THE LANCET
Digital Health

## Supplementary appendix

This appendix formed part of the original submission and has been peer reviewed. We post it as supplied by the authors.

Supplement to: Wang M, Xia C, Huang L, et al. Deep learning-based triage and analysis of lesion burden for COVID-2019: a retrospective study with external validation. *Lancet Digital Health* 2020; **2:** e506-15.

# Supplementary Materials

## Appendix 1 Algorithm development

Deep learning driven by big data is a sub-area of machine learning, and usually uses convolutional neural networks with a large number of parameters to build a model. We employed the architecture of U-Net to develop our deep learning-based triage and lesion burden analysis system.

### 1.1 Model Architecture

We used 2D U-Net[1] with a slight modification to predict a sigmoid-based confidence score map for the presence of opacity regions and Model architecture is illustrated in Figure E.1..

Three consecutive CT slices were fed into the three-channel structure, thereby incorporating some spatial information for training the 2D model[2]. In the downsampling path, the number of convolution filters in each block were 32, 64, 128, 256 and 512. In the upsampling path, the number of convolution filters in each block were 256, 128, 64 and 1. Moreover, we had introduced a residual connection structure as ResNet[3] in both the downsampling path and the upsampling path. This structure consisted of a 1x1 convolution with a stride of 1 and a maximum pooling layer with a stride of 2. The number of convolution filters in the residual connection structure was consistent with the corresponding block.

The output of the model is a confidence score mask of the same size as the middle slice of the input, with confidence scores ranging from 0-1 for each pixel. For a scan-level prediction score, we selected the highest confidence score on the output score map. We compared case-wise classification accuracy of pixel level highest confidence score versus the averaged pixel confidence score in a lesion region. The case triage performance was comparable between the two methods with highest pixel confidence score demonstrating a slightly higher AUC (diff = 0.007) result (Figure E.2).

### 1.2 Pre-processing

We randomly selected a slice from the training set and identified two slices adjacent to it, combined these three slices into a three-channel input in order and transformed the pixel values using a lung window (window level = -600, window width = 1500).

In the training phase, we cropped input data to a size of 3 × 448 × 448 pixels before passing it to our deep learning model to make sure of parallel processing that can achieve a higher computational efficiency. In the validation phase, a uniform input size of 448*448 was not necessary, as the U-Net structure can propagate forward input images of any size and therefore do not require a fixed input size. Thus, we entered CT images with its original size (3 × 512 × 512 pixels) into the model[4].

### 1.3 Training Details

A *train* set was used to train the model while a *val* set was used to select the hyper-parameters. We randomly initialized weights of our model. During training, data augmentation strategy partly followed He et al.[3], including random crops and pixel intensity augmentation. We used Adam with a learning rate of 0.0003 and a batch size of 96. The loss function was weighted cross entropy with a weight of 5 for positive pixels. After 50 training epochs, we began to evaluate model performance on the *val* set. If the loss of *val* set did not decrease on 20 consecutive epochs, we stopped training and chose the model with the least loss on the *val* set as the final model. Training was done on Nvidia GeForce RTX 2070 GPU with PyTorch framework.

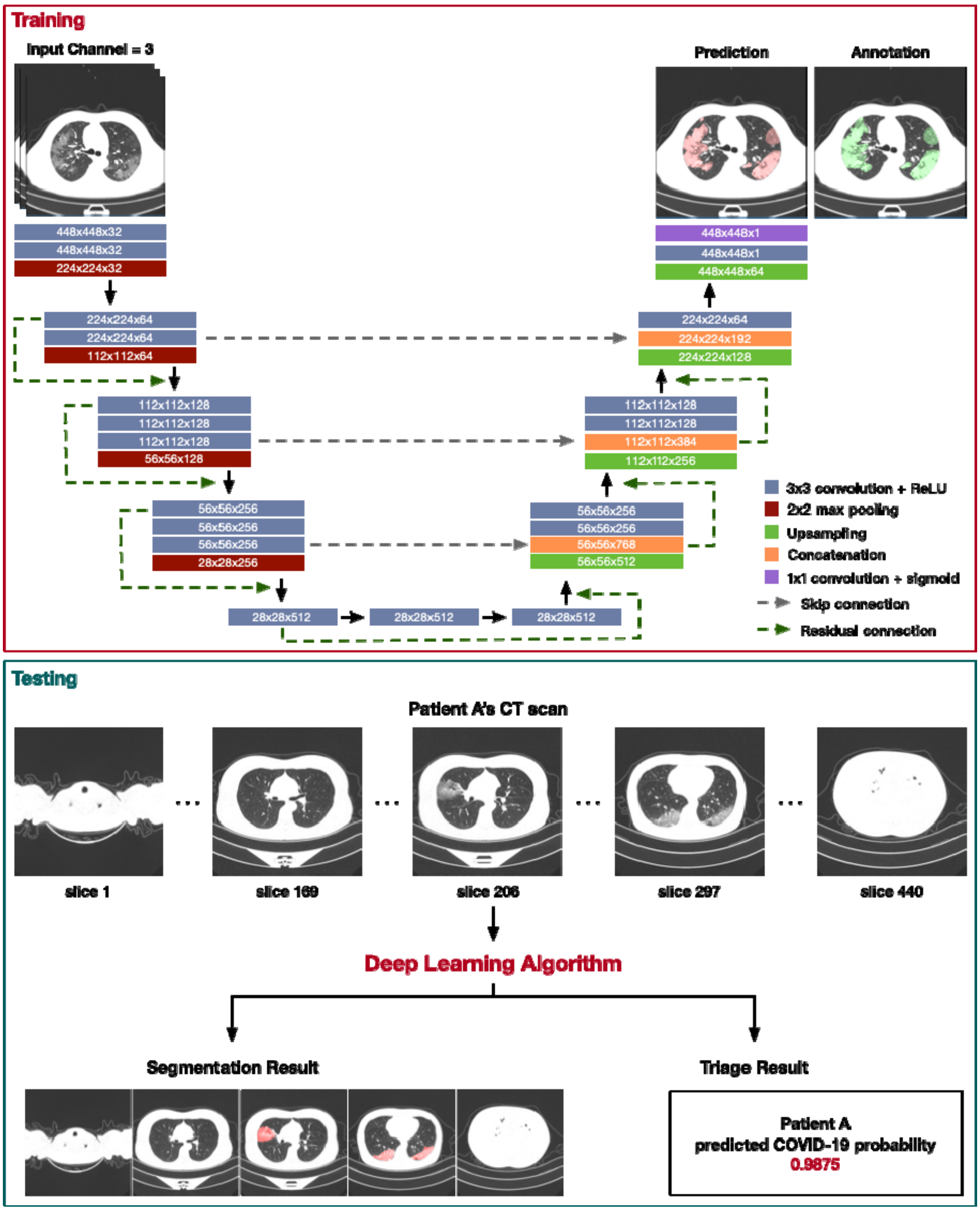

**Figure E. 1 Model architecture and algorithm training process**

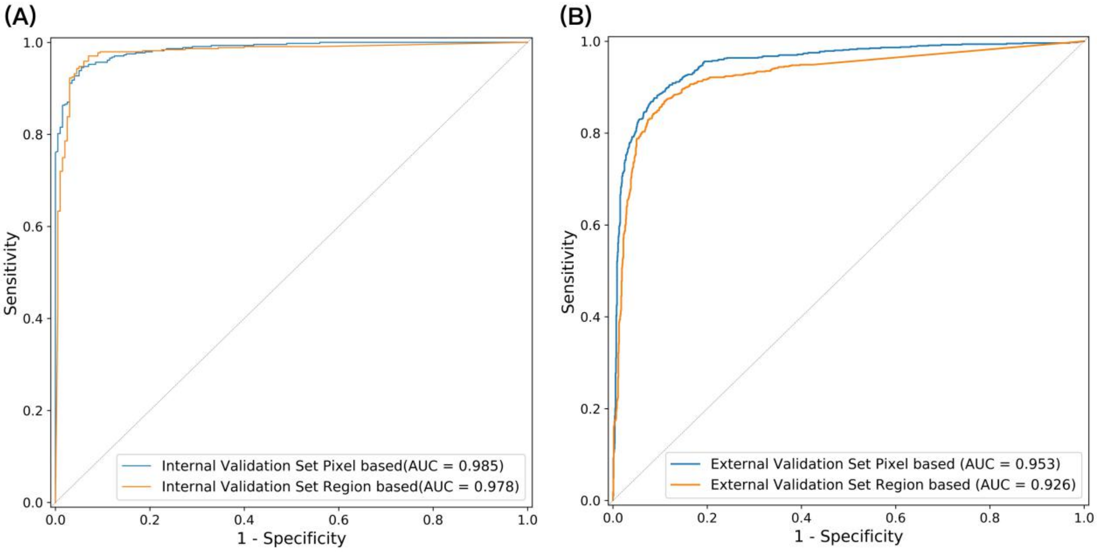

**Figure E. 2 ROC of pixel level highest confidence score vs ROC of averaged pixel confidence score in a lesion region (A. Internal validation set and B. External validation set)**

## Appendix 2 Data annotation and sample size estimation

### 2.1 Development set annotation

A group of radiologists annotated lung opacities with contours in the positive CT scans of the development set. Opacities referred to chest CT features related to COVID-19. According to the Chinese Clinical Guidance for COVID-19 Pneumonia Diagnosis and Treatment (7th edition) published by National Health Commission of the People's Republic of China[5], chest imaging features related to COVID-19 were defined as multiple small patchy shadows, interstitial changes, multiple ground-glass shadows and infiltrates shadows and pulmonary consolidation. For each CT scan from a COVID-19 patient, two radiologists were responsible for its annotation: one junior radiologist first annotated the slices by delineating the opacities and the second senior one revised and approved the annotations. Radiologists annotated the CT scans on InferScholar Centre (version 3.2), a research platform developed by Infervision which allowed viewing of medical imaging and image segmentation annotation.

During COVID-19 opacity annotation, radiologists were allowed to skip annotating slices adjacent to an already annotated slice if the slices had very similar opacity distribution. In case they encountered cases where opacities were too scattered to annotate, the first radiologist notified the second radiologist and if both agreed that manual annotation was almost impossible, the scan was excluded from the set. 105 scans were excluded due to pervasive lesions and difficulty in annotation. In total, 41 603 slices were annotated with 32 839 in the *train* set and 8 764 in the *val* set.

### 2.2 Sample size estimation for external validation

Minimum sample sizes were calculated prior to data collection for external triage performance validation and lesion burden analysis assessment. For the external validation of triage accuracy, at a target sensitivity of 90% and an estimated sensitivity of 95% we needed at least 231 positive cases; at a target specificity of 80% and an estimated specificity of 85% we needed at least 466 negative cases. As for the validation of lesion burden analysis, at a target kappa coefficient of 0.60 (the lower bound of "substantial agreement" by convention), an estimated kappa of 0.81 (the lower bound of "almost perfect agreement") and a 1:1 ratio of opacification increase versus no increase we needed at least 97 pairs of scans to evaluate the algorithm's accuracy at detecting volume increase in opacities. Sample size estimation above was done at α of 0.05 (two-tailed) and (1-β) of 0.8 using R (version 3.6.2).

### 2.3 Statistical analysis plan

The primary outcome measurements of triage accuracy are sensitivity and specificity and the primary outcome measurement of triage efficiency is the time difference between AI-aided triage and regular clinical workflow. To evaluate AI-aided triage accuracy, comparisons will be made between the triage sensitivity on the external validation set and the target sensitivity of 90% as well as between the triage specificity on the external validation set and the target specificity of 80%. For AI-aided triage efficiency, the time differences between regular clinical workflow and AI-aided triage under the two proposed triage pathways of scan-to-second-reader and scan-to-fever-clinician triage will be compared to 0 to evaluate whether AI could expedite the identification of suspected COVID-19 patients.

## Appendix 3 Definition of Suspected COVID-19 Cases in China

According to the Chinese Clinical Guidance for COVID-19 Pneumonia Diagnosis and Treatment (7th edition) published by National Health Commission of the People's Republic of China,[5] identifying suspected cases need to analyse epidemiological history and clinical manifestations comprehensively. A case can be identified as a suspected case, if the patients meets any one of the epidemiological history criteria and any two of the clinical manifestation, or all three clinical manifestations.

Epidemiological history:

1) Travel or residence history of Wuhan and surrounding areas, or other communities with documented COVID-19 positive cases within 14 days before the onset of illness;

2) History of contact with COVID-19-infected persons (positive for nucleic acid detection) within 14 days before the onset of illness.

3) History of contact with the patients presenting fever or respiratory symptoms, who travel to or reside in Wuhan and surrounding areas, or in other communities with documented COVID-19 positive cases within 14 days before the onset of illness.

4) Clustering onset (2 or more cases of fever and/or respiratory symptoms within 2 weeks in small areas such as home, office, school class, etc.)

Clinical manifestation:

1) Presenting with fever and/or respiratory symptoms.

2) With imaging features of COVID-19 pneumonia:

At the early stage of the disease, multiple small patchy shadows and interstitial changes appear, which are more obvious in the periphery of the lung. Then it developed into multiple ground-glass shadows and infiltrates shadows. In severe cases, pulmonary consolidation may occur. Pleural effusion is rare.

3) In the early stage of the disease, the total number of leukocytes was normal or decreased, and the lymphocyte count was normal or decreased.

## Appendix 4 Validation on Non-fever-clinic Patient Population

***Table E. 1:*** **AI's sensitivity on CT scans from Fangcang Hospital patients (all patients were confirmed by RT-PCR testing and clinically manifested as mild or asymptomatic cases).**

| | Fangcang Cases |
|---|---|
| Total No. of RT-PCR Confirmed Patients | 722 |
| Age | 49 (IQR: 37-57) |
| Gender | |
| Male | 367/722 (51%) |
| Female | 355/722 (49%) |
| Total No. of CT Scans | 761 |
| Positive scans | 618/761 (81%) |
| Negative scans | 143/761 (19%) |
| AI's Sensitivity on All CT Scans | 0.736 (95%CI: 0.720-0.752) |
| AI's Sensitivity on Positive CT Scans | 0.886 (95%CI 0.873-0.898) |

Data are n (%), unless otherwise stated. IQR: interquartile range.

***Table E. 2:*** **AI's specificity on absolutely non-COVID-19 cases collected from Tianyou Hospital and the Third People's Hospital of Shenzhen (all cases visited hospital and received chest CT examination for respiratory problems in October 2019)**

| | Absolutely Non-COVID-19 Cases |
|---|---|
| Total No. of Patients | 651 |
| Age | 57 (IQR: 43-68) |
| Gender | |
| Male | 405/651 (62%) |
| Female | 246/651 (38%) |
| Total No. of CT Scans | 686 |
| Positive scans | 652/686 (95%) |
| Negative scans | 34/686 (5%) |
| AI's Specificity on All Cases | 0.822 (95%CI: 0.808-0.836) |

Data are n (%), unless otherwise stated. IQR: interquartile range.

## Appendix 5 Three-dimensional video display of lung infection of an averaged COVID-19 patient

Of the 2086 CT scans from confirmed COVID-19 patients collected from Tongji Hospital, 946 patients were transferred to Tongji Hospital from other hospitals due to severe health conditions and the rest 1140 patients were admitted by Tongji Hospital on its own. To construct an average COVID-19 lung, we excluded those 946 transfer patients for they could over-represent the population with severe and critically severe cases and only used CT scans from the 1140 patients. We first used AI to segment lung opacities and then used a deep learning lung lobe segmentation algorithm developed previously by Infervision to segment lung lobes. We linearly co-registered the segmentation masks of lungs and lesions to produce an average lung of COVID-19 patients.[6, 7] From the video, we can observe that the CT features of COVID-19 pneumonia are bilateral and subpleural areas of opacification affecting mainly the lower lobes. Compared to the left lung, the area of infection in the right lung tends to be larger, which is consistent with the findings by Shi et al.[8]

## Appendix 6 CT Technical Information

All chest CT exams were performed with the patient in the supine position and with breath-holding following inspiration. The scanning range was from apex to the base of the lungs. For specific technical parameters of CT scan protocol of each hospital, please see Table E. 3.

***Table E. 3:*** **CT scanning parameters of five hospitals**

| | **Tongji Hospital** | **Tianyou Hospital** | **Xianning Hospital** | **Second Xiangya Hospital** | **Third People's Hospital of Shenzhen** |
|---|---|---|---|---|---|
| Tube Voltage, kVp | 110-120 | 120 | 120 | 130 | 120 |
| Tube Current, mAs | Automatic | Automatic | Automatic | Automatic | Automatic |
| Pitch | 0.8-1.375 | 0.562-1.75 | 0.03-1.5 | 1.5 | 0.8-1.0875 |
| Matrix | 512×512 | 512×512 | 512×512 | 512×512 | 512×512 |
| Reconstruction Slice Thickness, mm | 0.5-1.250 | 0.65 | 0.6 | 0.8 | 0.625-2.0 |

1 **Appendix 7 Clinical and Laboratory Information of COVID-19 Patients with pairs of CT Scans**

*Table E. 4:* **Characteristics of COVID-19 patients with pairs of CT scans**

| | All Patients (n=100) | No Volume Increase in Opacity (n = 48) | Opacity Volume Increase (n =52) |
|---|---|---|---|
| Median of age (IQR), years | 53 (41-65) | 57 (45-66) | 54 (43-65) |
| Sex | | | |
| Male | 50/100 (50%) | 22/48 (46%) | 28/52 (54%) |
| Female | 50/100 (50%) | 26/48 (54%) | 24/52 (46%) |
| Comorbidity | | | |
| Hypertension | 18/100 (18%) | 13/48 (27%) | 5/52 (10%) |
| Diabetes | 20/100 (20%) | 12/48 (25%) | 8/52 (15%) |
| Cardio-Cerebrovascular Diseases | 11/100 (11%) | 5/48 (10%) | 6/52 (12%) |
| Malignancy | 5/100 (5%) | 2/48 (4%) | 3/52 (6%) |
| Chronic Liver Disease | 3/100 (3%) | 1/48 (2%) | 2/52 (4%) |
| Tuberculosis | 2/100 (2%) | 1/48 (2%) | 1/52 (2%) |
| Signs and Symptoms | | | |
| Fever | 94/100 (94%) | 46/48 (96%) | 48/52 (92%) |
| Dry Cough | 60/100 (60%) | 32/48 (67%) | 28/52 (54%) |
| Shortness of Breath | 35/100 (35%) | 15/48 (31%) | 20/52 (38%) |
| Fatigue | 24/100 (24%) | 11/48 (23%) | 13/52 (25%) |
| Expectoration | 25/100 (25%) | 17/48 (35%) | 8/52 (15%) |
| Diarrhea | 26/100 (26%) | 12/48 (25%) | 14/52 (27%) |
| Anorexia | 35/100 (35%) | 22/48 (46%) | 13/52 (25%) |
| Myalgia | 10/100 (10%) | 5/48 (10%) | 5/52 (10%) |

| | | | |
|---|---|---|---|
| Headache | 2/100 (2%) | 1/48 (2%) | 1/52 (2%) |
| Nausea and Vomiting | 5/100 (5%) | 2/48 (4%) | 3/52 (6%) |
| Abdominal Pain | 2/100 (2%) | 1/48 (2%) | 1/52 (2%) |
| Pharyngalgia | 2/100 (2%) | 1/48 (2%) | 1/52 (2%) |
| Median of Onset of Symptom to Hospital Admission (IQR), days | 8 (6-10.5) | 8 (5-10) | 8 (6-13) |
| Median of During of Hospitalization (IQR), days | 25 (20-36) | 25 (20-33) | 28 (19-45) |

Data are n (%), unless otherwise stated. IQR: interquartile range.

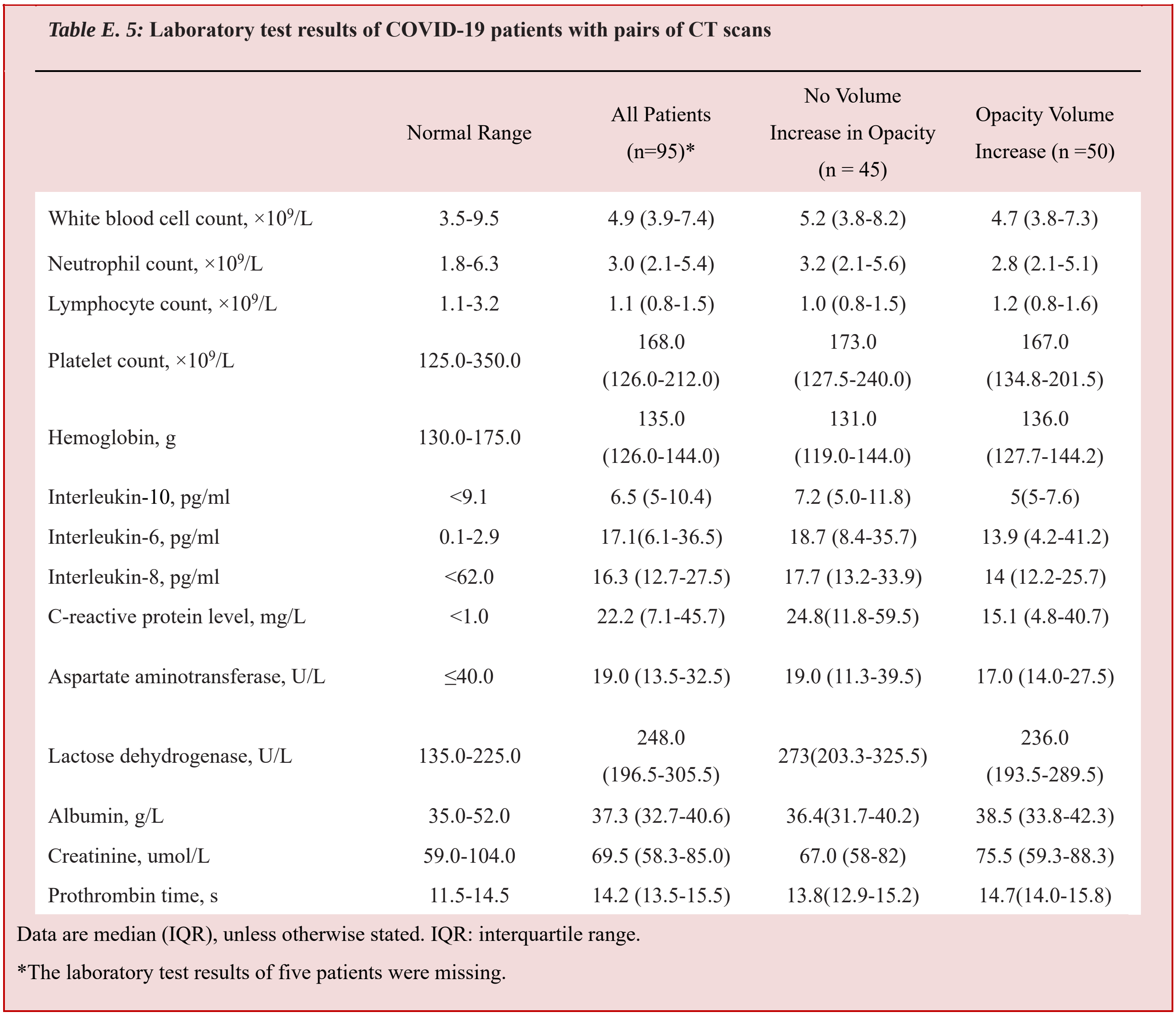

*Table E. 5:* **Laboratory test results of COVID-19 patients with pairs of CT scans**

| | Normal Range | All Patients (n=95)* | No Volume Increase in Opacity (n = 45) | Opacity Volume Increase (n =50) |
|---|---|---|---|---|
| White blood cell count, $\times10^9$/L | 3.5-9.5 | 4.9 (3.9-7.4) | 5.2 (3.8-8.2) | 4.7 (3.8-7.3) |
| Neutrophil count, $\times10^9$/L | 1.8-6.3 | 3.0 (2.1-5.4) | 3.2 (2.1-5.6) | 2.8 (2.1-5.1) |
| Lymphocyte count, $\times10^9$/L | 1.1-3.2 | 1.1 (0.8-1.5) | 1.0 (0.8-1.5) | 1.2 (0.8-1.6) |
| Platelet count, $\times10^9$/L | 125.0-350.0 | 168.0 (126.0-212.0) | 173.0 (127.5-240.0) | 167.0 (134.8-201.5) |
| Hemoglobin, g | 130.0-175.0 | 135.0 (126.0-144.0) | 131.0 (119.0-144.0) | 136.0 (127.7-144.2) |
| Interleukin-10, pg/ml | <9.1 | 6.5 (5-10.4) | 7.2 (5.0-11.8) | 5(5-7.6) |
| Interleukin-6, pg/ml | 0.1-2.9 | 17.1(6.1-36.5) | 18.7 (8.4-35.7) | 13.9 (4.2-41.2) |
| Interleukin-8, pg/ml | <62.0 | 16.3 (12.7-27.5) | 17.7 (13.2-33.9) | 14 (12.2-25.7) |
| C-reactive protein level, mg/L | <1.0 | 22.2 (7.1-45.7) | 24.8(11.8-59.5) | 15.1 (4.8-40.7) |
| Aspartate aminotransferase, U/L | ≤40.0 | 19.0 (13.5-32.5) | 19.0 (11.3-39.5) | 17.0 (14.0-27.5) |
| Lactose dehydrogenase, U/L | 135.0-225.0 | 248.0 (196.5-305.5) | 273(203.3-325.5) | 236.0 (193.5-289.5) |
| Albumin, g/L | 35.0-52.0 | 37.3 (32.7-40.6) | 36.4(31.7-40.2) | 38.5 (33.8-42.3) |
| Creatinine, umol/L | 59.0-104.0 | 69.5 (58.3-85.0) | 67.0 (58-82) | 75.5 (59.3-88.3) |
| Prothrombin time, s | 11.5-14.5 | 14.2 (13.5-15.5) | 13.8(12.9-15.2) | 14.7(14.0-15.8) |

Data are median (IQR), unless otherwise stated. IQR: interquartile range.

*The laboratory test results of five patients were missing.

**Appendix 8 Ethical approval**

This study was approved by the Institutional Review Board of each participating hospital:

Tongji Hospital, Tongji Medical College of Huazhong University of Science and Technology: TJ-IRB20200350;

Tianyou Hospital Affiliated to Wuhan University of Science and Technology: 20200205;

Xianning Central Hospital: [2020]001;

The Second Xiangya Hospital of Central South University: 2020[023];

The Third People's Hospital of Shenzhen, The Second Affiliated Hospital of Southern University of Science and Technology, National Clinical Research Centre for Infectious Diseases: 2020-0028).

## Appendix 9 Study protocol

### Study aim

In order to expedite chest CT based triage in fever clinics, we aim to develop a fully automated deep learning algorithm to alert suspected COVID-19 cases and analyse lesion burdens. By validating the algorithm on fever clinic cases across regions of varied COVID-19 prevalence, we aim to assess the clinical value of the developed algorithm in real-world scenarios.

### Study design

A deep learning algorithm based on annotated data from Tongji Hospital will be first developed to triage suspected COVID-19 cases and analyse lesion burdens. For detailed algorithm development and data annotation please see Appendix 1 and Appendix 2).

Since external validation is a more scientific way to evaluate diagnostic performance, we plan to collect data from fever clinics other than Tongji Hospital to assess the developed deep learning model. Fever clinics of Tianyou hospital, Xianning Hospital and Second Xiangya Hospital are selected because patients from these clinics represent populations of varied COVID-19 prevalence from high to low. For data inclusion, two weeks of consecutive patients who visited these fever clinics during COVID-19 outbreak will be collected to make sure that the minimum sample size requirement (for sample size estimation, please see Appendix 2.2) is met. The data exclusion criteria are: 1). patients ≤ 14 years old (i.e., children) and 2) repeated scans from the same patient. For data truthing, radiological reports will be used as the reference standards. Radiologists will be invited to classify patients into positive (with suspected COVID-19 imaging manifestations) and negative cases (without suspected COVID-19 imaging manifestations) based on the original radiological reports. After obtaining the external validation set, the performance of AI-aided triage accuracy and efficiency will be evaluated against the reference standard and compared to performance targets according to the statistical analysis plan in Appendix 2.3.

## References

1 Ronneberg er O, Fischer P, Brox T. U-Net: Convolutional Networks for Biomedical Image Segmentation. *MICCAI* 2015; 9351: 234-241.

2 Roth HR, Lu L, Liu, J, Yao J, Seff A, Cherry K, et al. Improving computer-aided detection using convolutional neural networks and random view aggregation. *IEEE Trans Med Imaging* 2016, 35(5), 1170-1181.

3 He K, Zhang X, Ren S, Sun J. Deep residual learning for image recognition. In Proceedings of the IEEE conference on computer vision and pattern recognition 2016; 770-778.

4 Long J, Shelhamer E, Darrell T. Fully convolutional networks for semantic segmentation. Proceedings of the IEEE conference on computer vision and pattern recognition. 2015: 3431-3440.

5 National Health Commission of the People's Republic of China. Chinese Clinical Guidance for COVID-19 Pneumonia Diagnosis and Treatment (7th edition). http://www.nhc.gov.cn/yzygj/s7653p/202003/46c9294a7dfe4cef80dc7f5912eb1989/files/ce3e6945832a438eaae415350a8ce964.pdf (accessed Mar 5, 2020).

6 Lowekamp BC, Chen DT, Ibáñez L, et al. The Design of SimpleITK. *Front. Neuroinform* 2013; 7:45.

7 Yaniv Z, Lowekamp BC, Johnson HJ, Beare R. SimpleITK Image-Analysis Notebooks: a Collaborative Environment for Education and Reproducible Research. *J Digit Imaging* 2018, 31, 290–303

8 Shi H, Han X, Jiang N, et al. Radiological findings from 81 patients with COVID-19 pneumonia in Wuhan, China: a descriptive study. *Lancet Infect Dis* 2020, 20(4), 425-434.