Medical & Biological Engineering & Computing (2022) 60:1763–1774
https://doi.org/10.1007/s11517-022-02568-2

ORIGINAL ARTICLE

Building a predictive model to identify clinical indicators for COVID-19 using machine learning method

Xinlei Deng[1] · Han Li[2] · Xin Liao[3] · Zhiqiang Qin[4] · Fan Xu[5] · Samantha Friedman[6] · Gang Ma[7] · Kun Ye[8] · Shao Lin[1,9]

Received: 10 November 2021 / Accepted: 25 March 2022 / Published online: 25 April 2022

Abstract

Although some studies tried to identify risk factors for COVID-19, the evidence comparing COVID-19 and community-acquired pneumonia (CAP) is inconclusive, and CAP is the most common pneumonia with similar symptoms as COVID-19. We conducted a case–control study with 35 routine-collected clinical indicators and demographic factors to identify predictors for COVID-19 with CAP as controls. We randomly split the dataset into a training set (70%) and testing set (30%). We built Explainable Boosting Machine to select the important factors and built a decision tree on selected variables to interpret their relationships. The top five individual predictors of COVID-19 are albumin, total bilirubin, monocyte count, alanine aminotransferase, and percentage of monocyte with the importance scores ranging from 0.078 to 0.567. The top systematic predictors for COVID-19 are liver function, monocyte increasing, plasma protein, granulocyte, and renal function (importance scores ranging 0.009–0.096). We identified five combinations of important indicators to screen COVID-19 patients from CAP patients with differentiating abilities ranging 83.3–100%. An online predictive tool for our model was published. Certain clinical indicators collected routinely from most hospitals could help screen and distinguish COVID-19 from CAP. While further verification is needed, our findings and predictive tool could help screen suspected COVID-19 cases.

Keywords COVID-19 · Predictor · Machine learning · Community-acquired pneumonia

1 Introduction

Coronavirus disease 2019 (COVID-19) is a serious clinical and public health challenge. The speed of the spread and severity of this new coronavirus are faster and greater than was the case for severe acute respiratory syndrome coronavirus. By February of 2022, the global cumulative cases and deaths of COVID-19 have been more than 411,000,000 and 5,810,000, respectively [1]. Thus far, the worldwide pandemic situation of COVID-19 has not been well-controlled.

Xinlei Deng and Han Li are equal contributors.

✉ Kun Ye
yezi5729@163.com

✉ Shao Lin
slin@albany.edu

[1] Department of Environmental Health Sciences, School of Public Health, University at Albany, State University of New York, Rensselaer, NY, USA

[2] Department of Hematology, Guangxi Academy of Medical Sciences & The People's Hospital Of Guangxi Zhuang Autonomous Region, Nanning, China

[3] Department of Scientific Research, Guangxi Academy of Medical Sciences & The People's Hospital Of Guangxi Zhuang Autonomous Region, Nanning, China

[4] Department of Respiratory, Guangxi Academy of Medical Sciences & The People's Hospital Of Guangxi Zhuang Autonomous Region, Nanning, China

[5] Guangxi Health Commission Key Laboratory of Ophthalmology and Related Systemic Diseases Artificial Intelligence Screening Technology & Research Center of Ophthalmology, Guangxi Academy of Medical Sciences & The People's Hospital Of Guangxi Zhuang Autonomous Region, Nanning, China

[6] Department of Sociology, University at Albany, State University of New York, Albany, NY, USA

[7] Department of Obstetrics and Gynecology, Guangxi Academy of Medical Sciences & The People's Hospital Of Guangxi Zhuang Autonomous Region, Nanning, China

[8] Department of Nephrology, Guangxi Academy of Medical Sciences & The People's Hospital Of Guangxi Zhuang Autonomous Region, Nanning, China

[9] Department of Epidemiology and Biostatistics, School of Public Health, University at Albany, State University of New York, Rensselaer, NY, USA

A key factor for effective prevention and control of this disease spread is to have early and rapid diagnosis. The clinical symptoms of early COVID-19 are not quite specific and it has similar symptoms as most forms of pneumonia. The real-time reverse transcription-PCR (RT-PCR) test is recommended as the gold standard test for diagnosis, but there are several limitations to this test. According to previous reports [2], the false negative rate (1-sensitivity) of this test could be as high as 62%. False negative results of RT-PCR testing mean that the patients with COVID-19 are easily misdiagnosed as having community-acquired pneumonia (CAP) because of their similar clinical characteristics. In addition, the cost of expensive testing is one of the reasons why RT-PCR testing cannot be widely adopted.

In order to improve the accuracy of diagnosis, computed tomography was used in several studies to differentiate COVID-19 from CAP because various types of pneumonias have certain typical imaging features [3]. However, imaging alone still made it difficult to differentiate COVID-19 from CAP. Additionally, large numbers of clinical indicators, which are routinely collected for each patient and stored in each hospital for a long time, have rarely been used for COVID-19 research. A few studies investigated the clinical laboratory data of patients with COVID-19 and found that some hematological parameters can be the warning indicators of patients that may experience more severe symptoms [4, 5]. These studies, however, are limited to COVID cases only without a control group and rely on few clinical indicators.

On the other hand, community-acquired pneumonia (CAP) is an ideal control group, which is defined as an acute lung infection acquired outside of the hospital setting. Among all pneumonia types, CAP is the most prevalent pneumonia, with a total incidence rate of 7.13 per 1000 person-years. In 2016, over 9.5 million cases occurred in China [6], and in the USA, approximately 5 million cases occur each year [7]. The most common pathogen of CAP is pneumoniae, followed by Mycoplasma, Chlamydia, H influenzae, and respiratory viruses. COVID-19 is one unique type of CAP, an infectious disease with a viral pathogen rather than a bacterial one and needs different treatments. In addition, both COVID-19 and CAP share similar symptoms of fever, cough, dyspnea, and chest radiograph of parenchymal or interstitial infiltration. Therefore, we used CAP in this study as the control group in order to identify the unique predictors of COVID-19.

The selection and analysis of many highly correlated clinical indicators also present challenges. In recent years, machine learning has been widely used in the medical field and has made breakthrough progress [8, 9]. Wang S et al. created a fully automatic deep learning system for COVID-19 diagnosis and prognosis using a chest computed tomography image [10]. In Wynants's review paper (2020), they included seven studies for COVID-19 predictions with clinical indicators [11]. For instance, Feng et al. (2021) only included 26 suspected and not confirmed COVID cases as the COVID outcome [12]. In Wu's study (2020), only 27 out of 105 COVID cases were truly confirmed as COVID cases [13]. Martine et al. (2020) used WHO guidelines to simulate 4096 artificial cases; their analyses relied on no real cases [14]. In another three studies, CT information from doctors' reports composed the predictors, in addition to the sample sizes being smaller [15–17]. In summary, previous studies suffered from methodological problems, including outcome misclassification, requiring CT or image reports, validation issues, and using non-representative controls.

This study fills these knowledge gaps and aims to identify important laboratory clinical markers of COVID-19 and uses machine learning models to distinguish patients with COVID-19 from CAP, which may assist in the screening, diagnosis, and monitoring of COVID-19 cases by evaluating large amounts of clinical indicators routinely collected.

2 Methods

2.1 Study design and patients

We used a case–control study design from February 16 to March 16, 2020, with patients diagnosed with COVID-19 as the cases and patients diagnosed with CAP as the controls. All patients with COVID-19 and CAP were directly enrolled from The People's Hospital of Guangxi Zhuang Autonomous Region, in Nanning, China, which is the largest and best hospital in Guangxi Region and represents the target population well in this region. This hospital was appointed by the Chinese government as the only designated hospital that treated all patients with COVID-19 in Guangxi in 2020. The patients who visit this hospital are also well representative of most ethnic and minority groups in China. The CAP patients were randomly selected from The People's Hospital of Guangxi Zhuang Autonomous Region and were frequency matched with the COVID-19 patients by gender and age.

The ethics committee of The People's Hospital of Guangxi Zhuang Autonomous Region approved this study and granted a waiver of informed consent from patients. Written informed consent was obtained from all controls.

2.2 Definitions of outcomes

The COVID-19 patients were diagnosed according to the guidelines by the National Health Commission of China and were confirmed by positive SARS-CoV-2 RNA with throat swab samples (Sansure Biotechnology, Changsha, Hunan, China) [18].

According to The Guidelines For Diagnosis And Treatment Of Community Acquired Pneumonia In Chinese Adults (version 2016), CAP is defined by acute symptoms, the presence of signs of a lower respiratory tract infection and new pulmonary infiltrate on a chest radiograph without other obvious causes [19]. SARS-CoV-2 RNA with throat swab samples must be negative to differentiate CAP from COVID-19.

2.3 Definitions of clinical indicators

We included 379 patients (62 patients with COVID-19; 317 patients with CAP). We extract 35 variables including demographic factors such as age and sex, and another 33 clinical indicators from laboratory results. These clinical indicators included albumin, total bilirubin, monocyte count, the coefficient of variation of red blood cell distribution width (RDW-CV), total protein, platelet count, neutrophil count, blood urea nitrogen (BUN), lymphocyte count, percentage of lymphocyte, percentage of monocyte, hemoglobin level, creatinine level, mean corpuscular volume, white blood cell count (WBC), globulin, alanine aminotransferase (ALT), percentage of neutrophil, red blood cell count (RBC), basophil count, percentage of acidophil, mean corpuscular hemoglobin (MCH), mean corpuscular hemoglobin concentration (MCHC), mean platelet volume, uric acid, creatine kinase, urinary occult blood, proteinuria, urobilinogen, bilirubin in urine, and urate crystal. These factors were categorized based on the clinical reference range in Chinese (please see Appendix 1).

We used Explainable Boosting Machine (EBM) to select important contributing factors from these clinical indicators and demographic factors including age and sex. Seventeen factors were selected, and only these selected factors were further grouped into nine categories based on the clinical information of their biological functions. These nine categories are (1) plasma protein including albumin, globulin, and total protein; (2) liver function including total bilirubin and ALT; (3) monocyte increasing including increasing monocyte count and increasing percentage of monocyte; (4) anemia indicators including RDW-CV, mean corpuscular volume, and hemoglobin level; (5) granulocyte including WBC, neutrophil count, and percentage of neutrophil; (6) renal function including creatinine level and BUN; (7) lymphocyte including lymphocyte count and percentage of lymphocyte; (8) platelet including platelet count; (9) urinalysis including proteinuria.

2.4 Statistical analysis

First, we randomly split the dataset into a training set (70%) and a testing set (30%) and adopted the ALLKNN under-sampling technique [20] to overcome an imbalance problem present in the training set. ALLKNN is one of the sampling strategies used when facing imbalanced datasets. Compared with other over-sampling or under-sampling methods, including SMOTE, SMOTEENN, SMOTETomek, and ADASYN, and random sampling, ALLKNN performed better when building the model. By adding a random term into the training set, we used a recently developed EBM to build a model based on the training set [21]. EBM was preferred because this model is derived from the generalized additive model (GAM); uses techniques from random forest and boosted tree models; and could be easier to interpret. The major differences between EBM and traditional GAMs include the following: (1) each feature function in EBM is determined using modern ML techniques, such as bagging and gradient boosting, with round-robin cycles; (2) EBM can automatically detect and include pairwise interaction terms and improves accuracy; and (3) EBM plots the feature function to examine the association between each variable and the outcome (see Appendix 2). Previous studies suggest that EBM performs better on health datasets than other established ML models, including the light gradient boosting model, regularized logistic regression, random forest, and xgboost [21]. Consistently, our previous study predicting congenital heart diseases also found that EBM had a better performance [22]. We also compared the prediction performance of EBM with other ML models in Appendix 3.

A random term was used to identify the threshold of relative importace score of the random effect [9, 23]. Hence, only the variables with importance scores greater than the random term were selected as important contributing variables. By using a fivefold cross-validation strategy and a grid search for hyper-parameters, we rebuilt the EBM model with only the selected variables as our final predictive model (final EBM). The optimal parameters included the out of bags, inner bags, learning rate, and the maximum and minimum number of tree leaves. The model performance was evaluated with receiver operating characteristic (ROC) curves based on both the training and testing sets and the calibration curve for the whole dataset.

After obtaining the predictive results from the final EBM, we chose the optimal cut-off value by using the point closest to the top-left part of the ROC curve with the perfect sensitivity or specificity.

$$\text{Optimal criterion} = \min((1 - \text{sensitivities})^2 + (1 - \text{specificities})^2)$$

We dichotomized the predictive values by using the optimal cut-off point. Then, the adjusted odds ratios (aOR) of contributing variables were calculated by multivariate logistic regressions using the dichotomized predictive values from the final EBM. We grouped these selected variables into nine categories by using average importance scores based on

the clinical information of biological functions they could provide. Therefore, the score of each category indicates the level of changed biological functions among patients with COVID-19.

To assess the inter-relationships among selected variables, a decision tree model was built on selected variables from EBM. The decision tree model was only used to explore the relationships among the selected variables and tried to find out the combinations of the predictors. This decision tree model also went through a grid search with fivefold cross-validation on the entire dataset. We also developed an online predictive tool (https://xdeng3.shinyapps.io/COVID-19/) for patients, clinicians, and other researchers to use.

3 Results

We selected the important contributing variables according to the importance scores with a random term (Fig. 1). The top 15 selected variables included albumin (0.567), total bilirubin (0.341), monocyte count (0.336), ALT (0.246), percentage of monocytes (0.197), RDW-CV (0.180), WBC (0.142), neutrophil count (0.139), creatinine level (0.125), globulin (0.102), BUN (0.100), lymphocyte count (0.096), percentage of lymphocyte (0.094), total protein (0.085), and percentage of neutrophil (0.078) (Fig. 1). In addition, the risk effect trends of the top 15 variables from the EBM were presented in Appendix 2. Based on the risk effect trends, we conclude that the typical characteristics of patients with COVID-19 may include the following six different categories: (a) low/normal levels of albumin and percentage of neutrophil; (b) low levels of total bilirubin, creatinine level, globulin, and total protein; (c) high levels of percentage of monocyte, BUN, and percentage of lymphocyte; (d) normal levels of RDW-CV, WBC, monocyte count, and neutrophil count; (e) low/high levels of lymphocyte count; and (f) negative ALT.

We grouped selected variables into nine categories by using average importance scores (Fig. 2). The average importance scores were used to indicate the level of changed biological functions among patients with COVID-19. The most prominent category was liver function, and the average weight or importance score was 0.096, followed by monocyte increasing (0.087), plasma protein (0.083), granulocyte (0.039), renal function (0.037), anemia (0.033), lymphocyte (0.031), platelet (0.024), and urinalysis (0.009) (Fig. 2).

Figure 3a shows the calibration of our predictive model. The predictive probability aligned with the observed probability well from 15 to 100%. There was a little underestimation when the probability was lower than 15%. In addition, the areas under the curves (AUC) of ROC for this model were 0.948 on the training set and 0.932 on the testing set (Fig. 3b). With both AUCs above 0.7 and very close to each other, we conclude that this predictive model overcame the over-fitting problem.

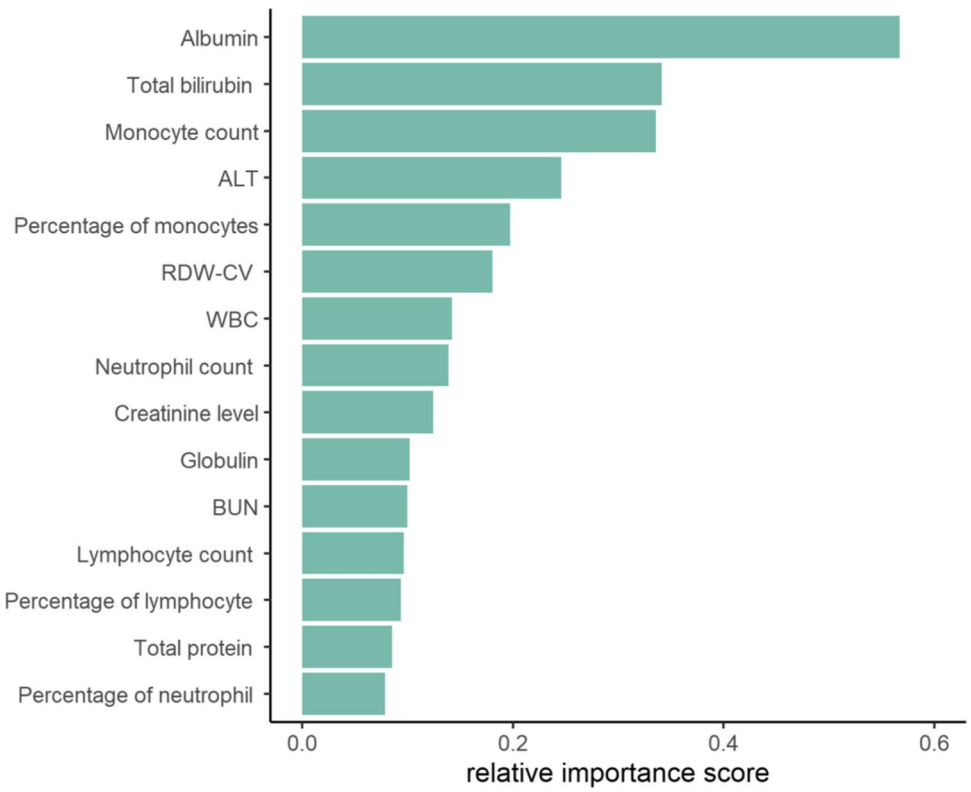

Fig. 1 Relative importance score of selected variables from EBM

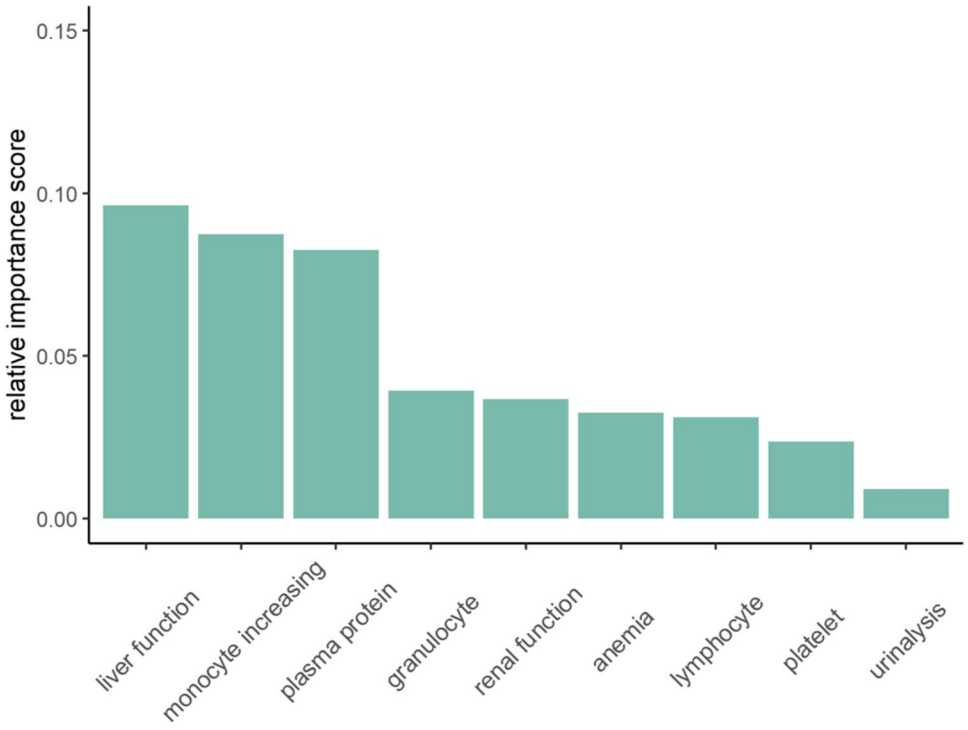

Fig. 2 Average relative importance score by nine groups

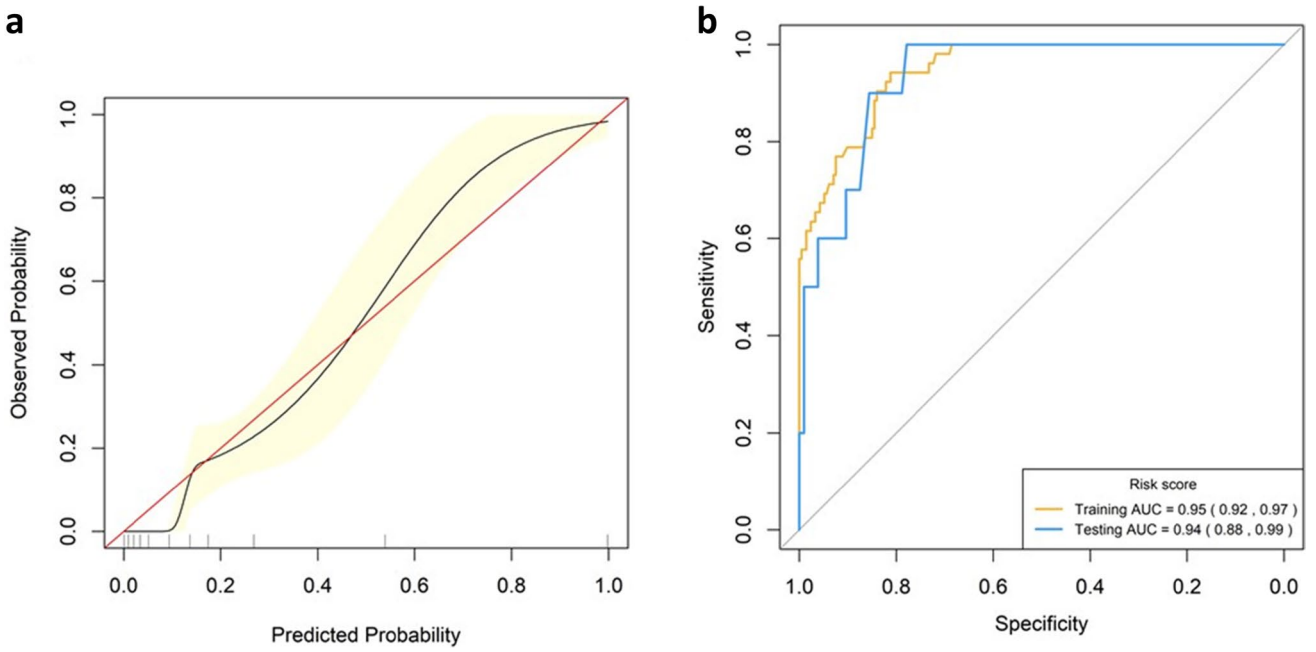

Fig. 3 Calibration curve and ROC curves for the predictive model. **a** The calibration curve. **b** The ROC curves

We calculated the adjusted OR of the clinical indicators from the logistic regression that used EBM predictions (Table 1). Albumin (< 40 g/L, aOR (95%CI) = 10.26(4.78,23.96)), globulin (< 20 g/L, aOR (95%CI) = 14.28(3.84,62.68)), creatinine level (< 59 umol/L, aOR (95%CI) = 10.26(1.00,3.49)), percentage of lymphocyte (> 40%, aOR (95%CI) = 6.51(1.51,32.39)), and percentage of monocyte (> 8%, aOR (95%CI) = 2.93(1.52,5.75)) were significantly associated with increasing odds of COVID-19. Monocyte count ($> 0.8 \times 10^9$/L, aOR (95%CI) = 0.08(0.03,0.18)), RDW-CV (> 15%, aOR (95%CI) = 0.09(0.03,0.24)), and percentage of neutrophil

Table 1 Adjusted OR of clinical indicators based on EBM predictions

Indicators	Levels	Observed number[a]		OR[b] (95%CI)
		CAP	COVID	
Albumin (g/L)	< 40	237	43	10.26 (4.78, 23.96)
	≥ 40	80	19	1.00 (Ref.)
Globulin (g/L)	< 20	4	14	14.28 (3.84, 62.68)
	≥ 20	313	48	1.00 (Ref.)
Monocyte count (10^9/L)	> 0.8	109	7	0.08 (0.03, 0.18)
	≤ 0.8	208	55	1.00 (Ref.)
RDW-CV (%)	> 15	70	3	0.09 (0.03, 0.24)
	≤ 15	247	59	1.00 (Ref.)
Creatinine level (umol/L)	< 59	90	28	10.26 (1.00, 3.49)
	59–104	29	2	0.08 (0.33, 3.52)
	> 104	198	32	1.00 (Ref.)
Percentage of lymphocyte (%)	< 20	145	10	0.91 (0.30, 2.80)
	> 40	20	16	6.51 (1.51, 32.39)
	20–40	152	36	1.00 (Ref.)
Percentage of monocyte (%)	> 8	188	45	2.93 (1.52, 5.75)
	≤ 8	129	17	1.00 (Ref.)
Percentage of neutrophil (%)	< 50	37	20	0.29 (0.07, 1.05)
	> 70	125	10	0.27 (0.08, 0.87)
	50–70	155	32	1.00 (Ref.)

[a]The observed number of patients with CAP or COVID

[b]Calculated based on the predictions from EBM

(> 70%, aOR (95%CI) = 0.27(0.08,0.87)) were found to be related to reduced odds of COVID-19. Due to limited sample size, other important contributing variables were not statistically significant or had infinite estimation from the multivariate logistic regression.

Figure 4 shows the inter-relationship among selected variables obtained from the decision tree model. Generally, five combinations showed a good ability to differentiate patients with COVID-19 from those with CAP. The first combination was globulin (< 20 g/L) and percentage of neutrophil (< 50%). This suggested that the prevalence of patients having these two indicators was 100%. The second combination was globulin (< 20 g/L), percentage of neutrophil (≥ 50%), percentage of monocyte (> 8%), and neutrophil count (≤ 7 × 109/L). The prevalence for patients within the second combination was 83.3%. The third combination was globulin (≥ 20 g/L), total bilirubin (< 5.1 umol/L), percentage of monocyte (≤ 8%), and lymphocyte count (> 4 × 109/L). The prevalence for patients within the third combination was 100%. The fourth combination was globulin (≥ 20 g/L), total bilirubin (< 5.1 umol/L), percentage of monocyte (> 8%), and hemoglobin level (≥ 105 g/L). The prevalence for patients within the fourth combination was 90%. The last combination was globulin (≥ 20 g/L), total bilirubin (≥ 5.1 umol/L), and albumin (40–55 g/L). The prevalence for patients within the last combination was 100%.

4 Discussion

4.1 Individual risk indicators of COVID-19

The top five most important contributing indicators are albumin, total bilirubin, monocyte count, ALT, and percentage of monocyte. Decreased serum albumin levels could be observed in acute and chronic infectious diseases [24]. Our study found that hypoalbuminemia was one of the top indicators for COVID-19, likely indicating higher albumin catabolism in patients with COVID-19 than CAP ($t = -2.34$, $P = 0.02$). Hypoalbuminemia was also described in a previous retrospective study of 99 patients with COVID-19 led by Liu etc. [4]. Liu's study found that the incidence of hypoalbuminemia among COVID-19 patients was as high as 98% and that hypoalbuminemia, the most common laboratory abnormality for patients with COVID-19, may be a vital predictor of disease severity [4]. Rod et al. identified 60 risk factors for COVID-19 severity by reviewing 17 articles and analyzing the consistency of the association between risk factors and a composite end-point of severe-fatal COVID-19, among which serum albumin was one of the factors with a high consistency of association [5].

Low levels of total bilirubin were also found to be one of the most important contributing variables for COVID-19 screen in our study. The evidence provided by prior studies is inconsistent. A few reports [25, 26] showed that

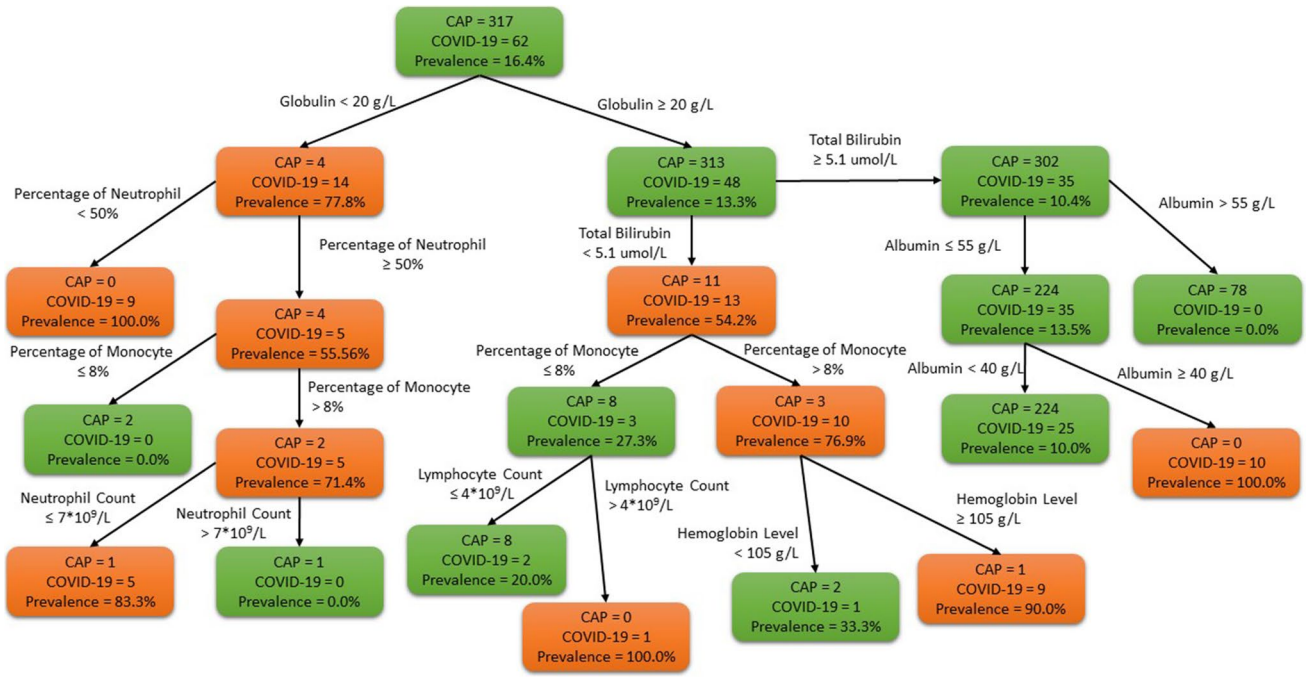

Fig. 4 The inter-relationship among selected clinical indicators

10.5–18% of patients with COVID-19 had a total bilirubin level elevated slightly at the time of admission, but most of the previous studies [27–29] demonstrated that the median of total bilirubin level in patients was in the normal range, and severe cases possessed higher levels than non-severe cases.

More importantly, a high percentage of monocyte and normal clinical levels of the monocyte count (compared to CAP) were found to be significant predictors of COVID-19 infection in our study. In the case of virus infection, the count and proportion of leukocytes might alter, including granulocytes, lymphocytes, and monocytes. Although the recognized hematologic abnormality of COVID-19 is currently lymphopenia, more literature has focused on the monocytes in the patients with COVID-19, especially the contribution of monocytes to immune response and the development of vaccines. Chen etc. analyzed the hematological changes of 113 deceased patients with COVID-19 and found the median monocyte count was in the normal range (0.4×10^9/L) [27]. In another study, there was no significant difference in the total number of monocytes between patients with COVID-19 and healthy individuals [30]. However, they discovered an increased proportion of activated monocytes in patients with COVID-19 by performing a detailed flow cytometric analysis of peripheral blood samples, which is consistent with our finding.

Our study also showed lower or normal ALT level in the patients with COVID-19 compared to CAP. However, some previous studies showed that 12–32% of the COVID-19 patients had elevated ALT levels at the time of admission [26, 29, 31]. Actually, these results are not contradictory. Our study focuses on the mean ALT level instead of the proportion of abnormal cases. All of the previous studies demonstrated the median or mean of total COVID-19 patients was in the normal range, which is consistent with our result.

4.2 Contributors grouped by system

To observe the predictive power of the joint effects of multiple predictors identified more intuitively and to improve the interpretability of the prediction model, we categorized and scored each variable according to systems. The top five prominent systematic predictive factors for COVID-19 are liver function, increasing monocyte, plasma protein, granulocyte, and renal function. These results might be related to the pathophysiological mechanism of COVID-19. Some recent studies [32, 33] demonstrated that liver damage in patients with severe COVID-19 may be caused by direct injury of SARS-CoV-2 through binding to the angiotensin-converting enzyme 2 (ACE2) receptor, the key receptor for SARS-CoV-2 cell entry, on cholangiocytes [34]. A cross-sectional study collected the liver function tests of 417 COVID-19 cases at admission and during hospitalization [32]. This study found that the presence of abnormal liver function testing after hospitalization for 2 weeks became more pronounced than that at admission; therefore, liver injury of the patient with COVID-19 was mainly related to medication used during hospitalization.

In addition, most prior studies identified significant morphological and functional differences in monocytes in COVID-19 [30]. Besides count and proportion, the morphological and function of monocytes varied in an obvious manner in the immune mechanism of COVID-19. Zhang etc. detected the phenotype of monocytes of COVID-19 patients by flow cytometry and found that those monocytes are expressed with CD11b, CD14, CD16, CD68, CD80, CD163, and CD206 and secrete IL-6, IL-10, and TNF-alpha [30]. At the same time, the patients with COVID-19 had larger monocytes under the microscope than the normal population. The possible mechanism is that the monocytes and macrophages in patients with COVID-19 may be infected by SARS-CoV-2, but after that, they can produce various cytokines and chemokines that can contribute to cytokine storm [35].

Furthermore, immunoglobulin is the most important component of globulin and along with albumin and total protein consists of plasma protein. Humoral immunity played a vital role in the immune responses during the SARS-CoV-2 infection. A prospective cohort followed 67 COVID-19 patients and found that patients with lower IgG titer had a higher rate of viral clearance [36]. As a result, we speculated that the decrease of globulin in our patients might be related to the depletion of immunoglobulin responsible for clearing the virus. As documented by other studies, a key difference between bacterial and viral pneumonia is found in the counts of granulocyte, which was also consistent with our result that granulocyte was one of the top five systematic indicators.

Finally, in accordance with our findings of elevated creatinine and BUN levels among the COVID cases [37], renal abnormalities and dysfunction have been reported to be closely associated with COVID-19, and the majority of patients with COVID-19 presented with proteinuria, hematuria, and acute kidney injury at the early stage of their infection [38]. More clinically important, the renal complications were reported to be related to high mortality among patients with COVID-19 [38].

4.3 Inter-relationship among risk factors

Our decision tree model identified five combinations of the inter-relationships among the selected variables. The most important combinations of the COVID-19 predictors with 100% prevalence at the end include (1) globulin < 20 g/L and percentage of neutrophil $< 50\%$ and (2) globulin ≥ 20 g/L, total bilirubin ≥ 5.1 umol/L, and albumin 40–55 g/L. Additionally, the combination with 90% prevalence comprises globulin ≥ 20 g/L, total bilirubin < 5.1 umol/L, percentage of monocyte $> 8\%$, and hemoglobin level ≥ 105 g/L; and the combination with 83.3% prevalence consists of globulin < 20 g/L, percentage of neutrophil $\geq 50\%$, percentage of monocyte $> 8\%$, and neutrophil count $\leq 7 \times 10^9$/L. These combinations are very convenient for physicians to use and could provide a key reference in early clinical screening. Hence, the predictive EBM model built in this study could provide a better prediction and minimize the potential misclassification error.

4.4 Machine learning models for COVID-19 screen

Our study aimed to help early screening for COVID-19 cases easier and affordable by using the existing and routinely collected clinical data, which could fill current knowledge gaps. In addition, there were some studies trying to use machine learning to help screen COVID-19, but these studies suffered from potential problems, including small sample sizes and outcome misclassification [12–14], requiring CT or image reports [15–17], validation issues [17], and non-representative controls [39]. To address the methodologic problems or uncertainty issues from the previous studies, we have used the methods to increase methodological validity as described below: (1) used all existing and routinely collected laboratory data to maximize feasibility and practical use of our findings; (2) reduced case misclassification bias by using clinically confirmed COVID-19 cases only; (3) used CAP patients as the controls; (4) in addition to using cross-validation for selecting the best combination of model hyperparameters, we split the data into training set and testing set and validated our model via testing set; and (5) published our model online at https://xdeng3.shinyapps.io/COVID-19/ for easy access, practical usage, or research purpose. Overall, the predictive probability from our model aligned with observed probability very well from 0.2 to 1.0, which demonstrated a good agreement between prediction and observation when the probabilities ranged between 20 and 100%. However, when the observed probability was less than 20%, our model had a little underestimation, which suggested the observed probability could be 2–7% higher than the predicted probability.

4.5 Clinical implications

Our study and model aimed to help the early screening for COVID-19 cases from CAP when the patient had initial symptoms including fever, respiratory symptoms, malfunctions of the liver or renal, or any suspected pneumonia symptoms. Clinicians could use our findings to screen cases earlier, tell if the patient should be quarantined based on fast blood tests, and then arrange these suspected cases for further diagnosis testing. It could help doctors' instructions on suspected patients before the RT-PCR testing results. Also, it could help patients self-determine if they should be quarantined or isolated based on the blood tests. To implement our model in practical settings, we developed

an online predictive tool for differentiating patients with COVID-19 from CAP (https://xdeng3.shinyapps.io/COVID-19/). On this website, probabilities of having COVID-19 infection will be calculated after entering the values of top predictors. In this way, early screening and quarantine of patients with COVID-19 could be achieved. Our findings may also help provide a basis to set the priority of arranging nucleic acid detection for suspected cases according to their clinical indicators.

4.6 Strengths and limitations

Our study had several strengths. First, we had innovatively used CAP patients as a control group for COVID-19. This control group had not been used in previous studies. Second, our study analyzed 35 routinely tested clinical indicators instead of just single, individual factors. All the indicators that we analyzed were the most used and collected routinely in clinical setting, which provide an effective tool by leveraging the resources and clinical markers routinely tested by clinical laboratories as the first step to screening COVID-19 infection from other common pneumonia. In addition, the advantages of machine learning models are that they are highly accurate and can easily handle multiple correlated variables. By building this diagnostic model, clinicians can screen patients with COVID-19 at an early stage, resulting in optimized use of test kits, helping early diagnosis, and early isolation or quarantine. Our study also had some potential limitations. A limitation that should be acknowledged was that we can only differentiate COVID-19 from CAP instead of all other respiratory diseases. However, CAP is the most common pneumonia which has similar symptoms as COVID-19. Another limitation was the small sample size we had. We would like to incorporate more cases in our model; however, due to the sensitivity of the COVID-19 data, it was difficult to obtain the same clinical indicators from previous studies. Additionally, as most prior literature did not involve similar clinical data as we did, it is difficult to compare and validate our findings. However, our finding on the associations of hypoalbuminemia, total bilirubin, monocyte, and ALT levels with COVID-19 was consistent with prior studies [4, 5, 25–27, 29–31]. In other words, we still identified some similar risk factors as found in previous studies even with the limited sample size, but we might have missed some other important risk factors due to our small sample size. Therefore, the findings obtained from our studies should be validated by future studies with larger sample sizes of COVID-19 cases. On the other hand, our predictive model could only help early screening for COVID-19 rather than make the early diagnosis, which is a common issue faced by most all predictive models.

5 Conclusion

We found that the top five contributing indicators of COVID-19 are albumin, total bilirubin, monocyte count, ALT, and percentage of monocyte, and these factors could help screen COVID-19 from CAP. Additionally, the top systematic predictors of COVID-19 are liver function, monocyte increasing, plasma protein, granulocyte, and renal function. We also identified the most optimal combinations of the selected predictors to interpret their interrelationships. These findings and innovative EBM methods may provide new directions to screen and distinguish suspected COVID-19 cases from common pneumonia.

Abbreviations CAP: Community-acquired pneumonia; COVID-19: Corona virus disease 2019 (COVID-19); RT-PCR: Real-time reverse transcription-PCR; RDW-CV: The coefficient of variation of red blood cell distribution width; BUN: Blood urea nitrogen; WBC: White blood cell count; ALT: Alanine aminotransferase; RBC: Red blood cell count; MCH: Mean corpuscular hemoglobin; ROC: Receiver operating characteristic; OR: Odds ratio; EBM: Explainable Boosting Machine

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/s11517-022-02568-2.

Acknowledgements We thank all the participants and doctors for their support in conducting this research.

Author contribution XL, SL, and KY contributed to the conception and design of the work. XL and HL contributed to the interpretation of data, drafted the work, and revised it; XD contributed to the acquisition, analysis, and drafting of the work. HL, ZQ, FX, KY, and GM contributed to the funding acquisition, data collection, and data curation. XL, HL, SF, and SL were major contributors to revising the manuscript. All authors read and approved the final manuscript.

Funding This work was financially supported by the Guangxi Critical Infectious Disease Center (2020281), Nanning Science and Technology Foundation (2018030), and the second batch of Guangxi medical high-level young and middle-aged discipline backbone training projects.

Data availability The data underlying this article cannot be shared publicly due to the privacy of individuals that participated in the study. The data will be shared on reasonable request to the corresponding author.

Declarations

Conflict of interest The authors declare no competing interests.

References

1. COVID-19 map - Johns Hopkins Coronavirus Resource Center. https://coronavirus.jhu.edu/map.html. Accessed 18 Aug 2021
2. Dramé M, TabueTeguo M, Proye E et al (2020) Should RT-PCR be considered a gold standard in the diagnosis of COVID-19? J Med Virol 92(11):2312–2313. https://doi.org/10.1002/jmv.25996

3. Dai WC, Zhang HW, Yu J et al (2020) CT imaging and differential diagnosis of COVID-19. Can Assoc Radiol J 71(2):195–200. https://doi.org/10.1177/0846537120913033
4. Liu Y, Yang Y, Zhang C et al (2020) Clinical and biochemical indexes from 2019-nCoV infected patients linked to viral loads and lung injury. Sci China Life Sci 63(3):364–374. https://doi.org/10.1007/s11427-020-1643-8
5. Rod JE, Oviedo-Trespalacios O, Cortes-Ramirez J (2020) A brief-review of the risk factors for covid-19 severity. Rev Saude Publica 54:60. https://doi.org/10.11606/S1518-8787.2020054002481
6. Sun Y, Li H, Pei Z et al (2020) Incidence of community-acquired pneumonia in urban China: a national population-based study. Vaccine 38(52):8362–8370. https://doi.org/10.1016/J.VACCINE.2020.11.004
7. File TM, Marrie TJ (2010) Burden of community-acquired pneumonia in North American adults. Postgrad Med 122(2):130–141. https://doi.org/10.3810/PGM.2010.03.2130
8. Hu K, lei Deng X, Han L, Xiang S, Xiong B, Pinhu L. Development and validation of a predictive model for feeding intolerance in intensive care unit patients with sepsis. Saudi J Gastroenterol Off J Saudi Gastroenterol Assoc. Published online 2021
9. Deng X, Thurston G, Zhang W et al (2021) Application of data science methods to identify school and home risk factors for asthma and allergy-related symptoms among children in New York. Sci Total Environ 770:144746. https://doi.org/10.1016/J.SCITOTENV.2020.144746
10. Wang S, Zha Y, Li W et al (2020) A fully automatic deep learning system for COVID-19 diagnostic and prognostic analysis. Eur Respir J 56(2):2000775. https://doi.org/10.1183/13993003.00775-2020
11. Wynants L, Van Calster B, Collins GS et al (2020) Prediction models for diagnosis and prognosis of covid-19: systematic review and critical appraisal. BMJ 369:26. https://doi.org/10.1136/BMJ.M1328
12. Feng C, Wang L, Chen X et al (2021) A novel artificial intelligence-assisted triage tool to aid in the diagnosis of suspected COVID-19 pneumonia cases in fever clinics. Ann Transl Med 9(3):201–201. https://doi.org/10.21037/ATM-20-3073
13. Wu J, Zhang P, Zhang L et al. Rapid and accurate identification of COVID-19 infection through machine learning based on clinical available blood test results. medRxiv. Published online April 6, 2020:2020.04.02.20051136. https://doi.org/10.1101/2020.04.02.20051136
14. Martin A, Nateqi J, Gruarin S et al (2020) An artificial intelligence-based first-line defence against COVID-19: digitally screening citizens for risks via a chatbot. Sci Rep 10(1):1–7. https://doi.org/10.1038/s41598-020-75912-x
15. Wang Z, Weng J, Li Z et al. Development and validation of a diagnostic nomogram to predict COVID-19 pneumonia. medRxiv. Published online April 6, 2020:2020.04.03.20052068. https://doi.org/10.1101/2020.04.03.20052068
16. Song CY, Xu J, He JQ, Lu YQ. COVID-19 early warning score: a multi-parameter screening tool to identify highly suspected patients. medRxiv. Published online March 8, 2020:2020.03.05.20031906. https://doi.org/10.1101/2020.03.05.20031906
17. Sun Y, Koh V, Marimuthu K et al (2020) Epidemiological and clinical predictors of COVID-19. Clin Infect Dis 71(15):786–792. https://doi.org/10.1093/CID/CIAA322
18. New coronavirus pneumonia treatment protocol. Published 2020. http://www.nhc.gov.cn/yzygj/s7653p/202002/8334a8326dd94d329df351d7da8aefc2/files/b218cfeb1bc54639af227f922bf6b817. Accessed 18 Aug 2021
19. Prina E, Ranzani OT, Torres A (2015) Community-acquired pneumonia. In: The Lancet, Vol 386. Lancet Publishing Group, pp 1097–1108. https://doi.org/10.1016/S0140-6736(15)60733-4
20. Lemaˆıtre Lemaˆıtre G, Nogueira F, Aridas char CK (2017) Imbalanced-learn: a python toolbox to tackle the curse of imbalanced datasets in machine learning, Vol 18. http://jmlr.org/papers/v18/16-365.html. Accessed 27 May 2021
21. Nori H, Jenkins S, Koch P, Caruana R. InterpretML: A unified framework for machine learning interpretability. Published online September 19, 2019. http://arxiv.org/abs/1909.09223. Accessed 23 Dec 2019
22. Qu Y, Deng X, Lin S et al (2021) Using innovative machine learning methods to screen and identify predictors of congenital heart diseases. Front Cardiovasc Med 8:797002. https://doi.org/10.3389/FCVM.2021.797002
23. Deng X, Zhang W, Lin S (2022) Package "APML" an approach for machine-learning modelling. https://cran.r-project.org/web/packages/APML/APML.pdf. Accessed 21 Jan 2022
24. Soeters PB, Wolfe RR, Shenkin A (2019) Hypoalbuminemia: pathogenesis and clinical significance. JPEN J Parenter Enteral Nutr 43(2):181. https://doi.org/10.1002/JPEN.1451
25. Chen N, Zhou M, Dong X et al (2020) Epidemiological and clinical characteristics of 99 cases of 2019 novel coronavirus pneumonia in Wuhan, China: a descriptive study. Lancet 395(10223):507–513. https://doi.org/10.1016/S0140-6736(20)30211-7
26. Guan W, Ni Z, Hu Y et al (2020) Clinical characteristics of coronavirus disease 2019 in China. N Engl J Med 382(18):1708–1720. https://doi.org/10.1056/nejmoa2002032
27. Chen T, Wu D, Chen H et al (2020) Clinical characteristics of 113 deceased patients with coronavirus disease 2019: retrospective study. BMJ 368:m1295. https://doi.org/10.1136/bmj.m1091
28. Bao J, Li C, Zhang K, Kang H, Chen W, Gu B (2020) Comparative analysis of laboratory indexes of severe and non-severe patients infected with COVID-19. Clin Chim Acta 509:180–194. https://doi.org/10.1016/j.cca.2020.06.009
29. Huang C, Wang Y, Li X et al (2020) Clinical features of patients infected with 2019 novel coronavirus in Wuhan, China. Lancet 395(10223):497–506. https://doi.org/10.1016/S0140-6736(20)30183-5
30. Zhang D, Guo R, Lei L et al. COVID-19 infection induces readily detectable morphological and inflammation-related phenotypic changes in peripheral blood monocytes, the severity of which correlate with patient outcome. medRxiv. Published online March 26, 2020:2020.03.24.20042655. https://doi.org/10.1101/2020.03.24.20042655
31. Yang W, Cao Q, Qin L et al (2020) Clinical characteristics and imaging manifestations of the 2019 novel coronavirus disease (COVID-19):a multi-center study in Wenzhou city, Zhejiang, China. J Infect 80(4):388–393. https://doi.org/10.1016/j.jinf.2020.02.016
32. Cai Q, Huang D, Yu H et al (2020) COVID-19: Abnormal liver function tests. J Hepatol 73(3):566–574. https://doi.org/10.1016/j.jhep.2020.04.006
33. Chai X, Hu L, Zhang Y et al. Specific ACE2 expression in cholangiocytes may cause liver damage after 2019-nCoV infection. bioRxiv. Published online February 4, 2020:2020.02.03.931766. https://doi.org/10.1101/2020.02.03.931766
34. Hoffmann M, Kleine-Weber H, Schroeder S et al (2020) SARS-CoV-2 cell entry depends on ACE2 and TMPRSS2 and is blocked by a clinically proven protease inhibitor. Cell 181(2):271-280.e8. https://doi.org/10.1016/j.cell.2020.02.052
35. Jamilloux Y, Henry T, Belot A et al (2020) Should we stimulate or suppress immune responses in COVID-19? Cytokine and anti-cytokine interventions. Autoimmun Rev 19(7):102567. https://doi.org/10.1016/j.autrev.2020.102567
36. Tan W, Lu Y, Zhang J et al. Viral kinetics and antibody responses in patients with COVID-19. medRxiv. Published online March 26, 2020:2020.03.24.20042382. https://doi.org/10.1101/2020.03.24.20042382

37. Yang X, Jin Y, Li R, Zhang Z, Sun R, Chen D (2020) Prevalence and impact of acute renal impairment on COVID-19: a systematic review and meta-analysis. Crit Care 24(1):356. https://doi.org/10.1186/s13054-020-03065-4
38. Pei G, Zhang Z, Peng J et al (2020) Renal involvement and early prognosis in patients with COVID-19 pneumonia. J Am Soc Nephrol 31(6):1157–1165. https://doi.org/10.1681/ASN.2020030276
39. Meng Z, Wang M, Song H et al. Development and utilization of an intelligent application for aiding COVID-19 diagnosis. medRxiv. Published online March 21, 2020:2020.03.18.20035816. https://doi.org/10.1101/2020.03.18.20035816

Xinlei Deng, MD, is currently a Ph.D. candidate from Department of Environmental Health Sciences, University at Albany. His research interests focus on machine learning modelling and predictive models.

Han Li, MD, is a clinical doctor from The People's Hospital of Guangxi Zhuang Autonomous Region. She was one of the frontline doctors that worked in the makeshift hospital for COVID-19.

Xin Liao, MD, is a clinical doctor from The People's Hospital of Guangxi Zhuang Autonomous Region and also was one of the frontline doctors that worked in the makeshift hospital for COVID-19.

Zhiqiang Qin, MD, PhD, is the director of the Department of Respiratory, The People's Hospital of Guangxi Zhuang Autonomous Region and specialize in community-acquired pneumonia.

Fan Xu, MD, PhD, is a clinical doctor from The People's Hospital of Guangxi Zhuang Autonomous Region and also was one of the frontline doctors that worked in the makeshift hospital for COVID-19.

Samantha Friedman, PhD, is an associate professor from Department of Sociology, University at Albany. Her research interests lie in the Urban Sociology and COVID-19.

Gang Ma, MD, is the vice-president of The People's Hospital of Guangxi Zhuang Autonomous Region. He has rich clinical experience and interpreted of the clinical significance of the results.

Kun Ye, MD, PhD, is the administrator of The People's Hospital of Guangxi Zhuang Autonomous Region. Her research focuses on the clinical medicine and published several COVID-19 papers.

Shao Lin, MD, PhD, is a professor in environmental health sciences. Her primary research interests are epidemiologic investigation of respiratory diseases, disaster research, and environmental health.