scientific reports

Check for updates

OPEN

Complement C3 identified as a unique risk factor for disease severity among young COVID-19 patients in Wuhan, China

Weiting Cheng[1], Roman Hornung[2], Kai Xu[3]✉, Cai hong Yang[3] & Jian Li[4]

Given that a substantial proportion of the subgroup of COVID-19 patients that face a severe disease course are younger than 60 years, it is critical to understand the disease-specific characteristics of young COVID-19 patients. Risk factors for a severe disease course for young COVID-19 patients and possible non-linear influences remain unknown. Data were analyzed from COVID-19 patients with clinical outcome in a single hospital in Wuhan, China, collected retrospectively from Jan 24th to Mar 27th. Clinical, demographic, treatment and laboratory data were collected from patients' medical records. Uni- and multivariable analysis using logistic regression and random forest, with the latter allowing the study of non-linear influences, were performed to investigate the clinical characteristics of a severe disease course. A total of 762 young patients (median age 47 years, interquartile range [IQR] 38–55, range 18–60; 55.9% female) were included, as well as 714 elderly patients as a comparison group. Among the young patients, 362 (47.5%) had a severe/critical disease course and the mean age was statistically significantly higher in the severe subgroup than in the mild subgroup (59.3 vs. 56.0, Student's t-test: $p < 0.001$). The uni- and multivariable analysis suggested that several covariates such as elevated levels of serum amyloid A (SAA), C-reactive protein (CRP) and lactate dehydrogenase (LDH), and decreased lymphocyte counts influence disease severity independently of age. Elevated levels of complement C3 (odds ratio [OR] 15.6, 95% CI 2.41–122.3; $p = 0.039$) are particularly associated with the risk of developing severe COVID-19 specifically in young patients, whereas no such influence seems to exist for elderly patients. Additional analysis suggests that the influence of complement C3 in young patients is independent of age, gender, and comorbidities. Variable importance values and partial dependence plots obtained using random forests delivered additional insights, in particular indicating non-linear influences of risk factors on disease severity. This study identified increased levels of complement C3 as a unique risk factor for adverse outcomes specific to young COVID-19 patients.

The pandemic caused by the coronavirus disease 2019 (COVID-19), which is associated with the severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2), has affected almost every corner of the world. As of November 10th, 2020, almost 50 million cases have been confirmed, including more than 1.2 million deaths according to a report by the World Health Organization (WHO)[1], where the numbers of cases and deaths are expected to continue to rise. The clinical spectrum of COVID-19 appears to be wide, encompassing asymptomatic infection, mild upper respiratory illness, neurological symptoms, renal and gastrointestinal complications, severe viral pneumonia with respiratory failure, multiple organ failure and even death[2–5]. Approximately 20–25% of patients will have a severe disease course.

Despite numerous studies showing a higher risk of severe COVID-19 in elderly patients, a substantial proportion of young patients also have an increased risk of developing a severe course. According to a report from the U.S. CDC, 47% of hospitalized patients are under the age of 65, as are 48% of those admitted to intensive care units (ICU)[6]. Although potential risk factors for mortality were reported to include advanced age, male gender,

[1]Oncology Department, Wuhan No.1 Hospital, Wuhan 430022, China. [2]Institute of Medical Information Processing, Biometry and Epidemiology, Ludwig-Maximilian-University Munich, Munich, Germany. [3]Department of Orthopedics, Tongji Hospital, Huazhong University of Science and Technology, Jiefang Avenue 1095, Wuhan 430030, Province Hubei, China. [4]Institute of Experimental Immunology, University Clinic of Rheinische Friedrich-Wilhelms-University, Bonn, Germany. ✉email: godocoto@163.com

presence of comorbidities, the development of a cytokine storm and an immunocompromised status[8,9], the risk factors for the development of a severe course specific to young patients (≤ 60 years old) remain under investigation. Of note, to date, a large proportion of studies applied logistic regression to analyze risk factors related to COVID-19 infection, assuming an underlying causal linear influence on the log odds[2–5]. However, given the intricate complexity of COVID-19 infection, statistical analysis with the consideration of non-linear relationships might provide more insightful information on COVID-19 related potential risk factors.

In an effort to fill these gaps, an important aim of this single-center study was to analyze clinical, demographic and treatment data of patients sequentially admitted into the Wuhan No.1 hospital, in an attempt to elucidate risk factors and main causes among young COVID-19 patients for experiencing a severe disease course. A further aim was to use the large amount of data available in this study to foster knowledge about general, age-independent, and non-linear relational risk factors for a severe disease course.

Methods

Study design and participants. This retrospective, single-center cohort study involved adult patients who were diagnosed with COVID-19 pneumonia between January 24th and March 27th, 2020, in the major government designated hospital in Wuhan: Wuhan No.1 Hospital. The date of the last follow-up was April 8th, 2020. The primary outcome was severity at the end of the study period. All patients were residents of Wuhan, and the diagnostic criteria of COVID-19 were based on the Diagnosis and Treatment Protocol for the 2019 Novel Coronavirus Pneumonia published by the National Health Commission of China.

The newly diagnosed patients were required to meet one of the following conditions: (1) positive signals of COVID-19 nucleic acids detected in fluorescent real-time RT-PCR; (2) viral gene sequencing showing a high degree of homology with the new coronavirus COVID-19. Patients with mild symptoms were required to meet the following conditions: (1) history of epidemiology; (2) fever or other respiratory symptoms; (3) CT image abnormalities typical of viral pneumonia. Patients with a severe condition met one of the following conditions: (1) shortness of breath, respiratory rate ≥ 30 breaths/min; (2) oxygen saturation (resting state) ≤ 93%; (3) PaO2/FIO2 ≤ 300 mm Hg. Critically ill patients were required to meet one of the following conditions: (1) respiratory failure requiring mechanical ventilation; (2) shock; (3) organ failure requiring ICU monitoring.

The following data were collected on admission: age, sex, symptoms from onset to hospital admission (fever, cough, dyspnea, myalgia, rhinorrhea, arthralgia, chest pain, headache, and vomiting), comorbidities (cardiovascular disease, chronic pulmonary disease, cerebrovascular disease and chronic neurological disorders, diabetes, malignancy, and smoking), vital signs (heart rate, respiratory rate, and blood pressure), laboratory values on admission (serum hemoglobin concentration, lymphocyte counts, platelet counts, diverse protein markers), treatment regime used for COVID-19 pneumonia (antiviral agents, antibacterial agents, and Chinese medicine), date of symptom onset, admission, virus testing, CT-scan, as well as condition improvement and living status. The study was approved by the Ethics Committee of Wuhan No.1 Hospital (No. 202008).

Treatment protocol for SARS-CoV-2 pneumonia. The treatment strategy for patients with COVID-19 pneumonia was based on the guidelines of the WHO[10], which included symptom relief, treatment of underlying diseases, prevention of superimposed bacterial infections, active prevention of complications such as sepsis and acute respiratory distress syndrome (ARDS) and support organ vital function in a timely fashion. Oxygen supplementation was provided for patients with desaturation by means of high flow oxygen via nasal prong, non-invasive and invasive mechanical ventilation, or extracorporeal membrane oxygenation (ECMO) if required.

Statistical considerations. The outcome in this study was whether or not the patients experienced a severe to critical course of disease. This outcome will be denoted "severe versus mild" in the following. All eligible variables were considered as covariates potentially influencing the outcome in the statistical analysis (supplement section "In-depth description of the statistical analysis flow"). Univariable analysis was performed using logistic regression analysis, where p values were adjusted for multiple testing by means of the Benjamini–Hochberg procedure. The popular multivariate imputation by chained equations (MICE) approach[11] was used to deal with missing values in the multivariable analyses, where 20 imputed data sets were used in each analysis ($m = 20$). The ratio of C-reactive protein (CRP) versus serum albumin (ALB) correlated very strongly with CRP, and the white blood cell count (WBC) correlated very strongly with absolute neutrophil count (ANC) ($\rho > 0.9$), which is why CRP and ANC were not considered in the multivariable analysis. The latter analysis was performed separately for the young patients and for all patients together (1) using logistic regression in combination with an automatic forward covariate selection procedure based on the Akaike information criterion (AIC)[12] applicable to multiply imputed data[13] and (2) random forest[14]. As a sensitivity analysis[15], the Bayesian information criterion (BIC)[16], which tends to select fewer covariates than the AIC, was also considered and backward selection was performed in addition, both when using the AIC and the BIC. The prediction performance of the models was estimated using 20 times repeated stratified K-fold cross-validation ($K = 3, 4, 5$), repeating the whole model selection process on each training set in each cross-validation iteration, excluding the corresponding test set[17]. Multiple imputation was performed separately on training and test sets[18]. As prediction performance measures the area under the curve (AUC) and the Brier score were used. Random forest was also used to rank the covariates with respect to their importance for prognosis via the AUC covariate importance values[19] and to estimate the influence forms of the covariates using partial dependence plots (PDPs)[20]. All statistical analyses were performed using the software R, version 3.6.3. All p values smaller than 0.05 were considered as statistically significant, where all statistical tests were performed two-sided. For further details and explanations, the interested reader is referred to the detailed description of the statistical analysis flow in the supplement (section "In-depth description of the statistical analysis flow").

Ethics, consent and permissions. The study was approved by the Ethics Committee of Wuhan No.1 Hospital (No. 202008). All methods were carried out in accordance with relevant guidelines and regulations.

Consent to publish. Informed consent was obtained from all patients for this study.

Results

Demographic and clinical features. The young cohort (age ≤ 60 years old) consisted of 762 patients (median age 47 years, interquartile ranges [IQR] 38–55, range 18–60; 55.9% female), who were admitted between January 2020 and March 2020 to Wuhan No.1 Hospitals. As shown in Table 1, 400 (52.5%) of the young patients had a mild condition during hospitalization (mild subgroup), while 362 (47.5%) developed a severe or critical disease course (severe subgroup). The mean age was statistically significantly higher in the severe subgroup than in the mild subgroup (59.3 vs. 56.0, Student's t-test: $p < 0.001$). 145 (19.8%) patients were affected by underlying diseases, hypertension (14.3%) being the most common one (Table 1). The median body temperature on admission was 36.6 °C (IQR 35.6–37.2), no difference was seen in median temperature between the two subgroups (mild vs. severe). The hospitalization in the severe subgroup was substantially longer than that in the mild subgroup (20.0 days [IQR 15.0–25.0] vs. 8.0 days [IQR 5.0–13.0]). The majority of patients received Chinese traditional medicine (93.4%), antiviral treatments (86.7%), and antibiotics (75.3%). In this cohort, the most frequently applied oxygen therapy was usual oxygen care (UOC) (78.2%). All key laboratory findings of this cohort are listed in Table 1. Additionally, an elderly cohort of patients (median age 69 years [IQR] 65–75, range 61–97, 52.8% female) was included, mainly for statistical comparisons with the young cohort.

Potential risk factors associated with disease severity in the young patients. In the univariable analysis, elevated level of complement C3, systemic immune-inflammation index (SII), CRP, serum amyloid A (SAA), lactate dehydrogenase (LDH), the ratio of CRP versus ALB, and the ratio of neutrophil versus lymphocyte (NLR) were statistically significantly associated with an increased risk for the development of a severe disease course of COVID-19 infection in the young patients (Table 2; full results, including statistically non-significant results, are shown in Supplementary Table 1). In contrast, decreased levels of ALB and lymphocyte (LYM) were statistically significantly correlated with the development of a severe disease course. SAA had the largest covariate importance value, both for young and elderly patients. The covariate importance values (Fig. 1) and the PDPs (Fig. 2, Supplementary Figs. 1 to 4) obtained using random forest analysis suggest that complement C3 is only prognostic in young patients.

Generally, the covariate importance values reveal that, while the number of relevant risk factors is larger for the elderly patients, the difference in importance between the most important risk factors and the remaining risk factors is more pronounced for the young patients. Supplementary Figs. 5 to 10 show the corresponding variable importance values and PDPs obtained for all patients, irrespective of age. Many of the PDPs indicate complex influence forms of the covariates. While SAA is again associated with the largest importance values here, the importance values of complement C3 are relatively small. The latter confirms that a higher level of complement C3 probably is a risk factor only associated with severity in young patients. After obtaining the latter result, additional analyses were performed in order to investigate whether the influence of complement C3 is different for specific subgroups of young patients (see supplement section "Subgroup analysis of the influence of complement C3 in young patients" for details). The results of these analyses did not suggest any relevant dependence of the influence of complement C3 in young patients on age, gender, and comorbidities, indicating that this risk factor is relevant for young patients independent of their specific characteristics (Supplement Figs. 11–13).

Multivariable statistical model-analysis for disease severity. The results of the multivariable analysis using logistic regression in combination with the forward selection algorithm and the AIC criterion showed that the risk for the development of a severe disease course for COVID-19 in young patients was higher for combinations of elevated levels of complement C3, increased SAA and SII, and reduced levels of LYM, platelet-lymphocyte ratio (PLR) and uric acid (UA) (Table 3). Applying the backward selection algorithm, UA was not selected in the multivariable model due to lack of importance, but gender, hypertension, thyroid related disease, and blood urea nitrogen (BUN) were selected instead, indicating potential importance to the risk of developing a severe disease course in young patients (Supplementary Table 2). However, a sensitivity analysis presented in the supplement (section "Model stability analysis") indicates that the results obtained using the forward selection algorithm (Table 3) may be more statistically stable. When using the BIC criterion instead of the AIC criterion, only complement C3 and SAA were shown to be relevant to the severity in these young patients (Supplementary Table 3).

The results of multivariable logistic regression for all patients (including young and elderly) using the AIC criterion and the forward selection algorithm differed partly from that obtained for the young patients. The risk of developing a severe disease course in all patients was high for increased levels of complement C3, SAA, BUN, LDH and immunoglobulin G (IgG) and decreased levels of ALB, PLR and immunoglobulin A (IgA) (Table 4). Note that the fact that complement C3 was included in the forward selection for all patients is very likely only due to its importance within the young cohort: In the univariable analysis complement C3 was only statistically significant for the young patients and, as seen in Fig. 1, the covariate importance value of complement C3 is only large for the young patients. As revealed by the covariate importance values obtained through the random forest analysis (Fig. 1), increased levels of SAA seem to be similarly associated with the development of a severe disease course in elderly patients as well as in young patients. When using the backward selection algorithm, LYM and blood platelet were selected in addition, indicating potential relevance associated with disease severity (Supplementary Table 4). Using the BIC criterion together with the forward selection algorithm, SAA, BUN,

	All patients	Young patients (< = 60 years)	Elderly patients (> 60 years)	*P* value
Clinical and demographic information				
Age (years)	60.00 (46.00–69.00)	47.00 (38.00–55.00)	69.00 (65.00–75.00)	
Gender: female				0.2500
Yes	803 (54.4)	426 (55.9)	377 (52.8)	
No	673 (45.6)	336 (44.1)	337 (47.2)	
Mild patients	718(48.6)	400(55.7)	318(44.3)	
Severe patients	758(51.4)	362(47.8)	396(52.2)	
Comorbidities				
Coronary heart disease				< 0.0001
Yes	124 (8.4)	19 (2.5)	105 (14.7)	
No	1352 (91.6)	743 (97.5)	609 (85.3)	
Diabetes				< 0.0001
Yes	200 (13.6)	51 (6.7)	149 (20.9)	
No	1276 (86.4)	711 (93.3)	565 (79.1)	
Hypertension				< 0.0001
Yes	441 (29.9)	109 (14.3)	332 (46.5)	
No	1035 (70.1)	653 (85.7)	382 (53.5)	
Thyroid related diseases				0.4751
Yes	32 (2.2)	19 (2.5)	13 (1.8)	
No	1444 (97.8)	743 (97.5)	701 (98.2)	
Initial symptoms				
Fever	973 (65.9)	529 (54.4)	444 (45.6)	
Cough	932 (63.1)	480 (51.5)	452 (48.5)	
Dyspnea	519 (35.2)	285 (54.9)	234 (45.1)	
Treatment				
Antiviral treatments	1232(83.5)	658(53.4)	574 (46.6)	
Chinese traditional medicine	1150 (77.9)	618 (53.7)	532 (46.3)	
Antibiotics	1071 (72.6)	543 (50.7)	528 (49.3)	
Oxygen therapy	1199 (81.2)	596 (49.7)	603 (50.3)	
Accu Troponin (μg/L)	0.0050 (0.0020–0.0120)	0.0020 (0.0010–0.0050)	0.0080 (0.0030–0.0190)	< 0.0001
Key laboratory findings				
ALB (g/L)	35.90 (32.50–39.00)	37.80 (34.90–40.20)	33.70 (30.70–36.60)	< 0.0001
ALT (U/L)	22.00 (14.00–35.00)	22.00 (14.00–37.00)	22.00 (15.00–33.00)	0.5136
ANC	3.30 (2.41–4.43)	3.07 (2.25–3.94)	3.62 (2.71–5.02)	< 0.0001
AST (U/L)	24.00 (19.00–34.00)	23.00 (18.00–32.00)	25.50 (20.00–36.00)	< 0.0001
Blood glucose (mmol/L)	5.40 (4.90–6.50)	5.10 (4.70–5.80)	5.90 (5.00–7.75)	< 0.0001
Blood platelet (10^9/L)	218.00 (170.00–280.75)	216.00 (171.00–275.00)	219.00 (168.00–286.00)	0.3292
BUN (mmol/L)	4.10 (3.30–5.30)	3.80 (3.10–4.50)	4.70 (3.70–6.40)	< 0.0001
CK (U/K)	58.00 (40.00–90.00)	57.00 (40.00–86.00)	59.00 (41.00–97.00)	0.3034
CK-MB (U/L)	7.00 (6.00–9.00)	7.00 (6.00–9.00)	8.00 (6.00–10.00)	0.0075
Complement C3 (g/L)	1.07 (0.92–1.23)	1.12 (0.98–1.26)	0.98 (0.88–1.18)	0.0076
Complement C4 (g/L)	0.27 (0.21–0.34)	0.29 (0.23–0.35)	0.24 (0.18–0.32)	0.0185
Cr (μmol/L)	62.00 (53.00–76.00)	60.00 (52.00–73.00)	65.00 (55.00–81.25)	< 0.0001
CRP	3.87 (3.00–21.80)	3.00 (3.00–12.10)	7.59 (3.00–39.10)	< 0.0001
D-Dimer (mg/L)	0.47 (0.23–1.10)	0.29 (0.18–0.55)	0.76 (0.37–2.02)	< 0.0001
Erythrocyte sedimentation rate	27.00 (15.00–48.00)	21.00 (12.00–36.00)	36.00 (19.00–59.00)	< 0.0001
Hemoglobin (g/L)	128.00 (118.00–139.00)	132.00 (123.00–143.00)	124.00 (114.00–135.00)	< 0.0001
IgA (g/L)	2.23 (1.61–2.82)	2.09 (1.58–2.62)	2.30 (1.78–3.06)	0.0761
IgG (g/L)	9.84 (8.21–12.00)	9.88 (8.17–11.83)	9.84 (8.23–12.50)	0.6426
IgM (mg/dl)	1.00 (0.79–1.37)	1.04 (0.81–1.37)	0.97 (0.74–1.33)	0.3179
LDH (U/L)	194.00 (158.00–262.00)	178.00 (150.00–231.00)	216.00 (170.50–300.50)	< 0.0001
LYM (10^9/L)	1.45 (1.01–1.92)	1.58 (1.15–2.00)	1.34 (0.91–1.74)	< 0.0001
MONO (10^9/L)	0.52 (0.40–0.68)	0.51 (0.39–0.66)	0.54 (0.41–0.70)	0.0045
Myohemoglobin (ng/mL)	37.15 (26.20–65.92)	27.60 (22.60–33.90)	49.20 (33.30–91.10)	< 0.0001
PLR	150.44 (111.06–213.47)	137.41 (104.98–191.12)	169.93 (119.89–242.33)	< 0.0001
Procalcitonin (ug/L)	0.05 (0.05–0.05)	0.05 (0.05–0.05)	0.05 (0.05–0.06)	< 0.0001
Continued				

	All patients	Young patients (< = 60 years)	Elderly patients (> 60 years)	*P* value
RBC	4.20 (3.87–4.58)	4.36 (4.02–4.67)	4.10 (3.72–4.41)	< 0.0001
SAA (mg/L)	58.40 (6.20–133.35)	38.75 (5.55–126.50)	76.55 (10.57–139.35)	0.0874
SII	481.19 (308.34–842.83)	389.75 (271.34–653.01)	590.55 (367.77–1069.11)	< 0.0001
UA (μmol/L)	286.00 (233.00–355.00)	285.00 (233.00–348.00)	287.50 (233.00–366.00)	0.5587
WBC (10^9/L)	5.58 (4.47–7.04)	5.38 (4.37–6.66)	5.86 (4.68–7.60)	< 0.0001
Ratios				
Ratio of CRP versus ALB	0.09 (0.08–0.50)	0.08 (0.07–0.25)	0.15 (0.08–1.00)	< 0.0001
Ratio of neutrophil versus lymphocyte	2.19 (1.48–3.44)	1.88 (1.35–2.66)	2.68 (1.77–4.54)	< 0.0001
Ratio of CRP versus ALB	0.09 (0.08–0.50)	0.08 (0.07–0.25)	0.15 (0.08–1.00)	< 0.0001
CT-scan				
Pulmonary consolidation				0.7776
Yes	52 (3.7)	26 (3.5)	26 (3.9)	
No	1359 (96.3)	716 (96.5)	643 (96.1)	
Ground glass opacity				0.8334
Yes	247 (17.5)	128 (17.3)	119 (17.8)	
No	1164 (82.5)	614 (82.7)	550 (82.2)	
Multiple patchy shadows				0.0219
Yes	1176 (83.3)	602 (81.1)	574 (85.8)	
No	235 (16.7)	140 (18.9)	95 (14.2)	

Table 1. Clinical and demographic characteristics, treatment and key laboratory findings. Metric covariates are reported as medians with interquartile ranges in the following form: 'Median (First Quartile-Third Quartile)'. Categorical covariates are reported as percentages in the following form: 'Absolute number (Percentage)'. Differences between the young and elderly patients were tested using the Wilcoxon test and Fisher's exact test for metric covariates and categorical covariates, respectively. The following abbreviations are used in the table: ALB: albumin, ALT: alanine aminotransferase, ANC: absolute neutrophil count, AST: aspartate aminotransferase, BUN: blood urea nitrogen, CK: creatine kinase, CK-MB: creatine kinase-MB, Cr: creatinine, CRP: C-reactive protein, LDH: lactate dehydrogenase, LYM: lymphocyte, MONO: monocyte, PLR: platelet-lymphocyte ratio, RBC: red blood cell, SAA: serum amyloid alpha, SII: systemic immune-inflammation index, UA: uric acid, WBC: white blood cell.

ALB, and LDH were selected, where the corresponding odds ratios (Supplementary Table 5) were very similar to those obtained in the model obtained using the AIC (Table 4). Applying the BIC criterion together with the forward selection algorithm delivered the same result.

In summary, the estimated prediction performances of the multivariable logistic regression models for all patients were better than those of the models obtained specifically for the young patients (Table 5). The random forest, which takes non-linear influences into account, performed best and the model selected using the BIC (which included fewer covariates) performed better than that selected using the AIC. Supplementary Table 6 provides an overview of which covariates were selected in each of the models obtained using the AIC and BIC criterion with forward and backward selection.

Discussion

COVID-19, caused by the SARS-CoV-2 virus, is a public health event that poses a serious threat to human health. The number of COVID-19 cases in young adults is higher than expected: According to the US CDC report almost half of the patients are younger than 65, and young patients have a substantial risk of developing a severe disease course. Facing the rapidly evolving circumstances caused by the COVID-19 pandemic, it is essential to prioritize medical resources by effectively conducting clinical stratification of COVID-19 young patients. Although several studies have identified risk factors for severity and mortality in COVID-19 patients[3,5,8], it remains critical to identify early and investigate potential risk factors specific for the development of a severe/critical disease course for young COVID-19 patients, because there could be strong differences in the functional state of the immune system between the young and elderly population[21]. In order to shed light on potential risk factors associated with a severe disease course in young patients, this study investigated clinical, demographic, treatment, and laboratory data from a group of COVID-19 patients using a set of comprehensive modern statistical methodologies. The univariable analysis suggested that potential risk factors for disease severity in young patients (age ≤ 60 years; n = 762) are in part different from that in elderly patients (age > 60 years; n = 714). Specifically, elevated levels of complement C3 and SAA were only statistically significantly associated with higher risks of a severe disease course in the young patients. In contrast, increased levels of ANC, aspartate aminotransferase (AST), BUN, creatine kinase (CK), creatinine (CR), D-dimer, myo-hemoglobin, and PLR, and decreased levels of red blood cells (RBC) had a statistically significant influence on this risk only for elderly patients. Even though SAA lacked statistical significance in the elderly subgroup, the covariate importance values (Fig. 1) suggest that SAA is an important risk factor irrespective of patients' age. While complement C4 missed significance in both age groups,

Variable	All patients		Young patients (<= 60 years)		Elderly patients (> 60 years)	
	Odds ratio [CI]	Adj. *p*	Odds ratio [CI]	Adj. *p*	Odds ratio [CI]	Adj. *p*
ALB (g/L)	0.9320 [0.9109, 0.9530]	< 0.0001	0.9501 [0.9170, 0.9831]	0.0317	0.9211 [0.8900, 0.9521]	< 0.0001
ANC	1.1055 [1.0549, 1.1611]	0.0002	1.0794 [0.9937, 1.1773]	0.2062	1.1032 [1.0408, 1.1741]	0.0062
AST (U/L)	1.0100 [1.0048, 1.0158]	0.0011	1.0062 [1.0005, 1.0130]	0.1762	1.0153 [1.0066, 1.0251]	0.0059
BUN (mmol/L)	1.1116 [1.0658, 1.1661]	< 0.0001	1.0461 [0.9923, 1.1270]	0.3516	1.1324 [1.0730, 1.2051]	0.0002
CK (U/K)	1.0023 [1.0011, 1.0036]	0.0011	1.0021 [1.0005, 1.0042]	0.0926	1.0023 [1.0007, 1.0041]	0.0196
Complement C3 (g/L)	2.0055 [0.6022, 6.9684]	0.3699	15.5808 [2.4111, 122.2841]	0.0392	0.3323 [0.0478, 1.9321]	0.3700
Cr (μmol/L)	1.0076 [1.0035, 1.0122]	0.0017	1.0046 [0.9979, 1.0121]	0.3963	1.0083 [1.0030, 1.0143]	0.0114
CRP	1.0119 [1.0079, 1.0161]	< 0.0001	1.0122 [1.0043, 1.0209]	0.0317	1.0110 [1.0064, 1.0161]	0.0001
D-Dimer (mg/L)	1.0808 [1.0386, 1.1358]	0.0017	1.0580 [0.9368, 1.2306]	0.6455	1.0770 [1.0319, 1.1374]	0.0078
LDH (U/L)	1.0027 [1.0018, 1.0038]	< 0.0001	1.0026 [1.0011, 1.0043]	0.0317	1.0027 [1.0015, 1.0040]	0.0002
LYM (10^9/L)	0.5842 [0.4926, 0.6908]	< 0.0001	0.5947 [0.4664, 0.7536]	0.0009	0.6016 [0.4698, 0.7662]	0.0003
Myohemoglobin (ng/mL)	1.0054 [1.0027, 1.0091]	0.0023	1.0022 [0.9918, 1.0134]	0.7994	1.0055 [1.0026, 1.0096]	0.0078
PLR	1.0015 [1.0006, 1.0024]	0.0040	1.0008 [0.9996, 1.0022]	0.3963	1.0017 [1.0005, 1.0031]	0.0196
Ratio of CRP versus ALB	1.3836 [1.2341, 1.5667]	< 0.0001	1.3994 [1.0999, 1.8321]	0.0430	1.3393 [1.1759, 1.5467]	0.0002
Ratio of neutrophil versus lymphocyte	1.1061 [1.0687, 1.1485]	< 0.0001	1.0835 [1.0256, 1.1539]	0.0416	1.1105 [1.0629, 1.1666]	0.0001
RBC (10^{12}/L)	0.7775 [0.6493, 0.9287]	0.0140	0.9856 [0.7644, 1.2707]	0.9337	0.6592 [0.5018, 0.8600]	0.0078
SAA (mg/L)	1.0065 [1.0032, 1.0100]	0.0006	1.0066 [1.0023, 1.0112]	0.0317	1.0059 [1.0008, 1.0114]	0.0639
SII	1.0003 [1.0002, 1.0004]	< 0.0001	1.0003 [1.0001, 1.0006]	0.0430	1.0003 [1.0001, 1.0004]	0.0063

Table 2. Univariable logistic regression for the outcome "severe versus mild". The *p* values were adjusted for multiple testing separately for the analysis of all patients, young patients, and elderly patients. Increased levels of complement C3 and SAA were associated with an increased risk for severity only in the young patient cohort. Supplementary Table 1 contains the results for all covariates.

the covariate importance values and PDPs suggest that this covariate, in addition to complement C3, could be a particularly strong risk factor in the young cohort, while its influence was not statistically significant in elderly patients. Further analysis showed that the significance of complement C3 in young patients was independent of the common risk factors age, gender and the presence versus absence of comorbidities. Risk factors reported previously in COVID-19 infection such as CRP[7], lactate dehydrogenase[3] and decreased lymphocyte[8] were validated both in young and elderly patient cohorts. Only the D-dimer and procalcitonin did not show significance in the young patient cohort. The PDPs confirmed the observed differences and revealed that the influences of many covariates on the log odds of disease severity of COVID-19 infection are strongly non-linear. However, in logistic regression it is assumed that these influences are linear, which is likely an important reason why the random forests outperformed the multivariable logistic regression models, both for the cohort of young patients and for all patients taken together.

The complement family is an important integral component of the innate immune response to viruses, not only protecting the body from infectious agents such as viruses and bacteria, but also playing a key role in promoting inflammatory processes triggering inflammatory cytokine storm[22]. The abnormal activation of various innate immune pathways, such as complement system, cytokines and thrombosis pathways, is considered as the driver of ARDS and may lead to multi-organ dysfunction[23,24]. The activation of the complement system also can be found in patients infected with coronaviruses, such as MERs-CoV, SARS-CoV-1 and SARS-CoV-2, which develop into ARDS[25]. When mice with complement C3 deficiency were infected with SARS-CoV, the infiltration of neutrophils and inflammatory monocytes in the lungs was strongly reduced, and the levels of cytokines and chemokines in the lungs and serum were decreased, as well as the incidence of respiratory failure. This suggests that the activation of the complement component C3 may aggravate the disease of SARS-CoV-related ARDS[24].

Furthermore, complement C3, involved in the function of innate immunity, has been shown to play a role in the recovery of COVID-19 patients[26], and critically low levels of this immune component have been shown to be connected with mortality following COVID-19 infection[27]. These results do not contradict the findings of our study. As the key initiator of innate immunity, complement C3 plays a major role in the activation of different immune cells including neutrophils and macrophages[25]. Critically low levels of complement C3 indicate an inability for the immune response to initiate, causing an immediate failure of anti-viral immune protection, whereas elevated levels of complement C3 may lead to excessive production of cytokine via diverse signaling pathways, causing a cytokine storm[25]. The majority of young patients (97.5%) in our cohort had a normal or elevated level of complement C3, reflecting the latter case. Young patients have fewer underlying diseases and are more immune-related compensated. This may be the reason why many indicators including D-dimer and procalcitonin do not change strongly, even when patients progress to severe COVID-19. However, young patients have more active immune function, which is why elevated levels of serum complement C3 may be a potential indicator of the severity of COVID-19 patients. A possible explanation why complement C3 did not have the

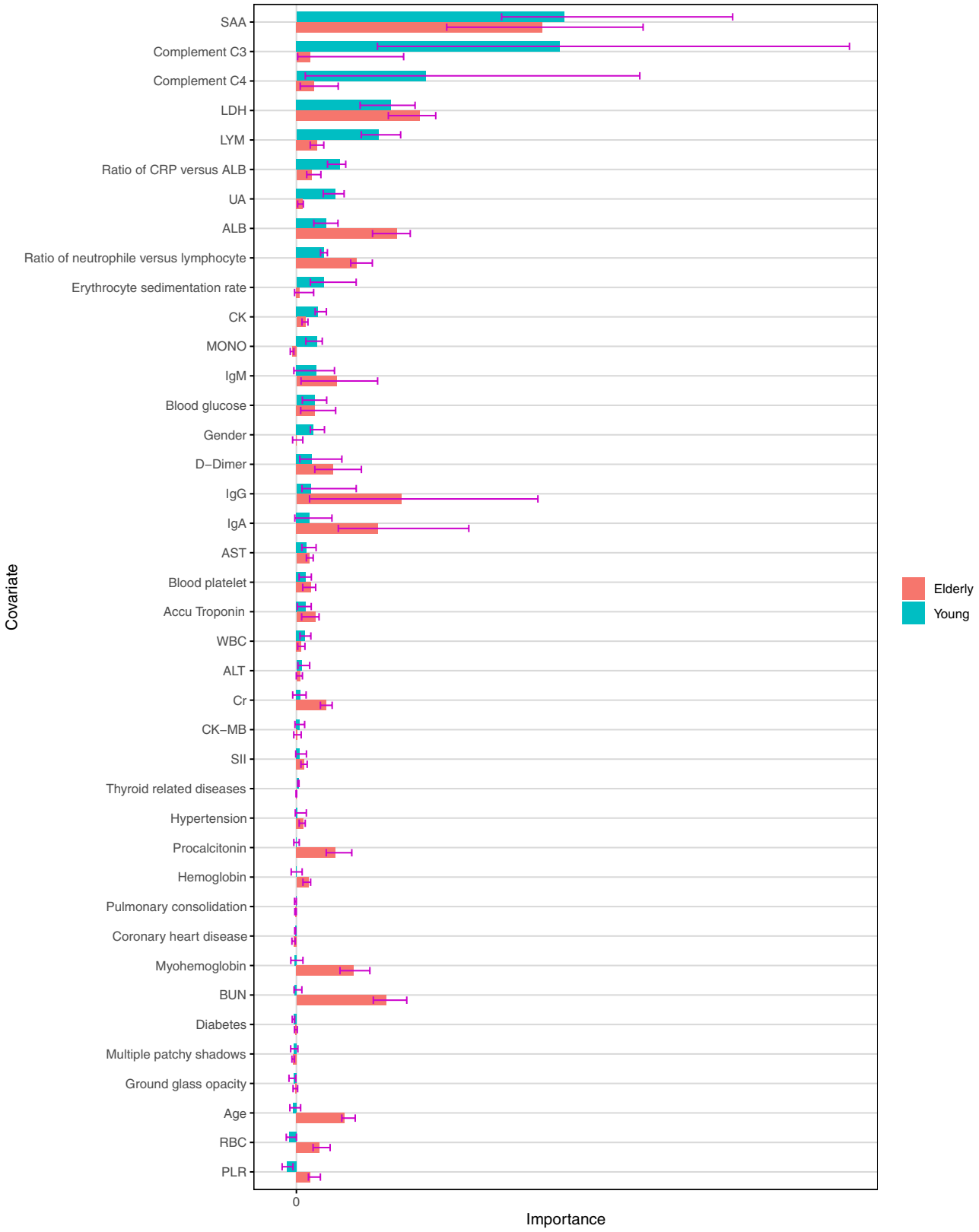

Figure 1. AUC variable importance with respect to predicting the outcome "severe versus mild" for young and elderly patients calculated using random forests. The larger the importance value of a covariate is, the greater the improvement in prediction performance by including this covariate in prognosis. Complement C3, C4, LYM, the ratio of CRP versus ALB, and UA seem to influence the prognosis of the development of severity mostly in young patients. The bars show the medians of the 20 importance values calculated using the 20 imputed data sets from the multiple imputation. The error bars illustrate the variabilities of the importance values: The lower/upper ends show the first/third quartiles of the 20 importance values, that is, 25% percent of the importance values lie below/above these values. To make the raw importance values comparable between young and elderly patients, both for the young and for the elderly patients, the raw importance values were divided by the means of all importance values with positive sign.

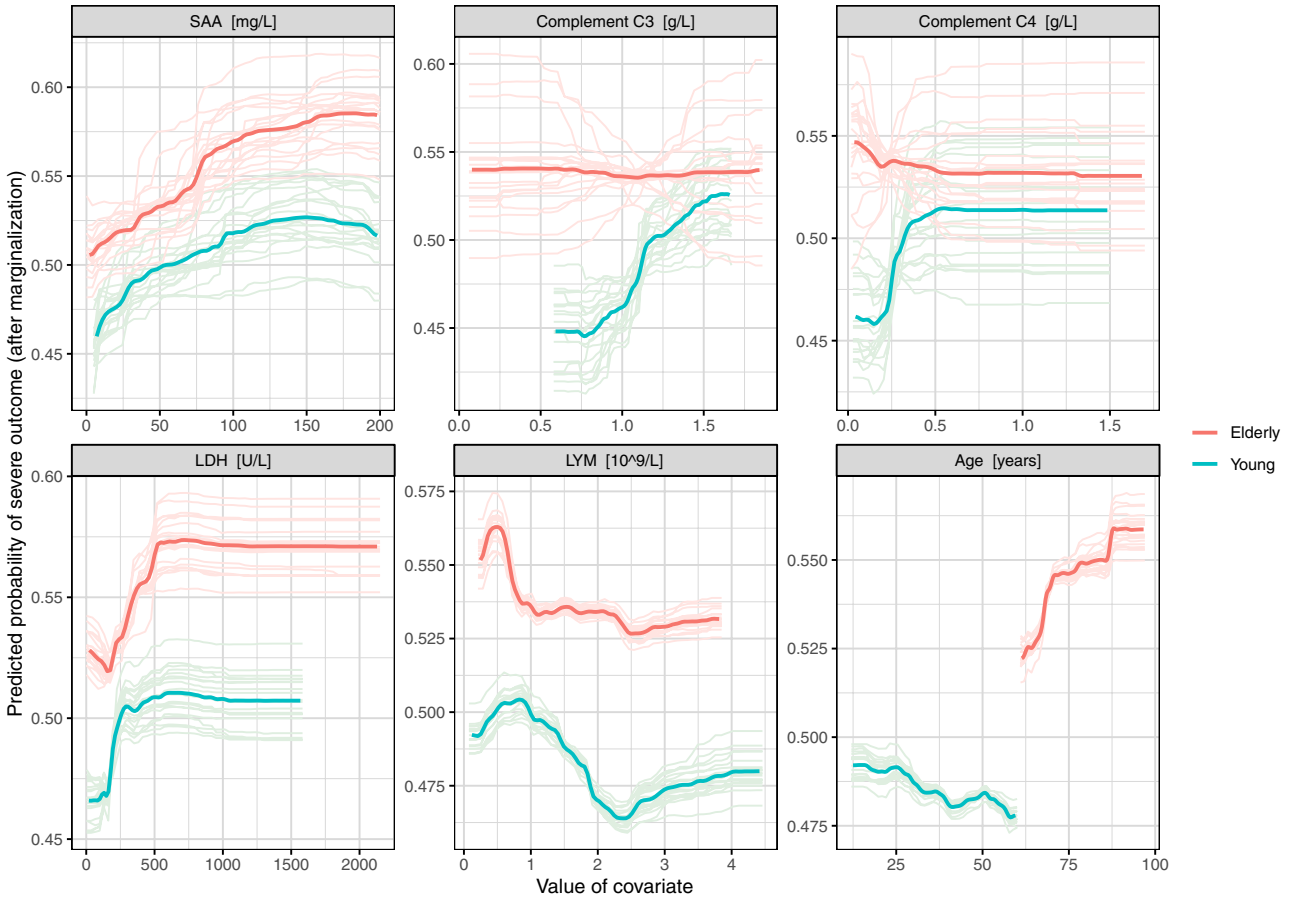

Figure 2. Partial dependence plots (PDPs) for young and elderly patients calculated using random forests. In simplified terms, a PDP shows the influence of a covariate on the outcome after adjusting for the influences of the other covariates. The PDPs for the five variables with largest AUC importance values in young patients and that for 'age' are shown. The light lines show the 20 individual PDPs calculated using the imputed data sets from the multiple imputation. The bold lines show averages over the 20 individual PDPs.

	Regression coefficient	Odds ratio
Intercept	−0.604671	–
Complement C3 (g/L)	1.621768	5.0620
SAA (mg/L)	0.005154	1.0052
LYM (10^9/L)	−0.477558	0.6203
PLR	−0.003868	0.9961
SII	0.000401	1.0004
UA (μmol/L)	−0.001553	0.9984

Table 3. Multivariable logistic regression models for the outcome "severe versus mild" in young patients selected using the AIC criterion and forward selection.

same prognostic role for elderly patients would be the immunosenescence caused by aging[21]. The hyperactivation of complement C3 alone does not suffice to induce activities of different immune cells.

This study possesses limitations aside from those inherent to all retrospective cohort studies such as their lack of causal inference. First, it is a single-center study featuring a limited number of cases. Second, the patient data were collected within 3 days after hospital admission, leading to missing data in a number of variables, which were imputed with a standard statistical approach.

	Regression coefficient	Odds ratio
Intercept	0.191987	–
Complement C3 (g/L)	0.506456	1.6594
IgA (g/L)	−0.134694	0.8740
IgG (g/L)	0.035108	1.0357
ALB (g/L)	−0.044892	0.9561
SAA (mg/L)	0.00574	1.0058
BUN (mmol/L)	0.06745	1.0698
LDH (U/L)	0.001668	1.0017
PLR	−0.000795	0.9992

Table 4. Multivariable logistic regression models for the outcome "severe versus mild" in all patients irrespective of age selected using the AIC criterion and forward selection.

		AUC			Brier score		
	K in cross-validation	Logistic regression—AIC	Logistic regression—BIC	Random forest	Logistic regression—AIC	Logistic regression—BIC	Random forest
Young patients	3	0.5512	0.5542	0.5962	0.2740	0.2584	0.2438
	4	0.5405	0.5454	0.5847	0.2764	0.2606	0.2462
	5	0.5327	0.5344	0.5838	0.2772	0.2640	0.2461
All patients	3	0.6069	0.6186	0.6310	0.2502	0.2425	0.2369
	4	0.6027	0.6140	0.6271	0.2525	0.2444	0.2378
	5	0.6059	0.6156	0.6276	0.2500	0.2433	0.2378

Table 5. Performances of the models measured using stratified K-fold cross-validation. For each value of K, the stratified K-fold cross-validation was repeated twenty times and the results averaged. Higher values of the AUC and smaller values of the Brier score are preferable.

In summary, this study conducted a comprehensive statistical analysis with a focus on non-linear relationships to identify risk factors and possible pathogenesis for the development of a severe disease course during COVID-19 infection in young patients. However, large-scale and multi-center analysis is needed to further build on the knowledge obtained.

Received: 11 July 2020; Accepted: 22 January 2021
Published online: 12 April 2021

References

1. World Health Organization. *Novel Coronavirus (2019-nCoV) Situation Reports. Weekly Epidemiological Update*. November 10, 2020; https://www.who.int/docs/default-source/coronaviruse/situation-reports/20201110-weekly-epi-update-13.pdf?sfvrsn=24435477_15.
2. Huang, C. *et al.* Clinical features of patients infected with 2019 novel coronavirus in Wuhan, China. *Lancet Lond. Engl.* https://doi.org/10.1016/S0140-6736(20)30183-5 (2020).
3. Zhou, F. *et al.* Clinical course and risk factors for mortality of adult inpatients with COVID-19 in Wuhan, China: A retrospective cohort study. *Lancet* **395**, 1054–1062 (2020).
4. Chen, N. *et al.* Epidemiological and clinical characteristics of 99 cases of 2019 novel coronavirus pneumonia in Wuhan, China: A descriptive study. *Lancet* **395**, 507–513 (2020).
5. Richardson, S. *et al.* Presenting characteristics, comorbidities, and outcomes among 5700 patients hospitalized with COVID-19 in the New York city area. *JAMA* https://doi.org/10.1001/jama.2020.6775 (2020).
6. US. CDC COVID-19 Response Team. *Severe Outcomes Among Patients with Coronavirus Disease 2019 (COVID-19)—United States.* https://www.cdc.gov/mmwr/volumes/69/wr/mm6912e2.htm (2020).
7. Wu, C. *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.* https://doi.org/10.1001/jamainternmed.2020.0994 (2020).
8. Ruan, Q. *et al.* Clinical predictors of mortality due to COVID-19 based on an analysis of data of 150 patients from Wuhan, China. *Intensive Care Med.* https://doi.org/10.1007/s00134-020-05991-x (2020).
9. Liang, W. *et al.* Development and validation of a clinical risk score to predict the occurrence of critical illness in hospitalized patients with COVID-19. *JAMA Intern. Med.* https://doi.org/10.1001/jamainternmed.2020.2033 (2020).
10. Clinical management of severe acute respiratory infection when novel coronavirus (nCoV) infection is suspected. https://www.who.int/publications-detail/clinical-management-of-severe-acute-respiratory-infection-when-novel-coronavirus-(ncov)-infection-is-suspected.
11. van Buuren, S. *et al.* Multivariate imputation by chained equations in R. *J. Stat. Softw.* **45**, 1–67 (2011).
12 Akaike, H. Information theory and an extension of the maximum likelihood principle. In *Selected Papers of HirotuguAkaike* (eds Parzen, E. *et al.*) 199–213 (Springer, Berlin, 1998).
13. Wood, A. M. *et al.* How should variable selection be performed with multiply imputed data?. *Stat. Med.* **27**, 3227–3246 (2008).
14. Breiman, L. Random forests. *Mach. Learn.* **45**, 5–32 (2001).

15. Thabane, L. *et al.* A tutorial on sensitivity analyses in clinical trials: The what, why, when and how. *BMC Med. Res. Methodol.* **13**, 92 (2013).
16. Schwarz, G. Estimating the dimension of a model. *Ann. Stat.* **6**, 461–464 (1978).
17 Hornung, R. *et al.* A measure of the impact of CV incompleteness on prediction error estimation with application to PCA and normalization. *BMC Med. Res. Methodol.* **15**, 95 (2015).
18 Wahl, S. *et al.* Assessment of predictive performance in incomplete data by combining internal validation and multiple imputation. *BMC Med. Res. Methodol.* **16**, 144 (2016).
19. Janitza, S., Strobl, C. & Boulesteix, A.-L. An AUC-based permutation variable importance measure for random forests. *BMC Bioinform.* **14**, 119 (2013).
20. Friedman, J. H. Greedy function approximation: A gradient boosting machine. *Ann. Stat.* **29**, 1189–1232 (2001).
21. Aw, D., Silva, A. B. & Palmer, D. B. Immunosenescence: Emerging challenges for an ageing population. *Immunology* **120**, 435–446 (2007).
22. Li, G. *et al.* Coronavirus infections and immune responses. *J. Med. Virol.* **92**, 424–432 (2020).
23. Campbell, C. M. *et al.* Will complement inhibition be the new target in treating COVID-19 related systemic thrombosis?. *Circulation* **141**, 1739–1741 (2020).
24. Ciceri, F. *et al.* Microvascular COVID-19 lung vessels obstructive thromboinflammatory syndrome (MicroCLOTS): An atypical acute respiratory distress syndrome working hypothesis. *Crit. Care Resusc.* **22**, 95–97 (2020).
25. Risitano, A. M. *et al.* Complement as a target in COVID-19?. *Nat. Rev. Immunol.* **20**, 343–344 (2020).
26. Xiao, Y. *et al.* Exploration of turn-positive RT-PCR results and factors related to treatment outcome in COVID-19: A retrospective cohort study. *Virulence* **11**, 1250–1256 (2020).
27. Zhao, Y. *et al.* Abnormal immunity of non-survivors with COVID-19: Predictors for mortality. *Infect. Dis. Poverty* **9**, 108 (2020).
28. https://www.medrxiv.org/content/10.1101/2020.07.24.20161414v1

Acknowledgements

The authors thank Alethea Charlton for making valuable language corrections. This manuscript has been released as pre-print at medRxiv, https://www.medrxiv.org/content/10.1101/2020.07.24.20161414v1[28].

Author contributions

Acquisition, analysis, or interpretation of data: W.C., K.X., R.H., J.L.; Drafting of the manuscript: W.C., R.H., J.L.; Statistical analysis: R.H., J.L.; Critical revision of the manuscript for important intellectual content: K.X.; Obtained funding: W.C., K.X., R.H.; Administrative, technical, or material support: W.C., K.X., CY; Conception design: W.C., K.X., J.L.; Supervision: W.C., K.X., J.L.

Funding

The study was funded by Sino-German Center for Research Promotion (SGC)'s rapid Response Funding for Bilateral Collaborative Proposals Between China and Germany in COVID-19 Related Research (Project No. C-0065), the Natural Science Foundation of Hubei Province (No. 2019CFB641), and by the German Science Foundation (DFG-Einzelförderung HO6422/1-2 to RH), and Core Funding of the medical faculty of University Bonn.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1038/s41598-021-82810-3.

Correspondence and requests for materials should be addressed to K.X.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.