



Generative adversarial network for automatic quantification of Coronavirus disease 2019 pneumonia on chest radiographs

Seung-Jin Yoo^a, Hyungjin Kim^b, Joseph Nathanael Witanto^c, Shohei Inui^{d,e}, Jeong-Hwa Yoon^f, Ki-Deok Lee^g, Yo Won Choi^a, Jin Mo Goo^{b,h}, Soon Ho Yoon^{b,c}

^a Department of Radiology, Hanyang University Medical Center, Hanyang University College of Medicine, Seoul, Republic of Korea

^b Department of Radiology, Seoul National University Hospital, Seoul National College of Medicine, Seoul, Korea

^c MEDICALIP Co. Ltd., Seoul, Korea

^d Department of Radiology, Graduate School of Medicine, The University of Tokyo, Tokyo, Japan

^e Department of Radiology, Japan Self-Defense Forces Central Hospital, Tokyo, Japan

^f Institute of Health Policy and Management, Medical Research Center, Seoul National University, Seoul, South Korea

^g Division of Infectious diseases, Department of Internal Medicine, Myongji Hospital, Goyang, Korea

^h Institute of Radiation Medicine, Seoul National University Medical Research Center, Seoul, Republic of Korea

ARTICLE INFO

Keywords:

COVID-19

Pneumonia

Chest X-Ray

Automatic quantification

GAN

ABSTRACT

Purpose: To develop a generative adversarial network (GAN) to quantify COVID-19 pneumonia on chest radiographs automatically.

Materials and Methods: This retrospective study included 50,000 consecutive non-COVID-19 chest CT scans in 2015–2017 for training. Anteroposterior virtual chest, lung, and pneumonia radiographs were generated from whole, segmented lung, and pneumonia pixels from each CT scan. Two GANs were sequentially trained to generate lung images from radiographs and to generate pneumonia images from lung images. GAN-driven pneumonia extent (pneumonia area/lung area) was expressed from 0% to 100%. We examined the correlation of GAN-driven pneumonia extent with semi-quantitative Brixia X-ray severity score (one dataset, $n = 4707$) and quantitative CT-driven pneumonia extent (four datasets, $n = 54$ –375), along with analyzing a measurement difference between the GAN and CT extents. Three datasets ($n = 243$ –1481), where unfavorable outcomes (respiratory failure, intensive care unit admission, and death) occurred in 10%, 38%, and 78%, respectively, were used to examine the predictive power of GAN-driven pneumonia extent.

Results: GAN-driven radiographic pneumonia was correlated with the severity score (0.611) and CT-driven extent (0.640). 95% limits of agreements between GAN and CT-driven extents were –27.1% to 17.4%. GAN-driven pneumonia extent provided odds ratios of 1.05–1.18 per percent for unfavorable outcomes in the three datasets, with areas under the receiver operating characteristic curve (AUCs) of 0.614–0.842. When combined with demographic information only and with both demographic and laboratory information, the prediction models yielded AUCs of 0.643–0.841 and 0.688–0.877, respectively.

Conclusion: The generative adversarial network automatically quantified COVID-19 pneumonia on chest radiographs and identified patients with unfavorable outcomes.

1. Introduction

Coronavirus disease 2019 (COVID-19) emerged as a significant threat to global health in 2020. Detection of viral RNA in nasopharyngeal or other upper respiratory tract specimens by reverse transcription-polymerase chain reaction (RT-PCR) is the gold standard for diagnosing COVID-19 [1]. Nonetheless, chest CT has been an important diagnostic

test for COVID-19 from the early stages of the pandemic due to the limited availability and time-consuming process of diagnostic RT-PCR test kits. Several studies reported a high sensitivity of chest CT for COVID-19 pneumonia [2–4]. Chest radiographs have lower sensitivity (59–89%) for COVID-19 than chest CT [5–7], but are more convenient, require more easily sterilizable equipment, are more affordable, and involve less radiation exposure than CT scans.

Abbreviations: CT, computed tomography; COVID-19, coronavirus disease 2019; RT-PCR, reverse transcription-polymerase chain reaction; GAN, generative adversarial network; AP, anteroposterior; LDH, lactate dehydrogenase; LOA, limits of agreement.

<https://doi.org/10.1016/j.ejrad.2023.110858>

Received 29 November 2022; Received in revised form 10 April 2023; Accepted 29 April 2023

Available online 12 May 2023

0720-048X/© 2023 Elsevier B.V. All rights reserved.

Currently, radiologic examinations are recommended to stratify patients with COVID-19 at risk for unfavorable outcomes [8]. A larger radiologic extent of pneumonia predicts worse outcomes in COVID-19 [4,9–12]. Notably, the radiologic extent was fully quantifiable on CT images using deep neural networks, highlighting the benefit of radiologic quantification in COVID-19. Deep neural networks were also applied to chest radiographs in COVID-19, but most research has focused on the automatic detection of COVID-19 [13–15], whereas only a small proportion of work has investigated semi-quantitative assessments [16–22]. If pneumonia extent is fully quantifiable on chest radiographs, similar to CT images, the role of chest radiographs in managing COVID-19 would be maximized because it would be possible to identify and monitor at-risk patients regardless of image readers' experience or variability in image interpretation across readers, particularly in resource-limited settings [8].

This study aimed to develop a generative adversarial network (GAN) to quantify COVID-19 pneumonia on chest radiographs automatically.

2. Materials and Methods

The Institutional Review Board of participating hospitals approved this retrospective study and waived the requirement for patient consent.

2.1. Study population and CT acquisition

We retrospectively collected 50,000 anonymized low-dose chest CT scans of non-COVID-19 patients at a single institution from January 2015 to June 2017 (mean age $\pm$ standard deviation, 59.3 ± 15.2 years; proportion of women, 52%). The CT scans were consecutively obtained

throughout the period from adult patients without any particular eligible criteria to encompass a full spectrum of chest radiologic manifestations in various thoracic diseases from normal to infectious or oncologic diseases. All chest CT scans were performed using one of the 10 following multi-detector CT scanners; Sensation 16, Somatom Definition, Somatom Definition Flash and Somatom Force (Siemens Healthineers, Erlangen, Germany); IQon spectral CT, Brilliance 64, Brilliance iCT 256, and Ingenuity (Philips Healthcare, Amsterdam, Netherlands); Aquilion One (Canon Medical Systems, Otawara, Japan); Discovery CT 750 HD (GE Healthcare, Milwaukee, WI, USA). All CT examinations were performed preferentially with a tube voltage of 120 kVp with automatic exposure control and a volume CT dose index less than 3.3 mGy. Axial images were reconstructed with a sharp reconstruction kernel at a slice thickness of 1.5 mm or thinner.

2.2. Preparation of training data

The training data included a total of 50,000 pairs of anteroposterior (AP) chest, lung parenchymal, and pneumonia projection images from chest CT images (Fig. 1). The AP projection images using native CT images simulated chest radiographs, and the AP projection images of segmented lung parenchyma and pulmonary lesions of CT images simulated lung and pneumonia radiographs. To segment lung parenchyma and pulmonary lesions, we used an *a priori*-developed deep-learning software, MEDIP PRO (<https://www.medicalip.com/medip>), which was found to have Dice similarity scores of 98.4–99.6% for lung parenchyma and 71.9–77.8% for COVID-19 pneumonia segmentation on CT images [4,23]. As the MEDIP PRO software was developed based on RT-PCR-proven COVID-19 patients' chest CT scans, pulmonary

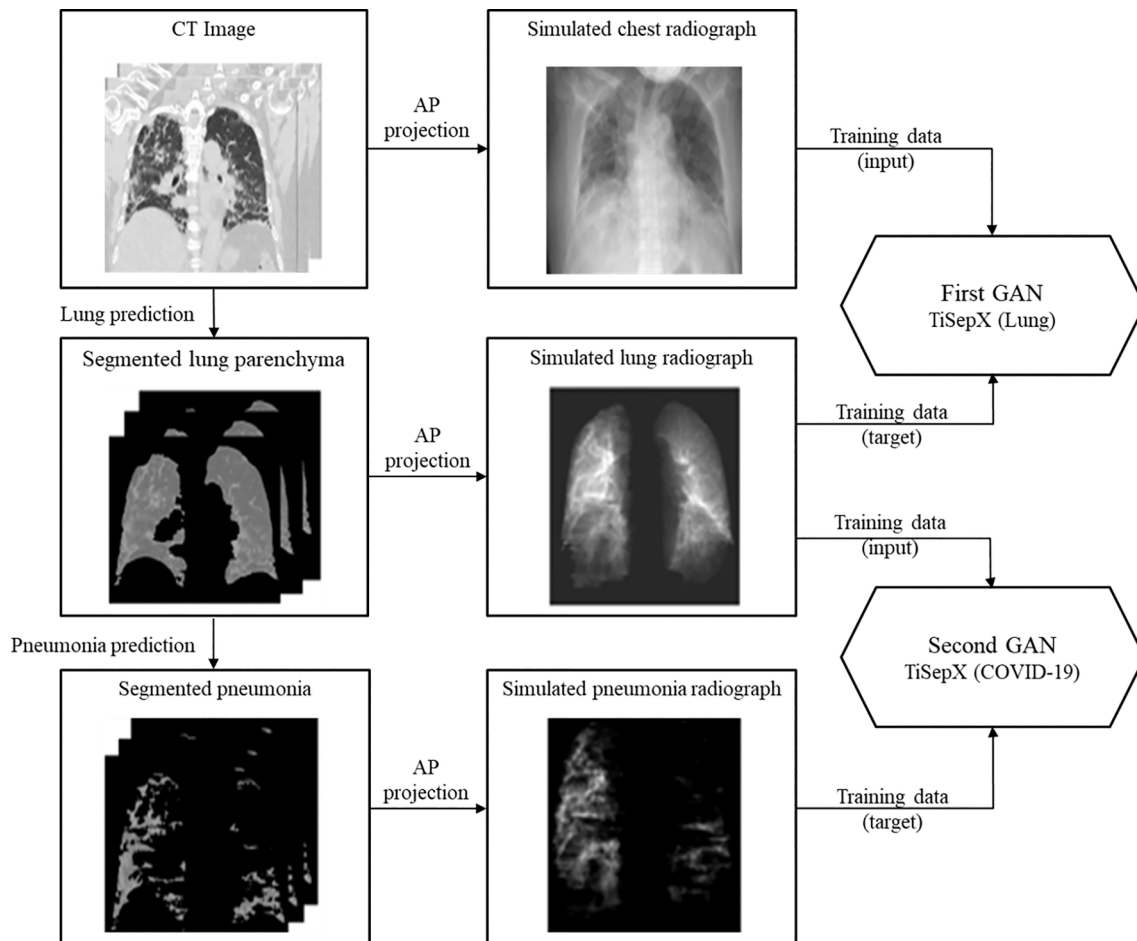


Fig. 1. Training system of lung and lesion image generation.

lesions similar to COVID-19 pneumonia such as GGO portion of the interstitial lung disease, pulmonary edema, pulmonary hemorrhage and pneumonic consolidation with peripheral distribution on chest CT images were segmented from non-COVID-19 CT scans (Supplementary Fig. 1).

2.3. Development of the deep neural network

To quantify COVID-19 pneumonia on chest radiographs, lung parenchyma and pneumonia need to be segmented from X-ray images. COVID-19 pneumonia preferentially manifests as ground-glass opacities in lower lung areas on CT or radiographic images and can be hardly found, particularly when overlapped with other soft tissue or bony structures on X-ray images. The mathematical approach [25] or conventional U-Net for segmentation [26] has limitations in segmenting retrocardiac and subdiaphragmatic lung parenchyma while removing overlying ribs, scapula, and soft tissue on X-ray images. Although this problem might be solved with changing the loss function, we applied a new approach with GAN for simultaneous bone suppression and lung field segmentation on chest radiograph. GAN is widely used for medical image synthesis as well as for image reconstruction, cross-modality synthesis, pseudo-healthy synthesis, and image analysis, such as image segmentation [27,28]. We assumed that GAN could a) generate lung parenchymal images containing retrocardiac/subdiaphragmatic lungs while excluding the overlying structures from X-ray images and b) efficiently extract COVID-19 pneumonia by generating pneumonia images from the lung parenchymal images as the overlying structures were already excluded.

One preparatory and two main networks were used for our task. As the spatial resolution of chest CT images is 512-by-512, the preparatory GAN [29] up-sampled the image resolution of the AP projection images from 512-by-512 to 2048-by-2048, equivalent to the resolution of chest radiographs. Two GANs, Pix2PixHD [30], then sequentially generated lung parenchymal radiographs from chest radiographs (first GAN) and pneumonia radiographs from lung parenchymal radiographs (second GAN) (<https://tiseptx.com/>; Fig. 1). Details of the network architecture, hyperparameters, hardware and software parameters are provided in Supplementary materials, Supplementary Table 1 and Supplementary Table 2, Supplementary Fig. 2.

2.4. Validation

We used publicly available chest radiographs of COVID-19 patients in a single institution in Italy which provide semi-quantitative severity score measured by dozens of radiologists (Brixia COVID-19 dataset, Italy) [31,32]. To validate the segmentation performance of our network, we used a publicly available dataset from Qatar University composed of 2913 chest radiographs of COVID-19 patients with ground truth masks of lung area and COVID-19 pneumonia area (COVID-QU-Ex Dataset, Qatar) [33–35]. We additionally validated the network-driven pneumonia extent in terms of the difference in quantification between CT scans and radiographs and in terms of clinical implications. The datasets for the former purpose consisted of chest CT scans and chest radiographs taken within 24 hours of CT scanning in patients with PCR-proven COVID-19 infection, while the datasets for clinical implications consisted of baseline chest radiographs and clinical information with a composite outcome (respiratory failure, intensive care unit [ICU] admission, or death). The following four datasets were used for the former validation: 375 patients from 17 institutions (Korean Imaging Cohort of Coronavirus Disease 2019, KICC-19, Korea) [36], 54 patients from a single secondary hospital (Myongji Hospital, Korea), 100 patients from a single institution in Japan [37], and 63 patients from the publicly available de-identified “COVID data save lives” dataset from HM Hospitales, Spain [38]. The latter clinical validation involved three datasets: 374 patients from the KICC-19 dataset, 1481 patients from Spain, and 243 patients from Hannover Medical School, Germany [39]. The details

of the datasets are presented in Fig. 2, the Supplementary materials, and Supplementary Table 3.

We used the KICC-19 dataset to measure structural similarity index metric (SSIM) between simulated lung mask by AP projection of CT and GAN-driven lung mask on chest radiograph and between simulated pneumonia mask by AP projection of CT and GAN-driven pneumonia mask. For qualitative evaluation, two experienced cardiothoracic radiologists with 35 and 5 years of experience with cardiothoracic imaging, respectively, visually inspected the randomly selected 100 chest radiographs from the external datasets and GAN-predicted pneumonia masks provided in DICOM and NIFTI file, respectively. Each radiologist scored every case regarding how satisfied the GAN predicted COVID-19 pneumonia on chest radiograph as follows: score 1, very dissatisfied; score 2, dissatisfied; score 3, somewhat dissatisfied; score 4, neutral; score 5, somewhat satisfied; score 6, satisfied; score 7, very satisfied.

2.5. Comparison with other networks for COVID-19 chest radiograph

We applied two previously developed convolutional neural networks for the prediction of airspace severity and geographic extent score of COVID-19 pneumonia on chest radiograph, COVID-Net CXR-S and COVID-Net S-Geo networks [16,17], in Brixia COVID-19 dataset, KICC-19 dataset, Myongji hospital dataset and Spain dataset and obtained the correlation with the pneumonia extent calculated by our developed GAN. COVID-Net CXR-S network predicts levels of airspace severity: level 1 = opacities in one or two lung zones; level 2 = opacities in three or more lung zones [16]. COVID-Net S-Geo network predicts the geographic extent of ground glass opacity and consolidation of each lung and scored as: 0 = no involvement; 1 = less than 25%; 2 = 25–50%; 3 = 50–75%; 4 = greater than 75% involvement [17]. The scores of each lung are added together and as a result, the geographic extent score ranges from 0 to 8 [17].

Additionally, we applied the COVID-Net S-Geo network [17] in Spanish, KICC-19, German datasets and estimated the performance of the networks using the area under the receiver operating characteristic curve (AUC) for predicting the composite outcome.

2.6. Statistical analysis

Correlation between semiquantitative severity score and AI-driven pneumonia extent (%) was assessed by Pearson's correlation coefficient. We used dice similarity coefficient (DSC) to calculate the similarity of ground truth and predicted mask of COVID-19 pneumonia on chest radiographs. To evaluate the agreement of the pneumonia extent (%) between the chest CT and radiographs within one day, for a comprehensive dataset, we assessed the mean relative measurement difference (RMD) of pneumonia extent (%) between chest CT and radiographs within 24 hours and presented a Bland-Altman plot with 95% limits of agreement (LOA) [40,41]. The LOA values were calculated using a linear mixed-effect model, taking into account inter-subject variance and inter-center variance [42]. The MEDIP PRO-derived pneumonia extent on CT images served as the reference standard. Pearson's correlation coefficients were also calculated to examine the linear relationship between the pneumonia extent (%) of chest CT and radiographic images from the comprehensive dataset.

Associations between patient characteristics and the composite outcome were tested using the Mann-Whitney *U* test or Fisher exact test, as appropriate. Multivariable logistic regression analysis was conducted to identify risk factors for the composite outcome after adjusting for demographic characteristics and laboratory findings [43,44]. The performance of network-driven pneumonia extent on chest radiographs was calculated using the AUC for predicting the composite outcome in the Spanish, KICC-19, and German datasets.

We built two clinical prediction models using multivariable logistic regression with maximum penalized likelihood after normalizing the variables to follow the distribution having a mean of 0 and variance of 1.

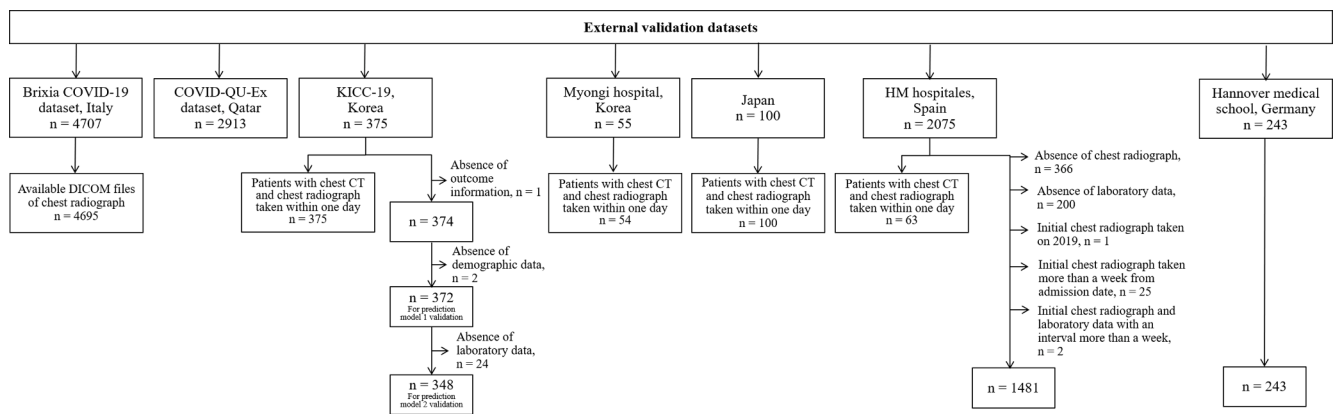


Fig. 2. Diagram of data selection in the external validation datasets. Abbreviation: KICC-19, Korean Imaging Cohort of Coronavirus Disease 2019.

The first model (model 1) included demographic information (age, sex, and comorbidities) and the GAN-driven radiographic extent, while the second model (model 2) included the neutrophil-to-lymphocyte ratio and lactate dehydrogenase (LDH) levels in addition to the variables included in model 1. Besides, to investigate the impact of the GAN-driven radiographic extent on models' performances we built the same clinical prediction models while excluding the pneumonia extent. Model 1 was intended for a resource-limited setting where blood tests are not available, while chest radiographs can be taken. The Spanish dataset had the largest number of COVID-19 patients with chest radiographs and laboratory findings ($n = 1481$), followed by the KICC-19 ($n = 374$), whereas the German dataset lacked age information and was excluded from clinical prediction modeling. The Spanish dataset served as the derivation dataset for modeling and was randomly dichotomized into training and internal validation datasets at a ratio of 3:1. The KICC-19 dataset was used to validate the clinical prediction models. We evaluated the performance of the clinical prediction models by assessing discrimination and calibration. Discrimination refers to how well the model differentiates the patients with the outcome from those without the outcome [45,46]. The AUC was calculated to evaluate the discrimination performance of the models. Calibration refers to the agreement between observed outcomes and predictions [45,46]. Calibration was analyzed using the calibration intercept and slope, by plotting the observed versus predicted probabilities, and by using the Spiegelhalter statistic [47,48]. A statistically significant value of the Spiegelhalter P value indicated poor calibration. The normalized GAN-driven pneumonia extent in the models was back-transformed to show how much odd ratios (ORs) for unfavorable outcomes increased as the percent extent increased.

All statistical analyses were performed using MedCalc Statistical Software version 20.027 (MedCalc Software Ltd, Ostend, Belgium; <http://www.medcalc.org>; 2022) and R software version 3.6.1 (<https://www.R-project.org>). A P value of less than 0.05 was considered to indicate statistical significance.

3. Results

3.1. Correlation coefficients and measurement difference of pneumonia extent

The basic characteristics of the validation datasets are shown in [Supplementary Table 3](#). Pearson's correlation coefficient between the semi-quantitative Brixia severity score and AI-driven extent of COVID-19 pneumonia on chest radiographs of the Italian dataset was 0.611 (95% confidence interval [CI], 0.593–0.629). A box plot of the semi-quantitative severity score and the GAN-driven extent of COVID-19 pneumonia on chest radiographs from the Italian dataset is shown in [Fig. 3](#). DSC of the COVID-QU-Ex dataset was $41.8 \pm 18.9\%$

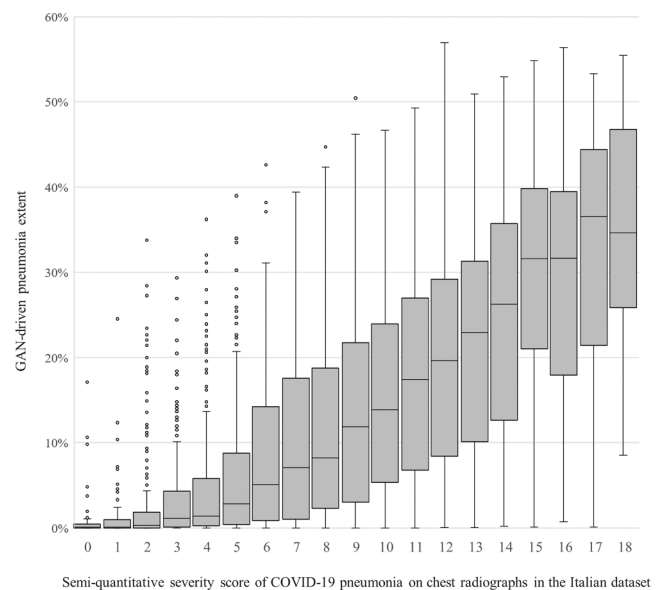


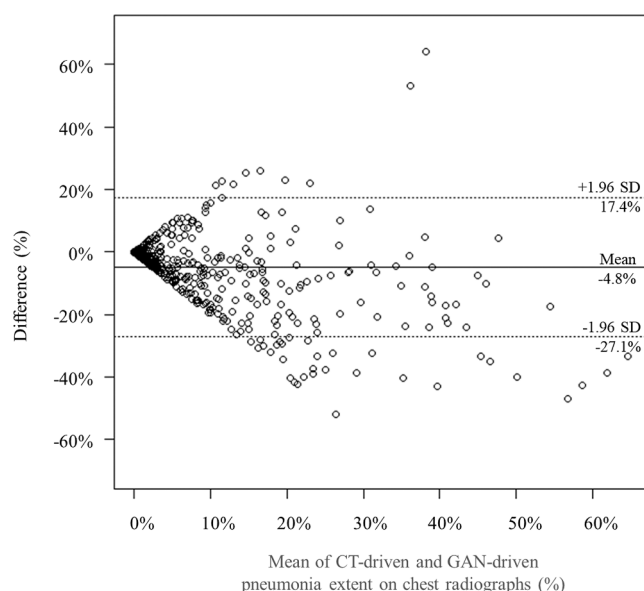
Fig. 3. Box plot of semiquantitative severity score of COVID-19 pneumonia and GAN-driven pneumonia extent of chest radiograph of Italian dataset. Boxes indicate interquartile range; horizontal lines, median; whiskers, 1.5 times the interquartile range and dots, extreme values.

([Supplementary Fig. 3](#)). The RMD between CT-driven pneumonia extent and GAN-driven radiographic extent was -4.8% ([Fig. 4](#)). The 95% LOA between the extents ranged from -27.1% to 17.4% . Pearson's correlation coefficient between the extents was 0.640 (95% CI, 0.590–0.685). [Fig. 5](#) shows representative images of the GAN-driven radiographic extent of COVID-19 pneumonia in the external validation dataset.

The mean and standard deviation of SSIM between simulated lung mask by AP projection of CT and GAN-driven lung mask on chest radiograph and between simulated pneumonia mask by AP projection of CT and GAN-driven pneumonia mask were 99.5 ± 0.3 and 98.9 ± 0.8 , respectively. In qualitative evaluation, the radiologists' mean and standard deviation of satisfaction score of COVID-19 pneumonia extraction of GAN on chest radiograph was 4.7 ± 2.2 and 4.6 ± 2.1 ([Supplementary Fig. 4](#)). 70% and 73% of the cases were scored equal or more than 4, suggesting neutral or satisfied segmentation result by each radiologist.

3.2. Comparison between participants with and without the composite outcomes

In the Spanish dataset ($n = 1481$), participants with the composite



Abbreviation: GAN, generative adversarial network

Fig. 4. Bland-Altman plot and limits of agreement for CT-driven and GAN-driven radiographic pneumonia extent taken on the same or within 1 day in the comprehensive external validation dataset. Abbreviation: GAN, generative adversarial network.

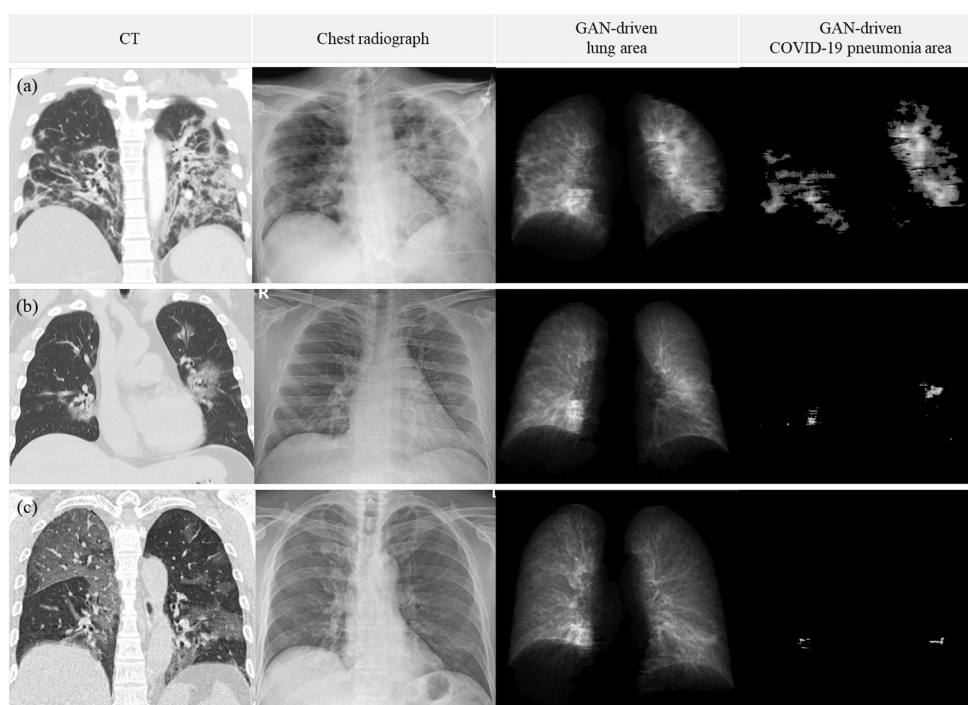
outcome (prevalence, 38%) were older than those without the

composite outcome (median age, 69 years versus 66 years, $P = .001$), were more likely to have a comorbidity ($P < .001$), and had a higher neutrophil-to-lymphocyte ratio (5.310 versus 3.730, $P < .001$), higher LDH levels (570 U/L versus 511 U/L, $P < .001$), and a larger GAN-driven radiographic pneumonia extent (7.2% versus 4.1%, $P < .001$) (Table 1). Similarly, in the KICC-19 dataset ($n = 374$; outcome prevalence, 10%), the significant risk factors for the composite outcome were older age (median age, 70 years versus 52 years, $P < .001$), male sex ($P = .01$), the presence of a comorbidity ($P < .001$), a higher neutrophil-to-lymphocyte ratio (5.097 versus 2.204, $P < .001$), higher LDH levels (644 U/L versus 346 U/L, $P < .001$) and larger radiographic pneumonia extent (3.3% versus 0.0%, $P < .001$) (Table 2). In the German dataset ($n = 243$; outcome prevalence, 78%), the significant risk factors for the composite outcome were higher LDH levels (395.0 U/L versus 287.5 U/L, $P < .001$) and larger radiographic pneumonia extent (14.7% versus 0.5%, $P < .001$) (Table 3).

3.3. Clinical prediction model for the composite outcome

The GAN-driven pneumonia extent alone provided AUCs of 0.614 (95% CI, 0.585–0.643), 0.806 (95% CI, 0.739–0.874), and 0.842 (95% CI, 0.778–0.906) in the Spanish internal validation, KICC-19, and German datasets, respectively (Fig. 6a). The ORs for unfavorable outcomes in each dataset were 1.05 (95% CI, 1.03–1.06), 1.18 (95% CI, 1.11–1.27), and 1.15 (95% CI, 1.10–1.21), respectively, indicating that risks increased 1.05, 1.18, and 1.15 times, each, per increase of pneumonia extent by 7.6%.

When the GAN-driven pneumonia extent was combined with demographic information (model 1), the pneumonia extent remained in the model along with comorbidities, whereas age and sex were



Abbreviation: GAN, generative adversarial network; KICC-19, Korean Imaging Cohort of Coronavirus Disease 2019

Fig. 5. GAN-driven radiographic extent of COVID-19 pneumonia compared with chest CT taken on the same day. (a) A 50-year-old male with COVID-19 pneumonia in the Spanish dataset. The pneumonia extent was 29.54% on chest CT and 23.09% on a chest radiograph. (b) A 41-year-old male patient in the KICC-19 dataset. The pneumonia extent was 2.4% on chest CT and 2.0% on a chest radiograph. Despite its ground glass-predominant pattern, the pneumonia was well detected on chest radiographs by GAN and highly correlated with the CT extent. (c) A 70-year-old male patient in the KICC-19 dataset. The pneumonia extent was 25.0% on chest CT and 0.2% on a chest radiograph. Extensive ground glass opacities in both lungs were detected on chest CT. However, on the chest radiograph, the lesions were subtle and hardly visualized. As a result, the GAN-driven pneumonia extent underestimated the CT-driven extent. Abbreviation: GAN, generative adversarial network; KICC-19, Korean Imaging Cohort of Coronavirus Disease 2019.

Table 1

Characteristics of Study Participants of the Spanish Dataset.

Variables	All Participants (n = 1481)	Participants with the Composite Outcome (n = 569)	Participants without the Composite Outcome (n = 912)	P Value
Age (y)*	67 (56–77)	69 (58–78)	66 (54–77)	0.001
Sex				0.08
Women	553 (37)	196 (34)	357 (39)	
Men	928 (63)	373 (66)	550 (61)	
Comorbidity†				<0.001
Absent	910 (61)	301 (53)	609 (67)	
Present	571 (39)	268 (47)	303 (33)	
Neutrophil-to-lymphocyte ratio*	4.270 (2.500–8.240)	5.310 (3.070–10.140)	3.730 (2.180–7.090)	<0.001
Lactate dehydrogenase (U/L)*	532 (416.7–699.4)	570 (443.7–753)	511 (400.1–661.2)	<0.001
Pneumonia extent on chest radiographs (%)	5.4 (1.4–11.0)	7.2 (2.9–12.9)	4.1 (0.8–9.8)	<0.001
Prevalence of the composite outcome‡	569 (38)	...	...	...

Data are number of patients with percentage in parentheses, unless otherwise specified.

* Data are medians, with the interquartile range in parentheses.

† Comorbidities included hypertension, type 2 diabetes mellitus, coronary artery disease, heart failure, cerebrovascular disease, and chronic kidney disease.

‡ The composite outcome included respiratory failure and death.

Table 2

Characteristics of Study Participants of the Korean Imaging Cohort of Coronavirus Disease 2019 External Validation Dataset.

Variables	All Participants (n = 374)	Participants with the Composite Outcome (n = 37)	Participants without the Composite Outcome (n = 337)	P Value
Age (years)*	54 (35–67)	70 (64–78)	52 (31–64)	<0.001
Sex				0.01
Women	195 (52)	11 (30)	184 (55)	
Men	178 (48)	26 (70)	152 (45)	
Comorbidity†				<0.001
Absent	247 (66)	13 (35)	234 (69)	
Present	127 (34)	24 (65)	103 (31)	
Neutrophil-to-lymphocyte ratio*	2.495 (1.510–4.135)	5.097 (3.749–11.312)	2.204 (1.481–3.763)	<0.001
Lactate dehydrogenase (U/L)*	358.5 (235–518)	644 (473–1027)	346 (223–466.5)	<0.001
Pneumonia extent (%) on chest radiographs	0.1 (0–0.2)	3.3 (0.7–9.4)	0.0 (0.0–1.3)	<0.001
Prevalence of the composite outcome‡	37 (9.9)	...	...	...

Age and sex was missing in one participant, respectively. Laboratory test results were not available in 24 patients. Data are number of patients with percentage in parentheses, unless otherwise specified.

* Data are medians, with interquartile range in parentheses.

† Comorbidities included hypertension, diabetes mellitus, coronary artery disease, COPD, cancer, chronic kidney disease, and chronic liver disease.

‡ The composite outcome included respiratory failure, ICU admission, and death.

Table 3

Characteristics of Study Participants of the German External Validation Dataset.

Variables	All Participants (n = 243)	Participants with the Composite Outcome (n = 189)	Participants without the Composite Outcome (n = 54)	P Value
Sex				0.11
Women	63 (26)	44 (23)	19 (35)	
Men	180 (74)	145 (77)	35 (65)	
Neutrophil-to-lymphocyte ratio*	5.000 (3.083–8.077)	5.417 (3.091–8.077)	4.091 (2.667–6.167)	0.25
Lactate dehydrogenase (U/L)*	365 (275–570)	395 (280–655)	287.5 (180–357.5)	<0.001
Pneumonia extent (%) on chest radiographs*	12.0 (1.9–22.6)	14.7 (5.7–25.3)	0.5 (0.0–3.9)	<0.001
Prevalence of the composite outcome†	189 (78)	...	...	...

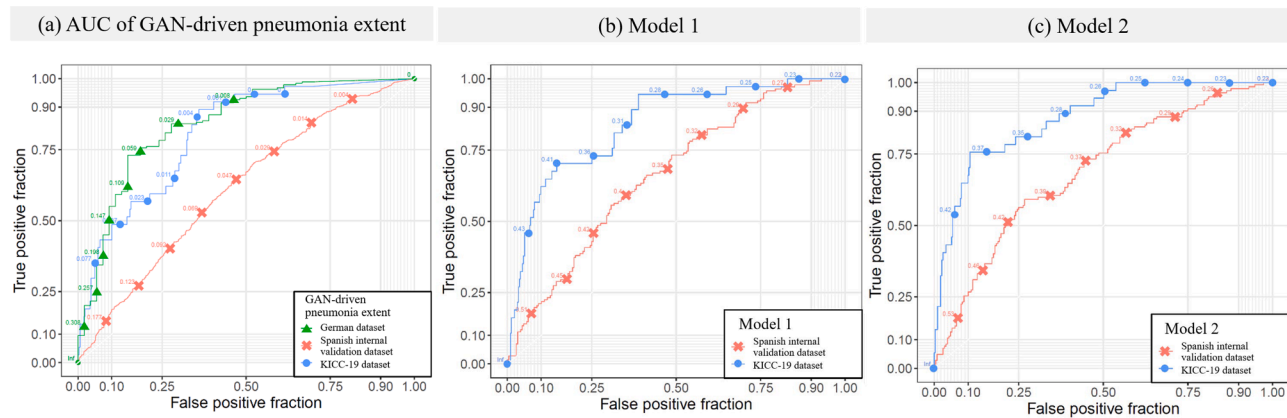
Data are number of patients with percentage in parentheses, unless otherwise specified.

* Data are medians, with the interquartile range in parentheses.

† The composite outcome included respiratory failure, ICU admission, and death.

insignificant. Prediction Model 1, which included pneumonia extent and demographic information, showed an AUC of 0.663 (95% CI, 0.608–0.719) in the Spanish dataset and 0.841 (95% CI, 0.775–0.906) in the KICC-19 dataset (Fig. 6b). Model 2, which integrated laboratory information, showed AUCs of 0.688 (95% CI, 0.633–0.743) and 0.877 (95% CI, 0.823–0.932) in the Spanish and KICC dataset, respectively

(Fig. 6c). In prediction Model 1, comorbidity (OR, 1.70; 95% CI, 1.31–2.21; $p < .001$) and GAN-driven pneumonia extent (OR, 1.33; 95% CI, 1.17–1.51; $p < .001$) were significant variables. In prediction Model 2, comorbidity (OR, 1.74; 95% CI, 1.34–2.28; $p < .001$), neutrophil-to-lymphocyte ratio (OR, 1.29; 95% CI, 1.13–1.49; $p < .001$), and GAN-driven pneumonia extent (OR, 1.26; 95% CI,



Abbreviation: GAN, generative adversarial network; KICC-19, Korean Imaging Cohort of Coronavirus Disease 2019

Fig. 6. Receiver operating characteristic curves of prediction models applied to the internal and external validation datasets. (a) Using GAN-driven pneumonia extent as the solitary variable, the areas under the receiver operating characteristic curve (AUCs) were 0.614, 0.806, and 0.842 in the internal validation, KICC-19, and German datasets, respectively. (b) For model 1, the AUCs were 0.663 and 0.841, (c) for model 2, the AUCs were 0.688 and 0.877 in the Spanish internal validation and KICC-19 datasets, respectively. Abbreviation: GAN, generative adversarial network; KICC-19, Korean Imaging Cohort of Coronavirus Disease 2019.

1.10–1.43; $p < .001$) were significant variables (Table 4 and Supplementary Table 4). Box plots of the model output according to the presence of unfavorable outcomes are provided in Supplementary Fig. 5. The exclusion of the GAN-driven pneumonia extent from the Model 1 and 2 significantly decreased models' performances in the Spanish and KICC dataset in general (Supplementary Table 5).

Table 4
Prediction Models for the Probability of the Composite Outcome in COVID-19 Patients.

	Model 1		
	Odds Ratio	P Value	Beta
Intercept	(95% CI)		Coefficient
	0.44 (0.35–0.55)	<0.001	−0.83
Age (y)	1.06 (0.92–1.21)	0.43	0.05
Sex, women vs. men	1.25 (0.96–1.62)	0.10	0.22
Comorbidity	1.70 (1.31–2.21)	<0.001	0.53
Extent of pneumonia on chest radiographs (%)	1.33 (1.17–1.51)	<0.001	0.29
	Model 2		
	Odds Ratio	P Value	Beta
Intercept	(95% CI)		Coefficient
	0.46 (0.36–0.58)	<0.001	−0.77
Age (y)	1.01 (0.88–1.15)	0.94	0.01
Sex, women vs. men	1.12 (0.86–1.46)	0.40	0.11
Comorbidity	1.74 (1.34–2.28)	<0.001	0.56
Neutrophil-to-lymphocyte ratio	1.29 (1.13–1.49)	<0.001	0.25
Lactate dehydrogenase	1.05 (0.93–1.20)	0.45	0.05
Extent of pneumonia on chest radiographs (%)	1.26 (1.10–1.43)	<0.001	0.23

All continuous variables were standardized using mean and standard deviation: $\frac{x - \text{mean}}{\text{standard deviation}}$

In terms of calibration, the first and second prediction models showed good calibration, with calibration slopes of 1.378 and 1.329, intercepts of 0.179 and 0.135, and Spiegelhalter P values over 0.05 in the internal validation dataset. However, in external validation with the KICC-19 dataset, the calibration slopes were 3.054 and 3.448, intercepts were −0.320 and 0.001, with Spiegelhalter P values under 0.05, indicating poor model calibration (Supplementary Table 6 and Supplementary Fig. 6).

3.4. Comparison with other networks for COVID-19 chest radiograph

The Pearson's correlation coefficient between our GAN-driven pneumonia extent and the COVID-Net CXR-S-driven severity score and COVID-Net S-Geo-driven pneumonia extent score respectively were: 0.563 (95% CI, 0.543–0.582) and 0.680 (95% CI, 0.665–0.696) in Brixia dataset; 0.486 (95% CI, 0.404–0.560) and 0.554 (95% CI, 0.480–0.620) in KICC-19 dataset; 0.695 (95% CI, 0.525–0.812) and 0.844 (0.744–0.907) in Myongi hospital dataset; 0.319 (95% CI, 0.077–0.525) and 0.560 (95% CI, 0.632–0.709) in Spain dataset.

The AUCs of COVID-Net S-Geo for predicting the composite outcome in Spanish, KICC-19, German dataset were 0.640 (95% CI, 0.614–0.664), 0.787 (95% CI, 0.724–0.827), 0.834 (95% CI, 0.781–0.879), respectively.

4. Discussion

Our generative adversarial network (GAN) was trained with virtual chest radiographs synthesized from non-COVID-19 chest CT scans and successfully applied to COVID-19 actual chest radiographs. The GAN-driven pneumonia extent on radiographs was correlated with the pneumonia extent observed on CT scans taken within 24 h but underestimated the pneumonia extent by 4.8% on average, with wide 95% LOA of the RMD (from −27.1% to 17.4%). This considerable difference could result from the inaccuracy of the quantification. Nevertheless, a certain degree of difference seems inevitable as COVID-19 lesions, which predominantly manifest as ground-glass opacities on CT, are often imperceivable on chest radiographs (Fig. 5). Indeed, underestimation also occurred in visual perceptions of COVID-19 radiographs: radiologists' visual assessments underestimated CT-proven COVID-19 lesions in half of the chest radiographs [49]. Furthermore, non-identical positioning and inspiration at scanning (supine position for CT scans; supine

or erect position for radiographs) can contribute to the difference between the extents. The temporal evolution of COVID-19 lesions between CT scans and radiographs may also have contributed to the difference.

The GAN-driven pneumonia extent on radiographs significantly differed according to the composite outcome and yielded odds ratios of 1.05–1.18 per percent in all three datasets. The three datasets had different prevalence rates of unfavorable outcomes: 10% of the KICC patients, 38% of the Spanish patients, and 78% of the German patients. Our model proved the prognostic value of chest imaging to stratify at-risk patients with COVID-19 regardless of how severely the target population was affected. Indeed, the incidence of unfavorable outcomes in COVID-19 can vary regionally according to where and when an outbreak occurs.

We tested the GAN-driven pneumonia extent alone and built two clinical prediction models for unfavorable outcomes. The discrimination of the two models and GAN-driven pneumonia extent for the composite outcome was generally poor (0.614–0.688) in the Spanish internal validation dataset, but fair to good (0.806–0.877) in the KICC-19 external validation dataset. The calibration was good in the former dataset, but poor in the latter dataset for both prediction models. This performance difference in each dataset is assumed to be caused by differences in disease severity in terms of unfavorable outcomes (38% versus 10%) and average pneumonia extent (5.4% versus 0.1%). In addition, the Spanish dataset is from the earliest days of the COVID-19 pandemic under constrained resources and limited experience in COVID-19, and it is likely that patients who were suffering from severe pneumonia had difficulty in receiving rapid and proper management. Therefore, we assumed that the Spanish internal validation dataset has high predictive difficulty. The prognosis of the patients in the lower disease severity group is more dependent on the presence or absence of chest radiograph abnormalities resulting in higher performance in our prediction models. Furthermore, in the Spanish dataset, demographic characteristics, laboratory findings, and the pneumonia extent showed considerable overlap between patients with unfavorable outcomes and those without in comparison to those in the KICC-19 and German datasets.

In the technical aspect, we choose to develop our GAN by training two Pix2PixHD networks (Fig. 1) instead of one Pix2PixHD network model with AP projection images of native CT images (simulated chest radiograph) as input and AP-projected CT images of pulmonary lesions (simulated pneumonia radiograph) as target output. When comparing the COVID-19 pneumonia extent predicted from CT by MEDIP PRO and chest radiograph by developed GANs, mean absolute error was higher in one Pix2PixHD network model than two Pix2PixHD networks model in all KICC-19, Myongji hospital, Spanish external validation datasets (one Pix2PixHD network model, 10.1%, 18.6%, 19.0%; two Pix2PixHD networks model, 7.8%, 10.3%, 9.7%).

In previous study comparing chest radiograph reporting systems in patients with COVID-19 infection, Brixia score, and visually assessed percentage opacification by radiologists showed similar results associated with the adverse outcomes and visually assessed percentage opacification of COVID-19 pneumonia were proven to be a predictor of adverse outcomes in patients with COVID-19 infection [12,50]. The correlation coefficient between our GAN-driven percent pneumonia extent and previously developed networks for COVID-19 pneumonia extent grading (COVID-Net S-Geo) was moderate to high in the four external validation datasets with various disease severity and nation (0.554–0.844). Additionally, AUCs of our GAN-driven pneumonia extent and COVID-Net S-Geo for predicting the composite outcome were similar in three datasets suggesting the acceptable performance of our GAN compared to the previously developed semi-quantitative convolutional neural network (Spanish, 0.614 vs 0.640; KICC-19, 0.806 vs 0.787; German, 0.842 vs 0.834). The GAN we developed makes it possible to quantify pneumonia automatically and rapidly on chest radiographs even in a pandemic situation where radiologists are in short supply, and as in previous studies, we proved that GAN-driven

pneumonia could predict adverse outcomes for COVID-19 patients.

There were studies in interest in developing artificial intelligence to score the severity of COVID-19 pneumonia on chest radiographs and trying to prove that severity scored by artificial intelligence correlates with adverse outcomes in COVID-19 patients [18–21]. However, in some of those studies, COVID-19 pneumonia in the subdiaphragmatic and retrocardiac areas were neglected. DSC of the COVID-QU-Ex dataset was low ($41.8 \pm 18.9\%$). In the visual review of the ground truth and predicted masks, we found the ground truth masks of COVID-QU-Ex dataset overestimated the COVID-19 pneumonia region and neglected the COVID-19 pneumonia in the bilateral subdiaphragmatic area and retrocardiac area, and we presumed this inaccuracy of ground truth resulted in low DSC. We overcame these easily neglectable areas and increased the accuracy of the training dataset by using the CT-based simulated radiographs. Our study has several limitations. First, the training dataset was from non-COVID-19 patients. Although we used a deep neural network dedicated to the segmentation of characteristic CT findings of COVID-19, the distribution and characteristics of non-COVID-19 lesions could differ from those of COVID-19 lesions. As a result, in the qualitative visual review of the predicted masks, there were some insufficient segmentations of peripheral pneumonia in the upper and mid lung zone. In non-COVID-19 patients, most of the peripheral opacities of bilateral upper and mid lung zone of simulated chest radiographs might be the soft tissue or scapular density. However, COVID-19 pneumonia is recognized for its typical peripheral distribution of pneumonia. Therefore, we presumed that the low accuracy of the peripheral pneumonia segmentation in bilateral upper and mid lung zone in chest radiographs by our GAN is due to the difference of patient groups of training and validation datasets. Second, we did not use any real chest radiographs in training as it was difficult to segment COVID-19 lesions in a sufficient number and impossible to prepare lung or pneumonia-specific radiographs from chest radiographs. Third, we could not fully adjust for laboratory parameters, such as D-dimer levels, due to missing or incomplete information. Fourth, there were heterogeneous patient characteristics across the datasets used to evaluate clinical implications, since they were from different nations and various institutions. Fifth, we did not examine the performance of GAN-driven pneumonia extent for the diagnosis of COVID-19, as the diagnostic use of chest radiographs for this purpose is not currently indicated. Sixth, since we develop and validate the GAN with datasets from the early days of the COVID-19 pandemic, we did not evaluate the predictive performance in patients with recent COVID-19 variant infection or vaccination. Further study to assess our GAN and prediction models in cohorts composed of patients with COVID-19 variants and with individual vaccination status information is necessary.

In conclusion, a generative adversarial network (GAN) could automatically quantify COVID-19 pneumonia on chest radiographs. GAN-driven pneumonia extent identified patients at risk for unfavorable outcomes alone or with demographic and clinical information regardless of how severely the target population was affected by COVID-19. GAN-driven pneumonia quantification guarantees further validation with a prospectively collected larger cohort. Nevertheless, our result supports the recommendation of chest imaging for initial risk stratification or monitoring of COVID-19 pneumonia in the triage of severe patients with COVID-19 pneumonia where access to CT scanning is limited.

Declaration of Competing Interest

The authors declare the following financial interests/personal relationships which may be considered as potential competing interests: [Soon Ho Yoon work as chief medical officer in MEDICALIP. Joseph Nathanael Wintanto is a machine learning researcher in the same company. Hyungjin Kim has a stock of MEDILCAIP. Those authors did not involve data/statistical analysis]. Other authors have no competing interest to declare.

Acknowledgements

The authors would like to acknowledge Eun-Young Kang, MD, PhD, and Woojin Lee, MD for their assistance in the qualitative analysis in this manuscript.

The authors would like to acknowledge Andrew Dombrowski, PhD (Compecs, Inc.) for his assistance in improving the use of English in this manuscript.

Appendix A. Supplementary data

Supplementary data to this article can be found online at <https://doi.org/10.1016/j.ejrad.2023.110858>.

References

- [1] N. Sethuraman, S.S. Jeremiah, A. Ryo, Interpreting Diagnostic Tests for SARS-CoV-2, *JAMA* 323 (22) (2020) 2249–2251.
- [2] T. Ai, Z. Yang, H. Hou, C. Zhan, C. Chen, W. Lv, Q. Tao, Z. Sun, L. Xia, Correlation of chest CT and RT-PCR testing in coronavirus disease 2019 (COVID-19) in China: A report of 1014 cases, *Radiology* (2020) 200642.
- [3] Y. Fang, H. Zhang, J. Xie, M. Lin, L. Ying, P. Pang, W. Ji, Sensitivity of chest CT for COVID-19: Comparison to RT-PCR, *Radiology* 200432 (2020).
- [4] S.-J. Yoo, X. Qi, S. Inui, H. Kim, Y.-J. Jeong, K.H. Lee, Y.K. Lee, B.Y. Lee, J.Y. Kim, K. N. Jin, J.K. Lim, Y.H. Kim, K.B. Kim, Z. Jiang, C. Shao, J. Lei, S. Zou, H. Pan, Y. Gu, G. Zhang, J.M. Goo, S.H. Yoon, Deep learning-based automatic CT quantification of coronavirus disease 2019 pneumonia: an international collaborative study, *J. Comput. Assist. Tomogr.* 46 (3) (2022) 413–422.
- [5] H.Y.F. Wong, H.Y.S. Lam, A.H. Fong, S.T. Leung, T.W. Chin, C.S.Y. Lo, M.M. Lui, J. C.Y. Lee, K.W. Chiu, T.W. Chung, E.Y.P. Lee, E.Y.F. Wan, I.F.N. Hung, T.P.W. Lam, M.D. Kuo, M.Y. Ng, Frequency and distribution of chest radiographic findings in patients positive for COVID-19, *Radiology* 296 (2) (2020) E72–E78.
- [6] G.D. Rubin, C.J. Ryerson, L.B. Haramati, N. Sverzellati, J.P. Kanne, S. Raoof, N.W. Schlager, A. Volpi, J.-J. Yim, I.B.K. Martin, D.J. Anderson, C. Kong, T. Altes, A. Bush, S.R. Desai, o. Goldin, J.M. Goo, M. Humbert, Y. Inoue, H.-U. Kauczor, F. Luo, P.J. Mazzone, M. Prokop, M. Remy-Jardin, L. Richeldi, C.M. Schaefer-Prokop, N. Tomiyama, A.U. Wells, A.N. Leung, The Role of Chest Imaging in Patient Management during the COVID-19 Pandemic: A Multinational Consensus Statement from the Fleischner Society, *Radiology* 296(1) (2020) 172–180.
- [7] S. Schiaffino, S. Tritella, A. Cozzi, S. Carriero, L. Blandi, L. Ferraris, F. Sardanelli, Diagnostic performance of chest X-Ray for COVID-19 pneumonia during the SARS-CoV-2 pandemic in lombardy, Italy, *J. Thorac. Imaging* 35 (4) (2020) W105–W106.
- [8] E.A. Akl, I. Blazić, S. Yaacoub, G. Frija, R. Chou, J.A. Appiah, M. Fatehi, N. Flor, E. Hitti, H. Jafri, Use of chest imaging in the diagnosis and management of COVID-19: A WHO rapid advice guide, *Radiology* 298 (2) (2021) E63–E69.
- [9] F. Liu, Q. Zhang, C. Huang, C. Shi, L. Wang, N. Shi, C. Fang, F. Shan, X. Mei, J. Shi, F. Song, Z. Yang, Z. Ding, X. Su, H. Lu, T. Zhu, Z. Zhang, L. Shi, Y. Shi, CT quantification of pneumonia lesions in early days predicts progression to severe illness in a cohort of COVID-19 patients, *Theranostics* 10 (12) (2020) 5613–5622.
- [10] B. Park, J. Park, J.-K. Lim, K.M. Shin, J. Lee, H. Seo, Y.H. Lee, J. Heo, W.K. Lee, J. Y. Kim, K.B. Kim, S. Moon, S. Choi, Prognostic implication of volumetric quantitative CT analysis in patients with COVID-19: A multicenter study in Daegu, Korea, *Korean J. Radiol.* 21 (11) (2020) 1256–1264.
- [11] D. Sun, X. Li, D. Guo, L. Wu, T. Chen, Z. Fang, L. Chen, W. Zeng, R. Yang, CT quantitative analysis and its relationship with clinical features for assessing the severity of patients with COVID-19, *Korean J. Radiol.* 21 (7) (2020) 859–868.
- [12] I. Au-Yong, Y. Higashi, E. Giannotti, A. Fogarty, J.R. Morling, M. Grainge, A. Race, I. Juurlink, M. Simmonds, S. Briggs, S. Cruickshank, S. Hammond-Pears, J. West, C. J. Crooks, T. Card, Chest radiograph scoring alone or combined with other risk scores for predicting outcomes in Covid-19, *Radiology* 302 (2) (2022) E11.
- [13] L. Wang, Z.-Q. Lin, A. Wong, Covid-net: A tailored deep convolutional neural network design for detection of covid-19 cases from chest x-ray images, *Sci. Rep.* 10 (1) (2020) 1–12.
- [14] R.M. Wehbe, J. Sheng, S. Dutta, S. Chai, A. Dravid, S. Barutcu, Y. Wu, D.R. Cantrell, N. Xiao, B.D. Allen, DeepCOVID-XR: An artificial intelligence algorithm to detect COVID-19 on chest radiographs trained and tested on a large US clinical Data Set, *Radiology* 299 (1) (2021) E167–E176.
- [15] E.J. Hwang, H. Kim, S.H. Yoon, J.M. Goo, C.M. Park, Implementation of a deep learning-based computer-aided detection system for the interpretation of chest radiographs in patients suspected for COVID-19, *Korean J. Radiol.* 21 (10) (2020) 1150–1160.
- [16] H. Aboutalebi, M. Pavlova, M.J. Shafiee, A. Sabri, A. Alaref, A. Wong, COVID-Net CXR-S: Deep convolutional neural network for severity assessment of COVID-19 cases from chest X-ray images, *Diagnostics* 12 (1) (2022) 25.
- [17] A. Wong, Z.-Q. Lin, L. Wang, A.G. Chung, B. Shen, A. Abbasi, M. Hoshmand-Kochi, T.Q. Duong, Towards computer-aided severity assessment via deep neural networks for geographic and opacity extent scoring of SARS-CoV-2 chest X-rays, *Sci. Rep.* 11 (1) (2021) 9315.
- [18] M.D. Li, N.T. Arun, M. Gidwani, K. Chang, F. Deng, B.P. Little, D.P. Mendoza, M. Lang, S.I. Lee, A. O'Shea, Automated assessment and tracking of COVID-19 pulmonary disease severity on chest radiographs using convolutional siamese neural networks, *Radiol. Artif. Intell.* 2 (4) (2020), e200079.
- [19] V.V. Danilov, D. Litmanovich, A. Proutski, A. Kirpich, D. Nefaridze, A. Karpovsky, Y. Gankin, Automatic scoring of COVID-19 severity in X-ray imaging based on a novel deep learning workflow, *Sci. Rep.* 12 (1) (2022) 12791.
- [20] S. Ebrahimi, F. Homayounieh, M.A.B.C. Rockenbach, P. Putha, T. Raj, I. Dayan, B.C. Bizzo, V. Buch, D. Wu, K. Kim, Q. Li, S.R. Digumarthy, M.K. Kalra, Artificial intelligence matches subjective severity assessment of pneumonia for prediction of patient outcome and need for mechanical ventilation: A cohort study, *Sci. Rep.* 11 (1) (2021) 858.
- [21] J. Mushtaq, R. Pennella, S. Laval, A. Colarieti, S. Steidler, C.M.A. Martinenghi, D. Palumbo, A. Esposito, P. Rovere-Querini, M. Tresoldi, G. Landoni, F. Ciceri, A. Zangrillo, F. De Cobelli, Initial chest radiographs and artificial intelligence (AI) predict clinical outcomes in COVID-19 patients: Analysis of 697 Italian patients, *Eur. Radiol.* 31 (3) (2021) 1770–1779.
- [22] T. Le Dinh, S.-H. Lee, S.-G. Kwon, K.-R. Kwon, COVID-19 chest X-ray classification and severity assessment using convolutional and transformer neural networks, *Appl. Sci.* 12 (10) (2022) 4861.
- [23] S.-J. Yoo, S.H. Yoon, J.H. Lee, K.H. Kim, H.I. Choi, S.J. Park, J.M. Goo, Automated lung segmentation on chest computed tomography images with extensive lung parenchymal abnormalities using a deep neural network, *Korean J. Radiol.* (2020).
- [25] B. van Ginneken, M.B. Stegmann, M. Loog, Segmentation of anatomical structures in chest radiographs using supervised methods: a comparative study on a public database, *Med. Image Anal.* 10 (1) (2006) 19–40.
- [26] M. Arbabshirani, A. Dallal, C. Agarwal, A. Patel, G. Moore, Accurate segmentation of lung fields on chest radiographs using deep convolutional networks, *SPIE2017*.
- [27] J.M. Wolterink, A. Mukhopadhyay, T. Leiner, T.J. Vogl, A.M. Bucher, I. Isgum, Generative Adversarial Networks: A Primer for Radiologists, *Radiographics* 41 (3) (2021) 840–857.
- [28] X. Yi, E. Walia, P. Babyn, Generative adversarial network in medical imaging: A review, *Med. Image Anal.* 58 (2019) 101552.
- [29] X. Wang, K. Yu, S. Wu, J. Gu, Y. Liu, C. Dong, Y. Qiao, C. Change Loy, Esrgan: Enhanced super-resolution generative adversarial networks. Proceedings of the European Conference on Computer Vision (ECCV), 2018.
- [30] T.-C. Wang, M.-Y. Liu, J.-Y. Zhu, A. Tao, J. Kautz, B. Catanzaro, High-resolution image synthesis and semantic manipulation with conditional gans, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2018, pp. 8798–8807.
- [31] A. Signoroni, M. Savardi, S. Benini, N. Adami, R. Leonardi, P. Gibellini, F. Vaccher, M. Ravanelli, A. Borghesi, R. Maroldi, D. Farina, BS-Net: Learning COVID-19 pneumonia severity on a large chest X-ray dataset, *Med. Image Anal.* 71 (2021) 102046.
- [32] A. Borghesi, R. Maroldi, COVID-19 outbreak in Italy: experimental chest X-ray scoring system for quantifying and monitoring disease progression, *Radiol. Med.* 125 (5) (2020) 509–513.
- [33] A.M. Tahir, M.E.H. Chowdhury, A. Khandakar, T. Rahman, Y. Qiblawey, U. Khurshid, S. Kiranyaz, N. Ibtehaz, M.S. Rahman, S. Al-Maadeed, S. Mahmud, M. Ezeddin, K. Hameed, T. Hamid, COVID-19 infection localization and severity grading from chest X-ray images, *Comput. Biol. Med.* 139 (2021) 105002.
- [34] T. Rahman, A. Khandakar, Y. Qiblawey, A. Tahir, S. Kiranyaz, S.B. Abul Kashem, M.T. Islam, S. Al Maadeed, S.M. Zugaheer, M.S. Khan, M.E.H. Chowdhury, Exploring the effect of image enhancement techniques on COVID-19 detection using chest X-ray images, *Comput. Biol. Med.* 132 (2021) 104319.
- [35] A. Degerli, M. Ahishali, M. Yamac, S. Kiranyaz, M.E.H. Chowdhury, K. Hameed, T. Hamid, R. Mazhar, M. Gabbouj, COVID-19 infection map generation and detection from chest X-ray images, *Health Information Science and Systems* 9 (1) (2021) 15.
- [36] S.H. Yoon, S.-Y. Ham, B.D. Nam, K.J. Chae, D. Lee, J.Y. Yoo, S.H. Bak, J.Y. Kim, J. H. Kim, K.B. Kim, Establishment of a nationwide korean imaging cohort of coronavirus disease 2019, *J. Korean Med. Sci.* 35 (46) (2020).
- [37] S. Inui, A. Fujikawa, M. Jitsu, N. Kunishima, S. Watanabe, Y. Suzuki, S. Umeda, Y. Uwabe, Chest CT findings in cases from the cruise ship diamond princess with coronavirus disease (COVID-19), *Radiol. Cardiothorac. Imaging* 2 (2) (2020), e200110.
- [38] Hmhospitales, Covid Data Save Lives, 2020. <https://www.hmhospitales.com/coronavirus/covid-data-save-lives/english-version>. (Accessed April 12 2021).
- [39] W. Hinrich B., L. Hans, G. Svetlana, M. Sabine K., B.H. Jan, V.-C. Jens, W. Frank K., H. Marius M., M. Bernhard C., COVID-19 Image Repository, 2020. https://figshare.com/articles/dataset/COVID-19_Image_Repository/12275009. (Accessed April 12 2021).
- [40] J.M. Bland, D.G. Altman, Statistical methods for assessing agreement between two methods of clinical measurement, *Lancet* 1 (8476) (1986) 307–310.
- [41] J.M. Bland, D.G. Altman, Measuring agreement in method comparison studies, *Stat. Methods Med. Res.* 8 (2) (1999) 135–160.
- [42] A.M. Euser, F.W. Dekker, S. le Cessie, A practical approach to Bland-Altman plots and variation coefficients for log transformed variables, *J. Clin. Epidemiol.* 61 (10) (2008) 978–982.
- [43] L. Yan, H.-T. Zhang, J. Goncalves, Y. Xiao, M. Wang, Y. Guo, C. Sun, X. Tang, L. Jing, M. Zhang, X. Huang, Y. Xiao, H. Cao, Y. Chen, T. Ren, F. Wang, Y. Xiao, S. Huang, X. Tan, N. Huang, B. Jiao, C. Cheng, Y. Zhang, A. Luo, L. Mombaerts, J. Jin, Z. Cao, S. Li, H. Xu, Y. Yuan, An interpretable mortality prediction model for COVID-19 patients, *Nature Machine Intelligence* 2 (5) (2020) 283–288.
- [44] F. Zhou, T. Yu, R. Du, G. Fan, Y. Liu, Z. Liu, J. Xiang, Y. Wang, B. Song, X. Gu, Clinical course and risk factors for mortality of adult inpatients with COVID-19 in Wuhan, China: A retrospective cohort study, *Lancet* (2020).

- [45] A.C. Alba, T. Agoritsas, M. Walsh, S. Hanna, A. Iorio, P.J. Devereaux, T. McGinn, G. Guyatt, Discrimination and calibration of clinical prediction models: Users' guides to the medical literature, *JAMA* 318 (14) (2017) 1377–1384.
- [46] E.W. Steyerberg, A.J. Vickers, N.R. Cook, T. Gerds, M. Gonen, N. Obuchowski, M. J. Pencina, M.W. Kattan, Assessing the performance of prediction models: A framework for traditional and novel measures, *Epidemiology* 21 (1) (2010) 128–138.
- [47] M.C. Tammemägi, K. Ten Haaf, I. Toumazis, C.Y. Kong, S.S. Han, J. Jeon, J. Commins, T. Riley, R. Meza, Development and validation of a multivariable lung cancer risk prediction model that includes low-dose computed tomography screening results: A secondary analysis of data from the national lung screening trial, *JAMA Netw. Open* 2 (3) (2019) e190204–e.
- [48] C.G. Walsh, K. Sharman, G. Hripcsak, Beyond discrimination: a comparison of calibration methods and clinical usefulness of predictive models of readmission risk, *J. Biomed. Inform.* 76 (2017) 9–18.
- [49] H. Choi, X. Qi, S.H. Yoon, S.J. Park, K.H. Lee, J.Y. Kim, Y.K. Lee, H. Ko, K.H. Kim, C.M. Park, Extension of coronavirus disease 2019 on chest CT and implications for chest radiographic interpretation, *Radiol. Cardiothorac. Imaging* 2 (2) (2020), e200107.
- [50] M. Balbi, A. Caroli, A. Corsi, G. Milanese, A. Surace, F. Di Marco, L. Novelli, M. Silva, F.L. Lorini, A. Duca, R. Cosentini, N. Sverzellati, P.A. Bonaffini, S. Sironi, Chest X-ray for predicting mortality and the need for ventilatory support in COVID-19 patients presenting to the emergency department, *Eur. Radiol.* 31 (4) (2021) 1999–2012.