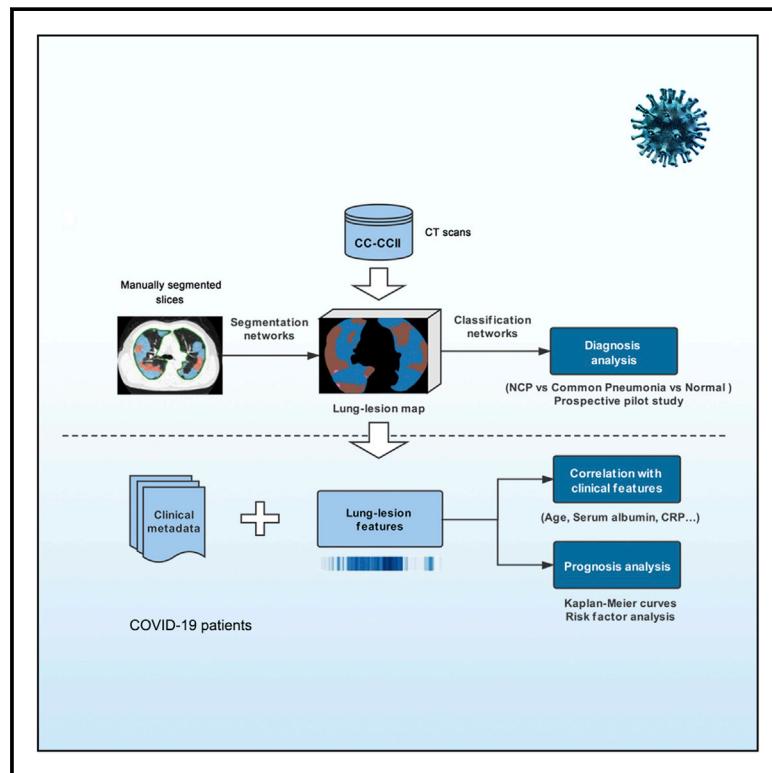


Clinically Applicable AI System for Accurate Diagnosis, Quantitative Measurements, and Prognosis of COVID-19 Pneumonia Using Computed Tomography

Graphical Abstract



Authors

Kang Zhang, Xiaohong Liu, Jun Shen, ..., Tianxin Lin, Weimin Li, Guangyu Wang

Correspondence

kang.zhang@gmail.com (K.Z.),
hejx@vip.163.com (J.H.),
lintx@mail.sysu.edu.cn (T.L.),
weimi003@yahoo.com (W.L.),
wangguangyu@mail.tsinghua.edu.cn (G.W.)

In Brief

Zhang et al. present an AI-based system, based on hundreds of thousands of human lung CT scan images, that can aid in distinguishing patients NCP versus other common pneumonia and can help to predict the prognosis of COVID-19 patients.

Highlights

- AI system that can diagnose COVID-19 pneumonia using CT scans
- Prediction of progression to critical illness
- Potential to improve performance of junior radiologists to the senior level
- Can assist evaluation of drug treatment effects with CT quantification



Resource

Clinically Applicable AI System for Accurate Diagnosis, Quantitative Measurements, and Prognosis of COVID-19 Pneumonia Using Computed Tomography

Kang Zhang,^{1,14,15,*} Xiaohong Liu,^{2,14} Jun Shen,^{3,14} Zhihuan Li,^{4,5,14} Ye Sang,^{6,14} Xingwang Wu,^{7,14} Yunfei Zha,^{8,14} Wenhua Liang,^{9,14} Chengdi Wang,^{4,14} Ke Wang,² Linsen Ye,¹⁰ Ming Gao,³ Zhongguo Zhou,¹ Liang Li,⁸ Jin Wang,¹⁰ Zehong Yang,³ Huimin Cai,⁵ Jie Xu,¹ Lei Yang,⁵ Wenjia Cai,⁵ Wenqin Xu,¹ Shaoxu Wu,³ Wei Zhang,³ Shanping Jiang,³ Lianghong Zheng,^{5,11} Xuan Zhang,² Li Wang,⁸ Liu Lu,^{5,11} Jiaming Li,^{5,11} Haiping Yin,¹² Winston Wang,¹ Oulan Li,⁵ Charlotte Zhang,⁵ Liang Liang,⁶ Tao Wu,⁶ Ruiyun Deng,^{1,11} Kang Wei,¹ Yong Zhou,¹ Ting Chen,² Johnson Yiu-Nam Lau,¹³ Manson Fok,¹ Jianxing He,^{9,*} Tianxin Lin,^{3,*} Weimin Li,^{4,*} and Guangyu Wang^{2,*}

¹Faculty of Medicine, Macau University of Science and Technology, Macau, China

²Department of Computer Science and Technology & BNRIst, Tsinghua University, Beijing, China

³Departments of Urology, Radiology, Emergency Medicine, and Respiratory Medicine, Sun Yat-Sen Memorial Hospital, Sun Yat-Sen University, Guangzhou, China

⁴Center for Translational Innovations and Department of Respiratory and Critical Care Medicine, West China Hospital, West China School of Medicine, Sichuan University, Chengdu, China

⁵Guangzhou Regenerative Medicine and Health Guangdong Laboratory, Guangzhou, China

⁶The First College of Clinical Medical Science, China Three Gorges University, Yichang, China

⁷Department of Radiology, The First Affiliated Hospital of Anhui Medical University, Hefei, China

⁸Department of Radiology, Department of Infection Prevention and Control, Renmin Hospital, Wuhan University, Wuhan, China

⁹Department of Thoracic Surgery/Oncology, the First Affiliated Hospital of Guangzhou Medical University, China State Key Laboratory and National Clinical Research Center for Respiratory Disease, Guangzhou, China

¹⁰Department of Radiology, and Liver Disease Center, Sun Yat-Sen Third Affiliated Hospital, Sun Yat-Sen University, Guangzhou, China

¹¹Guangzhou Kangrui AI Technology Co. and Guangzhou HuiBoRui Biological Pharmaceutical Technology Co., Ltd, Guangzhou, China

¹²The First People's Hospital of Yunnan Province, Kunming, China

¹³Department of Applied Biology and Chemical Technology, Hong Kong Polytechnic University, Hong Kong, China

¹⁴These authors contributed equally

¹⁵Lead Contact

*Correspondence: kang.zhang@gmail.com (K.Z.), hejx@vip.163.com (J.H.), liintx@mail.sysu.edu.cn (T.L.), weimi003@yahoo.com (W.L.), wangguangyu@mail.tsinghua.edu.cn (G.W.)

<https://doi.org/10.1016/j.cell.2020.04.045>

SUMMARY

Many COVID-19 patients infected by SARS-CoV-2 virus develop pneumonia (called novel coronavirus pneumonia, NCP) and rapidly progress to respiratory failure. However, rapid diagnosis and identification of high-risk patients for early intervention are challenging. Using a large computed tomography (CT) database from 3,777 patients, we developed an AI system that can diagnose NCP and differentiate it from other common pneumonia and normal controls. The AI system can assist radiologists and physicians in performing a quick diagnosis especially when the health system is overloaded. Significantly, our AI system identified important clinical markers that correlated with the NCP lesion properties. Together with the clinical data, our AI system was able to provide accurate clinical prognosis that can aid clinicians to consider appropriate early clinical management and allocate resources appropriately. We have made this AI system available globally to assist the clinicians to combat COVID-19.

INTRODUCTION

The outbreak of the 2019 novel coronavirus (SARS-CoV-2) began in early December of 2019 (Munster et al., 2020; Wang et al., 2020). The infection has a mean incubation period of 5.2 days and causes fever, cough, and other flu-like symptoms. It can affect multiple tissues and organ systems, and collectively the viral-induced disease is termed COVID-19. Many affected

patients develop pneumonia (called novel coronavirus pneumonia, NCP) and progress rapidly into severe acute respiratory failure with a very poor prognosis and high mortality (Guan et al., 2020; Huang et al., 2020a). Person-to-person transmission has been established (Chan et al., 2020; Phan et al., 2020; Rothe et al., 2020; Zhu et al., 2020), and the World Health Organization (WHO) has declared COVID-19 to be pandemic. Studies have shown that over 60% of patients died once they progressed



into the severe or critical illness stage (Guan et al., 2020; Huang et al., 2020a). Therefore, identifying risk factors and parameters that can allow the creation of an accurate prognosis predictive model and hopefully lead to improved clinical outcomes are critical in the planning of early intervention and intense monitoring in such a pandemic.

Currently, an individual with fever, cough, and flu symptoms would be screened first by clinical assessments, laboratory tests, and a chest X-ray to rule out pneumonia. If viral pneumonia is diagnosed, being able to diagnose NCP is of paramount importance for obvious public health reasons. A COVID-19 diagnosis is confirmed by a positive molecular PCR test. Chest computed tomography (CT) is an important tool in the diagnosis of lung diseases including pneumonia. CT scanning procedure has a faster turnaround time than a molecular diagnostic test performed in a standard laboratory, can provide more detailed information related to the pathology, and is better for the quantitative measurement of lesion size and the extent or severity of lung involvement, which may have prognostic implications (Shi et al., 2020). As seasonal flu also causes viral pneumonia, it is also important to differentiate NCP from the common flu or other types of pneumonia such as viral pneumonia and bacterial pneumonia. Therefore, an accurate CT-based artificial intelligence (AI) system may have the potential to assist in the early diagnosis for planning, monitoring and treatment, and establishing the reference for longitudinal follow ups (Huang et al., 2020b; Xie et al., 2020).

Recent new and exciting advances in the applications of AI in many healthcare areas (Esteva et al., 2019; Gulshan et al., 2016; Li et al., 2018; Norgeot et al., 2019; Poplin et al., 2018; Ravizza et al., 2019; Ting et al., 2017; Topol, 2019) have inspired innovations in the development of novel AI-based radiological diagnostic technology. Chen (Chen et al., 2020) reviewed various quantitative models of thin-section CT of the chest and showed effectiveness of quantitative tools in both accurate diagnosis and longitudinal follow-up. Another study showed that deep learning algorithms helped to identify head CT abnormalities that can assist the clinical triage process (Chilamkurthy et al., 2018). Recent studies demonstrated the potential of integrating AI into both the eye and childhood-disease diagnostic systems and found that it significantly improved clinical diagnostic efficiency and accuracy (Burlina et al., 2017; Kermany et al., 2018; Long et al., 2017; van Grinsven et al., 2016). Therefore, with a more accurate tool in CT scans, we hypothesized that an AI system can be established to accurately diagnose NCP, and this will assist radiologists and clinicians in the management of patients with symptoms suggestive of COVID-19 NCP.

Another urgent need is to identify the patients with higher risks of developing acute respiratory failure so that they can be monitored closely earlier and receive intervention treatment early. Otherwise, such patients will have a higher chance of eventually developing multi-organ failure associated with high mortality (Yang et al., 2020). As lesion characteristics, including number, size, and density of lesions and also the overall lung parenchyma, are indicators of lung damage and remaining lung reserve, we also tested the hypothesis as to whether an AI system can be established using both clinical data and CT parameters to generate an accurate clinical prognostic model, allowing

clinicians to plan for early monitoring and management of these patients. Accordingly, we have constructed a large CT dataset on NCP and other common pneumonia and normal controls, and we developed an AI diagnostic system for assisting the accurate diagnosis for application in an epidemic area and two non-epidemic areas in China. We also provided prognosis indications for patients with NCP by using a combination of CT and clinical parameters, with an aim to provide another tool to assist physicians (Figure 1).

RESULTS

Patient Characteristics and Image Datasets

A large CT dataset encompassing patient cohorts from the China Consortium of Chest CT Image Investigation (CC-CCII) was constructed, which consisted of a total of 617,775 CT images from 4,154 patients. CT images from 3,777 patients were employed to train and test our AI systems for differentiating NCP from other common pneumonia and normal controls (Figures 1 and S1; Table S1). The common pneumonia group consisted of viral pneumonia, bacterial pneumonia, and mycoplasma pneumonia, all of which are the most common causes of pneumonia in China. Our AI COVID-19 diagnostic system consisted of two models including a lung-lesion segmentation model and a diagnosis analysis model (Figures 1 and S2). We constructed a segmentation network trained with 4,695 manually segmented slices from NCP patients and common pneumonia patients. For the classification model, 361,221 CT images from 2,246 patients including 752 NCP patients, 797 common pneumonia patients, and 697 normal control patients were used for training (Table S1). Our AI system performance of diagnosis prediction was also tested in one retrospective cohort and three prospective pilot studies consisting of one in an epidemic area and two in non-epidemic areas in China. In addition, we have tested our AI diagnostic performance internationally using CT data generated outside China. For the CT quantitative and correlation analysis with clinical parameters, 843 hospital-admitted NCP patients with clinical metadata were included. Of these, 456 hospitalized patients with comprehensive CT images and clinical outcome information were subject to a prognosis prediction and survival analysis and the establishment of a clinical prognosis estimation (Figure 1).

Lung-Lesion Segmentation

A two-stage segmentation framework for accurately segmenting lung lesions from normal lung fields and background on raw CT slices with processing-time efficiency and accuracy was constructed (Figure S2). A total of 4,695 CT slice images were manually annotated at the pixel level into seven classes, including background, lung field, consolidation (CL), ground-glass opacity (GGO), pulmonary fibrosis, interstitial thickening, and pleural effusion.

To study our AI system performance on CT slice segmentation, several classic semantic segmentation tools as the backbone of our segmentation framework were tested, including U-net (Ronneberger et al., 2015), DRUNET (Devalla et al., 2018), FCN (Long et al., 2015), SegNet (Badrinarayanan et al., 2017), and DeepLabv3 (Chen et al., 2017). We evaluated system

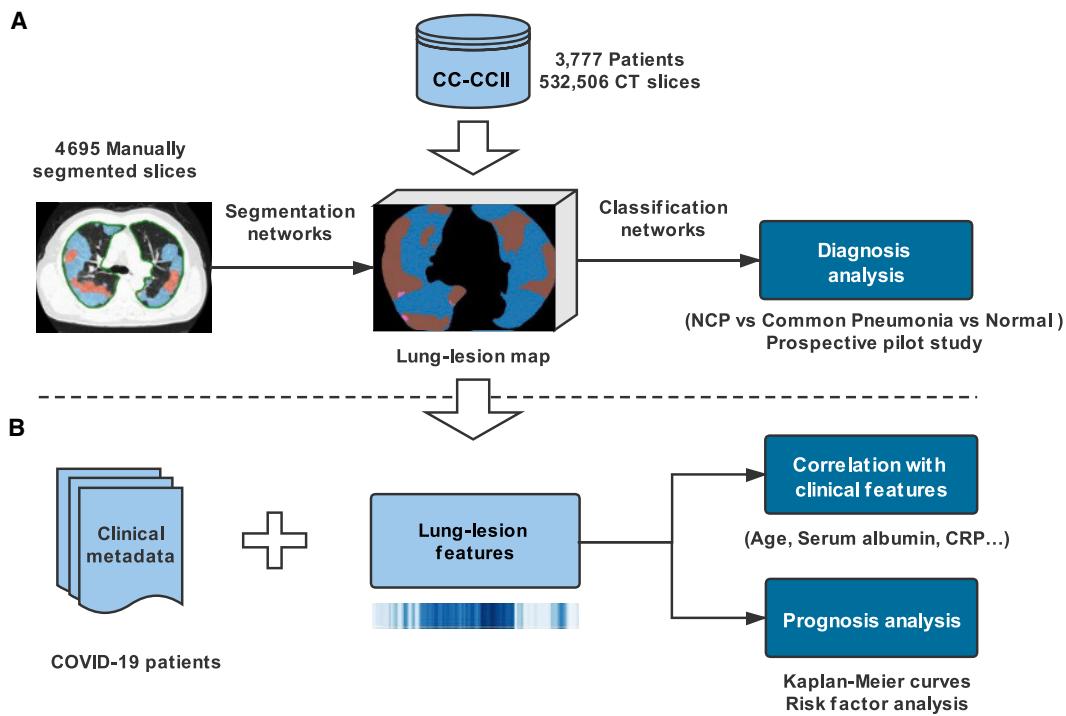


Figure 1. Our Proposed AI Framework for NCP Diagnosis and Prognosis Prediction

(A) A large CT dataset was constructed using the data from CC-CCII (532,506 CT images from NCP, common pneumonia, and normal controls). The NCP diagnosis system consisted of two models: a lung-lesion segmentation model and a diagnosis prediction model. We first trained a segmentation network with 4,695 manually segmented images from NCP and common pneumonia patients. The diagnosis classifier took as input the previous lung-lesion map and generated probability of three classes: NCP, common pneumonia, and normal controls with classification networks. A number of prospective pilot studies were also conducted to test our AI performance for clinical application.

(B) AI-assisted clinical prognosis estimation based on CT quantitative parameters and clinical metadata. A system for risk factor evaluation and Kaplan-Meier curve analysis for severe or critical illness as defined in the text was also implemented.

See also [Figures S1, S2, and S7](#) and [Table S1](#).

performance with two evaluation metrics including dice coefficient (DC) and pixel accuracy (PA) by a five-fold cross-validation test ([Table S3](#)). We adopt DeepLabv3 as the backbone for subsequent analyses for its better segmentation performance. Compared to human experts, our segmentation network obtained smoother and clearer lesion segmentation boundaries and archived a high accuracy ([Figures 2](#) and [S3](#)).

Accurate Diagnosis of NCP

To distinguish NCP from other common pneumonia and normal controls, we employed a diagnostic system based on a lung-lesion segmentation model and a diagnosis analysis model. The diagnosis classification took the lung-lesion map as an input generated by segmentation networks and utilized the normalized CT volumes for further diagnosis prediction. While real-world original scans contained noises and varied for different devices and human operations, our approach provided better generalization and interoperability during clinical implementations instead of end-to-end black-box networks.

We used a total of 40,880 slices from 260 patients including 83 NCP patients, 91 common pneumonia patients, and 86 normal controls to test our diagnosis classifier model as an in-

ternal validation ([Table S1](#)). Our system was able to differentiate NCP from other two classes (other common pneumonia and normal controls) with 92.49% accuracy, 94.93% sensitivity, 91.13% specificity, and an area under the receiver operating characteristic (AUROC) of 0.9797 (95% CI: 0.9665–0.9904) on an internal validation dataset. The overall performance for three-way classification obtained 92.49% accuracy and an AUROC of 0.9813 (95% confidence interval [CI]: 0.9691–0.9902) ([Figures 3A and 3B](#)).

AI System Performance Evaluation in Independent Chinese and International Cohorts

To address regional variations and general applicability of our AI diagnostic system, our AI system performance was tested in four other different regions using different datasets. Specifically, our AI performance was tested in a retrospective study in an external cohort from an epidemic area in Hubei, China (City of Yichang). In addition, we also tested our AI system performance in three other regions in a prospective fashion, with one cohort from the epicenter Hubei, China (City of Wuhan) and two cohorts from other non-epidemic areas in China (City of Hefei and City of Guangzhou).

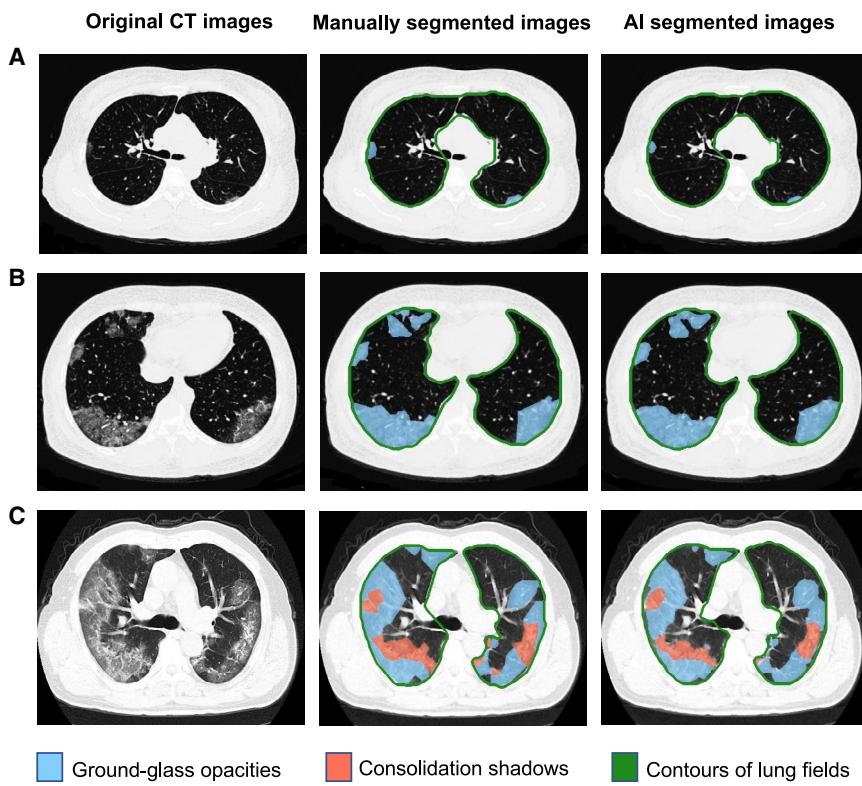


Figure 2. Performance of Our AI System on a Lesion Segmentation Task Shown in Three Examples

Left column: original CT slices from three NCP patients; middle column, manually segmented CT slices; right column, AI-based automated segmented CT slices.

Row (A) A CT slice with mild NCP lesions defined as small ground-glass opacities (GGO) of bilateral lung involvement.

Row (B) A CT slice with intermediate NCP lesions. Bilateral and predominantly peripheral lesions of GGO.

Row (C) A CT slice with severe NCP lesions. Bilateral and peripheral mixed lesions of GGO and consolidation shadows.

The severity level definitions are as follows: mild, defined as less than three GGO lesions of size less than 3 cm; intermediate, defined as a lesion area more than 25% of the entire lung field; severe, defined as a lesion area more than 50% of the entire lung field.

See also Figure S3 and Table S2.

In a retrospective study in the City of Yichang, Hubei Province, China, 155 NCP patients, 36 common pneumonia patients, and 17 normal controls who underwent CT imaging were enrolled in this study (Table S1). Our AI diagnostic system archived 90.70% accuracy, 92.15% sensitivity, 85.92% specificity, and AUROC of 0.9805 (95% CI: 0.9662–0.9899) for NCP versus all other groups and an accuracy of 89.92% and AUROC of 0.9805 (95% CI: 0.9662–0.9899) for the overall three-way classification (NCP, common pneumonia, and normal controls) (Figures 3C and 3D).

The first prospective pilot study was conducted in Wuhan, Hubei Province, China (the epicenter) from January 25 to March 18, 2020 (Table S1). Our AI diagnostic system achieved 91.20% accuracy, 94.03% sensitivity, 88.46% specificity, and AUROC of 0.9610 (95% CI: 0.9403–0.9785) for NCP versus all other groups together and an accuracy of 91.20% and AUROC of 0.9741 (95% CI: 0.9583–0.9856) for the overall three-way classification (NCP, common pneumonia, and normal controls) in this prospective study cohort (Figures 4A and 4B).

A second prospective pilot study was conducted in a non-epidemic region in Hefei, Anhui Province, China, from February 2 to March 25, 2020 (Table S1). Our AI diagnostic system archived 90.32% accuracy, 94.74% sensitivity, 89.19% specificity, and AUROC of 0.9700 (95% CI: 0.9500–0.9872) for NCP versus all other groups together and an accuracy of 91.76% and AUROC of 0.9776 (95% CI: 0.9630–0.9899) for the overall three-way classification (NCP, common pneumonia, and normal controls) in this second study prospective cohort (Figures 4C and 4D).

The third prospective pilot study was conducted in another non-epidemic region in Guangzhou, Guangdong Province (China) from February 8 to March 27, 2020 (Table S1). Our AI diagnostic system archived 84.78% accuracy, 90.00%

sensitivity, 84.15% specificity, and AUROC of 0.9512 (95% CI: 0.9124–0.9820) for NCP versus all other groups together and an accuracy of 89.67% and AUROC of 0.9755 (95% CI: 0.9545–0.9896) for the overall three-way classification (NCP, common pneumonia, and normal controls) (Figure 4E and 4F).

To validate our AI system's general applicability outside China, we obtained CT images from an open source and additional data from our collaborators in Ecuador (Table S1). Our AI diagnostic system achieved 84.11% accuracy, 86.67% sensitivity, 82.26% specificity, and AUROC of 0.905 (95% CI: 0.8421–0.9612) for NCP versus all other groups together and an accuracy of 85.05% and AUROC of 0.9381 (95% CI: 0.8944–0.9742) for the overall three-way classification (Figures S5A and S5B). Together, these five studies confirmed the high performance, accuracy, and general applicability of the AI diagnostic system both within China and internationally.

Evaluation of Drug Treatment Effects on AI-Based Lesion Quantitative Measurements

We also evaluated the effect of drug treatment on lesion size and volume changes using our AI-generated quantitative measurements. NCP patients undergoing three different drug treatment trials were analyzed. The enrollment criteria included a confirmed NCP diagnosis by a positive viral PCR test and no other prior treatment history. We quantified lung lesions (GGO and total lesion volumes) on CT scans and compared the differences between pre-treatment and post-treatment results for three administered experimental drugs. The CT scans demonstrated quantitative lesion changes, indicating the potential of

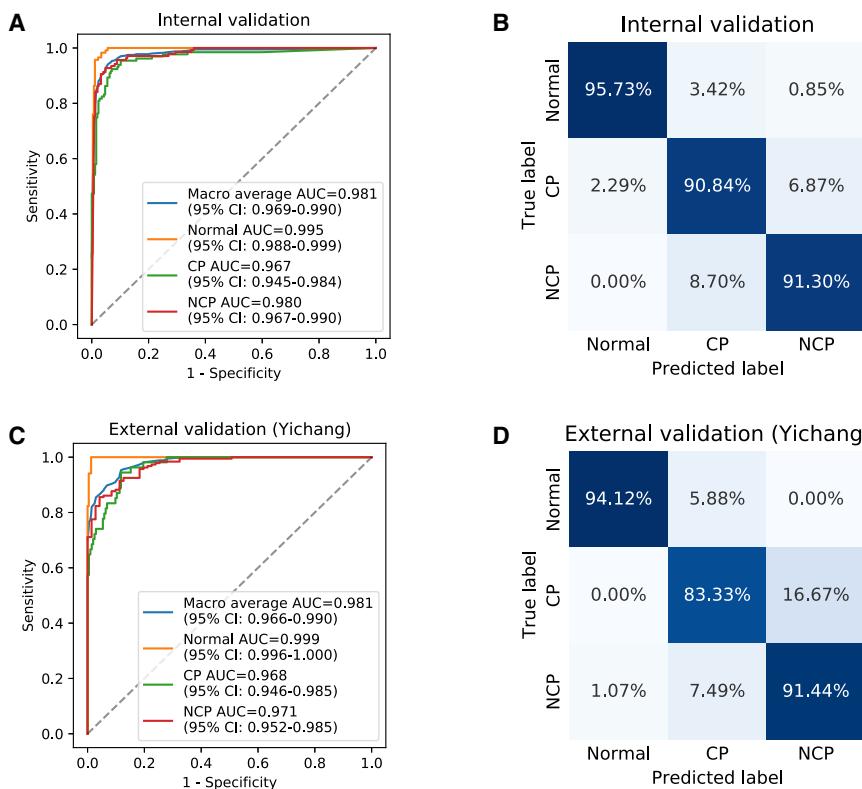


Figure 3. Performance of Our AI System in Identifying NCP Patients from Patients with Other Common Types of Pneumonia and Normal Controls

(A–D) Receiver operating characteristics (ROC) curves and normalized confusion matrices of multiclass classifications. The blue curve denoted macro-average area under the curve (AUC) of one (NCP) versus other two classes, including common pneumonia (CP) and normal controls (Normal). CI, confidence interval.

(A and B) AI system performance on internal validation data. (A) ROC curves. (B) Normalized confusion matrix. For three-way classification: accuracy = 92.49%, AUROC = 0.9813 (95% CI: 0.9691–0.9902). For NCP versus the rest: accuracy = 92.49%, sensitivity = 94.93%, specificity = 91.13%, AUROC = 0.9797 (95% CI: 0.9665–0.9904).

(C and D) AI performance on independent external validation data in Yichang (Hubei, China). (C) ROC curves. (D) Normalized confusion matrix. For a three-way classification: accuracy = 89.92%, AUROC = 0.9805 (95% CI: 0.9662–0.9899). For NCP versus the rest: accuracy = 90.70%, sensitivity = 92.51%, specificity = 85.92%, AUROC = 0.9712 (95% CI: 0.9516–0.9855). See also Figure S5.

our AI system in the evaluation of drug treatment efficacy in an objective quantitative way (Figure S4).

Comparison of the AI System against Practicing Radiologists

An independent test dataset of 18,392 CT slice images from 150 patients, including 40 NCP patients, 80 common pneumonia patients, and 30 normal control patients, was used to compare the AI system's performance with practicing radiologists in classifying NCP versus other pneumonia. We employed eight radiologists in two groups to participate in the study: four radiologists in the junior group with 5 to 15 years of clinical experience, and four radiologists in the senior group with 15 to 25 years of clinical experience. Ground truth was established based on a consensus from an independent group of four senior radiologists with 25 or more years of clinical experience.

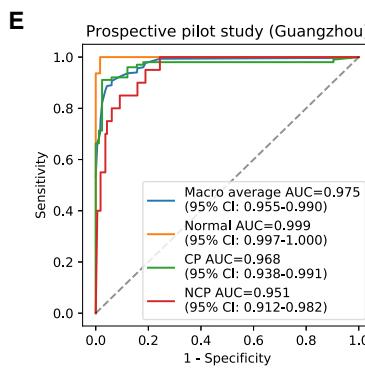
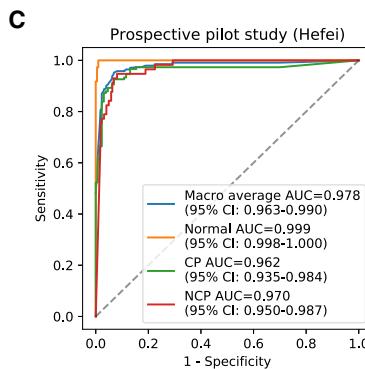
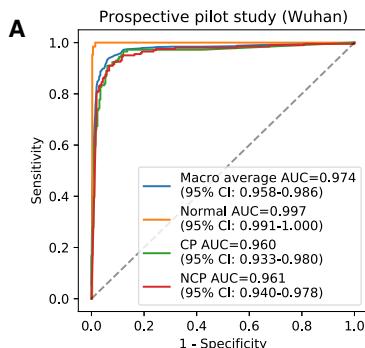
We then compared the performance between our AI system and radiologists on differentiating NCP from other common pneumonia and normal controls. Our AI system performance was overall superior to that of junior radiologists and comparable to mid-senior radiologists (Figures 5A and 5B). We used predicted errors, based on penalty scores (Figures 5C and S5C), to create a metric to evaluate and compare performance between our AI system and the radiologists. Our AI system yielded a weighted error of 9.29% compared to a range of weighted errors by the experts ranging from 7.14% to 19.15% with a mean of 13.55% (Figure 5C).

To investigate whether our AI system could help junior radiologists to improve their diagnostic performance, each junior radi-

ologist was given diagnosis probability on each patient by the AI system and was asked to make a diagnosis with the assistance of the AI-generated results. To avoid a potential memorization bias, the follow-up AI-assisted diagnostic test by junior radiologists was performed 2 weeks after the initial test. The performance was significantly improved compared to the previous one and was comparable to that of the senior radiologists (Figure 5).

Lung-Lesion Features and Correlations with Key Clinical Parameters

The ratio of total lesions (named lesions), CL, or GGO to an entire lung field was chosen as three lesion indicators (features). The correlations analysis between quantitative lesion features and clinical parameters were performed, and key clinical and biochemical markers were determined. Spearman's and Pearson's correlations were employed to study the association between quantitative lesion features on a CT scan and clinical parameters. C-reactive protein (CRP), age, serum lactic dehydrogenase (LDH), highest body temperature (Tmax), and neutrophil-to-lymphocyte ratio showed highly positive correlations with the lesion features. By contrast, degree of blood oxygen saturation, lymphocyte count, albumin, blood platelets, and Na⁺ showed highly negative correlations with the lesion features (Figures 6A–6D and S6A–S6E). We also designed a composite score (c-score) system using international guidelines to evaluate correlations between the lesion features and degree of lung and liver organ damage (see STAR Methods for more details), which showed a positive



B Prospective pilot study (Wuhan)

		True label		
		Normal	CP	NCP
True label	Normal	96.88%	1.56%	1.56%
	CP	0.00%	90.97%	9.03%
NCP	1.00%	9.45%	89.55%	
	Normal	CP	NCP	Predicted label

D Prospective pilot study (Hefei)

		True label		
		Normal	CP	NCP
True label	Normal	98.63%	1.37%	0.00%
	CP	1.34%	89.93%	8.72%
NCP	0.00%	12.28%	87.72%	
	Normal	CP	NCP	Predicted label

F Prospective pilot study (Guangzhou)

		True label		
		Normal	CP	NCP
True label	Normal	100.00%	0.00%	0.00%
	CP	0.99%	84.16%	14.85%
NCP	5.00%	10.00%	85.00%	
	Normal	CP	NCP	Predicted label

correlation between the lesion size and degree of organ damage (Figure 6E). Together, these data suggest lung lesions could not only directly correlate with the respiratory system function, but also with the clinical parameters of other organ systems, suggesting the interconnectivity between lung lesions and the health of other major organs and potential concomitant advancement of damages contributing to multiple organ failures. Obviously, we would like to emphasize that our current data only showed the correlation of these changes without any direct knowledge of the exact underlying pathogenetic mechanisms (e.g., secondary to respiratory failure or/and direct viral involvement).

Prognostic Factors and Development of a Prognostic Model for Critical Illness

Next, the clinical and radiological features that contributed to the progression to critical illness were further analyzed to develop an AI-assisted model to estimate the clinical prognosis. A c-score

Figure 4. Performance of the AI System in Prospective Pilot Studies in Four Independent Chinese Cohorts

(A–F) ROC curves and normalized confusion matrices of multiclass classifications. The blue curve denoted macro-average AUC of one (NCP) versus other two classes, including common pneumonia (CP) and normal controls (Normal).

(A and B) AI system performance on a cohort from an epidemic area in China (City of Wuhan). (A) ROC curves. (B) Normalized confusion matrix. For three-way classification: accuracy = 91.20%, AUROC = 0.9741 (95% CI: 0.9583–0.9856). For NCP versus the rest: accuracy = 91.20%, sensitivity = 94.03%, specificity = 88.46%, AUROC = 0.9610 (95% CI: 0.9403–0.9785).

(C and D) AI system performance on a cohort from a non-epidemic area in China (City of Hefei). (C) ROC curves. (D) Normalized confusion matrix. For three-way classification: accuracy = 91.76%, AUROC = 0.9776 (95% CI: 0.9630–0.9899). For NCP versus the rest: accuracy = 90.32%, sensitivity = 94.74%, specificity = 89.19%, AUROC = 0.9700 (95% CI: 0.9500–0.9872).

(E and F) AI system performance on a cohort from a non-epidemic area in China (City of Guangzhou). (E) ROC curves. (F) Normalized confusion matrix. For three-way classification: accuracy = 89.67%, AUROC = 0.9755 (95% CI: 0.9545–0.9896). For NCP versus the rest: accuracy = 84.78%, sensitivity = 90.00%, specificity = 84.15%, AUROC = 0.9512 (95% CI: 0.9124–0.9820).

based on quantitative lung-lesion features extracted by the AI system and clinical parameters (e.g., age, albumin, blood oxygen saturation, CRP) was created and applied to predict the clinical outcome defined by the progression (time from the initial hospital admission) to severe or critical illness (defined by death or the

clinical need for mechanical ventilation or to be transferred to the ICU). We implemented Light Gradient Boosting Machine (LightGBM) and Cox proportional-hazards (CoxPH) regression models for prognosis prediction. For interpreting the effects and relative contributions of the lung-lesion features and clinical parameters on prognosis prediction, we implemented an explainer Shapley Additive exPlanation (SHAP) (Lundberg et al., 2018a and 2018b).

As expected, the lesion features were identified as the most significant contributors in the clinical prognosis estimation. In addition, clinical parameters relating to respiratory function (oxygen saturation index and respiratory rate) and general clinical characteristics (age, the body temperature on admission, and Tmax) also contributed to the prediction of progression to severe or critical illness requiring an ICU admission. Interestingly, additional prognostic markers were also identified, including liver biochemistry markers (albumin, serum LDH, indirect bilirubin), coagulation markers (thrombin time, activated partial

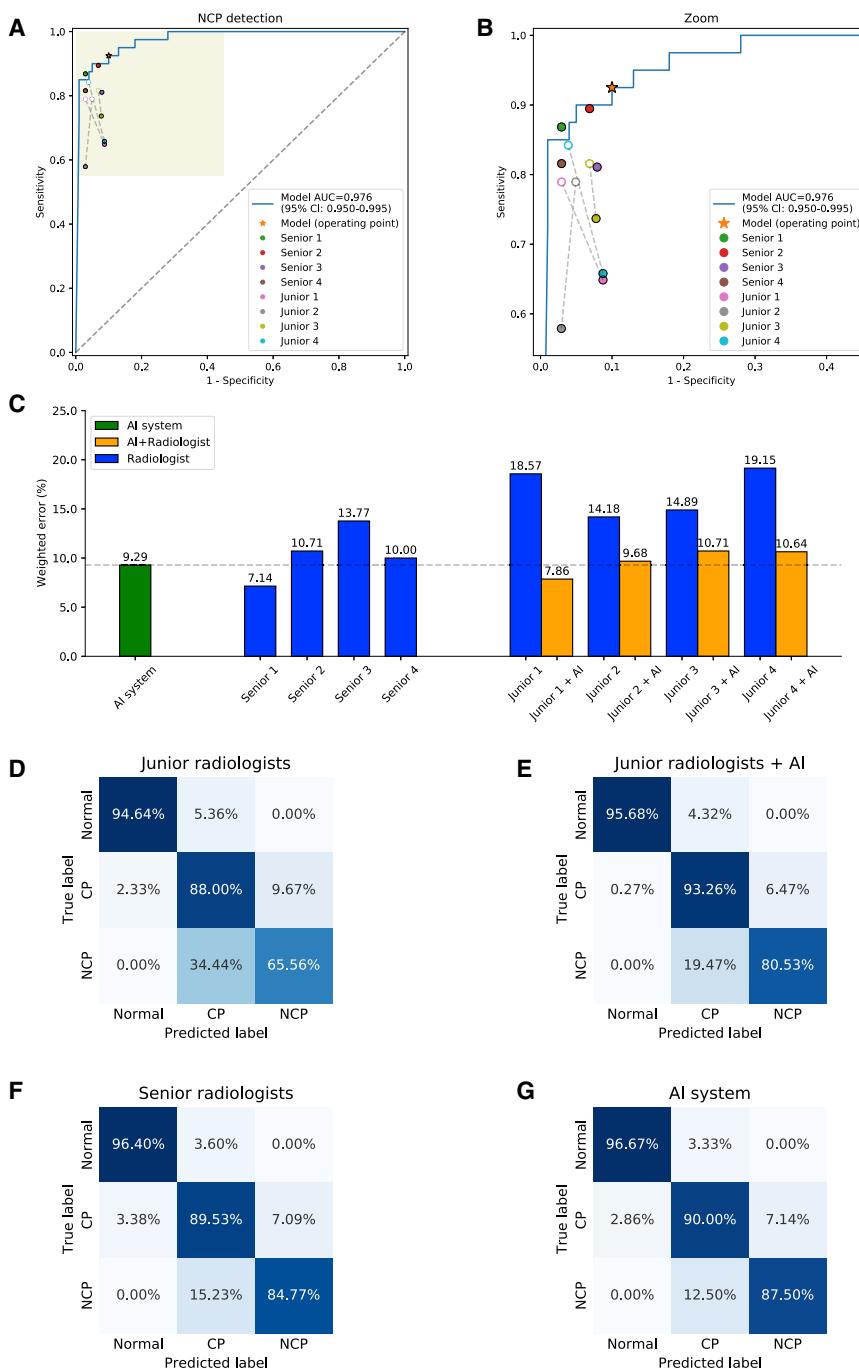


Figure 5. Comparisons of Diagnostic Performance between Our AI Model and Practicing Radiologists

(A and B) The performance of our AI system and eight practicing radiologists (four junior level and four senior level). ROC curve for diagnosis of NCP versus other classes. Filled dots denote junior and senior radiologists' performances, while the hollow dots denote the performance of junior group with AI assistance. Dashed lines linked the paired performance values of each junior radiologist.

(C) Weighted error results based on penalty scores (See Figure S5).

(D–G) Confusion matrices of multiclass classification. (D) Confusion matrix of the mean diagnostic performance of four junior radiologists.

(E) Confusion matrix of the mean diagnostic performance of four junior radiologists with AI assistance.

(F) Confusion matrix of the mean diagnostic performance of four senior radiologists.

(G) The AI system demonstrated performance comparable to that of senior practicing radiologists. Accuracy = 90.71%, sensitivity = 92.50%, specificity = 90.00%, AUROC = 0.9756 (95% CI: 0.9496–0.9948).

ical illness. Lung-lesion features alone gave an area under the curve (AUC) of 0.8479. When combined with clinical metadata, the performance of our model improved significantly to an AUC of 0.9093 (95% CI: 0.8775–0.9369), with a specificity of 80.00% and sensitivity of 86.71% (Figures 7A and 7B).

Kaplan-Meier curves using a c-score were generated based on these lung lesion and clinical parameters (Figure S6F). Based on the current size of the database, we stratified the patients into two groups: the high-risk group with a c-score of ≥ 0.5 and the low-risk group with a c-score of < 0.5 . The high-risk group (158 observations with 133 events) had a much lower survival probability compared to the low-risk group (274 observations with 37 events) with a highly significant statistical difference in medium survival ($p < 0.001$, log-rank test; Figure 7E). These results

thromboplastin time (APTT), platelet count), electrolyte and acid-base balance (Na^+ , K^+ , HCO_3^-), and markers of inflammation (CRP, lymphocyte count, neutrophil count) (Figures 7C and 7D; Table S4), suggesting the overall health of other tissues or organs and systemic homeostasis also played an important role in determining the clinical prognosis of these patients in terms of their evolving into severe/critical illness status.

AUROC curves were then generated to measure the impact of the NCP lesions and clinical metadata on progression to crit-

suggest that a combination of lung lesions and clinical metadata can contribute significantly to the prognosis prediction. It is also important to note that the curves will also provide estimates as to the time in days that critical care demands will be needed. Certainly, with a larger database in the future, our AI system may be able to provide an even more refined clinical prognostic model stratified into more groups with different levels of risks that are associated with different clinical prognoses.

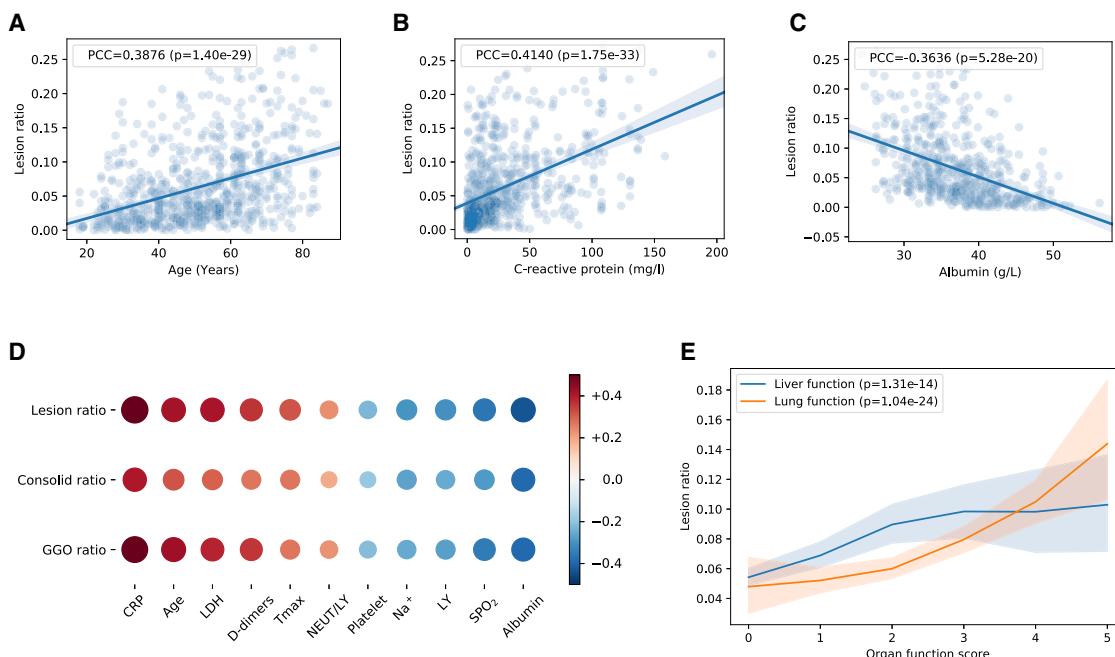


Figure 6. The Correlation of Lung-Lesion Features with Clinical Parameters

(A–C) Linear regression analysis comparing the volume lesion ratio and three correlated clinical parameters, including (A) age, (B) CRP, and (C) albumin.

(D) Correlation of three CT quantification features (volume ratio of GGO, CL, and total lesion) with clinical parameters. See [STAR Methods](#) for details.

(E) The correlations of the volume ratio of lesion and the c-scores for lung function and liver functions graded by physicians. All p values remained statistically significant after the Holm-Bonferroni adjustment. LDH, lactic dehydrogenase.

See also [Figure S6](#).

DISCUSSION

In this study, we described an AI system for the diagnosis of COVID-19 pneumonia based on chest CT images. The performance of our AI system was comparable to that of practicing radiologists with significant clinical experience and could assist and improve the performance of junior radiologists. This AI effort is driven by the desire to develop a system for rapid diagnosis of NCP to assist radiologists and clinicians in combating this pandemic. Such an AI system can also ease the significant demand for diagnostic expertise when the health system is overloaded in a pandemic situation or in remote areas. At present, our AI system is designed to assist radiologists and clinicians as an efficient first reading and/or screening tool as this may reduce patient waiting time, shortening diagnostic workflow time and therefore lessening a radiologist's overall workload and allowing them to respond quickly and more effectively in an emergency situation. The precise measurement of the different lung damage parameters in the CT scan will also allow an objective and quantitative measurement of the disease severity and has the potential to provide an objective and quantitative evaluation of the efficacy of a drug treatment, including antivirals and other immunomodulators, on the lung lesions.

The correlation of the lesion features in the CT scans as evaluated by our AI system and the clinical and biochemical evidence of disease severity based on other organ systems' laboratory parameters highlighted the pathogenesis interlink involving different organs of COVID-19. From a pathogenetic mechanistic perspec-

tive, our study can only show the correlation, and it remains to be determined through other pathogenesis studies whether this interlink is established directly through viral infection of these organs and/or secondary respiratory failures. The higher correlation with the lung lesions with outcomes compared to that of other clinical parameters did highlight the importance of lung damage as the key parameter in the overall prognostic implications. The association of age and outcome is consistent with recent reports that older age is a significant risk factor for poor outcomes (Huang et al., 2020a). The identification of the other parameters including CRP, serum sodium concentration, serum albumin, and platelet count as prognostic factors is consistent with the prognostic factors seen in patients with multi-organ failure. An increased CRP level may reflect the heightened systemic acute inflammation reflected in the severity of pulmonary inflammation, consistent with the prominent lung inflammation as observed in CT scans and postmortem studies. The association with liver and renal biochemistry with prognosis may be another reflection of the multi-organ failure observed in patients with severe or critical COVID-19, although whether there is direct viral involvement of these organs cannot be addressed in this study.

The development of a clinical prognostic model based on our AI system utilizing CT parameters and clinical data was an important advancement toward the use of AI as assisting clinical management. Based on our database, we were able to identify a c-score of ≥ 0.5 as the high-risk group in terms of the eventual progression into severe or critical illness resulting in ICU admission, mechanical ventilation, or death. Importantly, an estimated

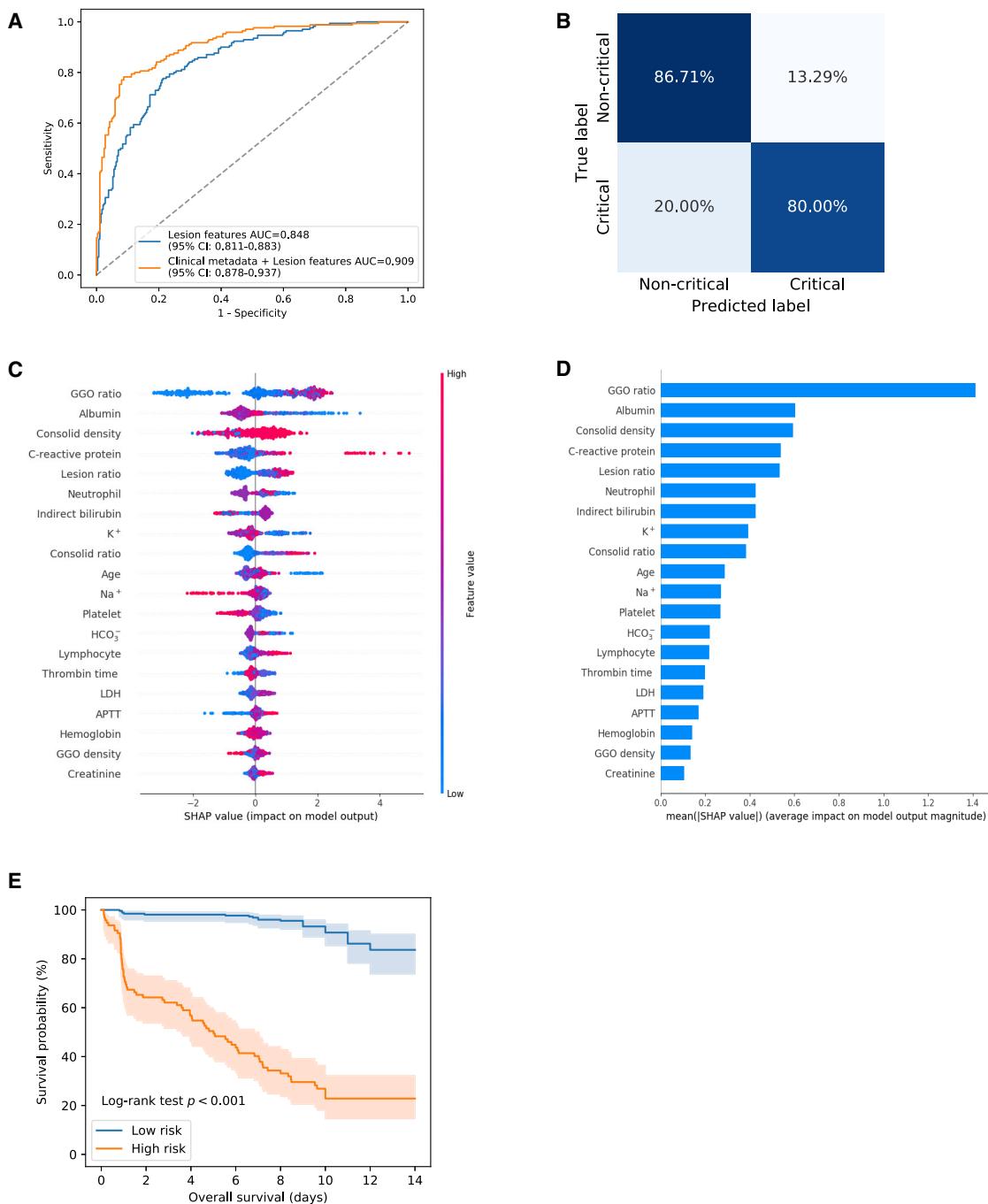


Figure 7. Risk Factors and Clinical Prognosis Analysis for Progression to Severe or Critical Illness

(A) The ROC curves for a binary classification of progression to critical illness stratified by lesion features and the combination of lesion features and clinical metadata.

(B) Corresponding normalized confusion matrix: sensitivity = 80.00%, specificity = 86.71%, AUROC = 0.9093 (95% CI: 0.8775–0.9369).

(C and D) Illustration of features contributing to progression to critical illness by SHAP values. (C) The relative contributions of CT and clinical parameters for prognosis prediction. Features on the right of the risk explanation bar pushed the risk higher, and features on the left pushed the risk lower. (D) The relative contribution of each of the CT or clinical parameters to predict the risk of progression to severe or critical illness.

(E) When the patients were stratified into high-risk ($c\text{-score} \geq 0.5$) and low-risk ($c\text{-score} < 0.5$) groups, Kaplan-Meier curves of progression to critical illness showed a distinct difference in survival probability in this cohort. APTT, activated partial thromboplastin time; C-reactive protein, CRP; ground-glass opacity, GGO.

See also Figures S4, S6, and S7 and Table S4.

time to this progression can also be provided. This information will assist clinicians in planning monitoring and allocating resources in the ICU, and in case of an epidemic, it will allow the healthcare system to have a few more days to react to the demand of resources.

In summary, an AI system that can accurately diagnose NCP and assist radiologists and physicians has been developed. A good correlation between the NCP lung lesions as reviewed by CT parameters and the clinical and biochemical markers of multiple organs was observed, highlighting that multi-organ failure were observed in COVID-19. Together with the clinical prognostic estimation function, this AI system can assist radiologists and physicians such as emergency room (ER) physicians and pulmonologists to accurately diagnose patients rapidly, and with the prognostic estimation generated, it can assist the physicians in determining the subset of patients that will require close monitoring and early intervention and/or support as needed and the estimated timing of these needs.

As this AI system may be of help to physicians and healthcare systems globally to better manage their patients during this pandemic, we are opening up this AI system to all radiologists and clinicians and hope that this can assist their management of patients, and the data that they input may further perfect our AI system, which we hope can evolve as a versatile tool for the global community to combat COVID and other emerging viral infections. For an example of the clinical deployment, see [Figure S7](#).

STAR★METHODS

Detailed methods are provided in the online version of this paper and include the following:

- [KEY RESOURCES TABLE](#)
- [RESOURCE AVAILABILITY](#)
 - Lead Contact
 - Materials Availability
 - Data and Code Availability
- [EXPERIMENTAL MODEL AND SUBJECT DETAILS](#)
 - Human Subjects
 - Experimental Data
- [METHODS DETAILS](#)
 - Image Labeling and Quality Control
 - AI versus Radiologists Comparison
 - Prospective Pilot Studies
 - Diagnosis System and Network Architectures
 - Evaluation of Drug Treatment Effects
 - Lung-Lesion Features and Clinical Parameters
 - Prognosis Analysis
- [QUANTIFICATION AND STATISTICAL ANALYSIS](#)

SUPPLEMENTAL INFORMATION

Supplemental Information can be found online at <https://doi.org/10.116/j.cell.2020.04.045>.

ACKNOWLEDGMENTS

We thank many physicians in China who generously donated their time and expertise for this project. We thank Drs. Roberto Hidalgo and Daniel Recalde

in Guayaquil, Ecuador for providing de-identified CT images from NCP patients, other common pneumonia patients, and normal patients. We thank members of Zhang, Lin, and Wang groups for their assistance and helpful discussions. We thank many volunteers and physicians for grading CT images and medical records. This study was funded by the National Key Research and Development Program of China (2019YFB1404804); National Natural Science Foundation of China (grants 61906105, 61872218, 61721003, 61673241, 81871890, and 91859203); a Macao FDCT grant (0035/2020/A); Guangzhou Regenerative Medicine and Health Guangdong Laboratory (2020GZR110306001); a Kunmin Science and Technology grant (2020-1-H-003); Special Item for Prevention and Control of COVID-19 Science and Technology, Guangdong Province; Tencent Charity Foundation; The three special items of Sun Yat-sen University for novel coronavirus, The Key Areas Research and Development Program of Guangdong (Grant no. 2018B010109006); and the Guangdong Provincial Clinical Research Center for Urinary Diseases, Recruitment Program of Leading Talent in Guangdong Province (2016LJ06Y375).

AUTHOR CONTRIBUTIONS

X.L., J.S., Z.L., Y.S., X.W., Y.C., W.Liang, C.W., K.Wang, L.Ye, M.G., Z.Z., L.Li, J.W., Z.Y., H.C., J.X., L.Yang, W.C., W.X., S.W., W.Z., S.J., L.Z., X.Z., L.W., L.Lu, J.L., H.W., W.W., O.L., C.Z., L.Liang, T.W., R.D., K.Wei, Y.Zhou, T.C., M.F., J.Y.-N.L., J.H., T.L., W.Li, G.W., and K.Z. collected and analyzed the data. K.Z. and G.W. conceived and supervised the project and wrote the manuscript with assistance from T.L., M.F., and J.Y.-N.L. All authors discussed the results and reviewed the manuscript.

DECLARATION OF INTERESTS

The authors declare no competing interests.

Received: March 4, 2020

Revised: April 6, 2020

Accepted: April 23, 2020

Published: May 4, 2020; corrected online: August 15, 2020

REFERENCES

- Badrinarayanan, V., Kendall, A., and Cipolla, R. (2017). Segnet: A deep convolutional encoder-decoder architecture for image segmentation. *IEEE Trans. Pattern Anal. Mach. Intell.* 39, 2481–2495.
- Burlina, P.M., Joshi, N., Pekala, M., Pacheco, K.D., Freund, D.E., and Bressler, N.M. (2017). Automated grading of age-related macular degeneration from color fundus images using deep convolutional neural networks. *JAMA Ophthalmol.* 135, 1170–1176.
- Chan, J.F., Yuan, S., Kok, K.H., To, K.K., Chu, H., Yang, J., Xing, F., Liu, J., Yip, C.C., Poon, R.W., et al. (2020). A familial cluster of pneumonia associated with the 2019 novel coronavirus indicating person-to-person transmission: a study of a family cluster. *Lancet* 395, 514–523.
- Chen, L.-C., Papandreou, G., Schroff, F., and Adam, H. (2017). Rethinking atrous convolution for semantic image segmentation. *arXiv:170605587*.
- Chen, A., Karwoski, R.A., Gierada, D.S., Bartholmai, B.J., and Koo, C.W. (2020). Quantitative CT Analysis of Diffuse Lung Disease. *Radiographics* 40, 28–43.
- Chilamkurthy, S., Ghosh, R., Tanamala, S., Biviji, M., Campeau, N.G., Venugopal, V.K., Mahajan, V., Rao, P., and Warier, P. (2018). Deep learning algorithms for detection of critical findings in head CT scans: a retrospective study. *Lancet* 392, 2388–2396.
- Davidson-Pilon, C. (2019). Lifelines: Survival analysis in Python. *Journal of Open Source Software* 4, 1317, 10.21105/joss.01317.
- Devalla, S.K., Renukanand, P.K., Sreedhar, B.K., Subramanian, G., Zhang, L., Perera, S., Mari, J.M., Chin, K.S., Tun, T.A., Strouthidis, N.G., et al. (2018). DRUNET: a dilated-residual U-Net deep learning network to segment optic

- nerve head tissues in optical coherence tomography images. *Biomed. Opt. Express* 9, 3244–3265.
- Efron, B. (1979). Bootstrap Methods: Another Look at the Jackknife. *Ann. Stat.* 7, 1–26.
- Esteva, A., Robicquet, A., Ramsundar, B., Kuleshov, V., DePristo, M., Chou, K., Cui, C., Corrado, G., Thrun, S., and Dean, J. (2019). A guide to deep learning in healthcare. *Nat. Med.* 25, 24–29.
- Force, A.D.T., Ranieri, V.M., Rubenfeld, G.D., Thompson, B.T., Ferguson, N.D., Caldwell, E., Fan, E., Camporota, L., and Slutsky, A.S. (2012). Acute respiratory distress syndrome: the Berlin Definition. *Jama* 307, 2526–2533.
- Guan, W.J., Ni, Z.Y., Hu, Y., Liang, W.H., Ou, C.Q., He, J.X., Liu, L., Shan, H., Lei, C.L., Hui, D.S.C., et al.; China Medical Treatment Expert Group for Covid-19 (2020). Clinical Characteristics of Coronavirus Disease 2019 in China. *N. Engl. J. Med.* 382, 1708–1720.
- Gulshan, V., Peng, L., Coram, M., Stumpe, M.C., Wu, D., Narayanaswamy, A., Venugopalan, S., Widner, K., Madams, T., Cuadros, J., et al. (2016). Development and validation of a deep learning algorithm for detection of diabetic retinopathy in retinal fundus photographs. *JAMA* 316, 2402–2410.
- Huang, C., Wang, Y., Li, X., Ren, L., Zhao, J., Hu, Y., Zhang, L., Fan, G., Xu, J., Gu, X., et al. (2020a). Clinical features of patients infected with 2019 novel coronavirus in Wuhan, China. *Lancet* 395, 497–506.
- Hara, K., Kataoka, H., and Satoh, Y. (2017). Learning Spatio-Temporal Features with 3D Residual Networks for Action Recognition. *arXiv:170807632*.
- Hara, K., Kataoka, H., and Satoh, Y. (2018). Towards good practice for action recognition with spatiotemporal 3D convolutions. In Proceedings of the 24th International Conference on Pattern Recognition (ICPR). <https://doi.org/10.1109/ICPR.2018.8546325>.
- Huang, P., Liu, T., Huang, L., Liu, H., Lei, M., Xu, W., Hu, X., Chen, J., and Liu, B. (2020b). Use of Chest CT in Combination with Negative RT-PCR Assay for the 2019 Novel Coronavirus but High Clinical Suspicion. *Radiology* 295, 22–23.
- Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., Ye, Q., and Liu, T.-Y. (2017). Lightgbm: A highly efficient gradient boosting decision tree. Proceedings of Advances in Neural Information Processing Systems (NIP 2017), 3146–3154.
- Kermany, D.S., Goldbaum, M., Cai, W., Valentim, C.C.S., Liang, H., Baxter, S.L., McKeown, A., Yang, G., Wu, X., Yan, F., et al. (2018). Identifying Medical Diagnoses and Treatable Diseases by Image-Based Deep Learning. *Cell* 172, 1122–1131.e9.
- Li, Z., He, Y., Keel, S., Meng, W., Chang, R.T., and He, M. (2018). Efficacy of a Deep Learning System for Detecting Glaucomatous Optic Neuropathy Based on Color Fundus Photographs. *Ophthalmology* 125, 1199–1206.
- Long, J., Shelhamer, E., and Darrell, T. (2015). Fully convolutional networks for semantic segmentation. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 3431–3440.
- Long, E., Lin, H., Liu, Z., Wu, X., Wang, L., Jiang, J., An, Y., Lin, Z., Li, X., Chen, J., et al. (2017). An artificial intelligence platform for the multihospital collaborative management of congenital cataracts. *Nat. Biomed. Eng.* 1, 1–8.
- Lundberg, S.M., Erion, G.G., and Lee, S.-I. (2018a). Consistent individualized feature attribution for tree ensembles. *arXiv:180203888*.
- Lundberg, S.M., Nair, B., Avilalala, M.S., Horibe, M., Eisses, M.J., Adams, T., Liston, D.E., Low, D.K., Newman, S.F., Kim, J., and Lee, S.I. (2018b). Explainable machine-learning predictions for the prevention of hypoxaemia during surgery. *Nat. Biomed. Eng.* 2, 749–760.
- Munster, V.J., Koopmans, M., van Doremalen, N., van Riel, D., and de Wit, E. (2020). A Novel Coronavirus Emerging in China - Key Questions for Impact Assessment. *N. Engl. J. Med.* 382, 692–694.
- Norgeot, B., Glicksberg, B.S., and Butte, A.J. (2019). A call for deep-learning healthcare. *Nat. Med.* 25, 14–15.
- Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., et al. (2019). PyTorch: An imperative style, high-performance deep learning library. *Adv. Neural Inf. Process. Syst.*, 8024–8035.
- Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., and Vanderplas, J. (2011). Scikit-learn: Machine learning in Python. *the Journal of machine Learning research* 12, 2825–2830.
- Phan, L.T., Nguyen, T.V., Luong, Q.C., Nguyen, T.V., Nguyen, H.T., Le, H.Q., Nguyen, T.T., Cao, T.M., and Pham, Q.D. (2020). Importation and Human-to-Human Transmission of a Novel Coronavirus in Vietnam. *N. Engl. J. Med.* 382, 872–874.
- Poplin, R., Varadarajan, A.V., Blumer, K., Liu, Y., McConnell, M.V., Corrado, G.S., Peng, L., and Webster, D.R. (2018). Prediction of cardiovascular risk factors from retinal fundus photographs via deep learning. *Nat. Biomed. Eng.* 2, 158–164.
- Ravizza, S., Huschto, T., Adamov, A., Böhm, L., Büsser, A., Flöther, F.F., Hinzmann, R., König, H., McAhren, S.M., Robertson, D.H., et al. (2019). Predicting the early risk of chronic kidney disease in patients with diabetes using real-world data. *Nat. Med.* 25, 57–59.
- Ronneberger, O., Fischer, P., and Brox, T. (2015). U-net: Convolutional networks for biomedical image segmentation. Proceedings of the International Conference on Medical Image Computing and Computer-Assisted Intervention, 234–241.
- Rothe, C., Schunk, M., Sothmann, P., Bretzel, G., Froeschl, G., Wallrauch, C., Zimmer, T., Thiel, V., Janke, C., Guggemos, W., et al. (2020). Transmission of 2019-nCoV Infection from an Asymptomatic Contact in Germany. *N. Engl. J. Med.* 382, 970–971.
- Shi, H., Han, X., Jiang, N., Cao, Y., Alwalid, O., Gu, J., Fan, Y., and Zheng, C. (2020). Radiological findings from 81 patients with COVID-19 pneumonia in Wuhan, China: a descriptive study. *Lancet Infect. Dis.* 20, 425–434.
- Singer, M., Deutschman, C.S., Seymour, C.W., Shankar-Hari, M., Annane, D., Bauer, M., Bellomo, R., Bernard, G.R., Chiche, J.D., Coopersmith, C.M., et al. (2016). The Third International Consensus Definitions for Sepsis and Septic Shock (Sepsis-3). *Jama* 315, 801–810.
- Ting, D.S.W., Cheung, C.Y., Lim, G., Tan, G.S.W., Quang, N.D., Gan, A., Hamzah, H., Garcia-Franco, R., San Yeo, I.Y., Lee, S.Y., et al. (2017). Development and Validation of a Deep Learning System for Diabetic Retinopathy and Related Eye Diseases Using Retinal Images From Multiethnic Populations With Diabetes. *JAMA* 318, 2211–2223.
- Topol, E.J. (2019). High-performance medicine: the convergence of human and artificial intelligence. *Nat. Med.* 25, 44–56.
- van Grinsven, M.J., van Ginneken, B., Hoyng, C.B., Theelen, T., and Sánchez, C.I. (2016). Fast convolutional neural network training using selective data sampling: Application to hemorrhage detection in color fundus images. *IEEE Trans. Med. Imaging* 35, 1273–1284.
- Wang, C., Horby, P.W., Hayden, F.G., and Gao, G.F. (2020). A novel coronavirus outbreak of global health concern. *Lancet* 395, 470–473.
- Xie, X., Zhong, Z., Zhao, W., Zheng, C., Wang, F., and Liu, J. (2020). Chest CT for Typical 2019-nCoV Pneumonia: Relationship to Negative RT-PCR Testing. *Radiology*, 200343.
- Yang, X., Yu, Y., Xu, J., Shu, H., Xia, J., Liu, H., Wu, Y., Zhang, L., Yu, Z., Fang, M., et al. (2020). Clinical course and outcomes of critically ill patients with SARS-CoV-2 pneumonia in Wuhan, China: a single-centered, retrospective, observational study. *Lancet Respir. Med.* S2213-2600(20)30079-5. [https://doi.org/10.1016/s2213-2600\(20\)30079-5](https://doi.org/10.1016/s2213-2600(20)30079-5).
- Zhu, N., Zhang, D., Wang, W., Li, X., Yang, B., Song, J., Zhao, X., Huang, B., Shi, W., Lu, R., et al.; China Novel Coronavirus Investigating and Research Team (2020). A Novel Coronavirus from Patients with Pneumonia in China, 2019. *N. Engl. J. Med.* 382, 727–733.

STAR★METHODS

KEY RESOURCES TABLE

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Deposited Data		
CT images and clinical data and codes	This paper	http://ncov-ai.big.ac.cn/download?lang=en
Software and Algorithms		
Pytorch v1.2	Paszke et al., 2019	https://pytorch.org
Deeplabv3	Chen et al., 2017	https://github.com/pytorch/vision
3D ResNet	Hara et al., 2018	https://github.com/kenshohara/3D-ResNets-PyTorch
lifelines v0.24.0	Davidson-Pilon, 2019	https://github.com/CamDavidsonPilon/lifelines
Lightgbm v2.2.3	Ke et al., 2017	https://github.com/microsoft/LightGBM

RESOURCE AVAILABILITY

Lead Contact

Further information and requests for resources should be directed to the Lead Contact, Kang Zhang (kang.zhang@gmail.com). All CT images data and metadata and codes generated in this study are available from the Lead Contact.

Materials Availability

This study did not generate new unique reagents.

Data and Code Availability

Chest CT images and clinical metadata and codes are deposited into the China National Center for Bioinformation at the website (<http://ncov-ai.big.ac.cn/download?lang=en>).

EXPERIMENTAL MODEL AND SUBJECT DETAILS

Human Subjects

CT images were collected from cohorts from the China Consortium of Chest CT Image Investigation (CC-CCII), which consists of Sun Yat-sen Memorial Hospital and Third Affiliated Hospital of Sun Yat-sen University, The first Affiliated Hospital of Anhui Medical University, West China Hospital, Nanjing Renmin Hospital, Yichang Central People's Hospital, Renmin Hospital of Wuhan University. Human subjects were deemed clinically appropriate for a chest CT scan during the management of the patients by the clinicians. The influence (or association) of age or gender were not taken into the exclusion criteria. Institutional Review Board (IRB)/Ethics Committee approvals were obtained in all the institutions and consent was obtained from all participants. The work was conducted in compliance with the Chinese CDC policy on reportable infectious diseases and the Chinses Health and Quarantine Law and in compliance with patient privacy regulations in China, and was adherent to the tenets of the Declaration of Helsinki.

Experimental Data

The CC-CCII contained a total of 617,775 CT slices of 6752 CT scans from 4154 patients. The study sample size was estimated by a standard AI training and validation approach. Patients were randomly assigned to a training set (80%), an internal validation set (10%) or a test set (10%). We used a total of 444,034 CT slices of 2,778 patients from the CC-CCII for training and internal validation of the AI diagnosis system, including 164,241 slices from 917 NCP patients, 183,933 slices from 983 pneumonia patients and 95,860 slices from 878 normal controls. Of these, NCP diagnosis was given when a patient had pneumonia with a confirmed reverse-transcriptase-PCR. The common pneumonia group include viral pneumonia (including adenoviral, influenza, and parainfluenza pneumonia), bacterial pneumonia, and mycoplasma pneumonia, all of which together are the most common causes of pneumonia, which were diagnosed based on standard clinical, radiological, culture/molecular assay results. We matched lesion severity levels between NCP and other common pneumonia by lesion volume ratio measurements (**Figure S5D**). All cohorts had viral pneumonia as a part of other common pneumonia diagnosis. The percentages of viral pneumonia were as following: initial training/validation/testing cohort, 48%; external validation cohort (City of Yichang), 27%; prospective cohort 1 (City of Hefei), 36%; prospective cohort 2 (City of Guangzhou), 38%; international cohort, 15%. CT scans from each patient were put into the corresponding dataset based on the pa-

tient assignment. We excluded patients without corresponding non-contrast CT scans. Scan sets without serial information or containing any motion artifacts or significant image resolution reductions were also excluded from the study. Details of patient characteristics and the inclusion and exclusion criteria are given in **Figure S1**. For the CT quantitative and correlation analysis with clinical parameters, 843 admitted NCP patients with clinical metadata were included. All the CT scans and clinical metadata used in our analysis were collected from patients at the time of hospital admission. Of these, 456 hospitalized patients with clinical outcome information were used in the prognosis estimation analysis. The endpoint in the study was defined as “severe or critical illness” which means admission to intensive care unit (ICU), on mechanical ventilation or death.

METHODS DETAILS

Image Labeling and Quality Control

To train and evaluate our semantic segmentation framework, a subset of 2,879 CT slices from NCP patients and a set of 1816 CT slices from other pneumonia patients were manually segmented at the pixel level. The annotation was done via polygons. The segmentation labels were selected as relevant pathological features for distinguishing NCP and other common pneumonia. The annotation included lung field, and five commonly seen categories of lesions including CL, GGO, pulmonary fibrosis, interstitial thickening and pleural effusion. There were 4,406 consolidation lesions, 10,544 ground-glass opacities and a total of 2,571 other three types of lesions annotated (**Table S2**). The segmentations were annotated and reviewed by five senior radiologists with 15 to 25 years of experience.

For the analysis of CT images from CC-CCII, all radiographs were initially screened for quality control by removing all low quality or unreadable scans. For all CT images, each image went through a tiered grading system consisting of two layers of trained graders of increasing expertise for verification and correction of image labels. Each image imported into the database started with a label matching the diagnosis of the patient. This first tier of graders who were radiologists with 5 to 15 years of clinical practice experience conducted initial quality control and excluded images containing severe artifacts or significant image resolution reductions. The presence or absence of lung lesions in CT images were recorded. The second tier of five independent radiologists with at least 25 years of clinical practice experience verified the true labels for each image randomly selected from 10% of all images, this group of senior radiologists also served on the consensus committee for the ground truth adjudicator in AI versus radiologist comparison.

AI versus Radiologists Comparison

For comparing the performance of our AI system with experienced practicing radiologists, an independent dataset which consisted of a total of 18,392 slices from 150 patients including 40 NCP patients, 80 common pneumonia patients and 30 normal controls were employed (**Table S1**). Eight practicing radiologists were enrolled to participate in the AI comparison study and were allocated into two groups: junior group, with 5 to 15 years of clinical experience; and senior group with 15 to 25 years of clinical experience. A weighted error based on penalty score was used to evaluate our AI system and the experts to reflect clinical performance. We set the misdiagnosing NCP to normal with a score of 2, as it may cause the most severe outcome as compared to misdiagnosing NCP as “other common pneumonia,” which had a score of 1. In addition, the scores of misdiagnosing the rest of classes were set as 1.

We further conducted a study to investigate the impact of the AI diagnostic system in aiding the performance of the four junior radiologists. During this re-reading of the same CT four weeks after the initial reading, the AI system readout was provided to the radiologists with lesion labeling of each slice and a final diagnostic probability at a patient level (three-way classifications) and the radiologist were asked to make a diagnosis again to access the impact of our AI system on the radiologists diagnosis and compared with the Ground Truth from the Senior radiologists

Prospective Pilot Studies

In the first prospective pilot study conducted in Wuhan (Hubei), we enrolled 201 consecutive NCP, 144 common pneumonia patients and 64 normal controls. We performed this study from Jan 25th to March 25th, 2020. Similarly, the second prospective pilot study was conducted in a non-epidemic central region in China with 41 consecutive NCP, 128 common pneumonia patients and 73 normal controls (Hefei from Feb 2nd to March 25th). The third prospective pilot study was conducted in another region in southern China with 20 consecutive NCP, 57 common pneumonia patients and 63 normal controls (Guangzhou from Feb 8th to March 27th, **Table S1**).

To further validate our AI system outside China, we also obtained CT images from a cohort of patients from Ecuador and data from an open source website (<https://radiopaedia.org/encyclopaedia/cases/all?lang=us>). This international cohort consisted of 40 NCP patients, 52 CP patients, and 10 normal controls.

This project was approved by the IRB of the respective institutions. Enrollment criteria included a diagnosis of viral pneumonia on clinical symptoms and signs, and standard laboratory tests. The CT screening was performed in all participants as a part of clinical management. A confirmative diagnosis of NCP was made by a molecular PCR test (**Table S1**).

Diagnosis System and Network Architectures

We constructed a computer-aided diagnosis (CAD) system for detecting COVID-19 patients which consisted of two models, the first lung-lesion segmentation model and the second diagnosis prediction model. The lung-lesion segmentation model took a raw CT scan as the input and produced a lung-lesion map as the output by segmentation networks, in which it generated one out of seven

classes at a pixel level, including background, lung fields, five lesions including CL, GGO, pulmonary fibrosis, interstitial thickening and pleural effusion. The entire scan of a patient was then cropped and transformed into a normalized CT volume map based on the lung-segmentation model. The diagnosis prediction model took the normalized lung-lesion CT volume map and produced a final prediction on whether a patient is normal, with NCP or other common pneumonia by classification networks.

Segmentation networks

As a CT scan usually consisted of many slices, ranging from around 50 (5mm) to 200 (1mm) in thickness, which was challenging for real-time application in clinical practice. To achieve the requirement of real-time segmentation, we constructed a fast and accurate segmentation framework to segment the lung field and lesions in the CT slice. We formulated our segmentation framework as a two-stage segmentation scheme which enabled the model to focus on the medically meaningful regions of input image which will reduce computation cost significantly (see Figure S2A for more details).

In the first stage, we down sampled the input image from 512×512 to 128×128 level and segmented the lung (lung field and lesions) from the image, as the patterns of lung can be learned at a relatively low resolution. In the second stage, we first calculated the bounding box with the lung field segmentation results. The bounding box was slightly larger than the lung field and contained all key regions with pathological features. Next, the key region from the original input image was cropped and resized to a 256×256 level as the input for the second stage segmentation model, which segmented all segmentation classes from the cropped image. The results from the second stage segmentation were transferred to the coordinates of the original input image to form a final segmentation mask. This method increased segmentation framework efficiency and was model agnostic, therefore it could be applied to any image segmentation models.

The two-stage segmentation framework with different backbones was adopted, in which we chose a range of classic semantic segmentation models to conduct extensive experiments, including U-net, DRUNET, FCN, SegNet and DeepLabv3. All images for training and evaluating semantic segmentation model were resized to 512×512 to balance the computation cost and accuracy. Batch normalization was utilized to accelerate the training procedure. The annotated labels were highly imbalanced, with lung field took up the majority of the labels. To solve this problem, we used the pixel-level weighted binary cross-entropy and dice loss as the loss function for optimization. During training, SGD optimizer was employed with an initial learning rate at 0.01, momentum at 0.9 and weight decay at 0.0001. The learning rate decays by a factor of 0.9 for every 5 epochs. The training batch size is 4.

Classification networks

The segmentation results of CT slices in a CT scan were stacked vertically to form a volume. The lung field region was then cropped from the volume and normalized to $64 \times 128 \times 128$ for depth, height and width respectively. Then the normalized volume is converted into one-hot representation for each pixel, deriving a tensor $7 \times 64 \times 128 \times 128$. The 3D classification network took the tensor as input and output the diagnosis probability of three categories: NCP, CP and normal controls.

The detailed structure of the 3D classification network was shown in **Figure S2B**, adapted from 3D ResNet-18 (Hara et al., 2017). The network used multiple 3D convolutional blocks with residual connections to continuously extract local and global contextual features, and the final predictions were calculated with a fully connected layer followed with a softmax activation function. For the three-way diagnosis decision, the model output the class with the maximum probability. The 3D classification network is trained with cross entropy loss between final predictions and ground truth labels. During training, we used Adam optimizer with an initial learning rate at 0.001. The learning rate decays by a factor of 0.1 for every 10 epochs. The training epoch is 20 in total. The training batch size is 8. The whole training, validation and testing procedures were also conducted with Pytorch (v.1.2.0) on NVIDIA GeForce 1080Ti graphical processing units (Paszke et al., 2019).

Evaluation of Drug Treatment Effects

We evaluated the effect of drug treatment on lesion size and volume changes using our AI-based quantitative measurements. NCP patients undergoing three different experimental drug treatment in observation trials were analyzed. Drug 1, drug 2 and drug 3 group enrolled 12, 8 and 22 NCP patients respectively. The enrollment criteria included a confirmed NCP diagnosis by a positive viral PCR test, and no other prior treatment history. We quantified lung lesions (GGO and total lesions volumes) on CT scans and compared the differences between pre-treatment and after treatment of the three experimental drugs. Dependent t test for paired samples was used for statistical analysis (**Figure S4**).

Lung-Lesion Features and Clinical Parameters

Clinical records of COVID-19 patients which consisted of patients' demographics and clinical data, including vital signs, symptoms and signs, as well as imaging studies and laboratory tests results on initial hospital admission and follow ups were manually annotated. Laboratory tests consisted of comprehensive blood biochemical analysis, blood gas analysis, liver and renal biochemistry indexes, cardiac function tests, coagulation tests, and certain serum protein levels and activities, including lactate dehydrogenase (LDH) and C-reactive protein (CRP), etc.

To investigate the correlations between CT quantitative features and clinical parameters, Pearson's and Spearman's correlation tests were performed. Three volume-level quantitative features were used, including GGO to lung volume ratio, CL to lung volume ratio, total lesion (GGO + CL) to lung volume ratio, since they were common lesions for the progressive stage of NCP patients. A correlation analysis between lung-lesion features and lung and liver damage assessment were conducted using a composite score system, as lung and liver functions were highly relevant to outcomes of COVID-19 patients based on previous study and international

guidelines (Force et al., 2012; Singer et al., 2016). We identified and quantified clinical parameters (including age, Tmax, RR, platelet, Na⁺, albumin, AST, LDH, CRP) that were highly correlated with lung lesions by a linear regression analysis. The resultant correlations were regarded as significant when P values was < 0.05 after correction with the Holm-Bonferroni method.

Prognosis Analysis

For an NCP prognostic and risk factor analyses, we denoted the “severe/critical illness” as the endpoint of the prognosis of COVID-19 patients. The status of “severe/critical illness” was defined and recorded as admission to an intensive care unit (ICU), the use of mechanical ventilation, or death. CT lesion features (only volume quantitative features) and clinical metadata features were used for a prognostic prediction with Gradient Boosting Decision Tree Algorithm (GBDT) as the classifier. The GBDT is a tree-based ensemble model, as each node in the tree can be converted to IF-THEN rules that are easily understandable, and the GBDT with default parameters by Python package Lightgbm was employed (Ke et al., 2017). A SHAP method was used to display the impact of relevant risk factors on prognostic prediction for critical illness. SHAP is a value explainable tool for tree-based models, which could efficiently and exactly compute local explanations and global explanations. The performance of a local explanation of SHAP for prognosis prediction with interpretability was also investigated. As an example, two patients from the critical illness and the non-critical illness group were used to show the effects of lung-lesion features and clinical parameters as the input risk factors for prognosis prediction (**Figure S7A** and **Figure S7B**). We used a five-fold cross-validation scheme for prognostic prediction. For each fold, we calculated a probability (c-score) for each patient in the test dataset (20%) using coefficient estimates from the training and validation dataset (70%:10%). Using a cut-off score of the c-score of 0.5, we were able to classify patients into high-risk group (c-score of ≥ 0.5) and low-risk group (c-score of < 0.5). With this stratification, the median survival times for the two groups were calculated by the Kaplan-Meier estimator and a log-rank test. The importance of each parameter for prognostic estimation were also estimated by fitting a multi-variable Cox proportional hazards model on CT quantitative lesion features and clinical parameters.

QUANTIFICATION AND STATISTICAL ANALYSIS

We evaluated semantic segmentation model performance with two evaluation metrices including Pixel Accuracy (PA) and Dice Coefficient (DC). The PA is the percentage of pixels that are classified correctly. The DC is twice the area of overlap between the predicted segmentation and the ground truth divide by the sum of areas of the predicted segmentation and the ground truth. ROC and AUC were used to assess model performance for each classification task. Confidence intervals (CI) of AUC were computed using bootstrapping approach with nonparametric, unstratified resampling of 1000 times (Efron., 1979). Sensitivity, specificity and accuracy were determined by the selected operating point. The operating point between a low false negative diagnostic rate (sensitivity) and a low positive rate (1 – specificity) were set at different thresholds accordingly. Pearson’s and Spearman’s correlation tests with Holm-Bonferroni method were employed for the statistical analyses. The training, validation and testing procedures of deep learning models were conducted with Pytorch (v.1.2.0). We used the Python scikit-learn library for data analysis and the Python matplotlib and seaborn libraries to plot graphs. The Python package lightgbm and lifelines were employed for prognostic prediction. The measurements of sensitivity, specificity, and accuracy were calculated by Python scikit-learn library (Pedregosa et al., 2011).

Supplemental Figures

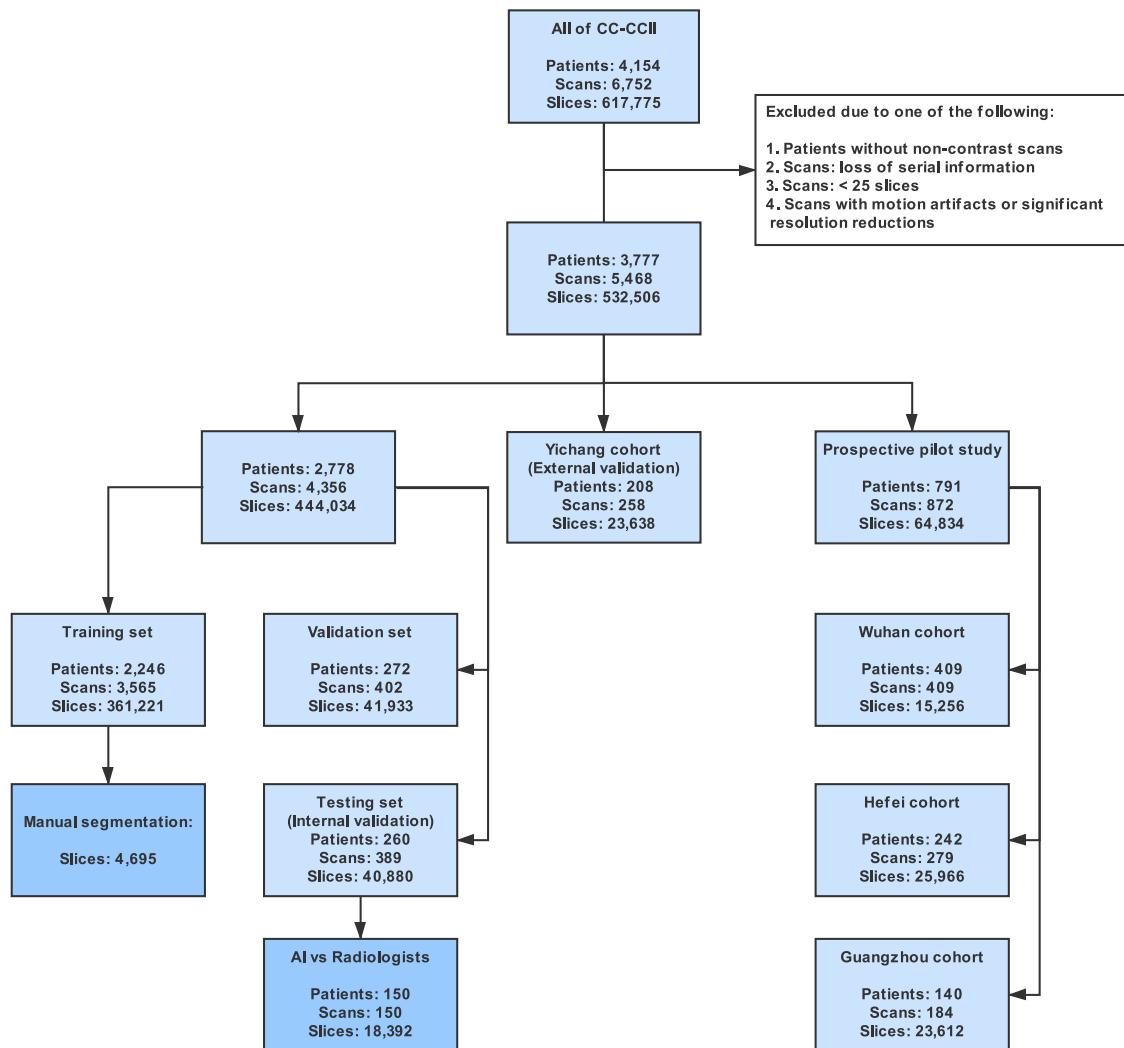


Figure S1. STARD Diagram Describing the CT Dataset Used for Our AI System from CC-CCII, Related to Figure 1
The exclusion criteria were also considered.

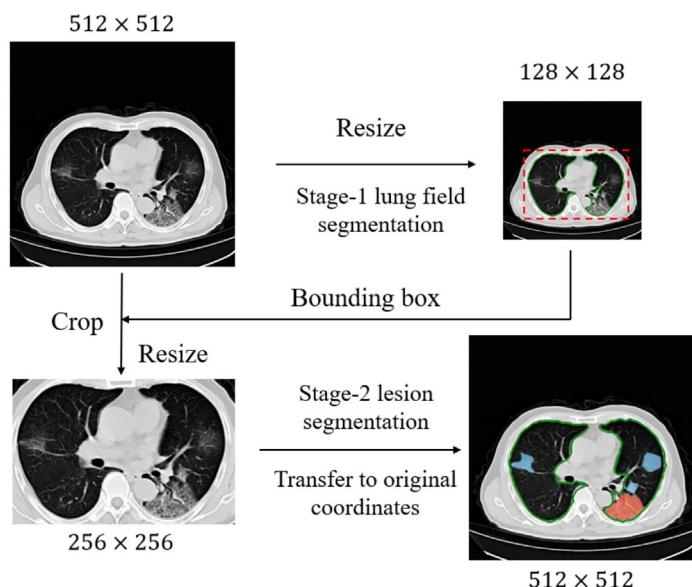
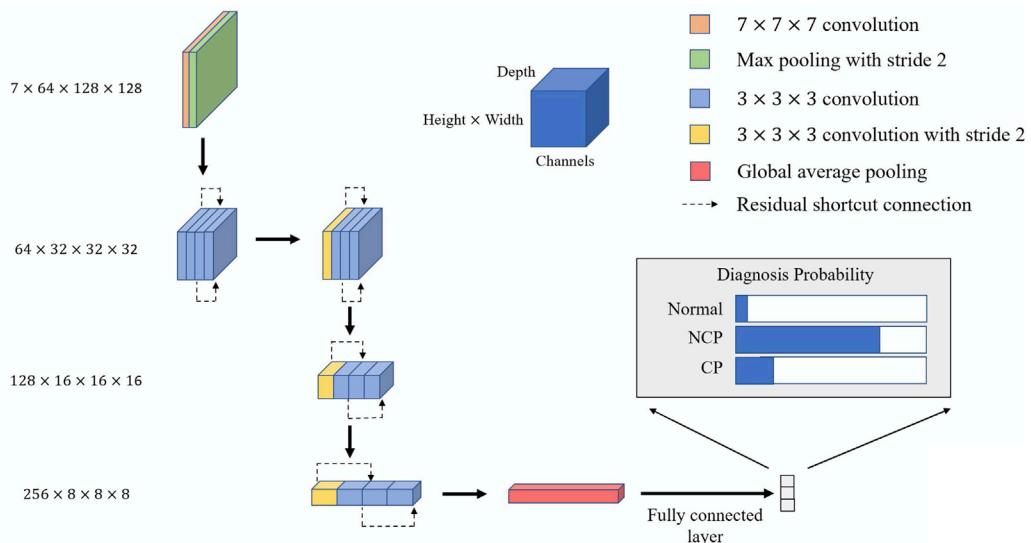
A**B**

Figure S2. Illustration of Network Architectures of the Proposed AI Diagnostic System, Related to Figure 1

(A) Two-stage segmentation module for acceleration. In the first stage, we down-sampled the input image to a 128×128 level and segmented the lung field from the image, as the patterns of lung fields were easily learned at a relatively low resolution. In the second stage, we first calculated the bounding box with the lung field segmentation results. The key region was cropped from the original input image and resized it to a 256×256 level as the input for the second stage segmentation model.

(B) The 3D classification networks used in our COVID-19 diagnosis system. For more details see [STAR Methods](#).

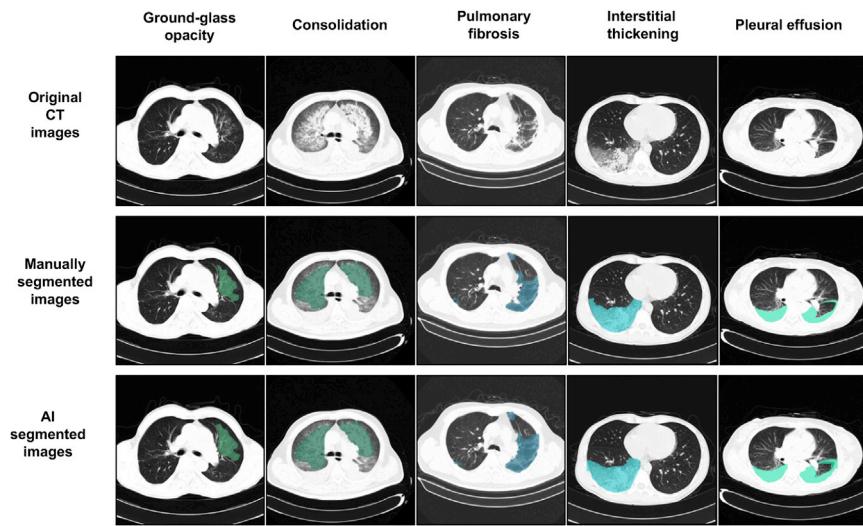


Figure S3. Segmentation Examples of Our Model for Lesion Segmentation Task, Related to Figure 2

Upper row, original CT slices of five types of lesions; middle row, manually segmented CT slices; lower column, AI-based automated segmented CT slices. The five columns represented CT slice with lesions of ground-glass opacity (GGO), consolidation, pulmonary fibrosis, interstitial thickening, and pleural effusion (from left to right).

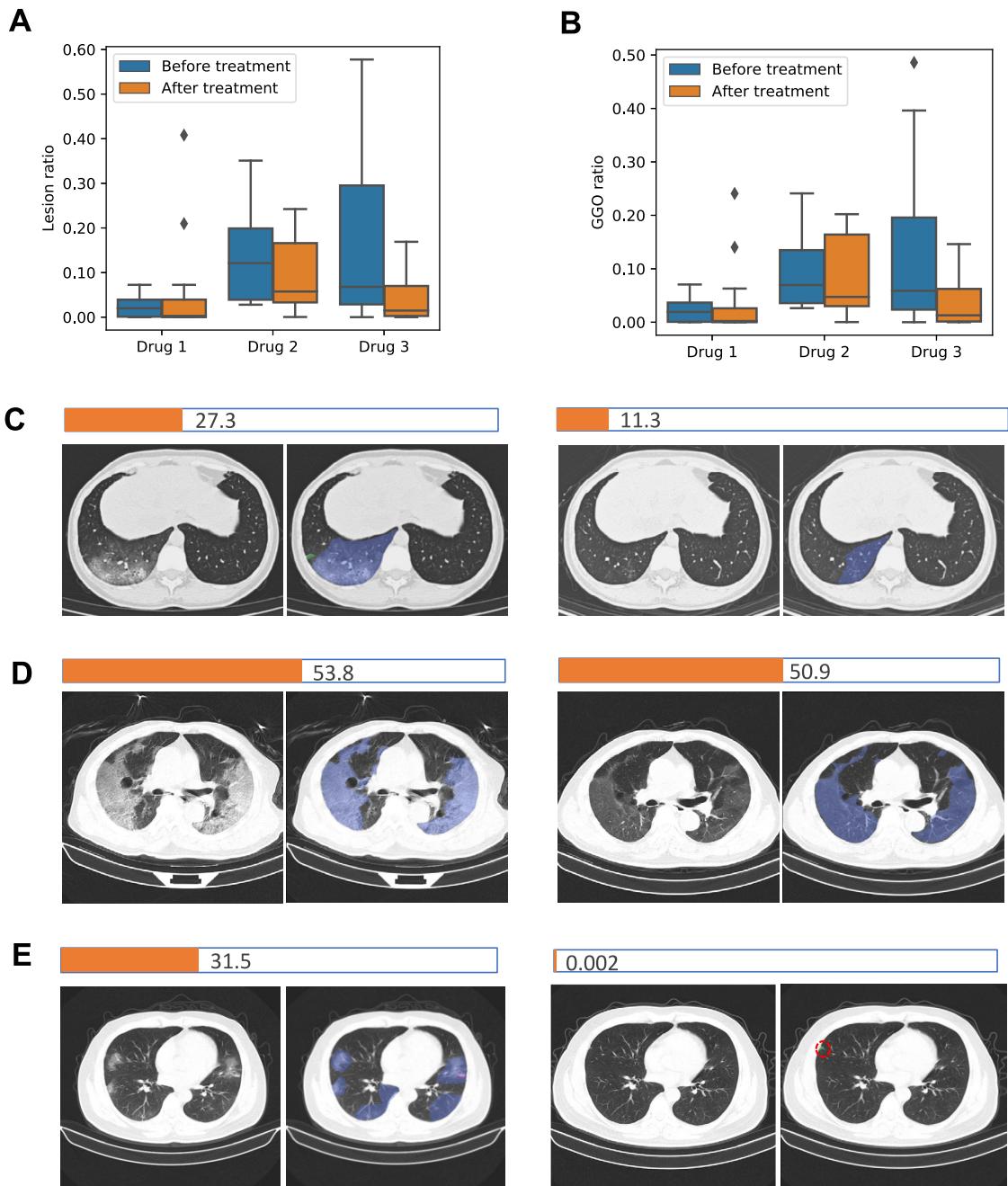


Figure S4. Evaluation of Drug Treatment Effects by AI-Based Lesion Quantitative Measurements, Related to STAR Methods

Comparative measurements of ground glass opacities (GGO) and total lesion (lesion) volume ratio before and after a drug treatment in three preliminary drug treatment observation trials (drug 1, 2 and 3).

(A and B) Bar graphs comparing lesion volume changes before and after treatment by three drugs.

(C-E) Image examples of lesion changes before treatment (left panels) and after treatment (right panels). The NCP total lesion area in the example slice of each patient was quantified as a horizontal bar. A typical image with lesions and corresponding AI segmentation was presented for each drug treatment. For the AI segmentation color code, blue, purple and green represented GGO, consolidation (CL) and pulmonary fibrosis, respectively. (C) A representative patient from the drug 1 group. (D) A representative patient from the drug 2 group. (E) A representative patient from the drug 3 group. A t test was used to measure statistical significance comparing before and after a treatment. The lesion change comparison before and after treatment was no statistically significant in the drug 1 group, whereas it was significant in the Drug 2 group ($p = 0.0345$) and the Drug 3 ($p = 0.00056$).

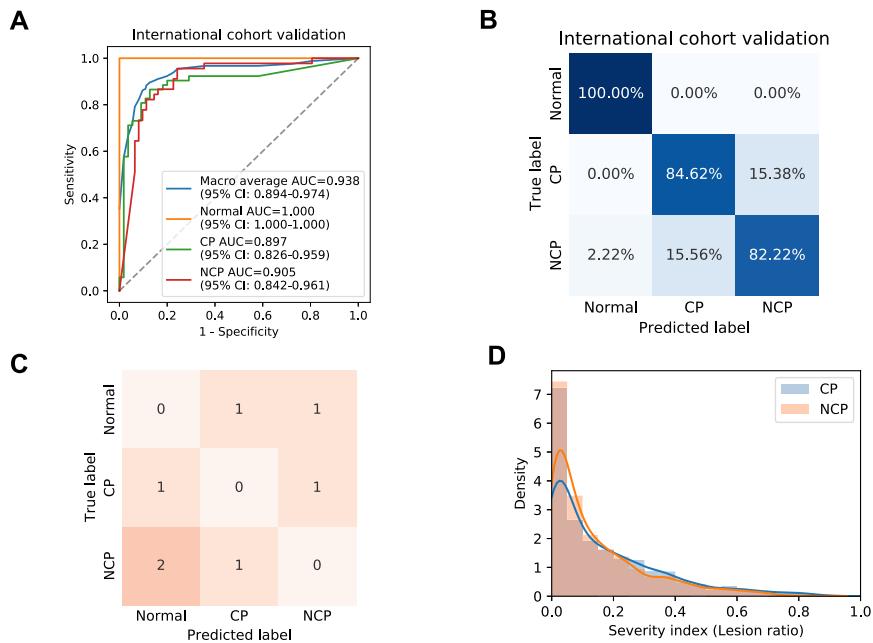


Figure S5. Evaluation and Diagnostic Performance of the AI System, Related to Figure 3

(A and B) AI performance in an independent international cohort. Receiver operating characteristic curves (ROC) and normalized confusion matrix of the model for detecting NCP patients from common pneumonia (CP) and normal controls. For three-way classification: accuracy = 85.05%, AUROC = 0.9381 (95% CI: 0.8944-0.9742). For NCP versus the rest: accuracy = 84.11%, sensitivity = 86.67%, specificity = 82.26%, AUROC = 0.9050 (95% CI: 0.8421-0.9612).

(C) Penalty scoring matrix.

(D) A distribution plot of the severity index (lesion volume ratios) between NCP patients and common pneumonia patients, which represented a severity level comparison between the two disease groups. The distribution difference between these two groups was evaluated by a statistical measurement of Jaccard Similarity (JS), which was the intersection divided by the union of distribution of two samples. The JS of the lesion ratios for CP and NCP patients was of 0.939, suggesting that the distributions of severity levels were similarly matched and would not generate a bias in diagnosis analysis.

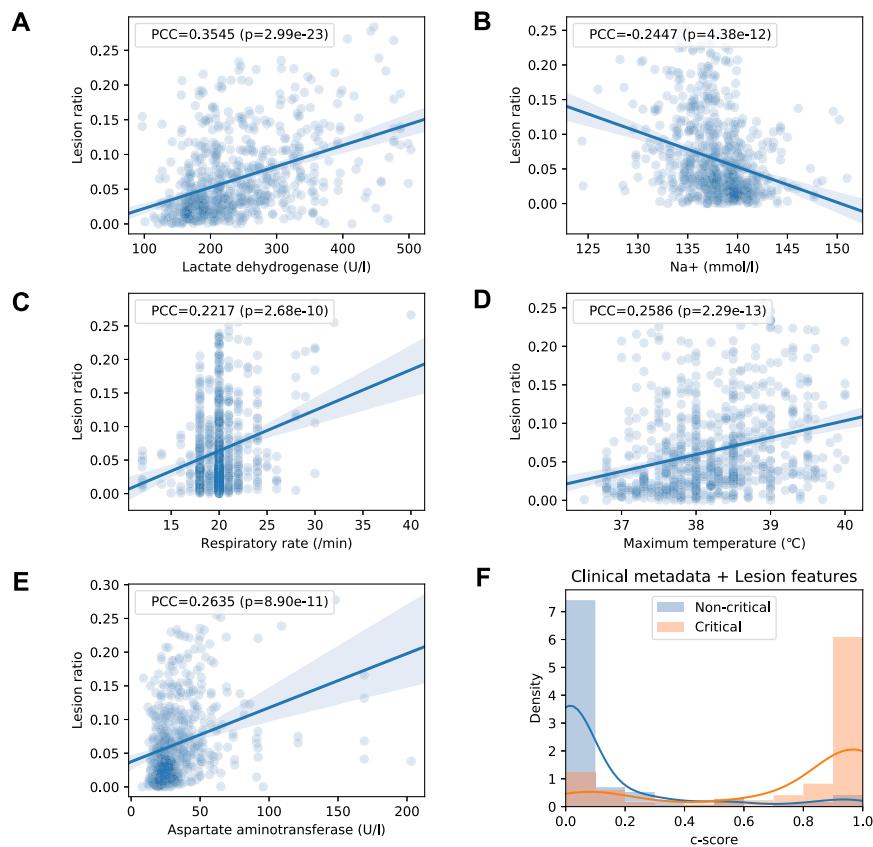
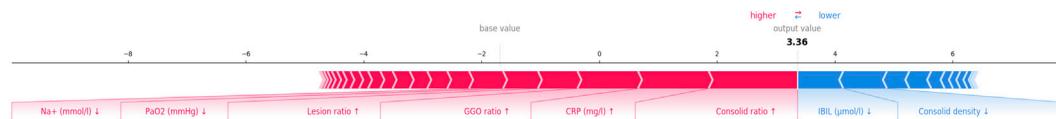


Figure S6. The Correlation of Lung-Lesion Features with Clinical Parameters and Progression of Disease, Related to Figure 6

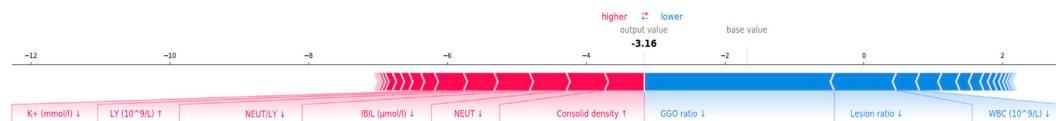
(A-E) Linear regression analysis comparing the volume lesion ratio and five correlated clinical parameters, including (A) serum lactate dehydrogenase (LDH), (B) Na⁺, (C) respiratory rate, (D) maximum body temperature, and (E) serum aspartate aminotransferase (AST). P -values were adjusted with the Holm-Bonferroni method.

(F) A density plot of the c-score for the prognosis prediction model used in STAR Methods.

A



B



C

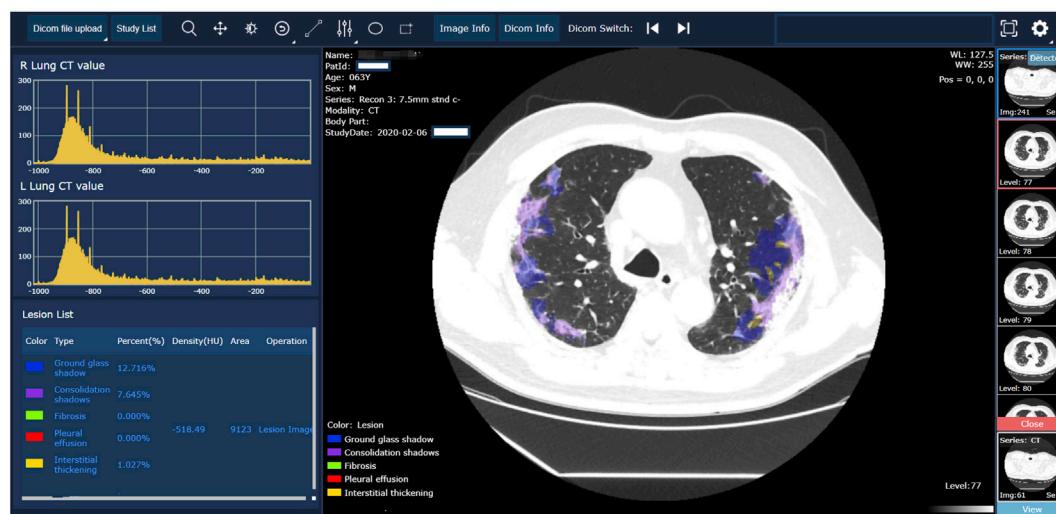


Figure S7. Illustration of Our AI System for Diagnosis and Clinical Prognosis Estimation of COVID-19 Patients during Clinical Deployment, Related to Figure 7

(A and B) Examples of clinical prognosis estimation. We selected two patients from the critical illness and the non-critical illness group to show interpretability of the effects of lung-lesion features and clinical parameters as the input risk factors for prognosis prediction. The effects of input from lung-lesion features and clinical parameters for risk prediction. Pink features pushed the risk higher (to the right) and blue features pushed the risk lower (to the left). (A) A patient from the critical illness group. (B) A patient from the non-critical illness group.

(C) Our system provided lesion segmentation of CT images and quantitative analysis of all the lesion types.