COVID-19 Pneumonia Severity Grading: Test of A Trained Deep Learning Model

Abstract

Purpose: To detect affected lung lobes and conduct severity grading of coronavirus disease 2019 (COVID-19) pneumonia on chest CT images using an artificial intelligence (AI) technique.

Materials and Methods: We used a deep learning model which was previously developed and trained to extract visual features from chest CT exams for the detection and severity grading of COVID-19 pneumonia. In this retrospective study, we tested this model with COVID-19 pneumonia cases in our institution. The numbers of affected lung lobes and severity grading values were compared for the AI method and manual method via the paired Chi-square test. The severity grading capability of the AI method was evaluated using receiver operating characteristic analysis.

Results: A total of 24 cases of confirmed COVID-19 were included (13 men and 11 women). The most frequent CT observation was bilateral ground-glass opacities with consolidation and more than one affected lung lobe. Most cases were mild cases. Compared with the manual method, the AI method presented excellent sensitivity (97.2%) and accuracy (80.8%) but poor specificity (57.1%) in detecting affected lung lobes and good ability (area under the curve=0.795, accuracy=91.6%) in severity grading of COVID-19. Additionally, the time consumed in checking the accuracy of the AI detected lesions within the whole lung was significantly shorter than that of severity assessment by the manual method ($t=9.434$, $p<0.001$).

Conclusion: The AI method with our model is useful in evaluating the severity grading of COVID-19 pneumonia.

Keywords: COVID-19; deep learning; artificial intelligence; computed tomography; severity grading

Introduction

The epidemic of severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) is still gradually spreading worldwide. Recently, the World Health Organization (WHO) declared coronavirus disease 2019 (COVID-19) a pandemic [1]. Nucleic acid testing or genetic sequencing is widely used as the gold standard method for confirmation of SARS-CoV-2 infection, but false-negative results from real-time reverse-transcription polymerase chain reaction (rRT-PCR) have been reported in several recent studies [2,3].

Although most COVID-19 cases are mild, the severe cases often experience dyspnea and hypoxemia one week after onset of the disease. Some patients develop acute respiratory distress syndrome (ARDS), thus admission to intensive care units (ICU) is required [4,5]. By April 8, the mortality of COVID-19 was 4.1% in China [6]. Therefore, identification of severe cases, especially those with higher risk of mortality, is highly important in managing COVID-19 patients.

Non-contrast chest CT plays a vital role in disease evaluation of COVID-19 pneumonia [7]. The typical chest CT features of this disease are bilateral ground glass opacification (with or without consolidation) with peripheral or subpleural distribution

[8-12]. For severe cases with COVID-19, typical findings are consolidation with bilateral multilobular and subsegmental distribution [8,9,13].

Artificial intelligence (AI) using deep learning technology has demonstrated great success in the field of medical imaging due to its prominent capability of feature extraction [14,15]. A recent study [15] showed that a machine learning approach using the convolutional neural networks model was able to distinguish COVID-19 from community-acquired pneumonia (CAP).

In this study, we aimed to detect affected lung lobes and conduct severity grading of COVID-19 on chest CT images using the AI technique.

Materials and Methods

This retrospective research was approved by the institutional review board of our institution, and written informed consent was waived because only de-identified data were analyzed, and no potential risk to patients was involved.

A. Patients and CT scans

From January 12, 2020, to March 14, 2020, 24 patients undergoing chest CT scans in our institution were confirmed with COVID-19. Diagnosis of COVID-19 was determined by positive results from rRT-PCR test of a patient throat swab of SARS-CoV-2 nucleic acid.

All patients underwent chest CT examinations on a UCT 528 scanner (United Imaging, Shanghai, China). The non-enhanced scan was performed in conventional helical mode with the following parameters: tube voltage=120 kVp; tube current=automatic milliamperes; helical pitch=1.375; slice thickness and interval=1.5 mm and 1.5 mm.

B. Manual Imaging Interpretation

Two radiologists (Y. Z. and H. W. W., with 13 and 20 years of experience in chest CT interpretation, respectively) blinded to patient information independently interpreted the images using a viewing console, and final decisions were reached by consensus. No negative cases were included in this study.

For each patient, the CT scans were evaluated based on the following features: (a) ground glass opacities, (b) consolidation, (c) number of lobes observed with ground-glass opacities or consolidation, and (d) overall lung “total severity score”. Other abnormalities (e.g., interlobular septal thickening, pleural effusion, emphysema, and thoracic lymphadenopathy) were also recorded. Ground-glass opacification was defined as hazily increased lung attenuation with visible margins of vessels and bronchi [16]. Consolidation was defined as opacification with blurred margins of vessels and bronchi and could be observed in the mediastinal window setting (window width=400 HU; level=20 HU) [16]. Thoracic lymphadenopathy was defined as lymph node with short-axis dimension larger than 10 mm. The overall lung “total severity score” was applied in reference to a previous study [10] based on the percentage of each lung lobe involvement.

The severity of COVID-19 pneumonia evaluated by manual work was classified according to the overall lung “total severity score”: mild (total severity score of 1-6), moderate (total severity score of 7-12), or severe (total severity score of >12). The time consumed for manual assessment of disease severity was recorded for each observer.

C. Artificial Intelligence: Model information and Precision Testing

JPAI (Hangzhou, China) supplied software and hardware support. This model for pneumonia analysis contains two components of training tasks: lung lobe segmentation and pneumonia lesion segmentation. Using advanced techniques such as imaging, deep learning, and transfer learning, the possible dominant and invisible lesions in pulmonary medical imaging were analyzed and predicted by combining low-level features such as morphology, texture, high-level deep learning features and human anatomy features. High precision of lung tissue segmentation (Fig. 1) is the premise and foundation for quantitative analysis of lung function and is beneficial to diagnosis of lung diseases. A standard U-net network structure was used in quantitative modeling of pneumonia (Fig. 2).

The training dataset of this model included 536 cases of COVID-19 confirmed by rRT-PCR and 721 cases of pneumonia with other types of infection. The chest CT images included scans from devices manufactured by GE Healthcare (Chicago, USA), Siemens (Erlangen, Germany), United Imaging (Shanghai, China), Philips (Amsterdam, Netherlands), and Neusoft (Shenyang, China).

In addition, chest CT scans from 125 cases of COVID-19 confirmed by rRT-PCR and 1000 negative cases were used to test the precision of this model, showing an excellent result with a sensitivity of 97% and specificity of 82% for lung lesion detection.

In this study, the preprocessed CT images were passed to the analysis system, and information on the lung infectious lesion detection and severity evaluation was output immediately. The time consumption for manual checking of the accuracy of the AI detected lesions within the overall lung was recorded.

D. Statistical Analysis

Commercial statistical analysis packages (version 22.0.0 SPSS, IBM) were used to analyze the measurements. The paired Chi-square test was used to statistically compare the two evaluation methods of manual work and AI using the number of affected lung lobes and the severity grading of COVID-19 pneumonia. The paired t-test was used to compare the time consumption of the two methods in evaluating the severity grading using manual work and manual checking of the results of the AI method. The ability of the AI method to conduct severity grading was evaluated using receiver operating characteristic (ROC) curve analysis. A p-value <0.05 was considered statistically significant.

Results

A total of 13 men and 11 women (age range=16-69 years; median age,=42 years) were included in this study (Table 1).

A. Manual Assessment

(1) Opacification Patterns and Distribution

In the 24 chest CT scans, ground-glass opacities (with or without consolidation) were observed in all patients, 10 patients (41.7%) had only ground-glass opacities (with no consolidation), and 1 patient (4.2%) demonstrated consolidation with small areas of ground-glass opacification (Table 2).

Six patients (25.0%) had one affected lobe, 5 patients (20.8%) had two affected lobes, 4 patients (16.7%) had three affected lobes, 2 patients (8.3%) had four affected lobes, and 7 patients (29.2%) had disease affecting all five lobes (Table 3). The most frequently affected lobes were the right lower lobe (19/24, 79.2%) and left lower lobe (18/24, 75.0%). Additionally, 18 patients had bilateral involvement and 6 patients had unilateral involvement.

Seven patients (29.2%) demonstrated interlobular septal thickening, 3 patients (12.5%) had bilateral pleural effusion, and 1 patient had emphysema (4.2%). No patient had thoracic lymphadenopathy (Table 2).

(2) Severity Assessment

The total lung severity score ranged from 1 to a maximum of 18, with a medium score of 3. For severity grading assessment of COVID-19 pneumonia, 19 cases (79.2%) were classified as mild pneumonia, 3 (12.5%) were classified as moderate, and 2 (8.3%) were classified as severe (Fig. 3, Table 4). The confidence of severity assessment between the two observers was high, with a kappa value of 0.871 ($p<0.001$).

The time consumption for manual assessment of disease severity was 52.1 ± 31.3 seconds for observer 1 and 37.0±9.5 seconds for observer 2.

B. Artificial Intelligence Assessment

(1) Opacification Distribution

The confidence of the affected-lung-lobe number assessed by the manual and AI methods was low, with a kappa value of 0.479 ($p<0.001$) (Table 3). Compared with the manual method, 14 cases had the same results for affected-lung-lobe number, 8 cases were over-evaluated by AI and 2 cases displayed missed detection. The sensitivity, specificity and accuracy of AI in detecting the affected lung lobes were 97.2%, 57.1% and 80.8%, respectively.

The reasons for over-evaluation included error in recognition as an infection lesion and error in location of the lung lobe. Recognition errors included fat tissue below diaphragm (1/24, 4.2%), pleural thickening (2/24, 8.3%) (Fig. 4), small vessel (1/24, 4.2%), and fibrotic lesion (1/24, 4.2%). The three cases of lung-lobe-location errors all occurred in the right lung, including mistaking the right upper lobe for the middle lobe (1/24, 4.2%), the right lower lobe for the middle lobe (1/24, 4.2%), and the right middle lobe for the lower lobe (1/24, 4.2%). The two cases of missed detection by AI were related to ground glass opacities (Fig. 5) with mean CT numbers of -770 HU and -804 HU.

(2) Severity Assessment

For the severity assessment of COVID-19 pneumonia, 19 cases (79.2%) were classified as mild pneumonia, 2 cases (8.3%) were classified as moderate, and 2 cases (8.3%) were classified as severe (Table 4). One case classified as mild pneumonia by manual assessment had a negative result when evaluated by AI, although ground glass opacity was detected within the right middle lung lobe by AI (Fig. 6).

The confidence of severity assessment between the manual method and AI was high, with a kappa value of 0.766 ($p<0.001$). If conducting severity grading by manual assessment as the gold standard, the ability of the AI method to evaluate the severity grading of COVID-19 pneumonia was good, with an area under the curve (AUC) of

0.795 and an accuracy of 91.6%. Compared with the manual method, all cases but two showed the same results of severity grading for AI. In addition to the negative case mentioned previously, another case had a total severity score of 7 and was classified as moderate grade by the manual method, whereas it was classified as mild grade by AI. However, all of the mild cases and the two severe cases had the same severity grading assessed by the manual method and AI.

The time consumption for checking the accuracy of AI in detecting lesions within the overall lung was 14.3±5.0 seconds, which is significantly shorter than severity assessment by the manual method ($p<0.001$).

Discussion

Deep learning has demonstrated superior performance in the field of radiology [3,14,15]. Previous studies have successfully applied deep learning techniques to detect pneumonia in chest images [14,17] and to further differentiate viral and other types of pneumonia in chest radiographs or CT imaging [15,18,19].

In this study, we evaluated a deep learning model for detection of COVID-19 pneumonia and identification of the grading of disease severity from chest CT images. The most frequent chest CT observation of COVID-19 was bilateral ground glass opacification with or without consolidation in the lower lung lobes, and most cases were mild cases as assessed by the manual method and AI. These findings were consistent with previous research [4,8-10].

The ability to detect the lung lobes affected by COVID-19 for this model was excellent in terms of sensitivity (97.2%) and accuracy (80.8%) but poor in terms of specificity (57.1%). The poor specificity in this study was attributed to two reasons. First, over-evaluation occurred because of error in recognition as infectious opacity and error in location of lung lobes. Because many different diseases show the same texture features, recognition error is still currently a challenge in the AI field [20]. Second, missed detection occurred because of poor contrast between ground glass opacification with notably low CT attenuation (-770 HU and -804 HU) and normal pulmonary parenchyma. Additionally, one case showed a detected lesion in the right middle lung lobe yet was given a negative result by the AI method. Therefore, with respect to the great risk of missed detection, we considered that this model might be not suitable for standalone use in positive case screening of COVID-19.

A recent study using the deep learning model showed high sensitivity and specificity of 95% and 96%, respectively, in detecting COVID-19 pneumonia [15]. In this study, a large number of 1296 cases of COVID-19 were included. In our study, a small number of 24 cases of COVID-19 were included, which might be the reason for the poor specificity. Because a test study of this model showed an excellent result with sensitivity of 97% and specificity of 82% for lung lesion detection, further research with more cases is needed.

Despite the poor specificity in this study, this model achieved superior performance in severity grading for COVID-19. Compared with the manual method, the AI method had perfect accuracy (91.6%) in severity grading for COVID-19 and was able to export the grading results more quickly. Additionally, the two methods show good

concordance (kappa value of 0.766, p<0.001). Moreover, the two severe cases assessed by the manual method received the same severity grading from the AI method. The existence of over-evaluation did not decrease the accuracy of severity grading of the disease for the AI method. The reason for this result might be the small volume of the mimicking tissues of fat tissue below diaphragm, pleural thickening and small vessels. Therefore, we suggest that the AI method could be used as a valuable warning system for severity grading of COVID-19, but more information is needed to draw robust conclusions.

There are several limitations in this study. One major limitation was the rather small sample size, especially the number of severe cases. Additionally, no control group of other diseases was used in our study. In one respect, our objective was to detect the infected lung lobes and evaluate the severity grading of COVID-19 using the AI method. Further studies with a large study population and other pulmonary diseases for differentiation and severity grading should be conducted in the future. In addition, as a gold standard in this study, severity grading by the manual method could have errors. However, the two radiologists were experienced in chest imaging interpretation, and consensus was reached on the final decision to decrease bias.

In conclusion, the AI method with our model is useful in evaluating severity grading of COVID-19 but is not suitable for screening of positive cases because of the great risk of missed detection.

Reference

1. WHO. WHO Director-General's opening remarks at the media briefing on COVID-19 - 11 March 2020. https://www.who.int/. Published March 11, 2020.

2. Xie X, Zhong Z, Zhao W, Zheng C, Wang F, Liu J. Chest CT for Typical 2019-nCoV Pneumonia: Relationship to Negative RT-PCR Testing. Radiology. 2020. Published Online. DOI: 10.1148/radiol.2020200343.

3. Li D, Wang D, Dong J, et al. False-Negative Results of Real-Time Reverse Transcriptase Polymerase Chain Reaction for Severe Acute Respiratory Syndrome Coronavirus 2: Role of Deep-Learning-Based CT Diagnosis and Insights from Two Cases. Korean J Radiol. 2020;21(4):505-508.

4. NHC. Pneumonia Treatment Program for New Coronary Virus Infection (Trial 7th Edition). National Health Commission of the People's Republic of China Web site. http://www.nhc.gov.cn/. Published March 4, 2020.

5. Wang C, Horby PW, Hayden FG, Gao GF. A novel coronavirus outbreak of global health concern. Lancet. 2020;395(10223):470-473.

6. NHC. Updates on the information of the new coronavirus pneumonia by 24:00 at March 20. http://www.nhc.gov.cn/. Published March 21, 2020.

7. Zu ZY, Jiang MD, Xu PP, et al. Coronavirus Disease 2019 (COVID-19): A Perspective from China. Radiology. Published Online. DOI:10.1148/radiol.2020200490.

8. Shi HS, Yu J, Zheng CS, et al. Radiological Diagnosis of New Coronavirus Infected Pneumonitis: Expert Recommendation from the Chinese Society of Radiology (First edition). Chin J Radiol, 2020; 54(00): E001-E001. DOI:

10.3760/cma.j.issn.1005-1201.2020.0001.
9. Huang C, Wang Y, Li X, et al. Clinical features of patients infected with 2019 novel coronavirus in Wuhan, China. Lancet. 2020; 395(10223):497-506.
10. Chung M, Bernheim A, Mei X, et al. CT Imaging Features of 2019 Novel Coronavirus (2019-nCoV). Radiology. 2020;295(1): 202-207.
11. Song FX, Shi NN, Shan F, et al. Emerging Coronavirus 2019-nCoV Pneumonia. Radiology. 2020;295(1):210-217.
12. Fernando K, Suhny A. The Many Faces of COVID-19: Spectrum of Imaging Manifestations. Radiology: Cardiothoracic Imaging. 2020. Published Online. DOI: 10.1148/ryct.2020200037.
13. Qian L, Yu J, Shi H. Severe acute respiratory disease in a Huanan Seafood Market worker: images of an early casualty. Radiology. 2020. Published Online. DOI: 10.1148/ryct.2020200033.
14. Liu K, Li Q, Ma JC, et al. Evaluating a Fully Automated Pulmonary Nodule Detection Approach and Its Impact on Radiologist Performance. Radiology: Artificial Intelligence. 2019; 1(3):e180084.
15. Li L, Qin L, Xu Z, et al. Artificial Intelligence Distinguishes COVID-19 from Community Acquired Pneumonia on Chest CT. Radiology. 2020:200905.
16. Kim H , Park CM , Woo S , et al. Pure and Part-Solid Pulmonary Ground-Glass Nodules: Measurement Variability of Volume and Mass in Nodules with a Solid Portion Less than or Equal to 5 mm. Radiology. 2013; 269(2):584-592.
17. Shan F, Gao Y, Wang J, et al. Lung infection quantification of COVID-19 in CT images with deep learning. arXiv preprint arXiv:2003.04655, 2020.
18. Kermany DS, Goldbaum M, Cai W, et al. Identifying Medical Diagnoses and Treatable Diseases by Image-Based Deep Learning. Cell. 2018;172(5):1122-1131.e9.
19. Shi F, Xia L, Shan F, et al. Large-Scale Screening of COVID-19 from Community Acquired Pneumonia using Infection Size-Aware Classification. arXiv preprint arXiv:2003.09860,2020.
20. Ma J, Song Y, Tian X, Hua Y, Zhang R, Wu J. Survey on deep learning for pulmonary medical imaging. Front Med. 2019. Published Online. DOI: 10.1007/s11684-019-0726-4.