running title: An application for aiding COVID-19 diagnosis

Development and utilization of an intelligent application for aiding COVID-19 diagnosis

Zirui Meng¹,a, Minjin Wang¹,a, Huan Song²,a, Shuo Guo¹, Yanbing Zhou¹, Weimin Li³,
Yongzhao Zhou³, Mengjiao Li¹, Xingbo Song¹, Yi Zhou¹, Qingfeng Li⁴, Xiaojun Lu¹,
Binwu Ying¹,*

¹ Department of Laboratory Medicine, West China Hospital, Sichuan University,
Chengdu, Sichuan Province, China

² West China Biomedical Big Data Center, West China Hospital, Sichuan University,
Chengdu, Sichuan Province, China

³ Department of Respiratory and critical care medicine, West China Hospital, Sichuan
University, Chengdu, Sichuan Province, China

⁴ Department of Clinical Laboratory, Public Health Clinical Center of Chengdu,
Chengdu, Sichuan Province, China.

“Zirui Meng, Minjin Wang, Huan Song. contributed equally to this work.

Correspondence author: Binwu Ying. No.37 Guoxue Alley, West China Hospital,
Sichuan University, Chengdu 610041, Sichuan Province, P. R China.

Tel: 86-028-85422751
Fax: 86-028-85422751
Email: binwuying@126.com

NOTE: This preprint reports new research that has not been certified by peer review and should not be used to guide clinical practice.
Development and utilization of an intelligent application for aiding COVID-19 diagnosis

ABSTRACT

Background: COVID-19 has been spreading globally since emergence, but the diagnostic resources are relatively insufficient.

Results: In order to effectively relieve the resource deficiency of diagnosing COVID-19, we developed a machine learning-based diagnosis model on basis of laboratory examinations indicators from a total of 620 samples, and subsequently implemented it as a COVID-19 diagnosis aid APP to facilitate promotion.

Conclusions: External validation showed satisfiable model prediction performance (i.e., the positive predictive value and negative predictive value was 86.35% and 84.62%, respectively), which guarantees the promising use of this tool for extensive screening.

Keywords: Corona Virus Disease 2019; diagnostic; machine learning; laboratory testing; Application.

BACKGROUND

Since the outbreak of Corona Virus Disease 2019(COVID-19) in Wuhan, in December 2019, the epidemic has spread rapidly all over the world, which consequently brought great challenges to global public health(1). As of March 14, a total of 142,539 cases had been confirmed worldwide, with many more requiring tests(2). Currently, accurate nucleic acid detection plays a pivotal role in diagnosis and prevention of
COVID-19, and reverse transcription polymerase chain reaction is the main method for nucleic acid detection (3). However, it is not feasible to use this time-and-labor-consuming approach for screening a large growing number of suspected patients and asymptomatic infected patients (4). Rapid and universal screening method is essential to save medical resources, improve diagnosis efficiency and avoid cross-infection.

Taking advantage of the emerging machine learning technique, which enables analysis of the existing multidimensional data by applying appropriate algorithms for feature expression and classification, we have potential to improve the accuracy of diagnosis. For instance, in a recent study published in Nature, deep learning-based statistical model was used to improve breast cancer treatment and survival through identifying breast cancer patients with a high risk of long-term recurrence (5). In this study, we analyzed a variety of basic laboratory examinations indicators that related to host reactions to generate a free COVID-19 diagnosis aid APP. Then, the accuracy of this app was validated in an independent cohort (6).

**METHODS**

**Patients and information collection**

We have complied with all relevant ethical regulations for work with human subjects and the studies were approved by the Clinical Trials and Biomedical Ethics Committee of West China Hospital, Sichuan University. Through the West China hospital, a regional medical center, we collected the information of confirmed COVID-19 cases from various regional medical institutions, between 20th December
2019 and 10th February 2020. Diagnostic criteria were: RT-PCR of respiratory or
blood samples was positive for SARS-CoV-2 (Severe Acute Respiratory Syndrome
Coronavirus 2) nucleic acid; or viral gene sequencing of respiratory or blood sample
was highly homologous with SARS-CoV-2. Controls were patients collected in the
same site and the same time period, who were free of COVID-19 but had a diagnosis
of viral pneumonia. Results of laboratory examinations on admission were collected
which included age, gender and 35 indicators of whole blood count, coagulation test
and biochemical examination. In addition, the independent validation cohort consisted
of samples collected prospectively from one single medical center, using similar data
collection strategy.

Feature selection and model construction

Factors under consideration are age, sex and indicators of basic laboratory
examinations(see Table 1). Using 70% samples from the derivation cohort (i.e., the
training set), we first performed feature selection using Least Absolute Shrinkage and
Selection Operator (LASSO) to select the valuable indicators of COVID-19. The main
idea of LASSO for feature selection is to construct a first-order penalty function,
which can shrink the regression coefficient (b) of each variable within a certain range,
and eliminate the feature with a coefficient of 0, and finally obtain an optimally
refined model(7). The penalty terms for LASSO are: sum (abs(b)) <= t. Then these
indicators were used to construct a primitive model by multivariate logistic regression.
The parameters were further adjusted through internal validation (i.e., the other 30%
samples, the testing set) to obtain the optimized final model.
Evaluation

In external validation cohort, we compared the predicted results, with the follow-up results to evaluate the diagnostic validity by positive predictive values, and negative predictive values. Finally, to facilitate further use in clinic setting, all model parameters and code were encapsulated as a visual application (APP), COVID-19 Diagnosis Aid APP.

Statistical analysis

Continuous variables are represented by the median (upper and lower quartiles). Categorical variables are expressed in terms of frequency. LASSO algorithm was used for feature selection and the model was constructed by multivariate logistic regression. The diagnostic performance of the model was assessed by AUC. Calibration curves and Hosmer-Lemeshow test were used to evaluate the degree of overestimation or underestimation of the model. DCA were used to measure the net clinical benefits. The LASSO algorithm was performed by “glmmnet” package. The logistic regression model was constructed by “glm” package. All statistical analyses were completed using R 3.5.0 version.

RESULTS

Participants and Clinical Characteristics

A total of 620 samples were included in the derivation cohort, among which 431 samples (211 COVID-19 vs. 220 Control) were in the training set and 189 samples (91 COVID-19 vs. 98 Control) were in testing set by simple randomization. The frequency of COVID-19 in the training set (48.96%) was not significantly different...
from that in the testing set (48.14%). We included 145 samples in the independent
validation cohort.

Model development and evaluation

Through Lasso regression screening and Multivariate logistic regression, 9
representative variables (age, Activated Partial Thromboplastin Time, Red Blood Cell
Distribution Width-SD, Uric Acid, Triglyceride, Serum Potassium, Albumin/globulin,
3-Hydroxybutyrate, Serum Calcium) with good identification value were selected and
constructed an optimized diagnostic model (see Figure 1). According to the suggestive
information from the model, among 145 samples that used for validation, 80 samples
were estimated as COVID-19 and 65 samples were considered COVID-19-free.

Compared with the status determined by nucleic acid detection (the golden standard),
we found 69 real COVID-19 samples were correctly identified by the app (69/80,
positive predictive value = 86.25%), while 55, out of the 65 COVID-19-free samples,
were confirmed to be negative (55/65, negative predictive value = 84.62%). The area
under curve (AUC) of the model were 0.890 and 0.872 in the testing set and
independent validation cohort, respectively (see Figure 2) (8). The calibration curve
performed well and Hosmer-Lemeshow test results P value was much greater than
0.05 (see Figure 3). The DCA quantitatively demonstrated high clinical net benefit
over the entire probability threshold (see Figure 4) (9).

Construction of application

The model was encapsulated as an APP, COVID-19 Diagnosis Aid APP. The risk of
COVID-19 can be obtained by inputting the required indicators into the APP. It will
be online in the android store and IOS APP Store soon for facilitating the further application (see Figure 5).

**DISCUSSION**

According to the latest epidemiological information of the World Health Organization (WHO), SARS-CoV-2 is spreading worldwide with an intensifying situation, and many countries including South Korea, Italy, and the United States have experienced outbreaks (10). Timely diagnosis, isolation and treatment are critical to control the global spread of COVID-19. However, nucleic acid testing is not always feasible or fully affordable in all regions. We therefore need more ways to realize the early and accurate community control and screening of suspicious people. To our knowledge, this is the first COVID-19 Diagnosis Aid APP based on laboratory tests for extensive screening. It has good accuracy and stability, which can help medical institutions to predict patients’ situation in advance and accurately invest the limited healthcare resources, so as to effectively control the spread of the epidemic.

Based on the optimized diagnostic model developed in current study, users can obtain the probability of COVID-19 according to the simple, easily available and reliable indicators from laboratory examinations. Given satisfiable model prediction performance (i.e., AUC was 0.872; positive predictive value and negative predictive value was reach 86.25% and 84.62%, respectively in independent validation cohort), the application of this app can provide reliable suggestion for further assessment of the index person, which has potential to reduce anxiety of public, as well as unnecessary hospital visit and nucleic acid testing.
It is undeniable that nucleic acid detection is one of the most reliable standards for the diagnosis of COVID-19, but we should also be aware of its shortcomings and deficiencies in practice. Nucleic acid testing requires professional technical platforms and physicians, and false negative results due to multiple reasons cannot be completely avoided\(^\text{11}\). Instead, laboratory indicators that identified in our study are accessible, and can be widely used in community hospitals by family doctors to complete the early stage of suspected population screening.

On the other hand, based on the technological revolution brought by the Internet, this APP software installed on personal mobile can not only help to discover COVID-19 in the most convenient way, but also timely monitor dynamic trends of the prevalence of the contagious disease through data transmission to the Disease Control Center or professional medical institution\(^\text{12, 13}\). This provides real-time data with predictive value, which can be reliable basis for more accurate health and epidemic prevention policies. Furthermore, this APP software is continuously upgradeable, with the possibility of extended scope of application. For example, we have added the data acquisition function which enabled the documentation of epidemiological and clinical features and a personalized reminder for necessary revisits. Also, with further implementation of communication function, it is possible to establish a direct pre-connection between individuals and medical centers, which may benefit patients by avoiding unnecessary hospital visits or shorten waiting time spend before admission\(^\text{14}\).

**CONCLUSION**
COVID-19 Diagnosis Aid APP can efficiently and accurately calculate the infection probability through simple and easily-obtained laboratory examination results. It will promote the screening of a large number of suspected people, save limited medical resources, optimize the diagnosis process, and it can constantly learn, adapt and upgrade, which are vital to control the global spread of COVID-19.

Ethics approval and consent to participate

The protocol of this study was approved by the West China Hospital, Sichuan University Medical Ethics Committee and conformed to the principles of the Declaration of Helsinki.

The registry number for clinical trial

ChiCTR2000030542

Consent for publication

Not applicable

Data Availability Statement

The datasets used and analysed during the current study are available from the corresponding author on reasonable request.

Competing interests

The authors declare that they have no competing interests

Funding

This study was supported by Science and Technology Department of Sichuan Province “2020YFS0004” and Science and Technology Project of West China
Authors' contributions

Zirui Meng analyzed the data and was a major contributor in writing the manuscript.

Minjin Wang designed the APP framework and was a major contributor in writing the manuscript. Huan Song analyzed the data and was a major contributor in writing the manuscript. Shuo Guo analyzed the data. Yanbing Zhou analyzed the data. Weimin Li analyzed patients' medical records. Yongzhao Zhou analyzed patients' medical records. Mengjiao Li collected the data. Xingbo Song verified the performance of the app. Yi Zhou verified the performance of the app. Qingfeng Li collected the data. Xiaojun Lu designed the functions of the app. Binwu Ying proposed research ideas and design research programs. All authors read and approved the final manuscript.

Acknowledgements

The authors thank Beijing Zhongben Tech.co., Ltd. for their technical contributions to this study.

REFERENCE


Table 1. Candidate of the model.

<table>
<thead>
<tr>
<th></th>
<th>Derivation cohort</th>
<th>Validation cohort</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Control</td>
<td>COVID-19</td>
</tr>
<tr>
<td><strong>SEX</strong></td>
<td>female(27.41%)</td>
<td>female(31.07%)</td>
</tr>
<tr>
<td><strong>AGE</strong></td>
<td>68.5(77.81)</td>
<td>46(55.73)</td>
</tr>
<tr>
<td><strong>PT</strong></td>
<td>11(11.75,13.25)</td>
<td>10.7(11.8,13.4)</td>
</tr>
<tr>
<td><strong>INR</strong></td>
<td>0.95(1.06,1.15)</td>
<td>0.94(1.04,1.16)</td>
</tr>
<tr>
<td><strong>APTT</strong></td>
<td>28.18(29.75,35.68)</td>
<td>25.60(27.90,31.40)</td>
</tr>
<tr>
<td><strong>TT</strong></td>
<td>17.10(18.00,19.13)</td>
<td>16.40(17.20,18.10)</td>
</tr>
<tr>
<td><strong>RBC</strong></td>
<td>3.13(3.76,4.13)</td>
<td>3.66(4.25,4.76)</td>
</tr>
<tr>
<td><strong>HGB</strong></td>
<td>93.25(110.50,126.50)</td>
<td>111.00(129.00,148.00)</td>
</tr>
<tr>
<td><strong>HCT</strong></td>
<td>0.29(0.35,0.39)</td>
<td>0.35(0.38,0.44)</td>
</tr>
<tr>
<td><strong>MCV</strong></td>
<td>88.13(92.20,96.30)</td>
<td>88.50(91.50,95.00)</td>
</tr>
<tr>
<td><strong>MCH</strong></td>
<td>29.00(29.90,31.20)</td>
<td>28.90(30.70,32.10)</td>
</tr>
<tr>
<td><strong>MCHC</strong></td>
<td>317.00(325.00,335.25)</td>
<td>320.00(334.00,345.00)</td>
</tr>
<tr>
<td><strong>RDW-SD</strong></td>
<td>44.58(47.45,51.13)</td>
<td>41.90(44.50,49.80)</td>
</tr>
<tr>
<td><strong>PLT</strong></td>
<td>135.75(186.00,230.75)</td>
<td>107.00(147.00,202.00)</td>
</tr>
<tr>
<td><strong>WBC</strong></td>
<td>5.73(7.39,9.02)</td>
<td>6.52(9.59,14.89)</td>
</tr>
<tr>
<td><strong>TBIL</strong></td>
<td>6.20(9.75,15.18)</td>
<td>8.50(12.20,17.50)</td>
</tr>
<tr>
<td><strong>DBIL</strong></td>
<td>2.48(3.80,6.28)</td>
<td>3.70(5.70,7.50)</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>----</td>
<td>----</td>
<td>----</td>
</tr>
<tr>
<td>IBIL</td>
<td>3.78(4.95,6.98)</td>
<td>4.00(6.10,9.10)</td>
</tr>
<tr>
<td>ALT</td>
<td>12.00(19.00,30.25)</td>
<td>18.00(27.00,55.00)</td>
</tr>
<tr>
<td>AST</td>
<td>18.75(23.00,36.00)</td>
<td>22.00(36.00,68.00)</td>
</tr>
<tr>
<td>TP</td>
<td>53.90(58.85,68.13)</td>
<td>60.80(65.90,72.10)</td>
</tr>
<tr>
<td>ALB</td>
<td>32.70(36.10,41.65)</td>
<td>30.70(34.10,39.00)</td>
</tr>
<tr>
<td>GLB</td>
<td>19.43(23.80,28.45)</td>
<td>27.50(31.80,35.90)</td>
</tr>
<tr>
<td>A/G</td>
<td>1.33(1.51,1.91)</td>
<td>0.90(1.11,1.32)</td>
</tr>
<tr>
<td>GLU</td>
<td>4.92(6.00,7.64)</td>
<td>6.31(7.33,9.67)</td>
</tr>
<tr>
<td>UREA</td>
<td>4.60(6.05,8.25)</td>
<td>4.20(6.20,10.40)</td>
</tr>
<tr>
<td>CREA</td>
<td>50.50(60.50,103.50)</td>
<td>60.00(75.00,94.00)</td>
</tr>
<tr>
<td>eGFR</td>
<td>60.87(89.49,97.73)</td>
<td>66.14(96.38,104.66)</td>
</tr>
<tr>
<td>Cys-C</td>
<td>0.90(1.06,1.69)</td>
<td>0.86(1.06,1.42)</td>
</tr>
<tr>
<td>URIC</td>
<td>154.75(219.00,291.25)</td>
<td>181.00(281.00,366.00)</td>
</tr>
<tr>
<td>TG</td>
<td>0.71(0.93,1.23)</td>
<td>0.98(1.17,1.80)</td>
</tr>
<tr>
<td>CK</td>
<td>41.75(56.00,86.75)</td>
<td>39.00(87.00,204.00)</td>
</tr>
<tr>
<td>NA</td>
<td>135.30(139.20,141.83)</td>
<td>131.40(135.50,139.10)</td>
</tr>
<tr>
<td>K</td>
<td>3.72(4.08,4.38)</td>
<td>3.32(3.76,4.09)</td>
</tr>
<tr>
<td>β-HBA</td>
<td>0.07(0.12,0.23)</td>
<td>0.08(0.23,0.38)</td>
</tr>
<tr>
<td>CA</td>
<td>2.06(2.15,2.21)</td>
<td>2.01(2.11,2.23)</td>
</tr>
</tbody>
</table>

Data are presented as n (%) for categorical variables and as median (interquartile range) for continuous variables.

PT, prothrombin time; INR, International Normalized Ratio; APTT, activated partial thromboplastin time; TT, thromboplastin time; RBC, red blood cell; HGB, hemoglobin; HCT, hematocrit; MCV, Mean Corpuscular Volume;
MCH, Mean Corpuscular Hemoglobin; MCHC, Mean Corpuscular Hemoglobin Concentration; RDW-SD, Red Blood Cell Distribution Width-SD; RDW-CV, Red Blood Cell Distribution Width-CV; PLT, Platelets; WBC, white blood cell; TBIL, total Bilirubin; DBIL, direct bilirubin; IBIL, indirect bilirubin; ALT, alanine aminotransferase;
AST, glutamic oxaloacetic transaminase; TP, Total Protein; Alb, albumin; ALB, albumin; GLB, globulin ratio; A/G, albumin globulin ratio; GLU, glucose; CREA, serum; eGFR, estimated Glomerular Filtration Rate; Cys-C, cystatin; URIC, uric acid; TG, Triglyceride; CHOL, cholesterol; CK, creatine kinase; NA, serum natrium; K, Serum Potassium; β-HBA, 3-Hydroxybutyrate; CA, Serum Calcium;

Figure legends

Figure 1(a) Coefficient diagram of LASSO variables. Each curve in the figure represents the trajectory of the coefficient of an independent variable. The ordinate is the value of the coefficient. The lower abscissa, $\lambda$, is the parameter that controls the severity of the penalty. The upper abscissa is the number of non-zero coefficients in the model under the penalty parameter. (b) Adjustment parameters in the LASSO model. The lambda is screened by 10 folds cross-validation. A dashed vertical line is drawn at one standard error (1-SE standard) of the minimum and minimum standards. Lambda.1se corresponds to a model with good performance but the fewest number of arguments.

Figure 2. Receiver operator characteristic curve. The AUC were 0.890 and 0.872 in derivation and validation cohort respectively.

Figure 3. Calibration curves. The 45° shaded line represents the ideal prediction and the prediction probability is consistent with the actual observation probability. The
Solid line represents the actual prediction of the model. The stapled histogram on the bottom line represents the distribution of patients' predicted probability. Abbreviations: $D_{xy}$ = Somer’s D rank correlation, $R^2$ = Nagelkerke-Cox-Snell-Maddala- Magee R-squared index, $D$ = Discrimination index, $U$ = Unreliability index, $Q$ = Quality index, $E_{\text{max}}$ = maximum absolute difference in predicted and calibrated probabilities, $S:z = \text{Spiegelhalter Z-Test}, S:p = \text{two-tailed p-value of the Spiegelhalter Z-test}$.

**Figure 4. Decision curve analysis.** The horizontal axis is the threshold probability of occurrence of COVID-19. The vertical axis shows the clinical benefits that patients may gain or lose using the APP. Dotted line: prediction model. Solid line: all patients were COVID-19. Horizontal line: all patients were not COVID-19.

**Figure 5. APP schematic.** The model can show the probability of COVID-19 after inputing the corresponding index into the the APP.
West China Hospital COVID-19 Screening Assistant

Fill in all test indicators - submit

Check the hospital: Please enter hospital name
Check date: Please select the inspection date
Contact Way: Please enter contact way

Personal data

My case

Feedback

Contact us

Settings

Diagnostic probability

Probability value: 86.20%

According to your test results, you are at higher risk of COVID-19. Do not be panic. Please observe your physical condition closely. It is recommended to wear a mask and consult a local doctor.

My case

2023-05-18 09:06:61  Probability value: 98.2%
Check the hospital: westChina
Check date: 2023-03-18
Contact Way: 15770917078
Gender: 女
Age: 23

2023-05-17 15:26:58  Probability value: 94.15%
Check the hospital: westChina
Check date: 2023-03-17
Contact Way: 15880016876
Gender: 女
Age: 23

2023-05-17 11:17:00  Probability value: 99.71%
Check the hospital: westChina
Check date: 2023-03-12
Contact Way: 15880016876
Gender: 女