Radiologist-supervised Transfer Learning

Improving Radiographic Localization of Pneumonia and Prognostication of Patients With COVID-19

Brian Hurt, MD, MS, Meagan A. Rubel, PhD, MPH,**
Evan M. Masutani, BS,† Kathleen Jacobs, MD,* Lewis Hahn, MD,**
Michael Horowitz, MD, PhD, Seth Kligerman, MD,**
*and Albert Hsiao, MD, PhD**

Purpose: To assess the potential of a transfer learning strategy leveraging radiologist supervision to enhance convolutional neural network-based (CNN) localization of pneumonia on radiographs and to further assess the prognostic value of CNN severity quantification on patients evaluated for COVID-19 pneumonia, for whom severity on the presenting radiograph is a known predictor of mortality and intubation.

Materials and Methods: We obtained an *initial CNN* previously trained to localize pneumonia along with 25,684 radiographs used for its training. We additionally curated 1466 radiographs from patients who had a computed tomography (CT) performed on the same day. Regional likelihoods of pneumonia were then annotated by cardiothoracic radiologists, referencing these CTs. Combining data, a preexisting CNN was fine-tuned using transfer learning. Whole-image and regional performance of the *updated CNN* was assessed using receiver-operating characteristic area under the curve and Dice. Finally, the value of CNN measurements was assessed with survival analysis on 203 patients with COVID-19 and compared against modified radiographic assessment of lung edema (mRALE) score.

Results: Pneumonia detection area under the curve improved on both internal (0.756 to 0.841) and external (0.864 to 0.876) validation data. Dice overlap also improved, particularly in the lung bases (R: 0.121 to 0.433, L: 0.111 to 0.486). There was strong correlation between radiologist mRALE score and CNN fractional area of involvement ($\rho = 0.85$). Survival analysis showed similar, strong prognostic ability of the CNN and mRALE for mortality, likelihood of intubation, and duration of hospitalization among patients with COVID-19.

Conclusions: Radiologist-supervised transfer learning can enhance the ability of CNNs to localize and quantify the severity of disease. Closed-loop systems incorporating radiologists may be beneficial for continued improvement of artificial intelligence algorithms.

Key Words: transfer learning, COVID-19, artificial intelligence, chest radiograph, chest computed tomography, patient outcomes, closed loop, radiograph

(*J Thorac Imaging* 2022;37:90–99)

Pneumonia and subsequent acute respiratory distress syndrome (ARDS) are the principal causes of death from COVID-19. Chest radiography and computed tomography (CT) play an important role in evaluating pulmonary involvement. As the pandemic has evolved, quantification of pneumonia severity has increasingly been sought as a marker of disease severity,[1–9] and standardized guides for reporting severity have emerged.[10] While CT provides exquisite details of the lung parenchyma, in the United States, it is primarily used as a problem-solving modality or to assess complications associated with COVID-19. In contrast, chest radiographs are often obtained during numerous time points throughout the course of disease.[11] Chest radiograph-based semiquantitative scoring metrics like the radiographic assessment of lung edema (RALE) have been shown to correlate with survival in ARDS,[12] found to be predictive for the likelihood of intubation and mortality, and proposed to help guide clinical management.[13–15]

Several investigators have begun to explore convolutional neural networks (CNNs) to assist with interpretation of chest radiographs. Many of the earliest approaches applied *whole-image* classification strategies based on findings extracted from radiologist reports,[16–19] and have recently applied these strategies to identify COVID-19.[2,5–9] While these studies have begun to show the diagnostic potential of CNNs, it is often difficult to interpret the reasons that the CNN makes a particular classification, a concept in machine learning known as a network's "explainability."[20] A lack of explainability currently limits the clinical utility of many algorithms. Various methods have been proposed to highlight areas of the image that are used by the CNN[21] post hoc, but these algorithms are often inconsistent or unreliable.[22]

More recently, *pixel-wise* segmentation CNNs have been proposed as an alternative strategy to whole-image classification. *Pixel-wise* segmentation CNNs provide

From the *Department of Radiology, University of California San Diego School of Medicine; and †Department of Bioengineering, University of California, San Diego, San Diego, CA.

A.H. is the senior author.

B.H. and M.A.R. are co-first authors, contributed equally to this work.

B.H. is supported by NIH T32EB005970 and Radiological Society of North America 304688-00001. E.M.M. is supported by NIH 5T32GM007198-44, American Heart Association Pre-Doctoral Fellowship 20PRE35180166, NIH T32GM007198, and NHLBI T32 HL105373.

The authors declare no conflicts of interest.

Correspondence to: Brian Hurt, MD, MS, Department of Radiology, University of California San Diego School of Medicine, 200 West Arbor Drive, MC 8756, San Diego, CA 92103 (e-mail: brhurt@health.ucsd.edu).

Supplemental Digital Content is available for this article. Direct URL citations appear in the printed text and are provided in the HTML and PDF versions of this article on the journal's website, www.thoracicimaging.com.

DOI: 10.1097/RTI.0000000000000618

natural explainability by directly localizing foci of pneumonia while achieving a diagnostic performance similar to whole-image classification CNNs.[23] Furthermore, segmentation CNNs benefit from *pixel-wise* labels that provide a more granular definition of ground truth. While labeling requires radiologists to participate in image annotation, it can allow radiologists to influence and directly teach CNNs to highlight areas of concern and enable CNNs to adapt to new data observed in the clinical environment. Transfer learning allows the CNN to incorporate knowledge from different but related source domains, and can produce highly accurate models from a smaller number of images than may be required to train a CNN from scratch.[24]

During the first wave of the pandemic in 2020, we began evaluation of a pixel-wise segmentation CNN for pneumonia detection[23] in our clinical environment.[25] We observed several flaws that were not captured in the summary statistics of performance. First, the CNN was not able to reliably detect pneumonias in the lung bases, especially the retrocardiac region behind the heart. Second, cardiothoracic radiologists easily identified smaller foci of pneumonia involving less than a whole lobe or entire lung from the clinical images, which the CNN could not identify. We thus considered the use of transfer learning to improve the performance of our CNN. We hypothesized that cardiothoracic radiologists could participate in the fine-tuning of CNNs by leveraging their ability to cross-reference findings between CT and radiographic images obtained on the same day. This might serve as a more reliable definition of ground truth for algorithm training. After performing transfer learning, we evaluated the performance of the *updated CNN* to detect viral pneumonia on patients with COVID-19 at our institution, testing its ability to prognosticate clinical outcomes as an additional benchmark of effectiveness.

MATERIALS AND METHODS

The first aim of this retrospective HIPAA-compliant and IRB-approved study sought to improve the ability of a previously trained U-net CNN (*initial CNN*) to detect and localize pneumonia on frontal chest radiographs.[23] This was accomplished by integrating a new data set, locally annotated by subspecialty cardiothoracic radiologists, through a process called transfer learning. The second aim was to assess the ability of this *updated CNN* to quantify severity of pneumonia, relative to visual scoring by subspecialty chest radiologists. The final aim assessed the effectiveness of the automated pneumonia quantification algorithm to prognosticate outcomes in patients with COVID-19.

Data and Annotations for Transfer Learning

Two data sets were used for transfer learning. First, we retrospectively curated an "internal data set" consisting of a consecutive series of 1466 frontal chest radiographs and paired chest CTs performed on the same day from patients 18 years or older from January 2020 to April 2020. No additional inclusion or exclusion criteria were used, to ensure inclusion of concurrent illnesses that occur in our local population. Foci of pneumonia were annotated on frontal radiographs based on findings on the corresponding CT. Examinations were split among 5 board-certified cardiothoracic radiologists with an average of 4.6 years (range: 2 to 12 y) postfellowship experience using in-house developed annotation software, which enabled pixel-wise probability assignment to each pixel of the image. No additional clinical information was available to the radiologist.

TABLE 1. Data Sources Used for Transfer Learning

	External	Internal	
	RSNA/NIH	Matched Cohort	COVID-19 Cohort
Radiographs	25,684	1466	203
Patients	11,171	1163	203
% AP	45%	73%	89%
% Men	56%	52%	56%
Mean age (range)	47 (1-92)	57 (18-98)	55 (19-100)
% PNA	22%	48%	86%
Application	PNA localization	PNA localization	Clinical evaluation

The updated convolutional neural network (CNN) was trained using a combination of radiographs and annotations, including an internal "matched" cohort of patients who underwent chest radiography and computed tomography (CT) on the same day, and an external data set.

AP indicates anterior-posterior; NIH, National Institute of Health; PNA, pneumonia; RSNA, Radiological Society of America.

Second, we obtained an "external data set" comprising 25,684 radiographs along with their bounding box annotations of pneumonia.[19,26] These same radiographs and annotations were also used in the training of a previous CNN,[23] which we refer to as the *initial CNN*.

Data were split by patient with ~80% used for training and 20% for evaluation. An overview of the data sources used for training and their data split for evaluation is provided in Table 1 and Figure 1.

Neural Network Training

To improve the performance of the CNN with the additional internal data, we had to solve 2 problems, which are conceptualized in Figure 2. First, we had to identify the optimal balance of external and internal data that would maximize the performance of the CNN. Second, we had to select between multiple potential loss functions that could optimize performance. We thus conducted a hyperparameter search, simultaneously searching across these two groups of variables, which produced 102 candidate CNNs. The details of the hyperparameter search are provided in the supplemental materials, Supplemental Digital Content 1 (http://links.lww.com/JTI/A205). Candidate CNNs were ranked based on area under the receiver-operating curve (AUC) and Dice similarity of overlap for their ability to detect and localize pneumonia (detailed further below) from the internal evaluation cohort. A single CNN with the highest AUC and Dice was selected from these candidates as the updated CNN for subsequent analysis.

An additional CNN was trained from scratch using only the internal data to provide an additional benchmark for comparison. This de novo CNN was identical in structure as the initial CNN, trained from random initial weights with the same loss function used to train the updated CNN. CNN training was carried out by a radiology resident (blinded) using a NVIDIA cloud cluster of 32 GV100s leveraging Kubernetes (Linux Foundation, https://www.kubernetes.io) running Ubuntu 18.04 (Linux Foundation, https://www.ubuntu.org) using the TensorFlow 2.0 library[27] for the Python 3.8 programming language (Python Software Foundation, https://www.python.org).

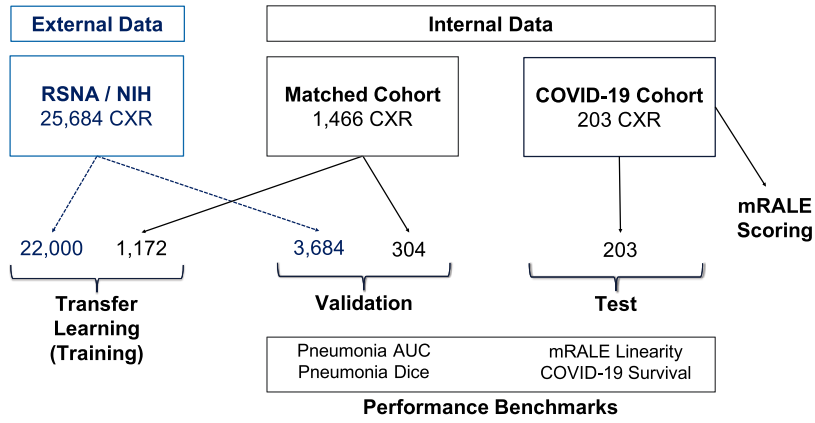

FIGURE 1. Data sources and performance benchmarks for CNN training, validation, and testing. We retrospectively obtained data from two cohorts of patients to first fine-tune a prior CNN, and then evaluate the CNN on patients with COVID-19 pneumonia. Technical performance of the algorithm was technically evaluated with receiver-operating characteristic area under the curve (ROC AUC) and Dice overlap of segmentations. Algorithm clinical performance was evaluated in the second patient population by assessing colinearity with radiologist-modified radiographic assessment of lung edema (mRALE) scores and survival analyses. full color online

Postprocessing and Quantification of Regional Severity

To quantify the severity of pneumonia, we applied post-processing to the resulting probability map generated by the CNNs. We created a separate CNN to segment the right and left lung (described in supplemental materials, Supplemental Digital Content 1, http://links.lww.com/JTI/A205), which we then used to divide the lungs into upper, middle, and lower lung zones. The probability map generated by the CNN was then multiplied by the lung zone masks to estimate regional involvement of pneumonia. We then constructed three metrics of severity: Maximum probability was defined as the maximum probability in each region; mean probability was defined as the mean within each region; and the fractional area was defined as the fraction of the region exceeding a probability of 50%. A detailed methodology is provided in the supplemental materials, Supplemental Digital Content 1 (http://links.lww.com/JTI/A205).

Evaluation of Pneumonia Detection and Localization

Whole-image pneumonia detection performance was evaluated on the *initial*, de novo, and *updated CNN*s using

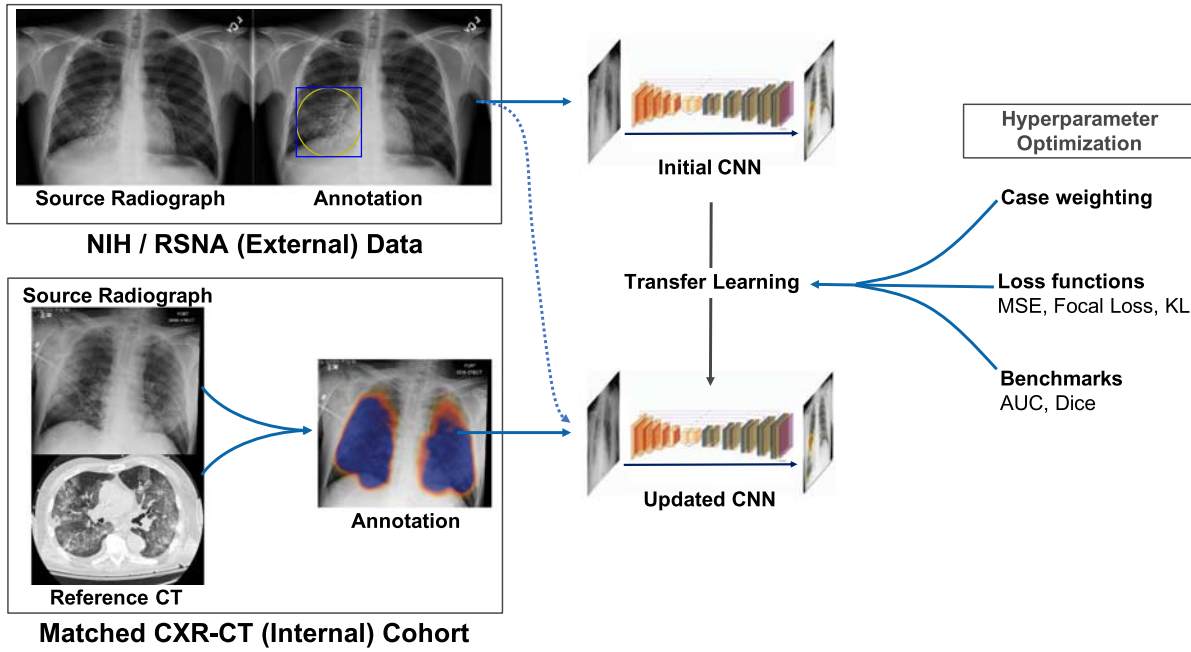

FIGURE 2. Transfer learning training strategy for CNN fine-tuning with enhanced ground truth. An initial CNN trained on external image data was refined on images and annotations of pneumonia from patients with chest x-ray and computed tomography that was obtained on the same day. Hyperparameters of loss function and training data that balanced multiple data sources were used to optimize the CNN's detection of pneumonia. KLD indicates Kullback–Leibler divergence; NIH, National Institute of Health; RSNA, Radiological Society of America. full color online

Copyright © 2021 Wolters Kluwer Health, Inc. All rights reserved.

sequestered validation cohort, comprising 304 internal and 3684 external radiographs. For each CNN, we compared AUCs for both internal and external data. Dice similarity was compared only on internal data with its higher quality ground truth annotation. Pneumonia detection ROCs were constructed by varying the threshold on the inferred probability maps, while setting a binary threshold on the ground truth annotations. Sensitivities, specificities, positive and negative predictive values, and accuracy were calculated at an operating point that equally maximized sensitivity and specificity (Youden *J* index).[28]

To assess regional performance, we additionally performed the same analyses as above, using only the 304 internal radiographs with high-quality ground truth annotations. We also evaluated the performance of our lung segmentation CNN with Dice similarity coefficient, comparing ground truth annotations to the inferred masks. Statistical analyses were performed using the SciPy package in python with 2-sided paired *t* tests and a type I error rate of 0.05. To compare AUC, we applied bootstrap sampling with 80% of the data to evaluate statistical significance between CNNs.

Prognostication in Patients With COVID-19

To assess the ability of the *updated CNN* to prognosticate hospital outcomes, we retrospectively obtained an additional independent sample of 1479 chest radiographs between March and July of 2020 from patients with RT-PCR-confirmed COVID-19 (Fig. 1). Each of the chest radiographs from this cohort was independently scored by 2 readers, evenly split among 5 cardiothoracic radiologists. The density and extent of the radiographic opacities were scored using a modified radiographic assessment of lung edema (mRALE) scoring system as previously described in Li et al.[1] The mRALE score is calculated based on visual assessment of the extent and density of airspace disease and range from 0 (normal chest radiograph) to a maximum of 24 (complete consolidation of both lungs). Inter-reader mRALE agreement between radiologists was assessed by linear Cohen κ.

Of these 1479 radiographs, 203 were performed on unique patients within the first 3 days of presentation or admission. None of these patients was included in algorithm training. Of these, 7% were obtained in the outpatient setting, 58% in the ER, 35% in the inpatient setting, 16% in the intensive care unit, and 12% were intubated. Dates of admission, discharge, intubation, and death were collected from the medical record. Kaplan-Meier curves for 3 outcomes (intubation, mortality, duration of hospitalization) and correlation analyses were performed on 203 COVID-19+ patients, using the radiographs taken within the first 3 days of presentation. mRALE scores were averaged between the 2 readers for each of these radiographs. Correlation between mean mRALE score with severity score (maximum probability, mean probability, and fractional area) was measured using the Pearson correlation coefficient. For survival analysis, mRALE scores were divided into 4 categories of severity (0 to 6, 7 to 12, 13 to 18, and 19 to 24), and each CNN severity score was divided into quartiles. Survival analyses were performed using the survival and survminer[29] packages in R. To assess the statistical significance of stratification between scores or quartiles, we conducted post hoc pairwise comparisons of each quartile using the log-rank statistic with Benjamini-Hochberg multiple test correction.

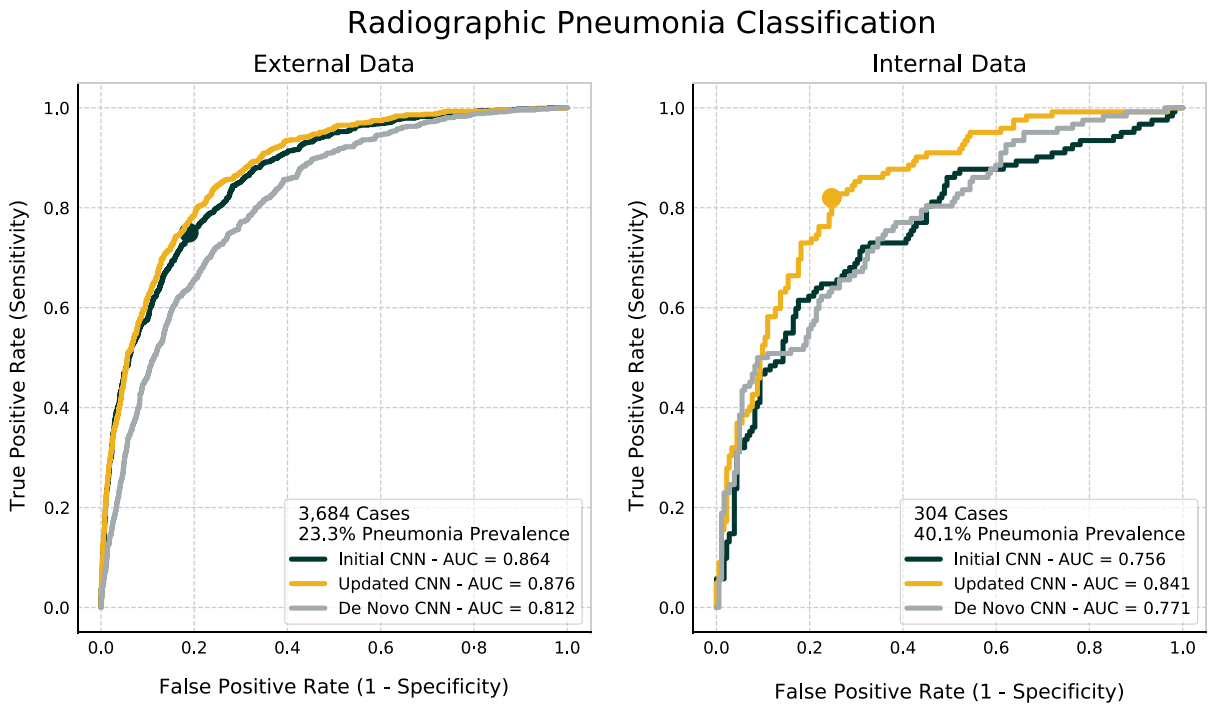

FIGURE 3. Improved performance of CNN pneumonia detection with transfer deep learning. The *updated CNN* (yellow) significantly outperformed the *initial CNN* (green) on both external (left) and internal (right) validation data sets. The AUC of pneumonia detection on the internal data set improved from 0.756 to 0.841 (right panel green to yellow; $P = 2.0e{-}4$) and from 0.864 to 0.876 on external image data (left panel green to yellow; $P = 3.8e{-}3$). Finally, the de novo *CNN* (gray), trained with only internal data, significantly underperformed the *updated CNN*. Operating points (circles) for *initial* and *updated CNN*s were defined by equally maximizing sensitivity and specificity (Youden *J* Index) applied to the external and internal data sets, respectively. full color online

TABLE 2. Performance of the CNN for Whole-image Detection of Pneumonia

	External (RSNA/NIH) 23% Pneumonia Prevalence		Internal (Matched Cohort) 40% Pneumonia Prevalence	
	Initial CNN*	Updated CNN	Initial CNN	Updated CNN†
AUC	0.864	0.876	0.756	0.841
	$P < 2.6e{-3}$		$P < 1.0e{-4}$	
Model probability threshold	0.64	0.71	0.64	0.71
Sensitivity	0.75	0.95	0.40	0.82
Specificity	0.81	0.52	0.92	0.75
Accuracy	0.80	0.62	0.71	0.78
Negative predictive value	0.91	0.97	0.70	0.86
Positive predictive value	0.55	0.38	0.77	0.69

After employing transfer learning, the convolutional neural network (CNN) showed significant improvement in AUC on internal and external validation data. CNN operating points were defined by Youden's index for *initial CNN* when applied to the external data set (*) and on the *updated CNN* using the internal data set (†). The *updated CNN* markedly improved in sensitivity with a modest loss in specificity when evaluating internal chest radiographs with the operating point defined by the Youden index. These CNN fine-tuning methods improved the overall negative predictive value and the overall accuracy in our clinical images.

NIH indicates National Institute of Health; RSNA, Radiological Society of America.

RESULTS

Selection of the Optimal CNN Algorithm

The top 10 candidate CNNs are listed in Supplemental Table 1 (Supplemental Digital Content 2, http://links.lww.com/JTI/A206). We selected the top performing candidate (30x Pixel-Weighted mean squared error) CNN and refer to this as our *updated CNN* for all subsequent analyses. The CNN with the maximum performance was optimized using a mean squared error loss function with a 30-fold weighting of pixels exceeding 20% on the ground truth pneumonia probability map. This CNN used an external training data mix of 1200 negative cases and 600 positive cases for each epoch of training.

Whole-Image Pneumonia Detection and Localization Performance

The *updated CNN* significantly outperformed the *initial CNN* for detection of pneumonia on both the internal and external validation data sets (Fig. 3, Table 2). AUC improved on the internal validation data set from 0.756 to 0.841 ($P < 1e{-4}$). Similarly, AUC on the external validation data set improved from 0.864 to 0.876 ($P = 2.6e{-3}$). In addition, pneumonia localization improved on the internal validation data set with a mean Dice improvement of 0.147 to 0.332 ($P < 1e{-3}$). The *updated CNN* also outperformed the de novo *CNN*, which had an AUC of 0.771 for internal data and 0.812 for external data. Comparisons on both data sets were statistically significant ($P < 1e{-7}$). Dice overlap for the de novo *CNN* was similar to the *updated CNN* on internal data, 0.295, without a statistically significant difference.

Regional Pneumonia Detection and Localization Performance

The lung segmentation CNN achieved a Dice mean and a SD of 0.869 ± 0.084, despite training on only 237 chest radiographs. On the portion of the internal data set reserved for validation (n = 304), AUC for detection of pneumonia improved from 0.739 to 0.812 for the right lung ($P = 1.0e{-3}$) and from 0.776 to 0.848 on the left lung ($P = 1.5e{-2}$). We observed the largest AUC improvement in the lower lung regions, from 0.747 to 0.808 ($P = 2.1e{-2}$) on the right and from 0.824 to 0.878 ($P = 3.7e{-2}$) on the left (see Table 3 for complete regional detection performance). Similarly, mean Dice scores for areas marked as involved with pneumonia improved from 0.154 to 0.333 ($P = 6.0e{-6}$) for the right lung and from 0.161 to 0.395 ($P = 1.6e{-2}$) in the left lung. We observed the biggest improvement in the lower lung regions, increasing from 0.121 to 0.433 ($P = 2.4e{-11}$) for the right

TABLE 3. Performance of the CNN for Regional Classification of Pneumonia on the Internal Data Set

	Updated CNN (95% CI)	Initial CNN (95% CI)	Mean Difference (95% CI), *P*
Lungs	0.841 (0.796-0.883)	0.756 (0.699-0.814)	0.085 (0.041-0.130), <1.0e−04
Right	0.812 (0.757-0.861)	0.739 (0.680-0.798)	0.072 (0.032-0.114), 1.0e−03
Upper	0.791 (0.709-0.870)	0.771 (0.686-0.852)	0.019 (−0.053-0.093), 6.1e−01
Middle	0.825 (0.770-0.874)	0.777 (0.717-0.836)	0.048 (0.008-0.089), 1.5e−02
Lower	0.808 (0.753-0.859)	0.747 (0.680-0.810)	0.061 (0.010-0.114), 2.1e−02
Left	0.848 (0.793-0.900)	0.776 (0.713-0.838)	0.072 (0.015-0.131), 1.5e−02
Upper	0.826 (0.750-0.892)	0.846 (0.768-0.912)	−0.020 (−0.077-0.037), 1.5e+00
Middle	0.871 (0.815-0.925)	0.824 (0.756-0.886)	0.047 (0.001-0.096), 4.4e−02
Lower	0.878 (0.833-0.917)	0.824 (0.768-0.881)	0.054 (0.003-0.106), 3.7e−02

The *updated CNN* significantly outperformed the *initial CNN* across nearly all lung regions, with the largest improvements occurring at the lung bases. For each region, the AUC and Dice confidence intervals were calculated for each model using a bootstrap method (10,000 iterations). From these distributions, pairwise mean AUC differences, confidence intervals (CI), and *P*-values (2-sided *t* test) were calculated. The regional CNN advantage was determined by the mean difference and the associated *P*-value.

Copyright © 2021 Wolters Kluwer Health, Inc. All rights reserved.

TABLE 4. Performance of the CNN for Regional Localization of Pneumonia on the Internal Data Set

	Updated CNN [IQR]	Initial CNN [IQR]	Mean Difference [IQR], *P*
Lungs	0.332 [0.075-0.503]	0.147 [0.000-0.285]	0.185 [0.000-0.339], 5.3e−08
Right	0.333 [0.026-0.552]	0.154 [0.000-0.244]	0.180 [0.000-0.332], 6.0e−06
Upper	0.395 [0.133-0.640]	0.161 [0.000-0.272]	0.234 [0.000-0.404], 9.6e−08
Middle	0.322 [0.000-0.685]	0.232 [0.000-0.524]	0.090 [0.000-0.229], 2.1e−01
Lower	0.343 [0.030-0.544]	0.197 [0.000-0.315]	0.147 [0.000-0.251], 2.6e−03
Left	0.433 [0.078-0.683]	0.121 [0.000-0.228]	0.312 [0.005-0.589], 2.4e−11
Upper	0.293 [0.000-0.649]	0.147 [0.000-0.245]	0.146 [0.000-0.441], 2.8e−02
Middle	0.381 [0.062-0.720]	0.242 [0.000-0.499]	0.139 [0.000-0.237], 1.6e−02
Lower	0.486 [0.267-0.723]	0.111 [0.000-0.077]	0.375 [0.075-0.636], 3.9e−15

The *updated CNN* significantly outperformed the *initial CNN* across nearly all lung regions with the largest improvements occurring at the lung bases, most notably a >4-fold increase at the left lung base. For each region and CNN, the mean Dice interquartile range (IQR) was calculated. Pairwise Dice differences, IQR, and *P*-values (2-sided *t* test) were calculated. The regional CNN advantage was determined by the mean Dice difference and the associated *P*-value.

lung and from 0.188 to 0.443 ($P < 3.9e{-}15$) for the left lung (see Table 4 for complete regional localization performance).

Exemplar cases are highlighted in Figures 4–7. Figure 4 illustrates the relationship between the radiologist's CT-aided annotation and updated CNN's inferred severity of pneumonia. Figure 5 illustrates the *updated CNN*'s improved sensitivity for foci of COVID-19 pneumonia in a patient who had a CT performed hours after the radiograph. Figure 6 illustrates the improvement in sensitivity of the *updated CNN* for more subtle opacities of COVID-19 pneumonia, as it blooms over several days. Figure 7 illustrates the regions of lung involvement inferred by the *updated CNN* in three additional individuals with COVID-19.

Severity Score and Survival Analysis

The inter-rater correlation for mRALE scores for 1479 radiographs scored by 5 cardiothoracic radiologists was substantial (linear Cohen κ, mean: 0.72). For the 203 radiographs that were obtained within 3 days of initial presentation, there was strong agreement between mRALE score and each of the metrics from the *updated CNN*: mean probability ($\rho = 0.86$, $P < 2.2e{-}16$), fractional area ($\rho = 0.85$, $P < 2.2e{-}16$), and maximum probability ($\rho = 0.64$, $P < 2.2e{-}16$) (Supplemental Table 2, Supplemental Digital Content 2, http://links.lww.com/JTI/A206).

As anticipated, survival analysis showed that patients with the lowest mRALE score had the best median survival, lowest probability of intubation, and shortest duration of hospital stay (Fig. 8). Patients with the highest mRALE score had the opposite result. CNN estimates of severity showed similar stratification. Notably, mean probability and fractional area both strongly stratified patients for all three clinical end points, though mRALE scores averaged between 2 radiologists was superior for prognosticating mortality. A low "maximum probability" estimated by the CNN was a strong predictor of immediate discharge without the need for hospitalization. A complete list of log-rank pairwise comparisons with Benjamini-Hochberg correction are provided in Supplemental Table 1 (Supplemental Digital Content 2, http://links.lww.com/JTI/A206).

DISCUSSION

In this study, we demonstrate the flexibility and plasticity of CNNs to learn from expert supervision by subspecialist cardiothoracic radiologists and show an improved ability to detect and localize pneumonia. We observed that the performance of the CNN trained initially only on external image data did not perform well on radiographs performed at our institution, as is often expected.[18] Similarly, the performance of the de novo CNN trained solely

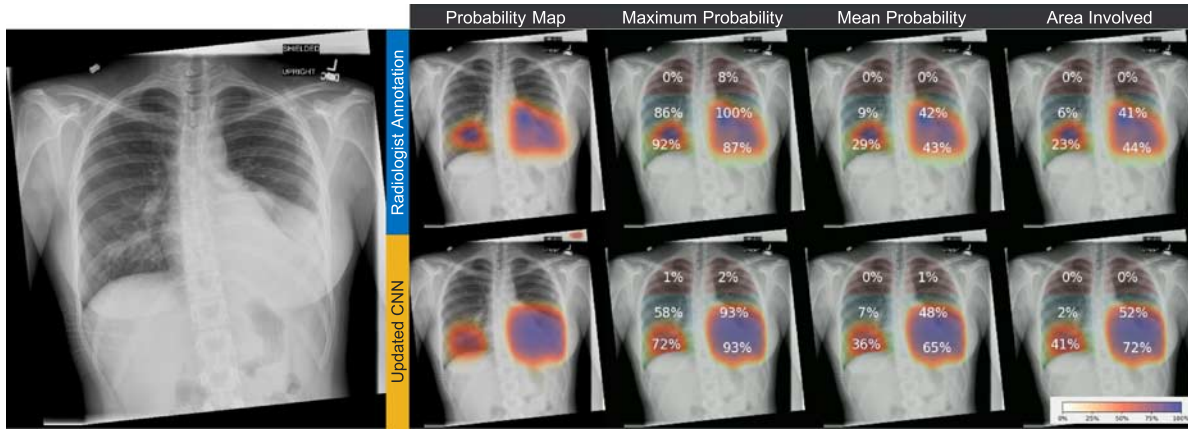

FIGURE 4. Quantification of the regional severity of pneumonia. Results are shown from a patient in the validation set. Manual annotations by a cardiothoracic radiologist (top row) closely matched the regions of pneumonia detected by the updated convolutional neural network (CNN) (bottom row). Regional quantitative measurements from manual radiologist annotation and the CNN were similar. full color online

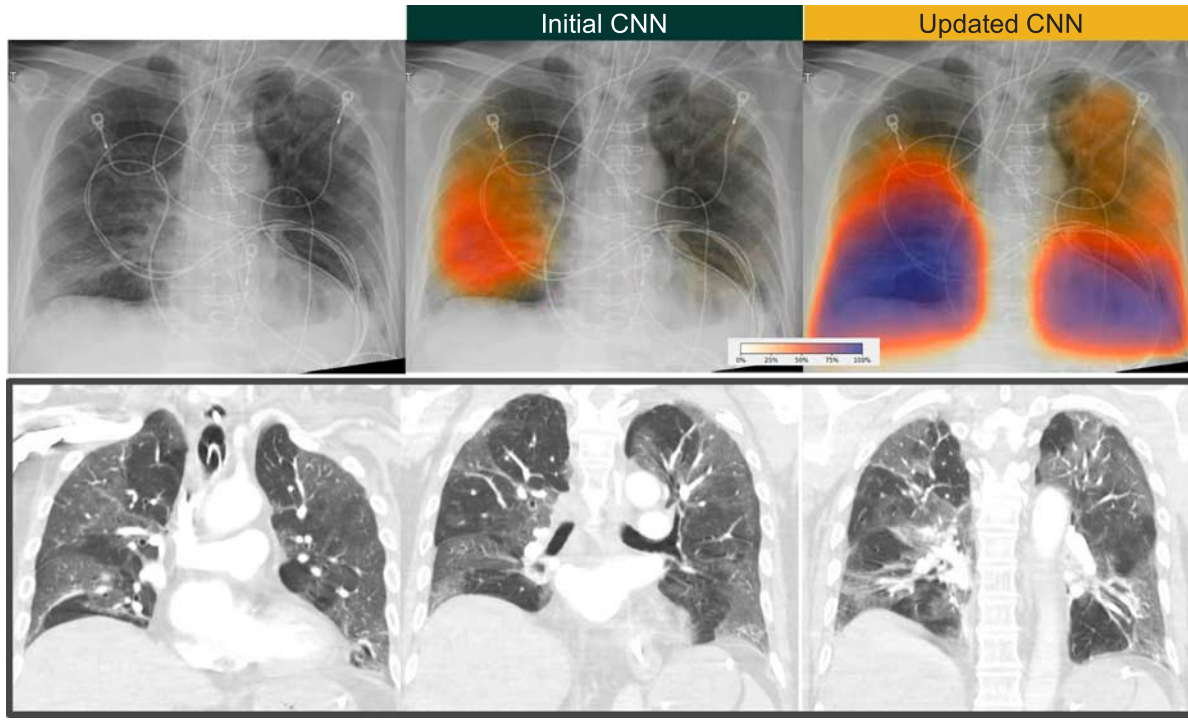

FIGURE 5. Improved pneumonia localization in a patient with COVID-19. Chest radiograph and coronal CT PE images in a 66-year-old male with a history of a cardiac transplant and PCR+ COVID-19, who presented with acute hypoxemic respiratory failure. The updated CNN (top right) better localizes areas of ground glass than the initial CNN (top middle), which are confirmed by CT performed several hours later (bottom row), which shows peripheral and basal predominant ground glass opacities consistent with COVID-19 pneumonia. full color online

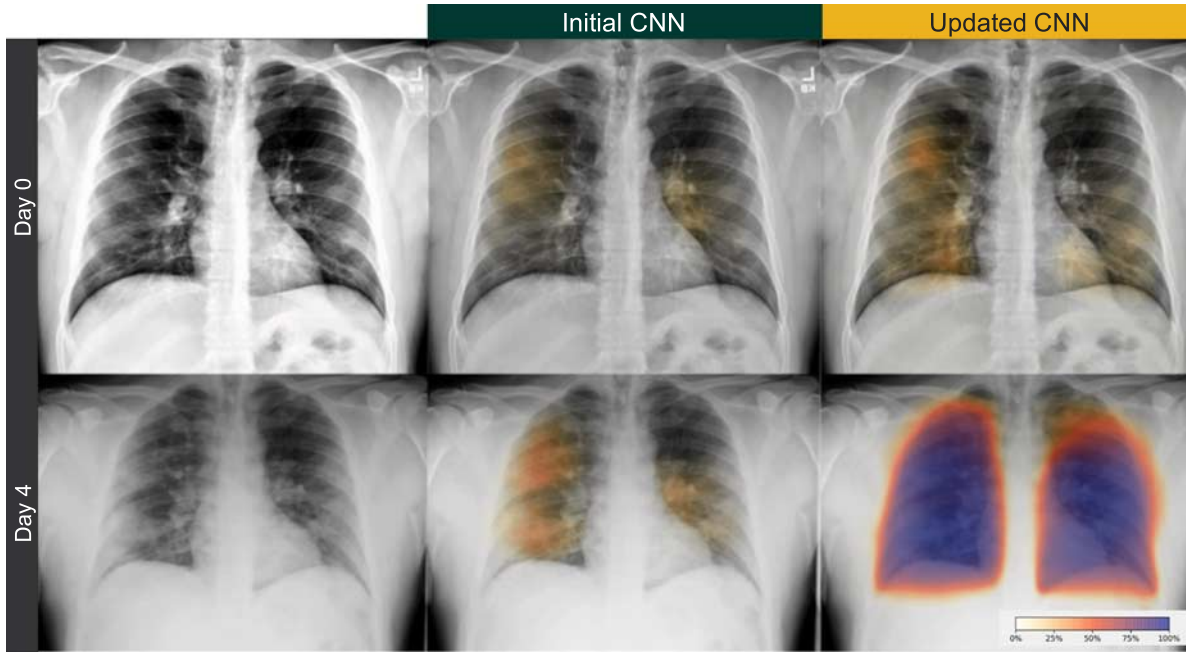

FIGURE 6. Longitudinal change in pneumonia in a patient with COVID-19. This 42-year-old man initially presented with nasal congestion, minimal cough, intermittent sweats, and no shortness of breath. COVID-19 RT-PCR was positive on day 0 and he was discharged to home self-isolation. The patient returned on day 4 with acute worsening of shortness of breath, fever, chills, myalgias, arthralgias, anosmia, cough, pleuritic chest pain, and was admitted with sepsis. The patient was discharged home on day 10. Subtle ill-defined opacities are present on the initial chest x-ray, which bloom considerably 4 days later, and are highlighted with greater certainty by the *updated CNN* algorithm. full color online

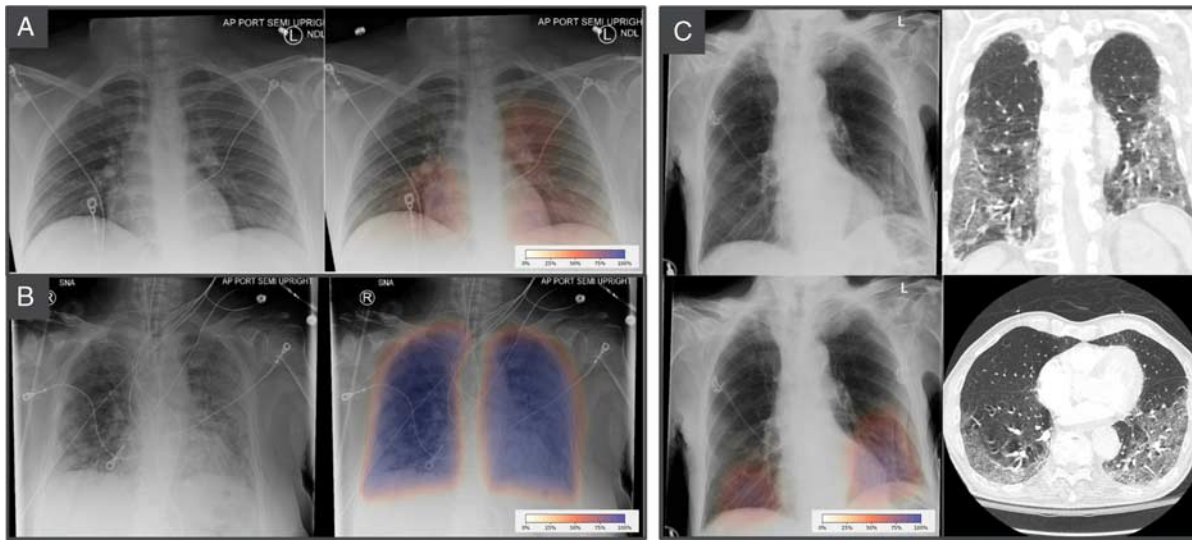

FIGURE 7. Updated CNN pneumonia localization on radiographs from three patients with COVID-19 pneumonia. A, Subtle bilateral perihilar and lower lung opacities detected with intermediate confidence by the updated CNN. B, Diffuse bilateral opacities in an intubated patient detected with high confidence by the updated CNN. C, A chest radiograph with peripherally predominant bilateral basal opacities, confirmed by CT 2 hours later. full color online

on a relatively small number of cases from our institution showed relatively weak performance. The optimal CNN was ultimately found leveraging a combination of both data sources. Interestingly, the de novo CNN showed greater performance on external data than internal data. We speculate that this was because the internal data included more

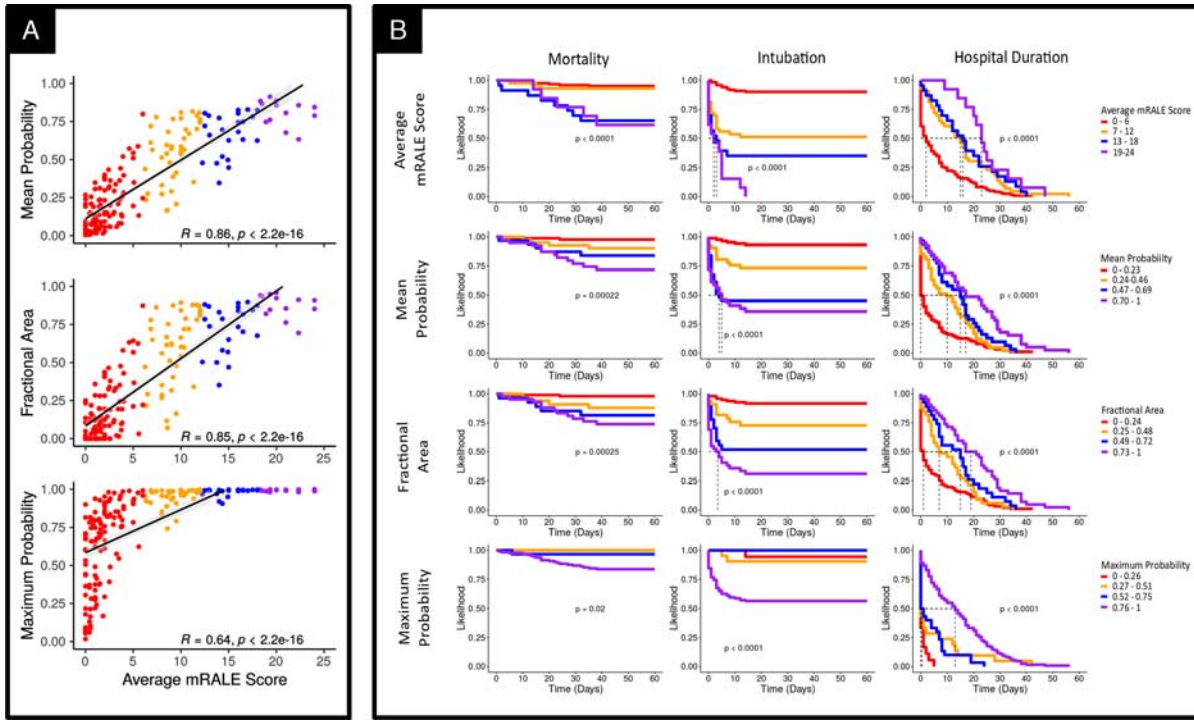

FIGURE 8. CNN pneumonia severity score and radiologist visual score of x-rays of patients with COVID-19. A, Correlation to radiologists' visual scoring: convolutional neural network (CNN) severity metrics (maximum probability, mean probability, and fractional area involvement) correlated well with visual scores. Modified radiographic assessment of lung edema (mRALE) scores are divided into colored quartiles. The mean probability and fractional area are linearly correlated with mRALE scores. Maximum probability shows a nonlinear relationship with mRALE. B, Survival analysis of patients with COVID-19 based on x-rays at initial presentation: stratifying patients based on radiographs obtained within the first 3 days of presentation or hospital admission strongly prognosticated mortality, likelihood of intubation, and duration of hospitalization. Visual mRALE score strongly separated patients for all 3 survival analyses. CNN severity measurements of disease severity also strongly separated patients. full color online

patients with subtle, smaller foci of pneumonia, which made the "internal" task more challenging. Other explanations for the difference in performance may include differences in equipment, image preprocessing, downsampling strategies, and quality of image annotation. There may have been differences in patient factors as well, including differences in demographics, body habitus, frequency of concurrent disease like cancer or heart failure, and types and severity of pneumonia.

Using a transfer learning approach, we were able to specifically improve the localization of lower lobe pneumonias, which were not well addressed by the *initial CNN*. In addition, training a de novo CNN showed inferior results compared with our *updated CNN,* suggesting that transfer learning may be a better approach for extending generalizability of CNN algorithms across institutions. Specifically, we highlighted how this transfer learning strategy can maximize performance of a CNN by combining and balancing the benefit of two distinct data sets: (a) a smaller number of chest radiographs with more precisely defined ground truth and (b) a larger volume of radiographs with less precisely defined ground truth. This strategy is made feasible because of our choice to use a *segmentation CNN* called a U-Net, which provides natural explainability through its production of image maps that can be readily interpreted by a supervising radiologist and engage this as a natural human–machine interface.[30]

Much of the existing literature has emphasized *classification* algorithms[16–19,31] and have shown impressive performance without explicit radiologist annotations, with AUCs for pneumonia detection ranging from 0.633 to 0.911.[16–19,31] Classification CNNs are an attractive approach because they do not require manual radiologist labeling and localization of the findings on chest radiograph, but generally require very large data sets on the order of hundreds of thousands of chest radiographs to achieve a high level of performance. However, they often lack clear explainability to their results, requiring *post-hoc* methods to reveal their rationale for classification.[32] Furthermore, it is unclear how classification approaches might benefit from radiologist supervision. In contrast, we show that by leveraging an alternative *segmentation* approach, it is possible to markedly improve performance of a pretrained CNN to perform better in our clinical environment after incorporating training with a modest number (1,172) of additional radiographs, while substantially increasing AUC on radiographs in our clinical environment from 0.756 to 0.841. This result highlights an opportunity for radiologists to participate in the tuning of CNN algorithms for clinical use. While the development of AI algorithms has been considered by many to be the domain of industry or research laboratories, these results suggest that radiologists may play an essential role in the training and tuning of CNNs for their local environments.

Using a segmentation strategy also yields other benefits, including the simultaneous quantification of disease. We show that with it, it is feasible to accomplish both detection and segmentation of pneumonia with a single segmentation CNN, which can be further leveraged to quantify disease severity. The performance of this strategy is comparable to the recently described dedicated algorithms for grading severity of pneumonia.[25] In addition, we find that measurements made through our CNN provide a strong prognostic value, particularly among patients with COVID-19 at our institution; they were able to stratify patients that required longer durations of hospitalization, required intubation, or ultimately succumbed to COVID-19. Furthermore, it is important to note that severity scoring of pneumonia is not routinely performed at most institutions as part of routine clinical practice. CNNs may fill new roles in diagnostic radiology as they are able to automatically track disease severity and prognosticate patient outcomes to assist in patient triage or management, as deployed into the clinical environment.[25,33]

The strategy outlined in this study is one of the several possible approaches to improve a pneumonia detection/localization CNN. Other ways to improve the CNN's performance may include preprocessing radiographs to exclude rib shadows,[34] altering the CNN architecture to additionally predict whole-image pneumonia likelihood or severity, and other transfer learning techniques such as differential CNN weight freezing during training. Whatever the technique, understanding how the data and the loss functions affect the training is pivotal to CNN improvement.

There are several limitations to this study and its proof of technical feasibility. First, the proposed algorithm does not incorporate clinical factors such as symptomatology, body temperature, or supporting laboratory findings, which are necessary for the diagnosis of pneumonia. Future algorithm improvements may benefit from integrating nonimaging clinical data. Second, our lung segmentation's performance does not approach that of similar CNN-based techniques.[35] In the future, our algorithm may be improved through using more training examples, using other CNN architectures, or non-CNN computer vision techniques that have proven effective in lung segmentation.[36] Additional improvements could include converting regional lung zone segmentations to the lobar anatomic correlates using lateral radiographs. Third, our algorithm was generated from patients at one academic institution in the United States and may benefit from additional data sources to ensure broad generalizability. Nevertheless, as emphasized earlier, we anticipate that continuous learning may become an important facet of this technology. It remains unclear how algorithms may improve through incorporation of multi-institutional data sets, fine-tuning that may be required to extend across regional populations, and control for technical differences. The strategy that we have highlighted here may be primarily beneficial for the latter caveat. Finally, we only explored survival analyses from a cross-section of COVID-19 patients at a single time point of their initial presentation. Longitudinal analyses incorporating chest radiographs and their temporal evolution may further improve prognostic value.

We successfully show that a transfer learning strategy incorporating radiologist-defined ground truth is feasible and can serve as an important strategy to improve CNN performance. This may be necessary for CNNs to perform effectively across new and constantly changing clinical environments. As we have observed from the COVID-19 pandemic, the practice of diagnostic radiology is dynamic and constantly evolving. To maximize their clinical value, artificial intelligence systems may benefit if designed to continuously learn from radiologist expertise.

ACKNOWLEDGMENTS

The authors would like to acknowledge research grant support from NSF Grant 2026809, UCOP TRDRP R00RG2480 and in-kind support from Microsoft AI for Health, NVIDIA and GroupWare.

REFERENCES

1. Li MD, Arun NT, Gidwani M, et al. Automated assessment and tracking of COVID-19 pulmonary disease severity on chest radiographs using Convolutional Siamese Neural Networks. *Radiol Artif Intell*. 2020;2:e200079.
2. Khan AI, Shah JL, Bhat MM. CoroNet: a deep neural network for detection and diagnosis of COVID-19 from chest x-ray images. *Comput Methods Programs Biomed*. 2020;196:105581.
3. Ni Q, Sun ZY, Qi L, et al. A deep learning approach to characterize 2019 coronavirus disease (COVID-19) pneumonia in chest CT images. *Eur Radiol*. 2020;30:6517–6527.
4. Minaee S, Kafieh R, Sonka M, et al. Deep-COVID: predicting COVID-19 from chest x-ray images using deep transfer learning. *Med Image Anal*. 2020;65:101794.
5. Oh Y, Park S, Ye JC. Deep learning COVID-19 features on CXR using limited training data sets. *IEEE Trans Med Imaging*. 2020;39:2688–2700.
6. Apostolopoulos ID, Mpesiana TA. Covid-19: automatic detection from x-ray images utilizing transfer learning with convolutional neural networks. *Phys Eng Sci Med*. 2020;43:635–640.
7. Vaid S, Kalantar R, Bhandari M. Deep learning COVID-19 detection bias: accuracy through artificial intelligence. *Int Orthop*. 2020;44:1539–1542.
8. Apostolopoulos ID, Aznaouridis SI, Tzani MA. Extracting possibly representative COVID-19 biomarkers from x-ray images with deep learning approach and image data related to pulmonary diseases. *J Med Biol Eng*. 2020:1–8.
9. Das D, Santosh KC, Pal U. Truncated inception net: COVID-19 outbreak screening using chest x-rays. *Phys Eng Sci Med*. 2020;43:915–925.
10. Litmanovich DE, Chung M, Kirkbride RR, et al. Review of chest radiograph findings of COVID-19 pneumonia and suggested reporting language. *J Thorac Imaging*. 2020;35:354–360.
11. Rubin GD, Haramati LB, Kanne JP, et al. The role of chest imaging in patient management during the COVID-19 pandemic: a Multinational Consensus Statement from the Fleischner Society. *Radiology*. 2020;158:106–116.
12. Warren MA, Zhao Z, Koyama T, et al. Severity scoring of lung oedema on the chest radiograph is associated with clinical outcomes in ARDS. *Thorax*. 2018;73:840–846.
13. Bahl A, Van Baalen MN, Ortiz L, et al. Early predictors of in-hospital mortality in patients with COVID-19 in a large American cohort. *Intern Emerg Med*. 2020;15:1485–1499.
14. Wu C, Chen X, Cai Y, et al. Risk factors associated with acute respiratory distress syndrome and death in patients with coronavirus disease 2019 pneumonia in Wuhan, China. *JAMA Intern Med*. 2020; 180:934–943.
15. Zhang L, Zhu F, Xie L, et al. Clinical characteristics of COVID-19-infected cancer patients: a retrospective case study in three hospitals within Wuhan, China. *Ann Oncol*. 2020;31: 894–901.
16. Irvin J, Rajpurkar P, Ko M, et al. CheXpert: a large chest radiograph dataset with uncertainty labels and expert comparison. 2019. Available at: http://arxiv.org/abs/1901.07031. Accessed August 31, 2020.
17. Rajpurkar P, Irvin J, Ball RL, et al. Deep learning for chest radiograph diagnosis: a retrospective comparison of the CheX-NeXt algorithm to practicing radiologists. *PLoS Med*. 2018; 15:e1002686.
18. Zech JR, Badgeley MA, Liu M, et al. Variable generalization performance of a deep learning model to detect pneumonia in chest radiographs: a cross-sectional study. *PLoS Med*. 2018;15:1–17.
19. Wang X, Peng Y, Lu L, et al. ChestX-ray8: hospital-scale chest x-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases. *2017 IEEE Conf Comput Vis Pattern Recognition(CVPR)*. 2017:2097–2106.
20. Holzinger A, Langs G, Denk H, et al. Causability and explainability of artificial intelligence in medicine. *Wiley Interdiscip Rev Data Min Knowl Discov*. 2019;9:e1312.
21. Reyes M, Meier R, Pereira S, et al. On the interpretability of artificial intelligence in radiology: challenges and opportunities. *Radiol Artif Intell*. 2020;2:e190043.
22. Arun NT, Gaw N, Singh P, et al. Assessing the validity of saliency maps for abnormality localization in medical imaging. *Radiol Artif Intell*. 2020:2. Available at: http://arxiv.org/abs/2006.00063.
23. Hurt B, Yen A, Kligerman S, et al. Augmenting interpretation of chest radiographs with deep learning probability maps. *J Thorac Imaging*. 2020;35:285–293.
24. Shin HC, Roth HR, Gao M, et al. Deep convolutional neural networks for computer-aided detection: CNN architectures, dataset characteristics and transfer learning. *IEEE Trans Med Imaging*. 2016;35:1285–1298.
25. Carlile M, Hurt B, Hsiao A, et al. Deployment of artificial intelligence for radiographic diagnosis of COVID-19 pneumonia in the emergency department. *J Am Coll Emerg Physicians Open*. 2020;1:1459–1464.
26. Shih G, Wu CC, Halabi SS, et al. Augmenting the national institutes of health chest radiograph dataset with expert annotations of possible pneumonia. *Radiol Artif Intell*. 2019;1:e180041.
27. Abadi M, Barham P, Chen J, et al. TensorFlow: a system for large-scale machine learning. *Proceedings of the 12th USENIX Symposium on Operating Systems Design and Implementation, OSDI 2016*. 2016.
28. Youden WJ. Index for rating diagnostic tests. *Cancer*. 1950;3: 32–35.
29. Kassambara A, Kosinski M, Biecek P, et al. Survminer: drawing survival curves using "ggplot2" (R package). version 0.4.3.
30. Ronneberger O, Fischer P, Brox T. U-net: convolutional networks for biomedical image segmentation. Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics). *MICCAI*. 2015;9351:234–241.
31. Shin H-C, Roberts K, Lu L, et al. Learning to read chest x-rays: recurrent neural cascade model for automated image annotation. 2016. Available at: http://arxiv.org/abs/1603.08486. Accessed August 31, 2020.
32. Ghosh A, Kandasamy D. Interpretable artificial intelligence: Why and when. *Am J Roentgenol*. 2020;214:1137–1138.
33. Annarumma M, Withey SJ, Bakewell RJ, et al. Automated triaging of adult chest radiographs with deep artificial neural networks. *Radiology*. 2019;291:196–202.
34. Yue Z, Goshtasby A, Ackerman LV. Automatic detection of rib borders in chest radiographs. *IEEE Trans Med Imaging*. 1995;14:525–536.
35. Kim M, Lee B-D. Automatic lung segmentation on chest x-rays using self-attention deep neural network. *Sensors*. 2021;21:369.
36. Peng T, Wang Y, Xu TC, et al. Segmentation of lung in chest radiographs using hull and closed polygonal line method. *IEEE Access*. 2019;7:137794–137810.