International Journal of Computer Assisted Radiology and Surgery (2021) 16:435–445
https://doi.org/10.1007/s11548-020-02299-5

ORIGINAL ARTICLE

Association of AI quantified COVID-19 chest CT and patient outcome

Xi Fang[1] · Uwe Kruger[1] · Fatemeh Homayounieh[2] · Hanqing Chao[1] · Jiajin Zhang[1] · Subba R. Digumarthy[2] · Chiara D. Arru[2] · Mannudeep K. Kalra[2] · Pingkun Yan[1]

Received: 20 September 2020 / Accepted: 10 December 2020 / Published online: 23 January 2021

Abstract

Purpose Severity scoring is a key step in managing patients with COVID-19 pneumonia. However, manual quantitative analysis by radiologists is a time-consuming task, while qualitative evaluation may be fast but highly subjective. This study aims to develop artificial intelligence (AI)-based methods to quantify disease severity and predict COVID-19 patient outcome.

Methods We develop an AI-based framework that employs deep neural networks to efficiently segment lung lobes and pulmonary opacities. The volume ratio of pulmonary opacities inside each lung lobe gives the severity scores of the lobes, which are then used to predict ICU admission and mortality with three different machine learning methods. The developed methods were evaluated on datasets from two hospitals (site A: Firoozgar Hospital, Iran, 105 patients; site B: Massachusetts General Hospital, USA, 88 patients).

Results AI-based severity scores are strongly associated with those evaluated by radiologists (Spearman's rank correlation 0.837, $p < 0.001$). Using AI-based scores produced significantly higher ($p < 0.05$) area under the ROC curve (AUC) values. The developed AI method achieved the best performance of AUC = 0.813 (95% CI [0.729, 0.886]) in predicting ICU admission and AUC = 0.741 (95% CI [0.640, 0.837]) in mortality estimation on the two datasets.

Conclusions Accurate severity scores can be obtained using the developed AI methods over chest CT images. The computed severity scores achieved better performance than radiologists in predicting COVID-19 patient outcome by consistently quantifying image features. Such developed techniques of severity assessment may be extended to other lung diseases beyond the current pandemic.

Keywords COVID-19 · Chest CT · Patient outcome · Severity scoring · Artificial intelligence

Introduction

The SARS-CoV-2 (COVID-19) outbreak at the end of 2019, which results from contracting an extremely contagious beta-coronavirus, has spread worldwide and is responsible for the latest pandemic in human history. Prior studies report frequent use of chest computed tomography (CT) in patients suspicious of pneumonia, including COVID-19 [1,4,10,14,16]. Chest CT is often recommended to assess disease severity and monitor progression in patients with moderate to severe pneumonia as well as to assess suspected complications. In sites with limited availability of reverse transcription polymerase chain reaction (RT-PCR) and high disease prevalence, chest CT is also used in diagnosis for patients with suspected COVID-19 [1,4,10,14,16].

Recent clinical studies with chest CT have reported that the qualitative scoring of pulmonary opacities can help assess severe COVID-19 pneumonia. Examples include Yang et al. [16], summing up individual scores of 0–2 from 20 lung regions. Many clinical studies focus on qualitative assessment and grading of pulmonary involvement in each lung lobe to establish disease severity [7,18]. For example, Li et al. [7] quantified the presence of consolidation from chest CT to derive a score between 0 and 5 for each of the five lobes. Summing all the scores together gives a total score in the range of [0, 25] indicating the severity of COVID-19 pneumonia. Their work [7] suggests that high CT severity scores

✉ Mannudeep K. Kalra
mkalra@mgh.harvard.edu

✉ Pingkun Yan
yanp2@rpi.edu

[1] Department of Biomedical Engineering, Center for Biotechnology and Interdisciplinary Studies, Rensselaer Polytechnic Institute, Troy, NY 12180, USA

[2] Department of Radiology, Massachusetts General Hospitals, Harvard Medical School, Boston, MA 02114, USA

(indicating extensive lobar involvement and consolidation) are associated with severe COVID-19 pneumonia.

In these clinical studies, radiologists visually assess the condition of each lobe and estimate the extent of opacities in each lung lobe to assign a severity score [7,15,18]. The scores are then added up to become the overall evaluation. However, such description is inconsistent, subjective, and suffers from intra- and inter-observer variations. To accurately quantify a patient's condition and to relieve clinicians' labor, an automated assessment of the extent of pulmonary parenchymal involvement and distribution of pulmonary opacities is useful since routine clinical interpretation of chest CT does not quantify the disease burden. To automate quantification of disease distribution and extent of pulmonary involvement, automatic segmentation of lung lobes, pulmonary opacities, and distribution of opacities within each lobe are desired. Tang et al. [11] employed threshold segmentation based on the Hounsfield units (HU) range of the ground glass opacity (GGO) to detect the severe disease on chest CT. He et al. [5] proposed a framework for joint lung lobe segmentation and severity assessment of COVID-19 in CT images. We hypothesized that AI-based automatic lung lobe segmentation and distribution of pulmonary opacities can help assess disease severity and outcomes in patients with COVID-19 pneumonia.

This work proposes an AI-assisted severity scoring method based on automated segmentation of the lung lobes and pulmonary opacities. We first analyze the correlation between scores obtained by AI and radiologists. We then show the statistics and data distribution of patients and derive the correlation between scores and patient outcomes. To quantify the correlation, we establish an evaluation scheme that uses three machine learning models for predicting the patient outcomes based on severity scores.

Data and annotation

Data sets

The deidentified data used in our work were acquired at two hospitals, i.e., Site A: Firoozgar Hospital (Tehran, Iran) and Site B: Massachusetts General Hospital (Boston, MA, USA). All the CT imaging data were from patients who underwent clinically indicated, standard-of-care, non-contrast chest CT.

Site A We reviewed medical records of adult patients admitted with known or suspected COVID-19 pneumonia from the site between February 23, 2020, and March 30, 2020. In the 117 patients with positive RT-PCR assay for COVID-19, three patients were excluded due to the presence of extensive motion artifacts on their chest CT. One patient was excluded due to the absence of ICU admission information. The radiologists annotated 105 CT volumes of the left 113 available scans excluding those with significant motion artifacts.

Table 1 Demographic statistics (mean ± SD, except for gender) for Site A dataset

ICU admission	Not admitted	ICU admitted
Gender (M:F)	43:28	27:12
Age (year)	56.7 ± 16.0	67.1 ± 16.7
Lym_r (%)	22.7± 8.3	15.2 ± 12.4
WBC	5831.0 ± 1848.9	8046.2 ± 4650.9
Lym	1244.7 ± 482.8	1013.6 ± 976.6

Table 2 Demographic statistics (mean ± SD, except for gender) for Site B dataset

ICU admission	Not admitted	ICU admitted
Gender (M:F)	18:21	24:25
Age (year)	79.5 ± 10.1	74.4 ± 9.1
Lym_r (%)	17.4 ± 12.8	12.6 ± 11.0
WBC	7587.0 ± 4527.5	10,625.5 ± 5941.8
Lym	999.7 ± 463.3	1274.7 ± 2033.3

Site B We reviewed medical records of adult patients admitted with COVID-19 symptom between March 24 and May 9, 2020. 125 RT-PCR positive admitted patients underwent unenhanced chest CT are selected to form this dataset. The radiologists labeled 88 patients with both ICU admission and mortality risk.

Demographic statistics Tables 1and 2 summarize the demographic data from Sites A and B, respectively. Several data variables (age, gender, white blood cell (WBC) count, lymphocyte count (Lym), lymphocyte-to-WBC ratio (Lym_r)) have significantly different values between patients with and without ICU admission.

Annotation of severity

Two thoracic subspecialty radiologists (one with 16 years of experience and the other with 14 years of experience) reviewed all CT images without the knowledge of clinical features, laboratory data, and patient outcomes. Chest CT images were reviewed on DICOM image viewer (MicroDicom DICOM Viewer, Sofia, Bulgaria) in lung windows (window width 1500 HU, window level-600 HU). The radiologists recorded the type of pulmonary opacities (ground glass opacities, mixed ground glass and consolidative opacities, consolidation, organizing pneumonia, nodular pattern, and ground glass opacities with septal thickening (crazy-paving pattern)). Extent of involvement of each lobe (right upper, right middle, right lower, left upper, and left lower lobes) by the pulmonary opacities was assessed using a previously

described scale (0: 0% lobar involvement; 1: < 5% involvement of lobar volume; 2: 5–25% involvement of lobe; 3: 26–50% lobar involvement; 4: 51–75% lobar involvement; 5: >75% lobar involvement) [10]. The two thoracic subspecialty radiologists reviewed all CT images independently. Any discordance between them was resolved by consensus readout of cases. Total extent of pulmonary opacities was estimated by adding the scores of all lobes (lowest score 0, highest possible score 25) [8,10,13].

AI-assisted severity assessment

Since pulmonary opacity (PO) is an important criterion in terms of the patient severity assessment, we further evaluated our AI-based quantification scores against radiologists' manually graded scores in patients' outcome prediction. In order to obtain explainable severity scores, we divide the AI-assisted procedure into two steps as same as radiologists. In the first step, we use deep learning-based method to automatically segment the lung lobes and pulmonary opacities. Then, severity scores of each patient are computed from the sizes of the lobes and pulmonary opacities. In this way, the severity scores are not dependent on the radiologists' annotation. Furthermore, we use the association of severity scores with patient outcome, ICU admission, and mortality risk of patients with COVID-19 pneumonia, as the criterion to evaluate the severity assessment between AI against radiologists.

Deep learning-based image segmentation

This work employs deep neural networks to segment both lungs, five lung lobes (left upper lobe, left lower lobe, right upper lobe, right middle lobe, right lower lobe) and pulmonary opacity regions of infection from non-contrast chest CT examinations. For network training, we semiautomatically labeled all five pulmonary lobes in 71 CT volumes from Site A using chest imaging platform [17]. A radiologist (M.K.K.) annotated the lung opacities slice by slice in 105 CT volumes from Site A. For lung lobe segmentation, we adopted the automated lung segmentation method proposed by Hofmanninger et al. [6]. Their work provides a trained U-net model for lung segmentation. The U-Net consists of an encoder with regular convolutions and max pooling layers, and a decoder that applies transposed convolutions along with regular convolutions. The network consists of 19 convolutional layers. Each convolutional block in the encoder and decoder uses two 3×3 convolutional layers. Finally, a 1×1 convolutional layer is applied to squeeze the number of feature map channels into 2. To improve the model performance, residual connections between the encoder and decoder were employed. Same as their pre-processing step, the intensity HU range is cropped into the window of $[-1024, 600]$ and then normalized into [0, 1]. The pre-trained model[1] was fine-tuned with a learning rate of 10^{-5} using our annotated data. During tuning, each slice was randomly cropped into patches with size 224×224. The tuned model was then applied to segment all the chest CT volumes.

Segmentation of pulmonary opacities was completed by our previously proposed method, Pyramid Input Pyramid Output Feature Abstraction Network (PIPO-FAN) [3] with publicly released source code.[2] The network integrates image pyramid and multi-scale feature analysis into one single end-to-end framework. It applies spatial pyramid pooling on one 2D slice to generate pyramid input and hierarchically fuses semantically similar features after convolutional blocks. Pyramid features were then adaptively fused via attention module to obtain the lesion map of the slice. In the pre-processing step, all slices were resampled into a fixed resolution of 256×256 pixels, and to improve the contrast of pulmonary opacity, the intensity HU range was cropped into a window of $[-1000, 200]$. In training, the learning rate was set to be 0.002. Softmax activation function with threshold 0.5 was used to obtain the infection area. Morphological operation was applied to refine the segmentation of lung lobes and pulmonary opacities. Using a 3×3 kernel which only connects the nearest neighbors to the center, we first perform the opening operation to remove small noisy segmentation and then apply the closing operation using the same kernel to generate a smooth segmentation.

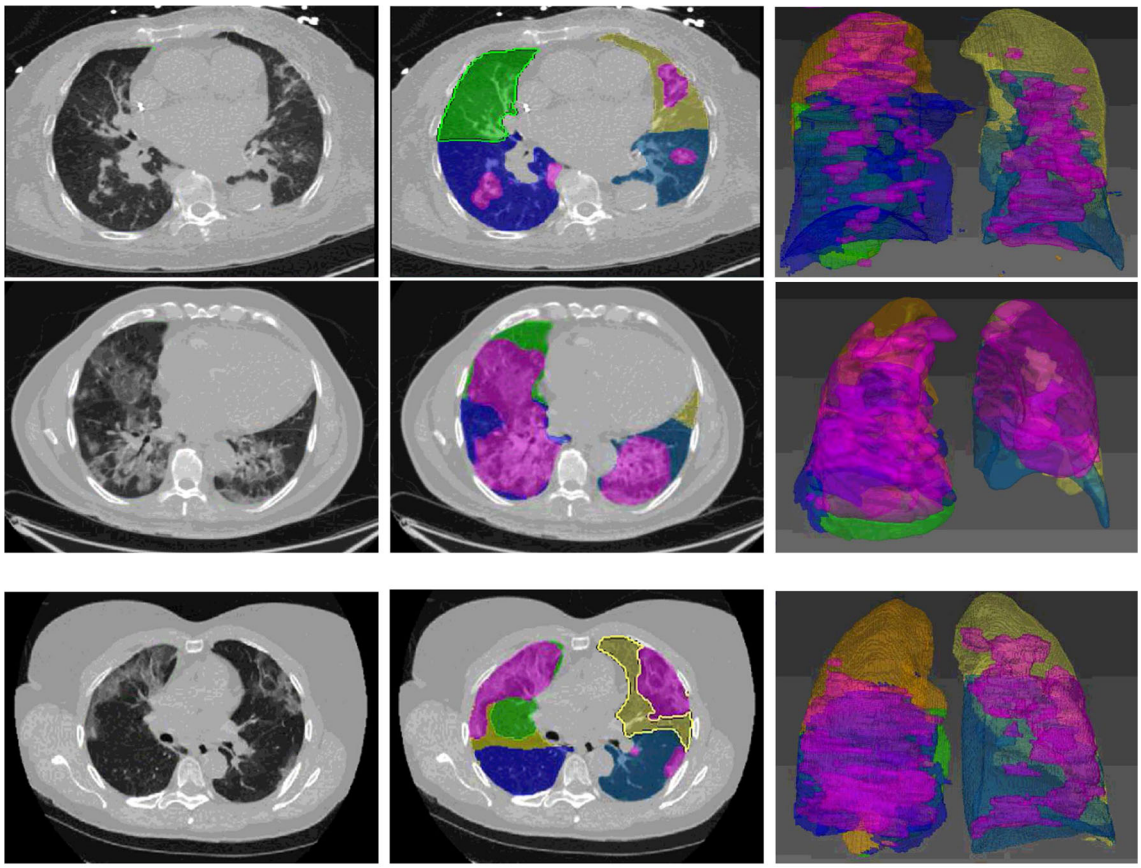
Figure 1 shows the segmentation results of lung lobes and pulmonary opacities. From axial and 3D view, we can see that the segmentation model can smoothly and accurately predict isolated regions with pulmonary opacities. Based on the area of pulmonary opacities, we then computed the ratio of opacity volume over lobes, which is a widely used measurement to describe the severity [11,20]. The dice similarity coefficient (DSC) was used to quantitatively evaluate the segmentation accuracy. For Site A dataset, we obtained DSC scores of 82.5% and 90.0% on lung opacity segmentation and lobe segmentation, respectively.

Severity scoring

Based on the segmentation process in the section "Deep learning-based image segmentation," the AI-based quantification scores are obtained based on the segmentation results with similar procedure, followed by radiologists (see section "Annotation of severity" for more details). First, RPO (ratio of pulmonary opacities) over lobe was calculated for each lobe and graded into 6 levels (0–5). Then, the final score for a patient was the sum of the scores of the 5 lobes which ranges from 0 to 25.

[1] https://github.com/JoHof/lungmask.

[2] https://github.com/DIAL-RPI/PIPO-FAN.



(a) COVID-19 CT image **(b)** Lobes and infection **(c)** 3D Visualization

Fig. 1 Segmentation results of lung lobes and pulmonary opacities. Areas colored in magenta indicate segmented lesions. Other colored areas represent the segmented lobes, with orange: right upper, green: right middle, navy: right lower, yellow: left upper and ocean: left lower

To show relationship between the severity scores and patient outcome, patients were divided into 4 groups (Group I, II, III and IV) based on their final outcome from mild to severe. Groups I and II included recovered patients without admission to ICU. Group I consisted of patients discharged from hospital within 7 days. Group II included patients with more than 7 days of hospital stay. Group III patients had ICU admission and recovered, whereas Group IV patients succumbed to COVID-19 pneumonia. We then divided the severity scores into four buckets. The patients were first divided into two groups using the mean severity score 15 assigned by the radiologists. Each group is then further divided by halving the buckets. We display the statistic of severity groups in different buckets and evaluate the correlation between severity buckets and patient outcome using mean absolute error (MAE).

Severity analysis

This work establishes an evaluation scheme to quantify the correlation between severity scores and patient outcome. The severity scores are used as an input to different machine learning models for predicting patient outcome. The AUC for this task indicates the correlation between scores and patient outcomes. Considering the variations introduced by different prediction targets, models and datasets, to obtain a more objective evaluation, we conducted experiments for both ICU admission and mortality prediction with 3 different models (i.e., support vector machine (SVM), random forest (RF) and logistic regression (LR)) on the two datasets. Radial basis function kernels were used to construct SVM models. Squared L2 norm is used as a regularization term for the SVM and LR models. The RF model has 300 trees, and the Gini index is used as the criterion for calculating information gain. We bootstrap AUCs with 1000 entries to obtain the 95% confidence interval.

In addition to the scores estimated by radiologists and out AI methods, we also include another two groups of scores, i.e., threshold-based scores, and the mixture of radiologists' scores and the AI-based scores. As HU $[-750, -300]$ corresponds to GGO regions [2,11], the method regards voxels within this threshold as pulmonary opacities and further cal-

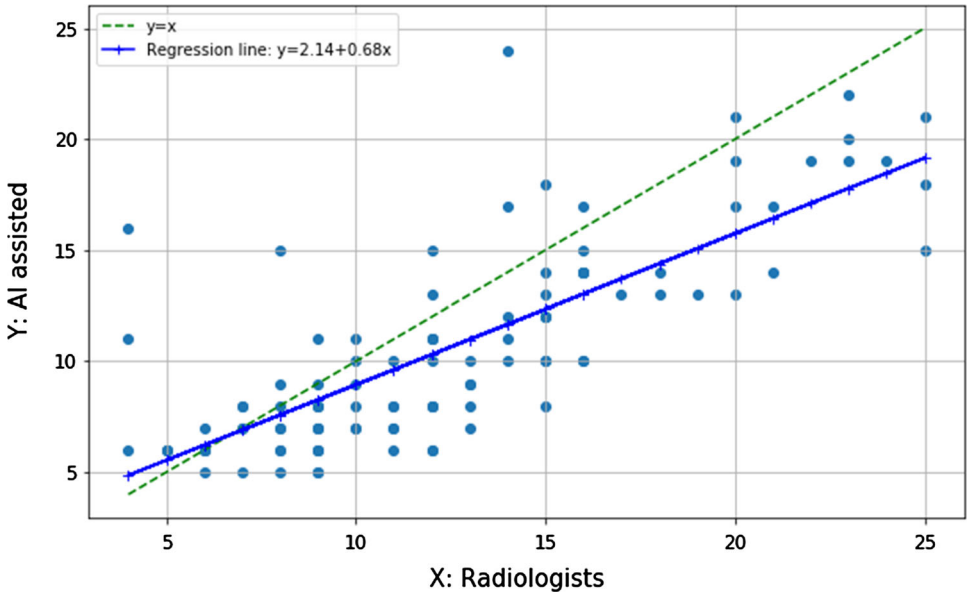


Fig. 2 Correlation between the severity scores assigned by radiologists and computed by our deep learning-based segmentation on Site A dataset

culates the severity scores. To investigate whether the scores of our AI method are complementary with those of the radiologists on patient outcome prediction, we merged their scores (denoted as AI + Radiologists). Specifically, while all other methods use 5 scores of 5 lobes as inputs of the prediction models, AI + Radiologists uses all 10 scores obtained by both the AI methods and the radiologists as inputs. The average results of the three models is used to compute AUC of each scoring method. We conducted one-tailed z-test [19] on ROC curves using scores from radiologists and other severity scoring methods.

Experimental results

This section presents the results of the developed techniques. We show the effectiveness of our proposed segmentation-based severity scoring on the two datasets separately through comparison with different severity scoring methods.

The results are summarized in three parts. In the first part, we computed the correlation coefficient between severity scores assigned by radiologists and computed by our deep learning-based segmentation to demonstrate a good consistency between AI and radiologists. In the second part, we display the statistics of number and proportion of patients in severity groups on different buckets based on section "Severity scoring." The number of patients in each group is overlaid on the corresponding segment. We use the mean absolute error (MAE) between severity score and patient severity group to evaluate the consistency between severity scores and patient outcome.

Finally, to further evaluate the association between the scores and patient outcome, we use the different groups of scores to do ICU admission and mortality prediction, respectively. We computed the AUC for mortality and ICU admission prediction using the severity scores to evaluate the association between scores and patient outcomes. Cross-dataset validation (training on Site A and testing on Site B; training on Site B and testing on Site A) was performed for each model on each group of severity scores to compute the AUC.

Correlation between AI and radiologists

The score for a patient is the sum of the scores of the 5 lobes which ranges from 0 to 25. Similar to the work of Li et al. [9], we evaluated Spearman's rank correlation and associated p-value to determine the strength of the correlation between scores computed by AI-assisted method and assigned by radiologists on Site A and Site B dataset. Figure 2 shows the correlation between the two types of severity scores. We obtained a considerable positive Spearman's rank correlation of 0.770 on Site A dataset ($p < 0.001$). Figure 3 shows the correlation between the two types of severity scores of 88 patients. We obtained a considerable positive Spearman's rank correlation of 0.837 on Site B dataset ($p < 0.001$), indicating that the AI-assisted prediction can obtain consistent results with radiologists.

Severity scores and patient outcome

Figures 4 and 5 display the statistics on number and proportion of 105 patients in Site A dataset falling into the four score buckets based on the computed severity scores, assigned by radiologists and our AI algorithm, respectively. As one can see from the figure, patient severity groups are positively

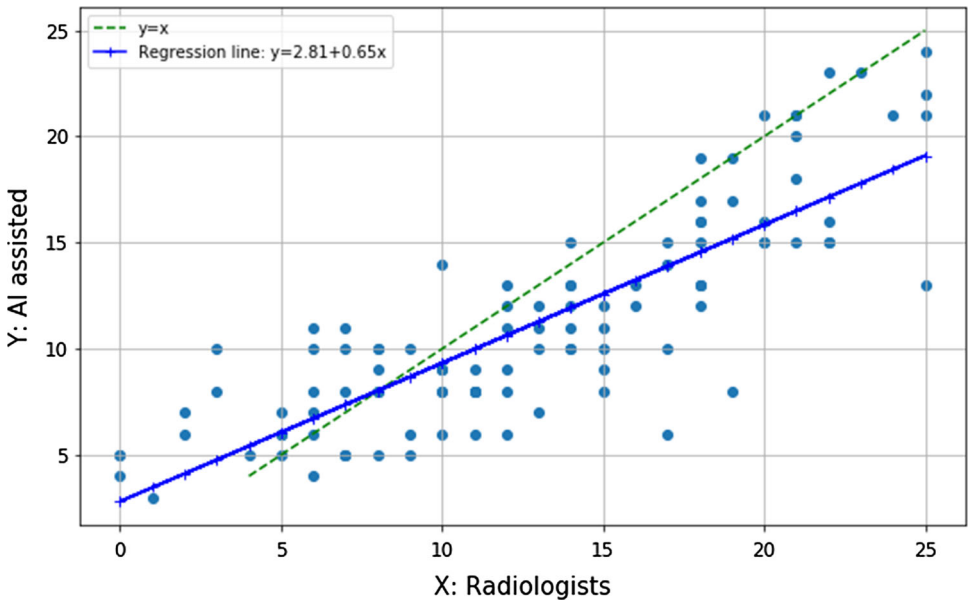


Fig. 3 Correlation between the severity scores assigned by radiologists and computed by our deep learning-based segmentation on Site B dataset

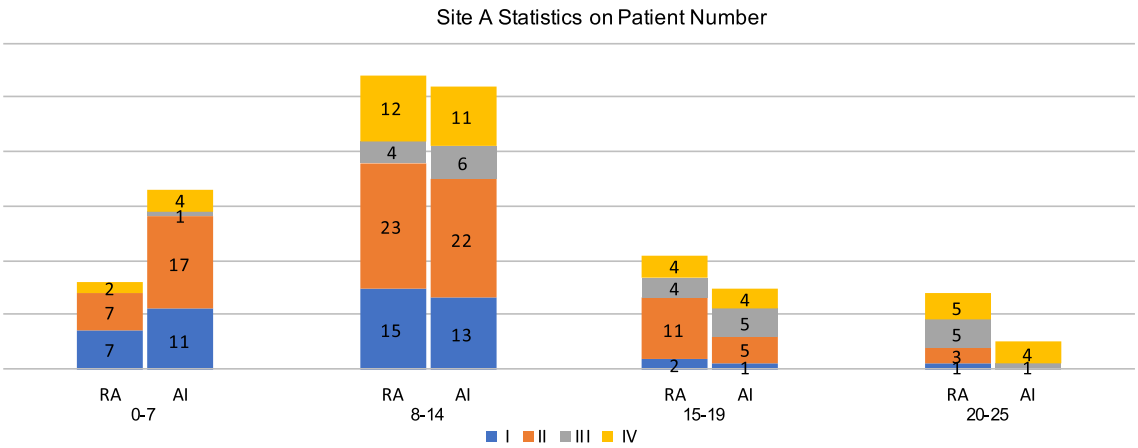


Fig. 4 For different buckets divided by severity scores obtained by AI and radiologists, values of number of patients from different patient groups (I, II, III, IV) are presented on Site A dataset. RA: Results of radiologists' annotation. AI: Results of the AI-based method. Heights of the bars represent the number

correlated with the severity scores buckets. In Fig. 5, we can see that for our AI method, no patients in Group III and Group IV are in the bucket [20,25]. For quantitative evaluation, we then computed the MAE between severity score buckets and patient severity groups. Each bucket of severity score is paired with one severity group. The MAE for radiologists and AI quantification computed using the number is 88 and 84, respectively. The MAE computed using the proportion is 3.51 and 2.65. Our AI method obtained comparable MAE when compared to the radiologists, i.e., the difference in the MAE scores is not statistically different.

Figures 6 and 7 display the statistics on number and proportion of 88 patients in Site B dataset falling into different categories based on the computed severity scores, given by radiologists and our AI algorithm, respectively. The MAE for radiologists and AI quantification computed using the number is 89 and 89, respectively. The MAE computed by the proportion is 3.88 and 3.38 for the radiologists and the AI, respectively. That means the AI-assisted method achieved comparable consistency with patient severity groups than radiologists. In Fig. 7, we can see that for our AI method, the proportion of the severe patients (Group III and Group IV) monotonically increases with the raising of the severity score, while the result of the radiologists has a drawback on the second score bucket.

Patient outcome prediction

Results on Site A

Table 3 summarizes the AUCs and 95% confidence intervals of three machine learning models on ICU admission and mor-

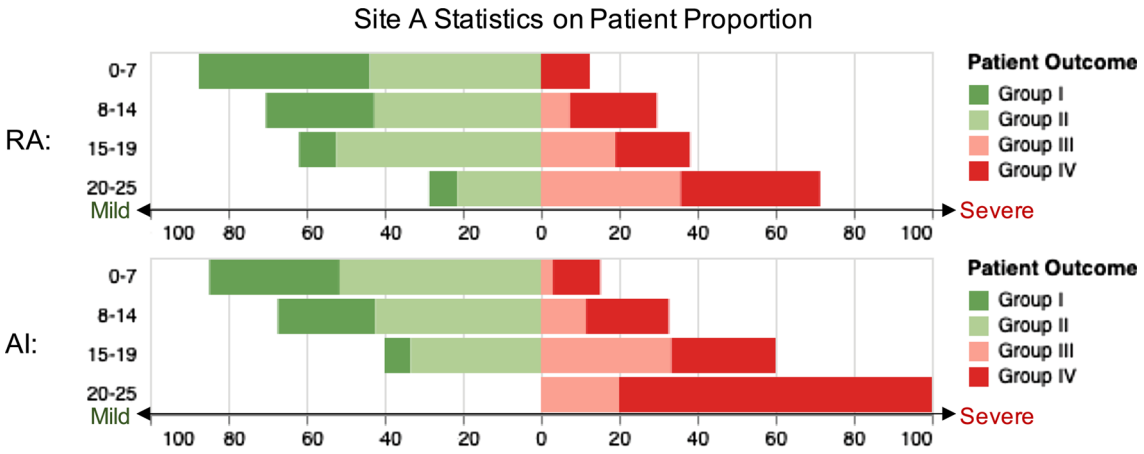


Fig. 5 For each severity score bucket shown in vertical, horizontal stacked bars present the proportion of patient number of one severity group to patient number of all severity groups of Site A in the bucket. Widths of the bars represent such proportion (mild patient: green for group I and light green for group II; severe patient: light red for group III and red for group IV)

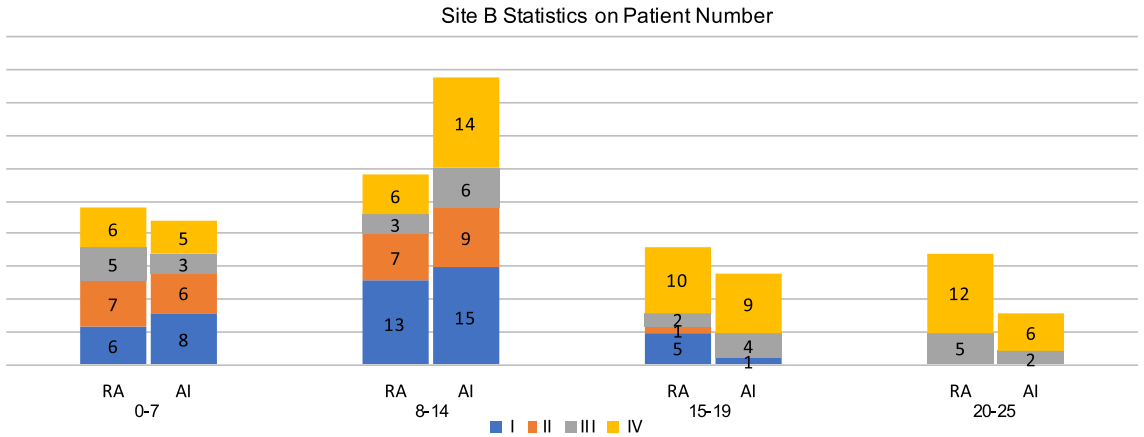


Fig. 6 For different buckets divided by severity scores obtained by AI and radiologists, values of number of patients from different patient groups (I, II, III, IV) are presented on Site B dataset. RA: Results of radiologists' annotation. AI: Results of the AI-based method. Heights of the bars represent the number

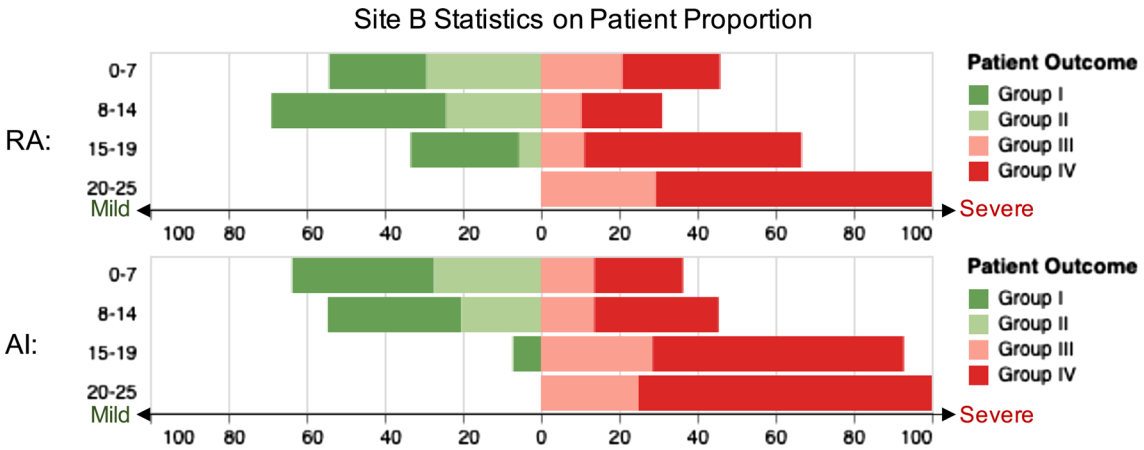


Fig. 7 For each severity score bucket shown in vertical, horizontal stacked bars present the proportion of patient number of one severity group to patient number of all severity groups of Site B in the bucket. Widths of the bars represent such proportion (mild patient: green for group I and light green for group II; severe patient: light red for group III and red for group IV)

Table 3 Comparison of AI with radiologists in predicting ICU admission and mortality prediction on Site A dataset

Outcome	Severity scoring	SVM		RF		LR	
		Mean	95% CI	Mean	95% CI	Mean	95% CI
ICU admission	Radiologists	0.689	(0.594, 0.778)	0.663	(0.562, 0.759)	0.642	(0.535, 0.739)
	Threshold [11]	**0.753**	(0.666, 0.830)	0.706	(0.607,0.796)	**0.787**	(0.713, 0.854)
	AI segmentation (AI)	0.733	(0.641, 0.816)	**0.778**	(0.698, 0.848)	0.723	(0.623, 0.810)
	AI + radiologists	0.726	(0.639, 0.804)	0.759	(0.638, 0.836)	0.681	(0.575, 0.778)
Mortality	Radiologists	0.639	(0.527, 0.743)	0.644	(0.531, 0.738)	0.566	(0.441, 0.674)
	Threshold [11]	0.605	(0.479, 0.718)	0.593	(0.474, 0.702)	0.614	(0.480, 0.741)
	AI segmentation (AI)	**0.720**	(0.626, 0.812)	**0.723**	(0.623, 0.820)	**0.670**	(0.545, 0.777)
	AI + radiologists	0.693	(0.589, 0.788)	0.719	(0.618, 0.807)	0.638	(0.505, 0.758)

tality. The three machine learning models were trained on Site B and tested on Site A. The bold values represent the best AUCs. We can see that AI achieved higher AUC than radiologists under all models and tasks. We further use the mean of the scores from the three models as a simple ensemble strategy to compute the AUC value for each severity scoring method. ROC curves on ICU admission and mortality prediction are shown in Fig. 8a, b. AI obtains best AUC 0.755 and 0.723 on ICU admission and mortality, respectively. One-tailed z-test is used to evaluate the statistical significance between radiologists and other scoring methods. Threshold-based method outperforms radiologists in ICU admission with $p = 0.031$ although performance in mortality prediction was not significantly different ($p = 0.426$). AI significantly outperforms radiologists with $p = 0.044 < 0.05$ in ICU admission and $p = 0.031 < 0.05$ on mortality.

Results on site B dataset

The same set of experiments were repeated on the Site B dataset. The three machine learning models are trained on Site A and tested in Site B. Table 4 summarizes the AUCs of different models with 95% confidence interval indicated by ICU admission and mortality on Site B dataset. The bold values represent the best AUCs. We can see that AI achieved best AUC under all models and tasks. ROC curves on ICU admission and mortality prediction are shown in Fig. 9a, b. AI obtains best AUC 0.813 and 0.741 on ICU admission and mortality, respectively. One-tailed z-test is used to evaluate the statistical significance between radiologists and other scoring methods. Radiologists outperform threshold-based method in ICU admission with $p = 0.016$; both methods had similar performance for mortality prediction ($p = 0.060$). AI significantly outperforms radiologists with $p = 0.022 < 0.05$ in ICU admission and $p = 0.045 < 0.05$ on mortality.

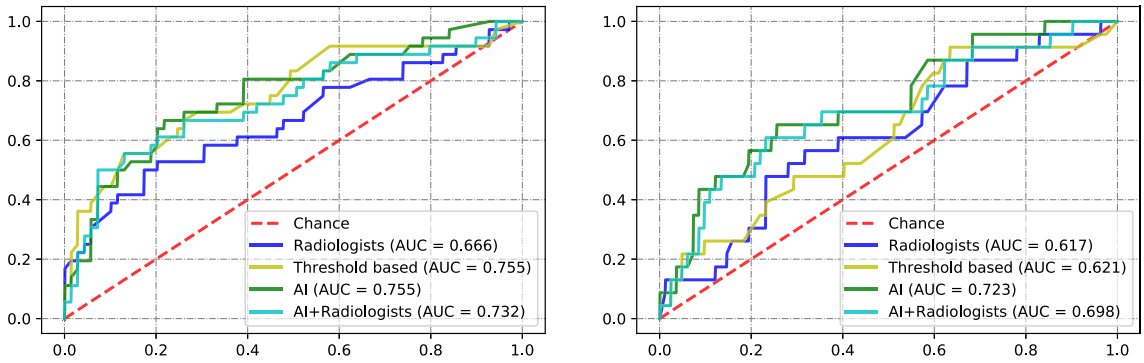
Discussion

In the current clinical radiology practice, radiologists from hospitals do not perform a quantitative or semiquantitative assessment of disease severity or distribution of pulmonary opacities. This lack of quantification is related to the fact that they are not trained and required to assign severity scores in patients with pneumonia. While in patients with cancer and focal lesions, radiologists measure and compare single or volumetric dimension of focal lesions, in patients with diffuse and ill-defined disease patterns found in pneumonia such measurements are not feasible and practical. As a result, radiology reports in patients with COVID-19 pneumonia are limited to semantic description of extent (such as diffuse, multifocal, or localized) and type of opacities rather than an assigned severity score. However, prior studies with subjective severity scores from both chest radiography and CT report on their ability to predict disease severity and patient outcome [12,16]. Threshold segmentation can detect coarse opacities with some degree, but it is known to be less accurate than AI-based methods.

Results analysis

The study confirms that AI-based severity scoring method yields AUCs in ICU admission and mortality prediction, on Site A and B datasets, that exceed those of radiologists. Considering the large differences in patient demographic statistics and clinical protocols at the two participating sites, the results highlight the robustness of the AI-based method and their generalization ability of the extracted scores.

The statistical test of differences between ROC curves using scores from AI and radiologists suggests that AI significantly outperforms radiologists (A $\rightarrow$ B: $p = 0.022$ on ICU admission, $p = 0.045$ on mortality; B $\rightarrow$ A: $p = 0.044$ on ICU admission, $p = 0.031$ on mortality). The results outline that the correlation between AI-assisted severity score and prognosis of patients has stronger correlation with patient

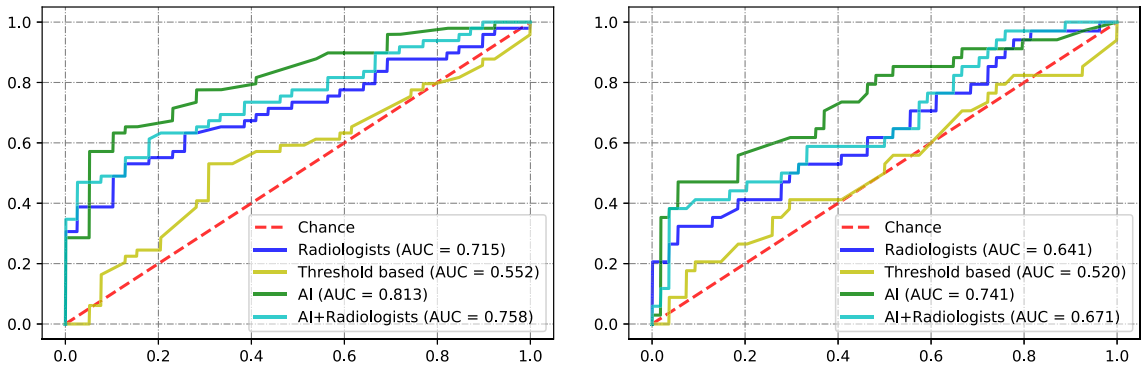


(a) ICU admission on Site A **(b)** Mortality on Site A

Fig. 8 Comparison of severity scoring methods under ROC curves on Site A

Table 4 Comparison of AI with radiologists in predicting ICU admission and mortality prediction on Site B dataset

Outcome	Severity scoring	SVM		RF		LR	
		Mean	95% CI	Mean	95% CI	Mean	95% CI
ICU admission	Radiologists	0.680	(0.581, 0.773)	0.695	(0.600, 0.782)	0.705	(0.608, 0.794)
	Threshold [11]	0.719	(0.624, 0.808)	0.703	(0.599, 0.796)	0.701	(0.602, 0.795)
	AI segmentation (AI)	**0.766**	(0.675, 0.851)	**0.757**	(0.665, 0.840)	**0.766**	(0.675, 0.851)
	AI + radiologists	0.725	(0.634, 0.811)	0.741	(0.657, 0.821)	0.725	(0.634, 0.811)
Mortality	Radiologists	0.566	(0.447, 0.679)	0.616	(0.504, 0.727)	0.670	(0.602, 0.795)
	Threshold [11]	0.584	(0.481, 0.685)	0.579	(0.470,0.679)	0.605	(0.505, 0.701)
	AI segmentation (AI)	**0.655**	(0.545, 0.766)	**0.676**	(0.578, 0.779)	**0.736**	(0.642, 0.828)
	AI + radiologists	0.603	(0.496, 0.719)	0.631	(0.530, 0.735)	0.716	(0.623, 0.810)



(a) ICU admission on Site B **(b)** Mortality on Site B

Fig. 9 Comparison of severity scoring methods under ROC curves on Site B

outcomes than radiologists, which could effectively improve the accuracy of severity scoring system.

We explored the potential of combining the scores of AI and radiologists to further improve the prognosis of COVID-19 pneumonia. Both score concatenation and sum were tested. However, compared with AI-based method, the combination of scores of AI and radiologists did not present a significant improvement. We also used investigated other machine learning models, decision tree (DT) and multilayer perceptron (MLP), to predict patient outcomes. The two models did not perform as well as the three models reported in this paper. Besides the choice of classification model, the segmentation quality of pulmonary lobes and opacities may be an additional factor that limits the performance of outcome prediction. More accurate image segmentation will give a more precise lung opacity quantification, which, in turn, may help to improve the performance of our AI method further. Also, ancillary imaging findings not

assessed with our segmentation tool such as coronary artery disease, cardiac enlargement, pleural effusions, co-existing bronchial wall thickening or emphysema, and mediastinal lymphadenopathy, may have affected the disease outcome, and thus decreased the performance of our models.

The AUC obtained for mortality prediction is lower than in ICU admission prediction. Since mortality is influenced by several factors including comorbidities, local treatment condition and disease exacerbation, mortality prediction is more challenging than ICU admission. However, the AI method can still persistently improve the AUC over radiologists for this task by 10%, which further demonstrate the effectiveness of scores extracted from AI.

Future directions

Although scores from radiologists may not be as accurate as AI, they may convey overall condition of severe patients. For example, although not included in our study, radiologists assessed findings beyond pulmonary opacities such as architectural distortion in lungs, pleural effusions, mediastinal/hilar lymphadenopathy, cardiac enlargement, and increased subcutaneous fat stranding or attenuation suggestive of anasarca and underlying fluid overload. Li et al. [9] showed that such findings can be learned by AI algorithms and help longitudinal disease evaluation on COVID-19 pneumonia.

For future work, we will continue to explore developing AI algorithms, which may efficiently incorporate the knowledge of radiologists with an overall evaluation of patient condition. It is also worth noting that the developed techniques of severity assessment may extend to other lung diseases beyond the current pandemic. As the world awaits the introduction of disease prevention and specific treatment for those infected by COVID-19 pneumonia, it is important that future versions of AI-based time-to-event measures assess the extent of chronic changes from the substantial patient population who had initial recovery but might have long lasting sequelae from their infection.

Conclusions

This paper has proposed an AI-assisted severity scoring method based on automatic segmentation of lung lobes and pulmonary opacities. The severity scores obtained by AI have shown to be consistent with those obtained by radiologists in the two datasets. We have further quantitatively evaluated the scores based on patient outcomes and demonstrated that the AI segmentation-based method was significantly more accurate than current severity scoring by radiologists only. The results suggest that AI can significantly improve the patient outcome prediction for patients with severe COVID-19 pneumonia. We believe such techniques have the potential to numerous clinical applications involving COVID-19 pneumonia.

Acknowledgements This work was partially supported by National Institute of Biomedical Imaging and Bioengineering (NIBIB) under award R21EB028001 and National Heart, Lung, and Blood Institute (NHLBI) under Award R56HL145172.

Compliance with ethical standards

Conflict of interest The authors declare that they have no conflict of interest.

Ethical approval The study received IRB approval at both participating sites. Need for informed consent was waived due to the retrospective nature of the study.

Informed consent There is no informed consent required for the work reported in this manuscript.

References

1. Ai T, Yang Z, Hou H, Zhan C, Chen C, Lv W, Tao Q, Sun Z, Xia L (2020) Correlation of chest CT and RT-PCR testing in coronavirus disease 2019 (COVID-19) in China: a report of 1014 cases. Radiology. https://doi.org/10.1148/radiol.2020200642
2. Chao H, Fang X, Zhang J, Homayounieh F, Arru CD, Digumarthy SR, Babaei R, Mobin HK, Mohseni I, Saba L, Carriero A, Falaschi Z, Pasche A, Wang G, Kalra MK, Yan P (2021) Integrative analysis for COVID-19 patient outcome prediction. Med Image Anal 67:101844. https://doi.org/10.1016/j.media.2020.101844
3. Fang X, Yan P (2020) Multi-organ segmentation over partially labeled datasets with multi-scale feature abstraction. IEEE Trans Med Imaging 39(11):3619–3629. https://doi.org/10.1109/TMI.2020.3001036
4. Harmon SA, Sanford TH, Xu S, Turkbey EB, Roth H, Xu Z, Yang D, Myronenko A, Anderson V, Amalou A, Blain M, Kassin M, Long D, Varble N, Walker SM, Bagci U, Ierardi AM, Stellato E, Plensich GG, Franceschelli G, Girlando C, Irmici G, Labella D, Hammoud D, Malayeri A, Jones E, Summers RM, Choyke PL, Xu D, Flores M, Tamura K, Obinata H, Mori H, Patella F, Cariati M, Carrafiello G, An P, Wood BJ, Turkbey B (2020) Artificial intelligence for the detection of COVID-19 pneumonia on Chest CT using multinational datasets. Nat Commun 11(1):4080. https://doi.org/10.1038/s41467-020-17971-2
5. He K, Zhao W, Xie X, Ji W, Liu M, Tang Z, Shi F, Gao Y, Liu J, Zhang J, Shen D (2020) Synergistic learning of lung lobe segmentation and hierarchical multi-instance classification for automated severity assessment of COVID-19 in CT images. arXiv preprint arXiv:2005.03832
6. Hofmanninger J, Prayer F, Pan J, Rohrich S, Prosch H, Langs G (2020) Automatic lung segmentation in routine imaging is primarily a data diversity problem, not a methodology problem. Eur Radiol Exp. https://doi.org/10.1186/s41747-020-00173-2
7. Li K, Wu J, Wu F, Guo D, Chen L, Fang Z, Li C (2020a) The clinical and chest CT features associated with severe and critical COVID-19 pneumonia. Investig Radiol 55(6):327–331. https://doi.org/10.1097/RLI.0000000000000672
8. Li L, Qin L, Xu Z, Yin Y, Wang X, Kong B, Bai J, Lu Y, Fang Z, Song Q, Cao K, Liu D, Wang G, Xu Q, Fang X, Zhang S, Xia J, Xia J (2020b) Artificial intelligence distinguishes COVID-19 from

community acquired pneumonia on chest CT. Radiology. https://doi.org/10.1148/radiol.2020200905
9. Li MD, Arun NT, Gidwani M, Chang K, Deng F, Little BP, Mendoza DP, Lang M, Lee SI, O'Shea A, Parakh A, Singh P, Kalpathy-Cramer J (2020c) Automated assessment and tracking of covid-19 pulmonary disease severity on chest radiographs using convolutional siamese neural networks. Radiol Artif Intell 2(4):e200079. https://doi.org/10.1148/ryai.2020200079
10. Shan F, Gao Y, Wang J, Shi W, Shi N, Han M, Xue Z, Shi Y (2020) Lung infection quantification of COVID-19 in CT images with deep learning. arXiv preprint arXiv:2003.04655
11. Tang Z, Zhao W, Xie X, Zhong Z, Shi F, Liu J, Shen D (2020) Severity assessment of coronavirus disease 2019 COVID-19 using quantitative features from Chest CT images. Phys Med Biol. https://doi.org/10.1088/1361-6560/abbf9e
12. Toussie D, Voutsinas N, Finkelstein M, Cedillo MA, Manna S, Maron SZ, Jacobi A, Chung M, Bernheim A, Eber C, Concepcion J, Fayad Z, Gupta YS (2020) Clinical and chest radiography features determine patient outcomes in young and middle age adults with COVID-19. Radiology 297:201754
13. Wong HYF, Lam HYS, Fong AHT, Leung ST, Chin TWY, Lo CSY, Lui MMS, Lee JCY, Chiu KWH, Chung T, Lee EYP, Wan EYF, Hung FNI, Lam TPW, Kuo M, Ng MY (2020) Frequency and distribution of chest radiographic findings in COVID-19 positive patients. Radiology. https://doi.org/10.1148/radiol.2020201160
14. Xie X, Zhong Z, Zhao W, Zheng C, Wang F, Liu J (2020) Chest CT for typical 2019-nCoV pneumonia: relationship to negative RT-PCR testing. Radiology 296:200343
15. Xiong Y, Sun D, Liu Y, Fan Y, Zhao L, Li X, Zhu W (2020) Clinical and high-resolution CT features of the COVID-19 infection: comparison of the initial and follow-up changes. Investig Radiol 55(6):332–339. https://doi.org/10.1097/RLI.0000000000000674
16. Yang R, Li X, Liu H, Zhen Y, Zhang X, Xiong Q, Luo Y, Gao C, Zeng W (2020) Chest CT severity score: an imaging tool for assessing severe COVID-19. Radiol Cardiothorac Imaging 2(2):e200047
17. Yip SS, Parmar C, Blezek D, Estepar RSJ, Pieper S, Kim J, Aerts HJ (2017) Application of the 3d slicer chest imaging platform segmentation algorithm for large lung nodule delineation. PLoS ONE 12(6):e0178944
18. Zhao W, Zhong Z, Xie X, Yu Q, Liu J (2020) Relation between chest CT findings and clinical conditions of coronavirus disease (COVID-19) pneumonia: a multicenter study. Am J Roentgenol 214(5):1072–1077. https://doi.org/10.2214/AJR.20.22976
19. Zhou XH, McClish DK, Obuchowski NA (2009) Statistical methods in diagnostic medicine, vol 569. Wiley, New York
20. Zhu X, Song B, Shi F, Chen Y, Hu R, Gan J, Zhang W, Li M, Wang L, Gao Y, Shan F, Shen D (2021) Joint prediction and time estimation of COVID-19 developing severe symptoms using chest CT scan. Med Image Anal 67:101824. https://doi.org/10.1016/j.media.2020.101824