of various categories.

Immune cell percentage	OR	log_2OR	95% CI lower	95% CI upper	*p-value*
(Intercept)	0.000	(10.812)	0.000	0.000	$p < 0.001$
Red blood cells	0.783	(0.245)	0.598	1.023	0.073
Blood leukocytes	1.063	0.061	1.010	1.124	0.026
Neutrophil percentage	1.130	0.122	1.110	1.151	$p < 0.001$
Immune cell absolute value					
(Intercept)	0.121	(2.111)	0.068	0.215	$p < 0.001$
Lymphocyte absolute value	0.215	(1.535)	0.147	0.310	$p < 0.001$
Monocyte absolute value	0.424	(0.859)	0.165	1.022	0.066
Neutrophil absolute value	1.415	0.347	1.330	1.510	$p < 0.001$
Electrolyte					
(Intercept)	271.269	5.603	0.448	188182.081	0.090
Serum magnesium	72.921	4.289	11.323	482.399	$p < 0.001$
phosphorus	0.170	(1.772)	0.086	0.331	$p < 0.001$
chlorine	0.840	(0.174)	0.791	0.892	$p < 0.001$
calcium	0.001	(6.881)	0.000	0.004	$p < 0.001$
Potassium	1.747	0.558	1.285	2.366	$p < 0.001$
sodium	1.154	0.143	1.079	1.234	$p < 0.001$
Renal function					
(Intercept)	0.014	(4.263)	0.009	0.022	$p < 0.001$
Cystatin C	3.782	1.330	2.683	5.485	$p < 0.001$
Urine red blood cells	1.001	0.001	1.000	1.002	0.006
Urine test					
(Intercept)	0.097	(2.334)	0.059	0.158	$p < 0.001$
Urea nitrogen	1.526	0.423	1.431	1.632	$p < 0.001$
Creatinine	0.995	(0.006)	0.991	0.998	0.001
Uric acid	0.991	(0.009)	0.989	0.993	$p < 0.001$
liver function					
(Intercept)	33.286	3.505	6.222	182.502	$p < 0.001$
Alanine aminotransferase	0.990	(0.010)	0.982	0.997	$p < 0.001$
Aspartate aminotransferase	1.016	0.016	1.006	1.026	0.001
albumin	0.797	(0.227)	0.764	0.829	$p < 0.001$
Total bilirubin	0.948	(0.053)	0.882	1.017	0.142
Direct bilirubin	1.288	0.253	1.122	1.485	$p < 0.001$
Total bile acid	0.904	(0.100)	0.875	0.930	$p < 0.001$
globulin	1.042	0.041	1.001	1.083	0.041
Alkaline phosphatase	1.007	0.007	1.003	1.012	0.001
Heart function					
(Intercept)	0.009	(4.709)	0.005	0.016	$p < 0.001$
Creatine kinase	0.995	(0.005)	0.993	0.998	0.001
Lactate dehydrogenase	1.013	0.013	1.010	1.015	$p < 0.001$
Creatine kinase isoenzyme	0.970	(0.031)	0.937	0.999	0.059
Myoglobin	1.012	0.012	1.008	1.017	$p < 0.001$
Hypersensitive troponin I	0.614	(0.488)	0.269	0.998	0.260
Coagulation index					
(Intercept)	0.001	(7.562)	0.000	0.002	$p < 0.001$
Prothrombin time	1.187	0.171	1.105	1.289	$p < 0.001$
Fibrinogen	1.277	0.244	1.090	1.585	0.014
Thrombin time	1.099	0.094	1.033	1.195	0.017
DD dimer	1.255	0.227	1.189	1.330	$p < 0.001$
Interleukin					
(Intercept)	0.058	(2.841)	0.046	0.073	$p < 0.001$
Interleukin-6	1.014	0.014	1.010	1.019	$p < 0.001$
C-reactive protein					
(Intercept)	0.017	(4.103)	0.011	0.023	$p < 0.001$
C-reactive protein	1.018	0.018	1.014	1.022	$p < 0.001$
Hypersensitive C-reactive protein	1.136	0.127	1.088	1.188	$p < 0.001$
Procalcitonin					
(Intercept)	0.076	(2.575)	0.064	0.090	$p < 0.001$
Procalcitonin	2.363	0.860	1.562	3.857	$p < 0.001$
carbohydrate					
(Intercept)	0.019	(3.966)	0.014	0.026	$p < 0.001$
Glucose	1.227	0.205	1.175	1.282	$p < 0.001$

Note: OR: odds ratio. log_2OR: log2(odds ratio). 95% CI lower: The lower of 95% confidence interval. 95% CI upper: The upper of 95% confidence interval.

this subpopulation were detected containing the expressed ACE2 with low expression level. In addition, we found that TMPRSS2 and FURIN could promote the binding of SARS-CoV-2 to ACE2, which were also mainly expressed in AT2 cells [15–17] (Fig. 4D ~ 4E). The average expression level of ACE2 and the cell percent expressing it in the old group (age: 55 years, 63 years, 57 years) were higher than the young group (age: 21 years, 22 years, 29 years) (Fig. 4F). Besides, we also found that higher percentage of cells in older patients express TMPRSS2 and FURIN (Fig. 4F), even though their expression levels in elder group were lower than the young group. The results of single-cell analysis partially explain the differences of the infection rate and mortality between people at different ages, which may not only be related to the expression of ACE2, TMPRSS2 and FURIN, but also the number of cells with the expression of these three key genes. However, in the analysis of the whole transcriptome from TCGA, there was

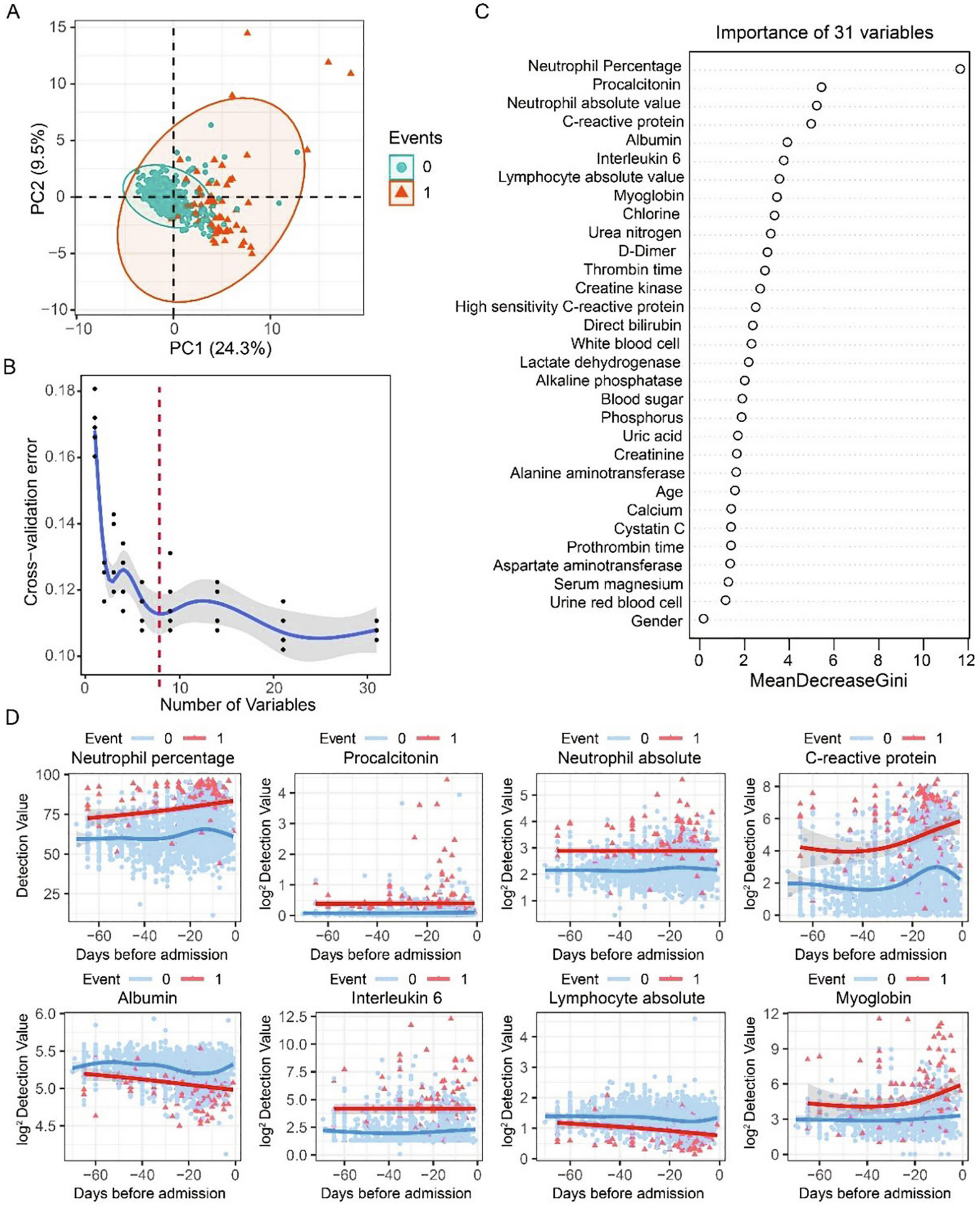
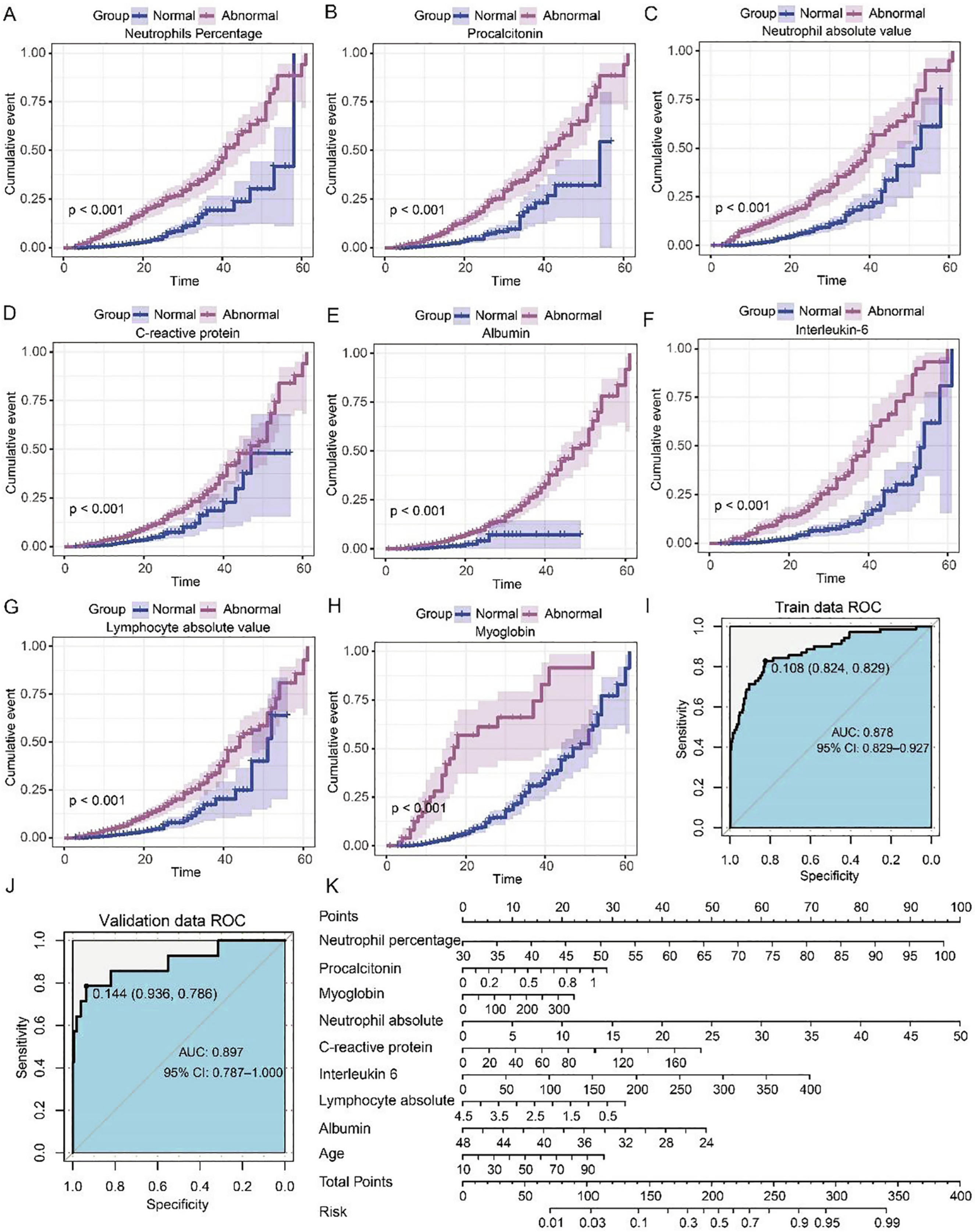


Fig. 2. Screening the most important prognostic indicators through random forest machine learning algorithms. A: Principal component analysis chart of age, gender and 29 laboratory test indexes. B: The 5 times 10-fold cross-validation curve shows the relationship between the model error and the number of variables used for fitting. C: The ranking of the importance of 31 prognostic indicators calculated based on the random forest algorithm. MeanDecreaseGini represents the influence of each variable on the heterogeneity of the observations on each node of the classification tree. The larger the value, the greater the importance of the variable. D: The overall dynamic changes of the eight most important prognostic indicators at different time points before the end event. Red and blue lines represent the fit curve. 1 represents the composite endpoint event group in which patients developed into critical illness or death or entered the ICU while 0 represents the group without composite endpoint events where patients have milder symptoms and better treatment effects. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)


A
Group Normal Abnormal
Neutrophils Percentage
Cumulative event
Time
p < 0.001
B
Group Normal Abnormal
Procalcitonin
Cumulative event
Time
p < 0.001
C
Group Normal Abnormal
Neutrophil absolute value
Cumulative event
Time
p < 0.001
D
Group Normal Abnormal
C-reactive protein
Cumulative event
Time
p < 0.001
E
Group Normal Abnormal
Albumin
Cumulative event
Time
p < 0.001
F
Group Normal Abnormal
Interleukin-6
Cumulative event
Time
p < 0.001
G
Group Normal Abnormal
Lymphocyte absolute value
Cumulative event
Time
p < 0.001
H
Group Normal Abnormal
Myoglobin
Cumulative event
Time
p < 0.001
0.00 0.25 0.50 0.75 1.00
0 20 40 60
I
Train data ROC
0.108 (0.824, 0.829)
AUC: 0.878
95% CI: 0.829–0.927
Sensitivity
Specificity
J
Validation data ROC
0.144 (0.936, 0.786)
AUC: 0.897
95% CI: 0.787–1.000
Sensitivity
Specificity
0.0 0.2 0.4 0.6 0.8 1.0
1.0 0.8 0.6 0.4 0.2 0.0
K
Points 0 10 20 30 40 50 60 70 80 90 100
Neutrophil percentage 30 35 40 45 50 55 60 65 70 75 80 85 90 95 100
Procalcitonin 0 0.2 0.5 0.8 1
Myoglobin 0 100 200 300
Neutrophil absolute 0 5 10 15 20 25 30 35 40 45 50
C-reactive protein 0 20 40 60 80 120 160
Interleukin 6 0 50 100 150 200 250 300 350 400
Lymphocyte absolute 4.5 3.5 2.5 1.5 0.5
Albumin 48 44 40 36 32 28 24
Age 10 30 50 70 90
Total Points 0 50 100 150 200 250 300 350 400
Risk 0.01 0.03 0.1 0.3 0.5 0.7 0.9 0.95 0.99

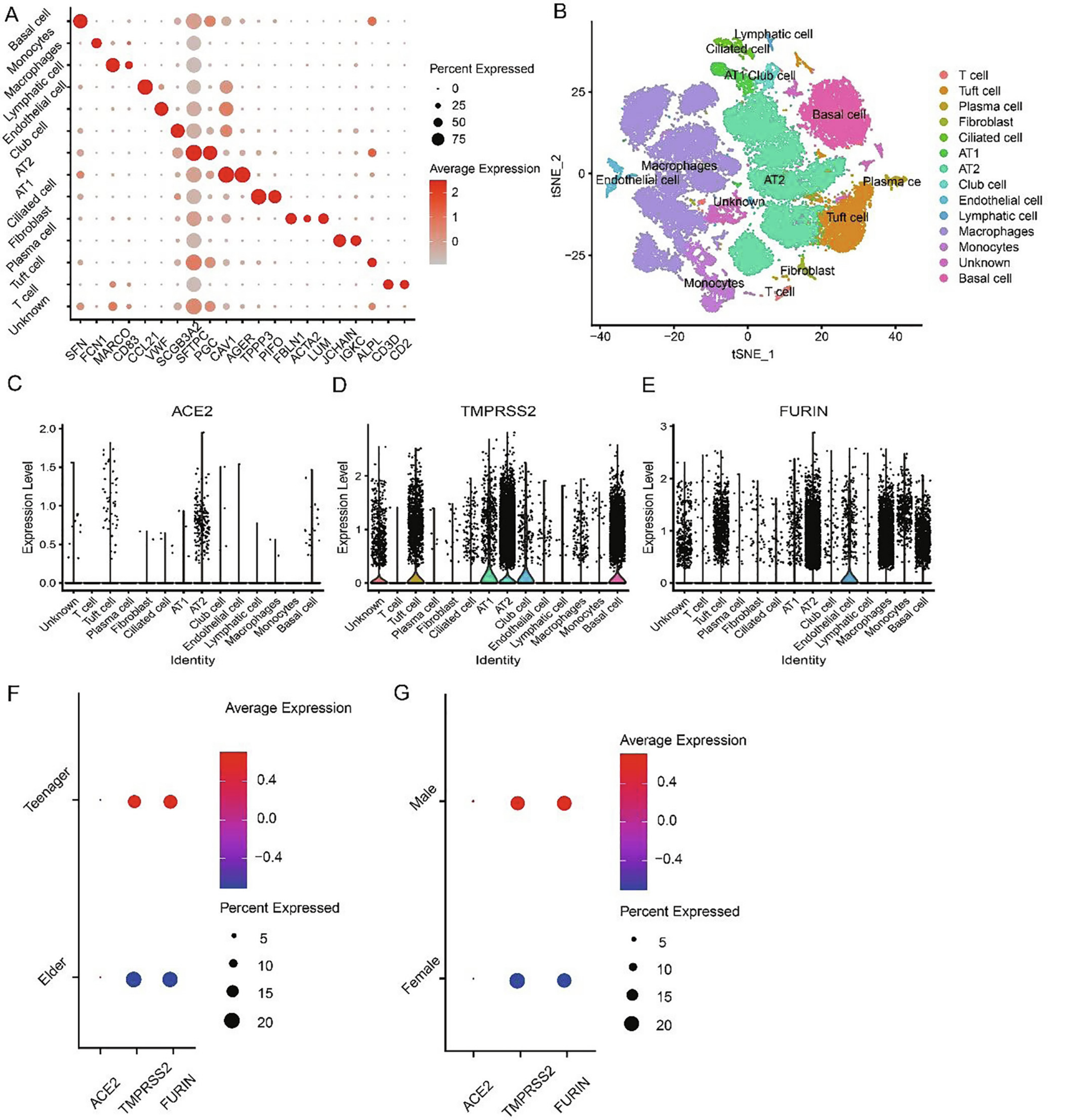


Fig. 4. Analysis of lung single cell transcriptomes of different ages and genders. A: The heat map shows the marker genes corresponding to different cell types in the lung. B: The UMAP cluster map shows the clustering of different cells in the lungs. C: Expression and distribution of ACE2 in different cell clusters in the lung. D: TMPRSS2 expression and distribution in different cell clusters in the lung. E: Expression and distribution of FURIN in different cell clusters in the lung. F: Differences in expression and proportion of ACE2, TMPRSS2 and FURIN in old and young people. G: Differences in expression and proportion of ACE2, TMPRSS2 and FURIN in men and women.

Fig. 3. A joint prognostic model that can be used for clinical decision-making. A ~ H: The cumulative event rate of patients with abnormal and non-abnormal groups of 8 important prognostic indicators. The normal reference range is as follows: neutrophil percentage: 40%-75%, procalcitonin: 0–0.05 ng/ml, neutrophil absolute value: 1.8–$6.3*10^{-9}$ /L, C reactive protein: 0–4 mg/L, albumin: 40–55 g/L, interleukin 6: 0–5.90 pg/mL, lymphocyte absolute value: 1.1–$3.2*10^{-9}$ /L, myoglobin: 0–80 ng/ml. I: ROC curve of the joint model in the training set. J: Validate the ROC curve of the joint model in the concentration. K: The nomogram shows the model prediction in all 611 samples detecting 8 indicators at the same time. The line segment corresponding to each variable is marked with a scale, which represents the range of possible values of the variable, and the length of the line segment reflects the contribution of this factor to the ending event. Point in the Fig represents the individual score corresponding to each variable under different values. Total Point represents the total score of the individual scores after the values of all variables are added up. The risk probability represents the patient's probability that a composite endpoint event will occur.

no significant difference in ACE2, TMPRSS2, and FURIN in population older than 60 years old and younger than 60 years old (Fig. S2). We inferred that the tissue-wide transcriptome data masked subtle differences in key molecules of different ages, which also highlighted the advantages of single-cell transcriptome data. Considering gender, the infection rate of men was only 1.58% higher than women (Fig. 1) while the proportion of men who developed to critical illness and death was almost twice that of women (Table 1), which might also attribute to higher ACE2 expression level [18]. To verify whether there is such a difference, we compared the expression levels and cell ratios of ACE2, TMPRSS2 and FURIN of different genders based on single cell RNA sequencing. As shown in Fig. 4G, compared with females, the expression levels of ACE2, TMPRSS2 and FURIN were higher in males, and the proportion of cells expressing ACE2 and TMPRSS2 were also higher. In the bulk transcriptome, the expression of TMPRSS2 in males was significantly higher than that in female, which was consistent with the results derived from single cell level. No significant differences between ACE2 and FURIN of different genders was detected. Based on these results, the reasons of higher infection and mortality rates in male were illustrated at molecular level.

4. Discussion

It is urgent and necessary to find effective methods to predict and monitor the critical illness and death of COVID-19 patients. Regarding various clinical statistical indicators of COVID-19 patients that have been achieved in previous studies, they have some inherent flaws. For example, limited sample size and deficiency of detailed laboratory examination results made it difficult to determine the main contribution of multiple testing indicators.

In our study, the most susceptible people are concentrated between 51 and 70 years old and the average age of critically ill patients and deceased patients is higher (Table 1). Consistent with previous extensive reports, the elderly is the main population of COVID-19 [13]. In addition, some studies have reported that the difference between infection and death in elderly and young people may be related to the expression of ACE2 receptor in the body [19]. To further verify this conclusion at single cell level and bulk tissue. Our results show that the elderly at the single-cell level seems to express more ACE2 and the proportion of cells expressing ACE2 is higher than that of the young (Fig. 4F), no significant difference was observed at the bulk transcriptome level (Fig. S2). The same uncertainty appears in the results of TMPRSS2 and FURIN genes (Fig. S2). Therefore, although this may partially illustrate the difference between critical illness and death in elderly and young people, more sufficient evidences still lack to fully explain. According to that, we propose that the difference in the outcome of the elderly and young people is related to the weakened immunity and more comprehensive underlying diseases accompanied with increasing age. Our research has proved that patients with underlying diseases have a higher critical illness ratio and mortality (Table 1, $p < 0.001$ and Table S1). From the perspective of gender, compared with the 39.7% female infection rate in the United States [14], the infection rate of men is only 1.58 percentage points higher than that of women (Fig. 1F). However, the proportion of men who develop critical illness and death doubles compared with women (Table 1 and Table S1), which may be also attributed to higher ACE2 expression [18]. It is better to believe that the expression level difference of ACE2, TMPRSS2 and FURIN genes is the cause of different critical illness rate and death rate in different genders rather than different ages, because these can be strongly supported by the data analysis results at the single cell and bulk transcriptome level (Fig. 4G and Fig. S2). In terms of comorbidities, hypertension, diabetes, coronary heart disease, tumors, and chronic obstructive pulmonary disease are prone to critical illness and clinical outcomes (Table 1 and Table S1), which is consistent with previous reports [14,20]. From the patient's treatment outcome, the number of patients cured and improved reached 2644 and 281 respectively (Fig. 1B) and the number of deaths was only 66, accounting for 2.17% (Fig. 1B). This shows that timely and active medical treatment is essential to curb the mortality of COVID-19 patients.

A large number of disorders are presented in COVID-19 patients with death and critical illness regarding laboratory indicators. Compared with mild and severe ill patients, those critical and dead ones have obvious abnormalities in the immune system, kidney function, liver function, heart function, blood coagulation indexes and inflammatory factors. In order to facilitate clinical monitoring and supervision, we further found the 8 most important prognostic indicators. Among them, the increase of neutrophil percentage, neutrophil absolute value and interleukin-6 indicated that inflammation and inflammatory storm are some of the main manifestations of critical symptoms and death. The decrease of lymphocyte absolute value represents a decrease in the immunity of critically ill and dead patients, resulting in the inability to defend against the combined infection and sepsis represented by the increase in procalcitonin and C-reactive protein. In addition, the disorders of myoglobin and albumin are related to impaired heart and liver function, suggesting that many important organs of critical and dead patients have been damaged. Compared with mild and severe patients, the values of these eight indicators are always in a higher state before the end event, and patients with abnormal indicators are more likely to have composite endpoint events (Fig. 2D and Fig. 3). The model constructed by the combined age and gender of the patients and the eight detection indicators has good accuracy in the training set and validation set with the AUC values for 0.878 and 0.897, respectively (Fig. 3I ~ J). Finally, we establish a clinically useful regression equation and nomogram to predict the risk probability of developing critical illness and death (Fig. 3K). We believe that this model has practical significance for the prediction and monitoring of COVID-19 patients.

In summary, we performed a statistical analysis of 3044 COVID-19 patients to find the eight most important prognostic factors (neutrophil percentage, procalcitonin, neutrophil absolute value, C-reactive protein, albumin, interleukin-6, lymphocyte absolute value and myoglobin) of COVID-19, and constructed a model to predict the prognosis of patients, which is of great significance for the management and monitoring of COVID-19. Moreover, through reanalyzing public lung single-cell and bulk transcriptome data, we suggest that compared with different ages, different genders have different critical illness rates and mortality are more likely to be attributed to differences in key genes such as ACE2, TMPRSS2, and FURIN.

However, our study still has many limitations. First, our established prediction model still lacks an effective validation from external queues, which may result in over-fitting of the model to a certain extent. Therefore, in the further study, we suggest that it is important to integrate multiple queues for modeling and validation. One model can only withstand validation from multiple external cohorts, it can be applied to a complex COVID-19 patient population. Second, our model construction is mainly based on random forest and logistic regression algorithm, which may have certain deficiencies. In fact, a comprehensive comparison of the results of the multiple algorithms will deepen our impression of the key prognostic factors and models. Support vector machine, Adaptive Boosting, neural network and artificial intelligence algorithms are good choices. Third, our current study has not been able to analyze the genetic background of these differences in laboratory test indicators. This is mainly because we lack the genetic information of these patients. We believe that the integration of laboratory indicators and genetic information such as genomes, transcriptome and proteome will greatly broaden our understanding of COVID-19. We also hope that future studies will pay attention to the output of these data.

5. Statement of Ethics

All COVID-19 patients were implemented in accordance with the New Coronavirus Pneumonia Diagnosis and Treatment Plan (7th edition) issued by the National Health Commission of China. According to the WHO/International Severe Acute Respiratory and Emerging Infection Consortium case record form for severe acute respiratory infections, baselines of participants, epidemiological and clinical manifestations, laboratory findings and outcome data were extracted from electronic medical records. Major basic information (i.e., age, sex, the highest historical classification, preliminary diagnosis, discharge diagnosis and discharge conditions) were collected except for patients' personal information (e.g., name and ID) and comorbidities were also included in clinical symptoms. The research was approved by the Research Ethics Commission of Huoshenshan hospital. Written informed consent was obtained from each patient.

Funding

Key Foundation of Wuhan Huoshenshan Hospital (2020 [18]), Key Research& Development Program of Jiangsu Province (BE2018713), Medical Innovation Project of Logistics Service (18JS005), The Foundation of Jiangsu Population Association (JSPA2019017) and Independent research projects of provincial public welfare research institutes (BM2018033-1).

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

We thank our colleagues for their suggestions and criticisms on the manuscript.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.csbj.2021.01.042.

References

[1] Gudbjartsson DF, Helgason A, Jonsson H, et al. Spread of SARS-CoV-2 in the Icelandic population. N Engl J Med 2020;382(24):2302–15.
[2] Baud D, Qi X, Nielsen-Saines K, Musso D, Pomar L, Favre G. Real estimates of mortality following COVID-19 infection. Lancet Infect Dis 2020 Jul;20(7):773.
[3] Wang YS, Kang HYJ, Liu XF, Tong ZH. Combination of RT-qPCR testing and clinical features for diagnosis of COVID-19 facilitates management of SARS-CoV-2 outbreak. J Med Virol Jun 2020;92(6):538–9.
[4] Fang Y, Zhang H, Xie J, Lin M, Ying L, Pang P, et al. Sensitivity of chest CT for COVID-19: comparison to RT-PCR. Radiology 2020 Aug;296(2):E115–7.
[5] Wu H, Zhu H, Yuan C, et al. Clinical and immune features of hospitalized pediatric patients with coronavirus disease 2019 (COVID-19) in Wuhan, China. JAMA Netw Open Jun 1 2020;3(6):e2010895.
[6] Yang Q, Xie L, Zhang W, et al. Analysis of the clinical characteristics, drug treatments and prognoses of 136 patients with coronavirus disease 2019. J Clin Pharm Ther May 25, 2020.
[7] Zou X, Fang M, Li S, et al. Characteristics of liver function in patients with SARS-CoV-2 and chronic HBV co-infection. Clin Gastroenterol Hepatol 2020 Jun 15: S1542–3565(20)30821-1
[8] Connors JM, Levy JH. COVID-19 and its implications for thrombosis and anticoagulation. Blood 2020;135(23):2033–40.
[9] He B, Zhong A, Wu Q, et al. Tumor biomarkers predict clinical outcome of COVID-19 patients. J Infect 2020 Sep;81(3):452–82.
[10] Reyfman PA, Walter JM, Joshi N, et al. Single-cell transcriptomic analysis of human lung provides insights into the pathobiology of pulmonary fibrosis. Am J Respir Crit Care Med 2019 Jun 15;199(12):1517–36.
[11] Stuart T, Butler A, Hoffman P, et al. Comprehensive integration of single-cell data. Cell 2019;177(7):1888–1902 e21.
[12] Becht E, McInnes L, Healy J, et al. Dimensionality reduction for visualizing single-cell data using UMAP. Nat Biotechnol 2018. https://doi.org/10.1038/nbt.4314.
[13] Davies NG, Klepac P, Liu Y, et al. Age-dependent effects in the transmission and control of COVID-19 epidemics. Nat Med 2020 Aug;26(8):1205–11.
[14] Richardson S, Hirsch JS, Narasimhan M, et al. Presenting characteristics, comorbidities, and outcomes among 5700 patients hospitalized With COVID-19 in the New York City Area. JAMA 2020 May 26;323(20):2052–9.
[15] Hoffmann M, Kleine-Weber H, Schroeder S, et al. SARS-CoV-2 cell entry depends on ACE2 and TMPRSS2 and is blocked by a clinically proven protease inhibitor. Cell 2020;181(2):271–280.e8.
[16] Walls AC, Park YJ, Tortorici MA, Wall A, McGuire AT, Veesler D. Structure, function, and antigenicity of the SARS-CoV-2 spike glycoprotein. Cell 2020;181(2):281–292.e6.
[17] Wrobel AG, Benton DJ, Xu P, et al. SARS-CoV-2 and bat RaTG13 spike glycoprotein structures inform on virus evolution and furin-cleavage effects. Nat Struct Mol Biol 2020 Aug;27(8):763–7.
[18] Takahashi T, Wong P, Ellingson M, et al. Sex differences in immune responses to SARS-CoV-2 that underlie disease outcomes. medRxiv 2020 Jun 9:2020.06.06.20123414.
[19] Bunyavanich S, Do A, Vicencio A. Nasal gene expression of angiotensin-converting enzyme 2 in children and adults. JAMA 2020 Jun 16;323(23):2427–9.
[20] Liu J, Liu Y, Xiang P, et al. Neutrophil-to-lymphocyte ratio predicts critical illness patients with 2019 coronavirus disease in the early stage. J Transl Med 2020 May 20;18(1):206.