Radiology: Artificial Intelligence

Automated Assessment and Tracking of COVID-19 Pulmonary Disease Severity on Chest Radiographs using Convolutional Siamese Neural Networks

Authors/Affiliations:

Matthew D. Li, MD, Nishanth Thumbavanam Arun, Mishka Gidwani, BS, Ken Chang, MSE, Francis Deng, MD, Brent P. Little, MD, Dexter P. Mendoza, MD, Min Lang, MD, MSc, Susanna I. Lee, MD, PhD, Aileen O'Shea, MD, Anushri Parakh, MD, Praveer Singh, PhD, Jayashree Kalpathy-Cramer, PhD*

From the Athinoula A. Martinos Center for Biomedical Imaging (M.D.L., N.T.A., M.G., K.C., P.S., J.K.C.), Department of Radiology (F.D., M.L.), Division of Thoracic Imaging and Intervention (B.P.L, D.P.M.), Division of Abdominal Imaging (S.I.L., A.O., A.P.), and MGH and BWH Center for Clinical Data Science (J.K.) of the Massachusetts General Hospital, Harvard Medical School, Boston, MA, USA.

Address:

Athinoula A. Martinos Center for Biomedical Imaging, 149 13th Street, Charlestown, MA 02129.

Corresponding Author:

*Jayashree Kalpathy-Cramer, Athinoula A. Martinos Center for Biomedical Imaging, 149 13th Street, Charlestown, MA 02129. E-mail: kalpathy@nmr.mgh.harvard.edu

Funding: Research reported in this publication was supported by a training grant from the National Institute of Biomedical Imaging and Bioengineering (NIBIB) of the National Institutes of Health under award number 5T32EB1680 and by the National Cancer Institute (NCI) of the National Institutes of Health under Award Number F30CA239407 to K. Chang. This study was supported by National Institutes of Health grants U01 CA154601, U24 CA180927, and U24 CA180918 to J. Kalpathy-Cramer. This research was carried out in whole or in part at the Athinoula A. Martinos Center for Biomedical Imaging at the Massachusetts General Hospital, using resources provided by the Center for Functional Neuroimaging Technologies, P41EB015896, a P41 Biotechnology Resource Grant supported by the National Institute of Biomedical Imaging and Bioengineering (NIBIB), National Institutes of Health. GPU computing resources were provided by the MGH and BWH Center for Clinical Data Science.

Manuscript Type: Original Research; Thoracic Imaging

Word Count: 3142

This article is being published in press, and this version is the accepted, unedited, version of the manuscript.

Automated Assessment and Tracking of COVID-19 Pulmonary Disease Severity on Chest Radiographs using Convolutional Siamese Neural Networks

Article Type: Original Research; Thoracic Imaging

SUMMARY

A convolutional Siamese neural network-based algorithm can calculate a continuous radiographic pulmonary disease severity score in COVID-19 patients, which can be used for longitudinal disease evaluation and clinical risk stratification.

Key Points

- A Siamese neural network-based severity score correlates with radiologist-annotated pulmonary disease severity on chest radiographs from patients with COVID-19 (r=0.86 (95% CI 0.80-0.90) and r=0.86 (95% CI 0.79-0.90) in internal and external test sets respectively).
- The direction of change in the severity score in follow-up radiographs is concordant with radiologist assessment (ρ=0.74 (95% CI 0.63-0.81)).
- The admission chest radiograph severity score can help predict subsequent intubation or death within three days of admission (receiver operating characteristic area under the curve=0.80 (95% CI 0.75-0.85)).

Abbreviations:

COVID-19 – coronavirus disease 2019; CXR – chest radiograph; RT-PCR – reverse transcriptase-polymerase chain reaction; AP – anterior-posterior; mRALE score – modified Radiographic Assessment of Lung Edema score; PXS score - pulmonary x-ray severity score; AUC – area under the curve; CI – confidence interval.

ABSTRACT

Purpose: To develop an automated measure of COVID-19 pulmonary disease severity on chest radiographs (CXRs), for longitudinal disease tracking and outcome prediction.

Materials and Methods: A convolutional Siamese neural network-based algorithm was trained to output a measure of pulmonary disease severity on CXRs (pulmonary x-ray severity (PXS) score), using weakly-supervised pretraining on ~160,000 anterior-posterior images from CheXpert and transfer learning on 314 frontal CXRs from COVID-19 patients. The algorithm was evaluated on internal and external test sets from different hospitals (154 and 113 CXRs respectively). PXS scores were correlated with radiographic severity scores independently assigned by two thoracic radiologists and one in-training radiologist (Pearson r). For 92 internal test set patients with follow-up CXRs, PXS score change was compared to radiologist assessments of change (Spearman ρ). The association between PXS score and subsequent intubation or death was assessed. Bootstrap 95% confidence intervals (CI) were calculated.

Results: PXS scores correlated with radiographic pulmonary disease severity scores assigned to CXRs in the internal and external test sets (r=0.86 (95%CI 0.80-0.90) and r=0.86 (95%CI 0.79-0.90) respectively). The direction of change in PXS score in follow-up CXRs agreed with radiologist assessment (ρ=0.74 (95%CI 0.63-0.81)). In patients not intubated on the admission CXR, the PXS score predicted subsequent intubation or death within three days of hospital admission (area under the receiver operating characteristic curve=0.80 (95%CI 0.75-0.85)).

Conclusion: A Siamese neural network-based severity score automatically measures radiographic COVID-19 pulmonary disease severity, which can be used to track disease change and predict subsequent intubation or death.

Introduction

The role of diagnostic chest imaging continues to evolve during the COVID-19 pandemic. According to American College of Radiology guidelines, while chest CT is not recommended for COVID-19 diagnosis or screening, portable chest radiographs (CXRs) are suggested when medically necessary (1). The Fleischner Society has stated that CXRs can be useful for assessing COVID-19 disease progression (2) and one study found that 69% of these patients have an abnormal baseline CXR (3).

While radiographic findings are neither sensitive nor specific for COVID-19, with findings overlapping other infections and pulmonary edema, CXRs can be useful for assessing pulmonary infection severity and evaluating longitudinal changes. However, there is substantial variability in the interpretations of CXRs by radiologists, as has been demonstrated for pneumonia (4–6). In addition, commonly used disease severity categories on chest radiographs, such as “mild,” “moderate,” and “severe,” are challenging to reproduce as the thresholds between these categories are subjective.

One possible solution to these challenges is to train a convolutional Siamese neural network to estimate radiographic disease severity on a continuous spectrum (7). Siamese neural networks take two separate images as inputs, which are passed through twinned neural networks (8,9). The Euclidean distance between the final two layers of the networks can be calculated, which serves as a measure of distance between the two images with respect to the imaging features being trained on, such as disease features. If an image-of-interest is compared pairwise to a pool of “normal” images, the disease severity can be abstracted to the median of those Euclidean distances.

In this study, we hypothesized that a convolutional Siamese neural network-based algorithm could be trained to yield a measure of radiographic pulmonary disease severity on frontal CXRs (pulmonary x-ray severity (PXS) score). We evaluated the algorithm performance

on internal and external test sets of CXRs from patients with COVID-19. We also investigated the association between the admission PXS score and subsequent intubation or death.

Materials and Methods

This Health Insurance Portability and Accountability Act-compliant retrospective study was reviewed and exempted by the Institutional Review Board of Massachusetts General Hospital (Boston, MA), with waiver of informed consent.

Chest Radiograph Data

To train our model, we used a publicly available CXR data set, CheXpert, from Stanford Hospital, Palo Alto (10), for pretraining and a CXR data set from COVID-19 positive patients for subsequent training (Figure 1A). Additional COVID-19 CXR datasets were assembled for model testing and analysis of longitudinal change.

CheXpert contains 224,316 CXRs, with annotations for image view, which we used to filter for AP radiographs only, as suspected or confirmed COVID-19 positive patients tend to be imaged more frequently in the AP projection in emergency rooms and hospitals. CheXpert also includes a partition for training and validation, and after filtering for only AP images, the training and validation sets used for pre-training contained 161,590 and 169 images, respectively. For each image in this dataset, there are multiple radiology report-derived annotations that represent pulmonary parenchymal findings, including “lung opacity,” “lung lesion,” “consolidation,” “pneumonia,” “atelectasis,” and “edema.” For the purpose of creating a binary label for model pre-training, we considered any image with at least one of these annotations (labeled positive or uncertain) to have an abnormal lung label. All other images were considered to have normal lungs (irrespective of lines and tubes, cardiomegaly, and other findings). 81% of training images had abnormal lung labels (Supplemental Table 1).

To assemble COVID-19 CXR datasets, we obtained raw DICOM data for CXRs at a large urban quaternary-care hospital in the United States (Massachusetts General Hospital [Boston, MA]), from COVID-19 positive patients (confirmed by nasopharyngeal swab RT-PCR). The COVID-19 training set contained 314 admission CXRs from consecutive unique patients hospitalized at least in part April 1-10, 2020, randomly partitioned 9:1 for training and validation (282:32 images). The COVID-19 internal test set contained 154 admission CXRs from consecutive unique patients hospitalized at least in part March 27-31, 2020. One hospitalized patient with COVID-19 from this time period was excluded from the test set due to prior pneumonectomy. There was no overlap between training and test set patients. Among the COVID-19 internal test set patients, 92 underwent a follow-up CXR within 12 days of admission. The DICOM data for these follow-up radiographs were also obtained for longitudinal analysis For DICOMs containing more than one frontal image acquisition, the standard frontal CXR image without postprocessing was selected, with the best positioning available (selected by M.D.L., postgraduate year 4 in-training radiologist). Most of these studies were in AP projection, as extracted from the DICOM metadata (Supplemental Table 2). Intubation and mortality data were collected from the medical record by two investigators blinded to CXR findings (A.O. and A.P., radiologists in fellowship training). We also obtained raw DICOM data for 113 consecutive admission CXRs associated with unique patients hospitalized at least in part on April 15, 2020 at a community hospital in the United States (Newton-Wellesley Hospital [Newton, MA]), from COVID-19 positive patients (confirmed by nasopharyngeal swab RT-PCR), which served as an external test set.

Radiologist Scoring of Pulmonary Disease Severity on Chest Radiographs

To provide a reference standard assessment of disease severity on CXRs, we used a simplified version of the Radiographic Assessment of Lung Edema (RALE) score (11). This grading scale

was originally validated for use in pulmonary edema assessment in acute respiratory distress syndrome (ARDS) and incorporates the extent and density of alveolar opacities on CXRs. The grading system is relevant to COVID-19 patients as the CXR findings tend to involve multifocal alveolar opacities (3) and many hospitalized COVID-19 patients develop ARDS (12). In our study, we use a modified RALE (mRALE) score. Each lung is assigned a score for the extent of involvement by consolidation or ground glass/hazy opacities (0=none; 1=<25%; 2=25-50%; 3=50-75%; 4=>75% involvement). Each lung score is then multiplied by an overall density score (1=hazy, 2=moderate, 3=dense). The sum of scores from each lung is the mRALE score (examples in Supplemental Figure 1). Thus, a normal CXR receives a score of 0, while a CXR with complete consolidation of both lungs receives the maximum score of 24. mRALE differs from the original RALE score in that the lungs are not divided into quadrants.

Using the mRALE scoring system, two in-training radiologists (M.D.L. and F.D., both postgraduate year 4) independently annotated each image in the COVID-19 training set. Two fellowship-trained thoracic radiologists (B.P.L., 11 years of experience; D.P.M., 2 years of experience) and an in-training radiologist (M.D.L. for the internal test set and F.D. for the external test set) independently annotated each image in the COVID-19 internal and external test sets. The reference standard mRALE score for each image is the average of the raters. Annotator instructions and viewing conditions are in the Supplemental Materials. Inter-rater correlations between each of the raters were evaluated.

Radiologist Assessment of Longitudinal Change

The same raters who assessed the COVID-19 internal test set also evaluated the 92 internal test set patients with follow-up CXRs. For each longitudinal image pair, the raters independently assigned the label: decreased, same, or increased pulmonary disease severity (see Supplemental Materials for annotator viewing conditions). The majority change label was assigned with two or more votes for one label.

Convolutional Siamese Neural Network Training

A convolutional Siamese neural network architecture takes two separate images as inputs, which are separately passed through identical subnetworks with shared weights (schematic in Figure 1A, see Supplemental Materials for image pre-processing details) (8,9). We built such a network using DenseNet121 (13) as the underlying subnetwork with initial pre-training on ImageNet, as this architecture had empirically performed well for classification tasks in the CheXpert study (10). The Euclidean distance D_w between the subnetwork outputs, $G_w(X_1)$ and $G_w(X_2)$, given image input vectors X_1 and X_2, is calculated from the equation ($D_w(X_1, X_2) = \|G_w(X_1) - G_w(X_2)\|_2$) (9).

We used a two-step training strategy, that involves pre-training with weak labels on the large CheXpert data set using the contrastive loss function (8), followed by transfer learning to the relatively small COVID-19 training set using mean square error loss, using the assigned mRALE scores as disease severity labels. The contrastive loss function teaches the model the difference between abnormal and normal lungs, while the mean square error loss teaches the model a representation of difference in mRALE scores. Details regarding the training strategy are in the Supplemental Materials. The code is available at *https://github.com/QTIM-Lab/PXS-score.* For comparison, models were also trained using only the first or second training steps.

Calculating the Pulmonary X-Ray Severity (PXS) Score

After training the Siamese neural network, when two CXR images are passed through the subnetworks, the Euclidean distance calculated from the subnetwork outputs can serve as a continuous measure of difference between the two CXRs, with respect to pulmonary parenchymal findings. Thus, to evaluate a single image-of-interest for pulmonary disease severity, an image can be compared to a pool of *N* images without a lung abnormality (schematic in Figure 1B). We created a pool of normal images using all cases labeled with “No

Finding" from the CheXpert validation set (*N*=12, ages 19-68 years, 7 women; Supplemental Materials). Using the Siamese neural network, the Euclidean distance is calculated between the image-of-interest and each of the *N* normal images, and the median Euclidean distance is calculated. This median Euclidean distance is the Pulmonary X-Ray Severity (PXS) score.

Occlusion sensitivity maps for visualizing Siamese neural network outputs

We used an occlusion sensitivity approach (14) to visualize what portions of the input images were important to the Siamese neural network for calculating the PXS score. See the Supplemental Materials for details.

Statistical Analysis

We used Chi-square and Mann-Whitney tests, Pearson correlation (r), Spearman rank correlation (ρ), linear Cohen's kappa (κ), Fisher's exact test for odds ratios, and bootstrap 95% confidence intervals where appropriate (details in Supplementary Materials). The threshold for statistical significance was considered *a priori* to be $P<0.05$.

Results

COVID-19 Data Set Characteristics

There was no significant difference in age, sex, or mRALE scores between the training set and internal test set; patients in the external test set were significantly older than in the training and internal test sets, but there was no significant difference in sex or mRALE scores (Table 1). For the 468 patients from the combined training and internal test sets, 134 patients were intubated or died within 3 days of hospital admission. The age and mRALE scores were significantly higher in these patients (Table 2).

mRALE Score Inter-Rater Correlation

The correlation between the mRALE scores assigned by the radiologist raters was similar in the COVID-19 datasets (r=0.84-0.88, P<0.001 in all cases; see Supplemental Materials for details).

Siamese Neural Network-based PXS Score Correlates with mRALE score

In the internal test set, the Siamese neural network-based PXS score correlated with the average mRALE score assigned, which is a measure of radiographic pulmonary disease severity (r=0.86 (95% CI 0.80-0.90), P<0.001) (Figure 2A). In the external test set, the PXS score also correlated with the average mRALE score assigned (r=0.86 (95% CI 0.79-0.90), P<0.001) (Figure 2B). Using an occlusion sensitivity map-based approach, we show that the network focuses its attention on pulmonary opacities (Figure 2C). Pre-training improved model performance (Table 3; Supplemental Materials).

Longitudinal Change Assessment with the PXS Score

Of the internal test set patients with available longitudinal CXRs, according to the assigned majority vote change labels, 24 (26%), 19 (21%), and 44 (48%) of patients showed a decrease, no change, or increase in pulmonary disease severity respectively. Five patients (5%) did not receive majority votes (i.e. the three raters each voted differently; examples in Supplemental Figure 2) and were omitted from further analysis, which reflects subjectivity in the interpretation of heterogeneous CXRs. The inter-rater reliability between the three raters for assigning change labels was moderate (linear Cohen's κ=0.58, 0.59, 0.57).

The change in PXS score between two longitudinally acquired images correlates with the majority vote change label (ρ=0.74 (95% CI 0.63-0.81), P<0.001) (Figure 3A). For patients labeled with decreased disease severity, 18 (75%) were associated with decreased PXS score. For patients labeled for increased disease severity, 43 (98%) were associated with increased PXS score. For patients labeled for no change, the mean PXS score change is 0.1 (standard

deviation ± 1.3). Illustrative examples of longitudinal change assessment are shown in Figure 3B. In cases labeled for no change but with an PXS score absolute change >1, variations in inspiratory effort and positioning seem to account for the PXS change (examples shown in Supplemental Figure 3).

Association Between PXS Score and Intubation or Death

The PXS score was significantly higher on admission CXRs of patients with COVID-19 who were intubated or dead within 3 days of admission from our training and internal test sets, compared to those who were not intubated (median PXS score 7.9 versus 3.2, P<0.001) (Figure 4A). Importantly, the PXS score algorithm is not trained on outcomes data. Of the 134 patients who were intubated or died within 3 days of admission, 76 were intubated or died on the admission day and 31, 12 and 15 patients on hospital days 1, 2, and 3 respectively. A higher PXS score is associated with a shorter time interval before intubation or death in these patients (ρ=0.25, P=0.004) (Figure 4B).

Given these findings, we used the PXS score as a continuous input for prediction of intubation or death within 3 days of hospital admission. For the 437 patients without an endotracheal tube present on the admission CXR, the receiver operating characteristic area under the curve (AUC) was 0.80 (bootstrap 95% CI 0.75-0.85) (Figure 4C). The PXS threshold can be set at different levels to obtain different test characteristics, which also be expressed as odds ratios (Table 4).

Discussion

Front-line clinicians estimate the risk for clinical decompensation in patients with COVID-19 using a combination of data, including epidemiologic factors, comorbidities, vital signs, lab values, and clinical intuition (12,15). The chest radiograph can help contribute to this assessment, but manual assessment of severity is subjective and requires expertise. In this

study, we designed and trained a Siamese neural network-based algorithm to provide an automated measure of COVID-19 disease severity on chest radiographs in hospitalized patients, the Pulmonary X-ray Severity (PXS) score. The PXS score correlates with a manually annotated measure of radiographic disease severity in internal and external test sets, and the direction of change in PXS score for longitudinally acquired radiographs is concordant with radiologist assessment. For patients with COVID-19 presenting to the hospital with an admission chest radiograph, the PXS score can help predict subsequent intubation or death.

The automatic PXS score can potentially be rapidly scaled and deployed, which has important clinical applications in the COVID-19 pandemic, particularly in countries like the United States or under-resourced settings where CXRs are frequently acquired, while CT studies are relatively rarely obtained. For example, in the emergency room, clinicians must decide whether or not a patient is safe to discharge home. By setting the PXS score threshold in favor of sensitivity for prediction of intubation or death, the score can be used to help with such decisions. Additionally, PXS score can potentially be used to improve existing and newCOVID-19 machine learning models that account for other variables like vital signs, lab values, and co-morbidities (16). Other potential applications include radiologist workflow optimization, where CXRs with more severe findings can be interpreted earlier, and hospital resource management, where the PXS score can help with resource allocation (e.g. prediction of future ventilator need).

Various grading systems have been developed to measure respiratory disease severity on chest imaging, including for pulmonary edema in ARDS (11), severe acute respiratory infection (17), parainfluenza virus-associated infections (18), and pediatric pneumonia (19). A manual radiographic grading system for COVID-19 lung disease severity has been associated increased odds of intubation (20). These studies use manually annotated features from chest imaging to predict outcomes, such as mortality, need for intensive care, and other adverse events. However, barriers to adoption of these systems include inter-rater reliability and learning curve for users. In our study, raters assessing longitudinal change showed only moderate inter-

rater agreement. Our automated Siamese neural network-based approach addresses these challenges.

Deep learning-based algorithms have been applied to CXRs extensively, but primarily for disease detection, such as for pneumonia and tuberculosis (21,22), as well as for COVID-19 localization on CXR images (23). However, due to the nature of chest radiography, there are limits to the sensitivity and specificity of this modality for COVID-19 detection (3). There is a relative paucity of research using deep learning for disease severity assessment on CXRs. Automated evaluation of pulmonary edema severity on CXRs has been explored using a deep learning model that incorporates ordinal regression of edema severity labels in training (no, mild, moderate, or severe edema) (24). These severity labels were extracted from associated radiology reports, but are inherently noisy given the variability in interpretation of the CXRs (25,26). This problem of noisy labels extends beyond pulmonary edema to any disease process where there is subjectivity in interpretation. Our Siamese neural network-based approach mitigates the label noise via transfer learning on data labeled with mRALE, a more fine-grained scoring system which showed high agreement between raters in our study. In addition, pre-training of the Siamese neural network on public data with weak labels helped boost performance.

There are limitations to this study. First, patients in this study were from urban areas of the United States, which may limit the external generalizability of this algorithm to other locations. However, given that the model was able to generalize to a second hospital (community hospital vs quaternary care center) with similar performance, the model seems robust. The generalizability between two hospitals also suggests the model is reasonably robust to image acquisition technique, including differences in x-ray machinery, beam penetration, and technologist technique. Second, abnormal patient positioning and respiratory phase may introduce variability, that may impact the algorithm performance. However, since the algorithm explicitly learns to assess radiographic disease severity, quality control is relatively simple as

the PXS score can be compared visually to what is expected on sample studies. Third, our algorithm was trained using predominantly AP chest radiographs, as AP positioning is more common than posterior-anterior images among patients with COVID-19. This may limit the generalizability of the algorithm model for posterior-anterior (PA) radiographs, though future testing on PA test sets is required. Fourth, the longitudinal images were presented to raters as side-by-side JPEG image pairs for convenience, which could be less accurate than if the studies were viewed in PACS.

We developed an automated Siamese neural network-based pulmonary disease severity score for patients with COVID-19, with the potential to help with clinical triage and workflow optimization. With further validation, the score could be incorporated into clinical treatment guidelines to be used together with other clinical and lab data. The score could be validated for association/prediction with other outcomes, like oxygen saturation. Beyond the COVID-19 pandemic, this automated severity score could also be modified and applied to other continuous disease processes manifesting on chest radiographs, like pulmonary edema, interstitial lung disease, and other infections.

Acknowledgments: The authors thank Jeremy Irvin for sharing the CheXpert pre-processing script.

References

1. ACR Recommendations for the use of Chest Radiography and Computed Tomography (CT) for Suspected COVID-19 Infection | American College of Radiology. https://www.acr.org/Advocacy-and-Economics/ACR-Position-Statements/Recommendations-for-Chest-Radiography-and-CT-for-Suspected-COVID19-Infection. Accessed March 27, 2020.

2. 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;201365.

3. Wong HYF, Lam HYS, Fong AH-T, et al. Frequency and Distribution of Chest Radiographic Findings in COVID-19 Positive Patients. Radiology. 2019;201160.

4. Albaum MN, Hill LC, Murphy M, et al. Interobserver reliability of the chest radiograph in community-acquired pneumonia. Chest. 1996;110(2):343–350.

5. Loeb MB, Carusone SBC, Marrie TJ, et al. Interobserver Reliability of Radiologists' Interpretations of Mobile Chest Radiographs for Nursing Home-Acquired Pneumonia. J Am Med Dir Assoc. 2006;7(7):416–419.

6. Neuman MI, Lee EY, Bixby S, et al. Variability in the interpretation of chest radiographs for the diagnosis of pneumonia in children. J Hosp Med. 2012;7(4):294–298.

7. Li MD, Chang K, Bearce B, et al. Siamese neural networks for continuous disease severity evaluation and change detection in medical imaging. NPJ Digit Med. 2020;3(1):48.

8. Bromley J, Bentz Jw, Bottou L, et al. Signature Verification Using a "Siamese" Time Delay Neural Network. Int J Pattern Recognit Artif Intell. 1993;7(4):669–688.

9. Hadsell R, Chopra S, LeCun Y. Dimensionality Reduction by Learning an Invariant Mapping. 2006 IEEE Comput Soc Conf Comput Vis Pattern Recognit - Vol 2. IEEE; 1735–1742. http://ieeexplore.ieee.org/document/1640964/. Accessed June 9, 2019.

10. Irvin J, Rajpurkar P, Ko M, et al. CheXpert: A Large Chest Radiograph Dataset with Uncertainty Labels and Expert Comparison. 2019. http://arxiv.org/abs/1901.07031. Accessed January 4, 2020.

11. 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(9):840–846.

12. Zhou F, Yu T, Du R, et al. Clinical course and risk factors for mortality of adult inpatients with COVID-19 in Wuhan, China: a retrospective cohort study. Lancet. 2020;395(10229):1054–1062.

13. Huang G, Liu Z, van der Maaten L, Weinberger KQ. Densely Connected Convolutional Networks. Proc - 30th IEEE Conf Comput Vis Pattern Recognition, CVPR 2017. Institute of Electrical and Electronics Engineers Inc.; 2016;2017-January:2261–2269. http://arxiv.org/abs/1608.06993. Accessed March 29, 2020.

14. Zeiler MD, Fergus R. Visualizing and Understanding Convolutional Networks. 2013. http://arxiv.org/abs/1311.2901. Accessed December 7, 2019.

15. Phua J, Weng L, Ling L, et al. Intensive care management of coronavirus disease 2019 (COVID-19): challenges and recommendations. Lancet Respir Med. 2020.

16. Wynants L, Van Calster B, Bonten MMJ, et al. Prediction models for diagnosis and prognosis of covid-19 infection: Systematic review and critical appraisal. BMJ. 2020;369.

17. Taylor E, Haven K, Reed P, et al. A chest radiograph scoring system in patients with severe acute respiratory infection: A validation study. BMC Med Imaging. 2015;15(1).

18. Sheshadri A, Shah DP, Godoy M, et al. Progression of the Radiologic Severity Index predicts mortality in patients with parainfluenza virus-associated lower respiratory infections. PLoS One. 2018;13(5):e0197418.

19. McClain L, Hall M, Shah SS, et al. Admission chest radiographs predict illness severity for children hospitalized with pneumonia. J Hosp Med. 2014;9(9):559–564.

20. Toussie D, Voutsinas N, Finkelstein M, et al. Clinical and Chest Radiography Features Determine Patient Outcomes In Young and Middle Age Adults with COVID-19. Radiology. 2020;201754.

21. Rajpurkar P, Irvin J, Zhu K, et al. CheXNet: Radiologist-Level Pneumonia Detection on Chest X-Rays with Deep Learning. 2017. http://arxiv.org/abs/1711.05225. Accessed January 4, 2020.

22. Lakhani P, Sundaram B. Deep learning at chest radiography: Automated classification of pulmonary tuberculosis by using convolutional neural networks. Radiology. 2017;284(2):574–582.

23. Hurt B, Kligerman S, Hsiao A. Deep Learning Localization of Pneumonia. J Thorac Imaging. 2020;1.

24. Liao R, Rubin J, Lam G, et al. Semi-supervised Learning for Quantification of Pulmonary Edema in Chest X-Ray Images. 2019. http://arxiv.org/abs/1902.10785. Accessed January 4, 2020.

25. Kennedy S, Simon B, Alter HJ, Cheung P. Ability of Physicians to Diagnose Congestive Heart Failure Based on Chest X-Ray. J Emerg Med. 2011;40(1):47–52.

26. Hammon M, Dankerl P, Voit-Höhne HL, et al. Improving diagnostic accuracy in assessing pulmonary edema on bedside chest radiographs using a standardized scoring approach. BMC Anesthesiol. 2014;14:94. Accessed January 4, 2020.

27. Sabottke CF, Spieler BM. The Effect of Image Resolution on Deep Learning in Radiography. Radiol Artif Intell. 2020;2(1):e190015.

28. Mason D. SU-E-T-33: Pydicom: An Open Source DICOM Library. Med Phys. 2011;38(6Part10):3493–3493.

29. The OpenCV Library | Dr Dobb's. https://www.drdobbs.com/open-source/the-opencv-library/184404319. Accessed April 12, 2020.

30. Kingma DP, Ba J. Adam: A Method for Stochastic Optimization. 2014. http://arxiv.org/abs/1412.6980. Accessed June 16, 2019.

Table 1. Summary of dataset characteristics and radiologist mRALE scores. N, Number; Q1-Q3, Quartile 1 to Quartile 3 (i.e. interquartile range).

	Internal Dataset (Quaternary Care Hospital)				External Dataset (Community Hospital)	
	All	**Training/ Validation Set**	**Internal Test Set**	**p-value[a]**	**External Test Set**	**p-value[b]**
Admission CXRs, *N*	468	314	154		113	
Age (years), *median (Q1-Q3)*	57 (43-72)	56 (43-72)	59 (44-73)	0.2	74 (59-84)	<0.001*
Sex, *N women (%)*	192 (41%)	132 (42%)	60 (39%)	0.6	54 (48%)	0.2
mRALE, *median (Q1-Q3)*	4.0 (2.0-7.5)	4.0 (1.5-8.0)	4.0 (2.1-6.9)	0.9	3.3 (1.3-6.7)	0.1
mRALE, *N (%)*						
mRALE = 0	28 (6%)	20 (6%)	8 (5%)		7 (6%)	
0 < mRALE ≤ 4	213 (46%)	143 (46%)	70 (45%)		61 (54%)	
4 < mRALE ≤ 10	164 (35%)	105 (33%)	59 (38%)		30 (27%)	
mRALE > 10	63 (13%)	46 (15%)	17 (11%)		15 (13%)	

[a]p-value for comparison of internal test set with training/validation set.
[b]p-value for comparison of external test set with internal dataset (all).
*statistically significant.

Table 2. Patient and CXR characteristics stratified by outcome for the combined training and internal test set data (N = 468). N, Number; Q1-Q3, Quartile 1 to Quartile 3 (i.e. interquartile range).

	Intubated or dead within 3 days of hospital admission	**Not intubated or dead within 3 days of hospital admission**	**p-value**
Patients, *N (% total)*	134 (29%)	334 (71%)	
Age (years), *median (Q1-Q3)*	60 (50-72)	56 (42-72)	0.049*
Sex, *N women (% subgroup)*	50 (37%)	142 (43%)	0.4
mRALE, *median (Q1-Q3)*	9.0 (5.0-12.2)	3.0 (1.5-5.7)	<0.001*

*statistically significant.

Table 3. Siamese neural network model performance was improved by pre-training on CheXpert using weak labels (abnormal versus normal lung) followed by training using COVID-19 CXRs annotated for lung disease severity (mRALE score).

	Model performance with only CheXpert training (weak labels for abnormal vs normal lung)		**Model performance with only COVID-19 training set training (mRALE annotations)**		**Model performance with CheXpert pre-training and COVID-19 training set training**	
	Pearson r (95% CI)	**Spearman ρ (95% CI)**	**Pearson r (95% CI)**	**Spearman ρ (95% CI)**	**Pearson r (95% CI)**	**Spearman ρ (95% CI)**
Internal test set	0.66 (0.55-0.75)	0.70 (0.60-0.77)	0.81 (0.74-0.86)	0.77 (0.69-0.83)	0.86 (0.80-0.90)	0.84 (0.77-0.88)
External test set	0.50 (0.37-0.62)	0.57 (0.42-0.69)	0.87 (0.80-0.91)	0.75 (0.63-0.83)	0.86 (0.79-0.90)	0.78 (0.67-0.85)

Table 4. Admission radiograph PXS scores in hospitalized patients with COVID-19 (without endotracheal tube on admission CXR, N = 437) are associated with increased odds ratios for subsequent intubation or death within 3 days of admission. N, number.

PXS score threshold	Patients with PXS above threshold, total *N (% total)*	Odds Ratio	p-value
≥ 2	417 (95%)	2.9	0.2
≥ 4	209 (48%)	5.9	<0.001*
≥ 6	120 (27%)	6.8	<0.001*
≥ 8	60 (14%)	12.1	<0.001*

*statistically significant.

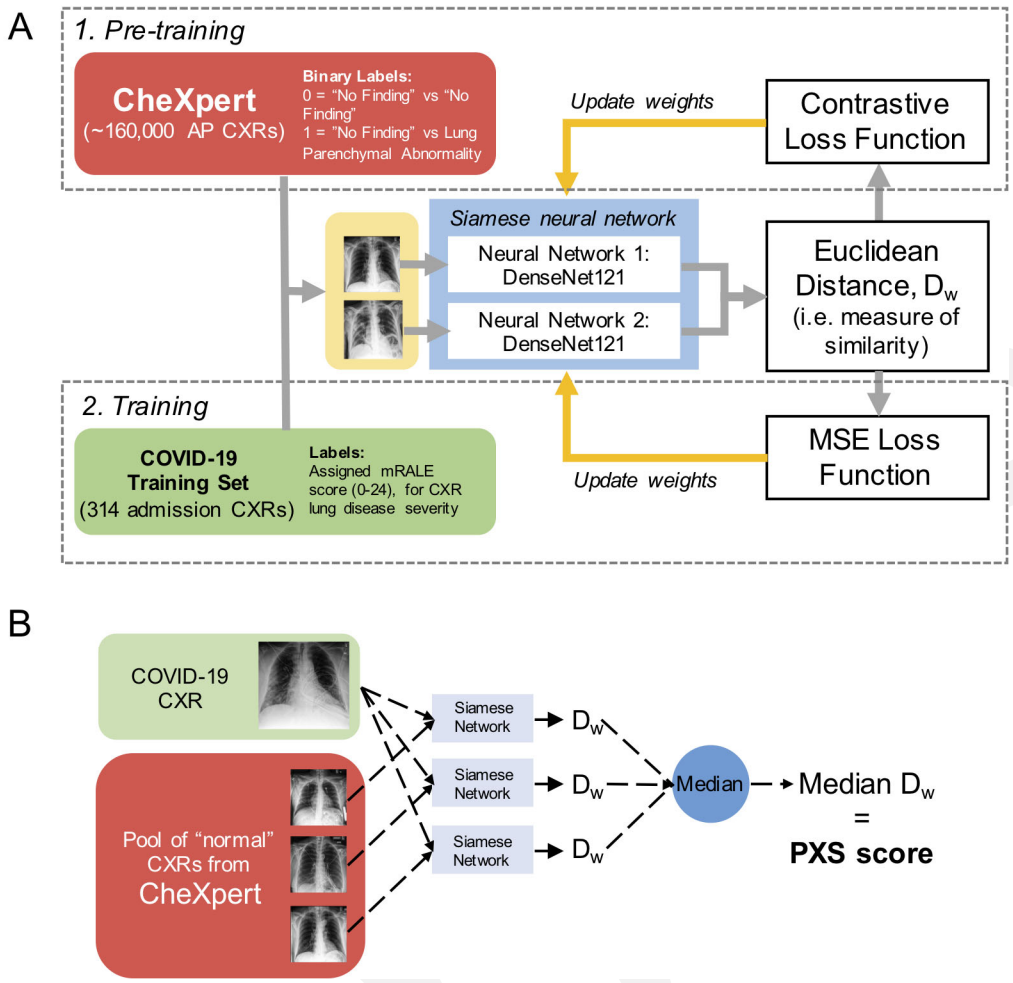


Figure 1: A, Schematic for training the convolutional Siamese neural network-based algorithm used to calculate the Pulmonary X-Ray Severity (PXS) score, a continuous measure of radiographic pulmonary disease severity in COVID-19 patients. The network is pre-trained with chest radiographs (CXRs) from CheXpert (10) using binary lung disease presence labels and then trained on CXRs from a COVID-19 training set using annotations for modified Radiographic Assessment of Lung Edema (mRALE) scores. B, Schematic for calculating the PXS score, which is calculated by comparing the image-of-interest pairwise with a pool of normal CXRs from CheXpert. D_w = Euclidean distance; MSE loss = mean square error.

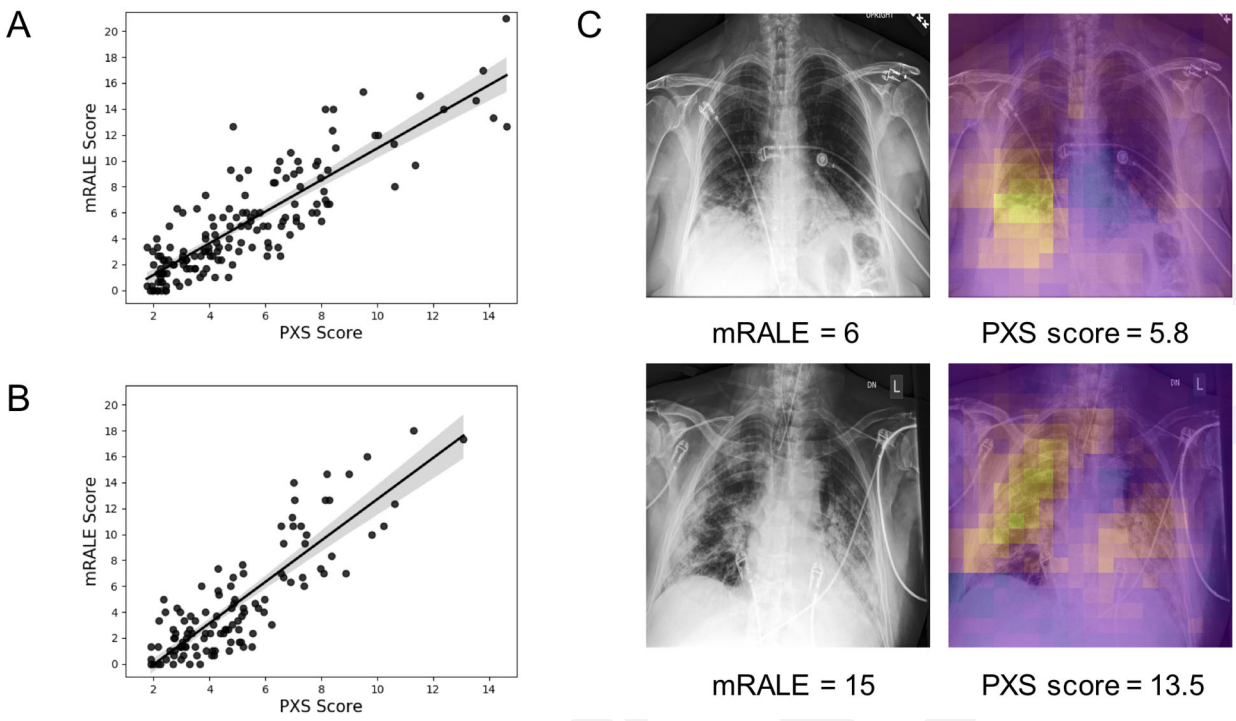
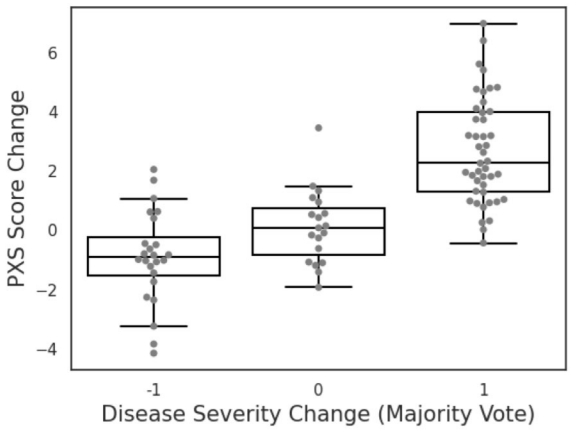
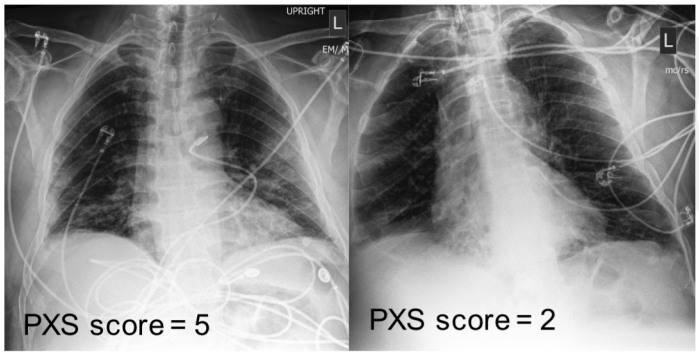


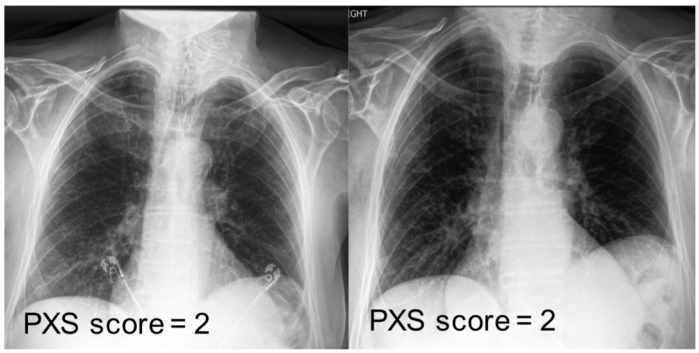
Figure 2: Siamese neural network-based Pulmonary X-Ray Severity (PXS) score is a measure of radiographic pulmonary disease severity in patients with COVID-19. A and B, Scatterplots show, in a 154-patient internal test set (A) and 113-patient external hospital test set (B), the PXS score correlates with the modified Radiographic Assessment of Lung Edema (mRALE) score, a measure of pulmonary disease severity on chest radiographs (p=0.86, P<0.001 and p=0.86, P<0.001, respectively) (linear regression 95% confidence interval shown in the scatterplots). C, Occlusion sensitivity map-based approach shows that the Siamese neural network is focusing on pulmonary opacities. Yellow areas indicate parts of the image important to the neural network.

A


PXS Score Change
6
4
2
0
−2
−4
-1
0
1
Disease Severity Change (Majority Vote)

B


PXS score = 5
PXS score = 2


PXS score = 2
PXS score = 2

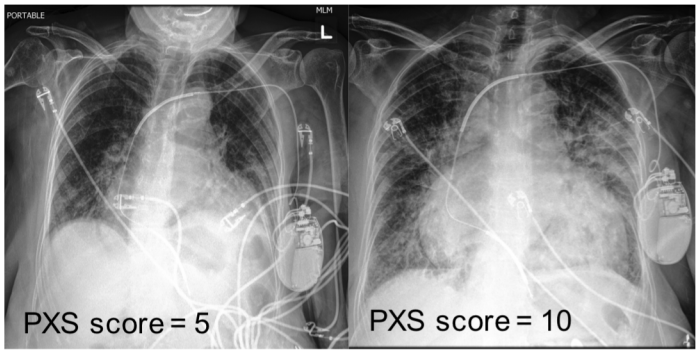

PXS score = 5
PXS score = 10

Figure 3: Siamese neural network-based Pulmonary X-Ray Severity (PXS) score can be used to assess longitudinal change in radiographic disease severity over time in COVID-19 patients. A, Boxplot shows the PXS score correlates with majority vote change in pulmonary disease severity (ρ=0.74, P<0.001), where -1, 0, and 1 indicate decreased, unchanged, and increased severity in longitudinal chest radiograph pairs, assigned by three independent raters (2 thoracic radiologists, 1 in-training radiologist). The boxplot boxes indicate the median and interquartile range (IQR), with whiskers extending to points within 1.5 IQRs of the IQR boundaries. B, Examples of PXS score evaluation of longitudinal change in three patients with COVID-19.

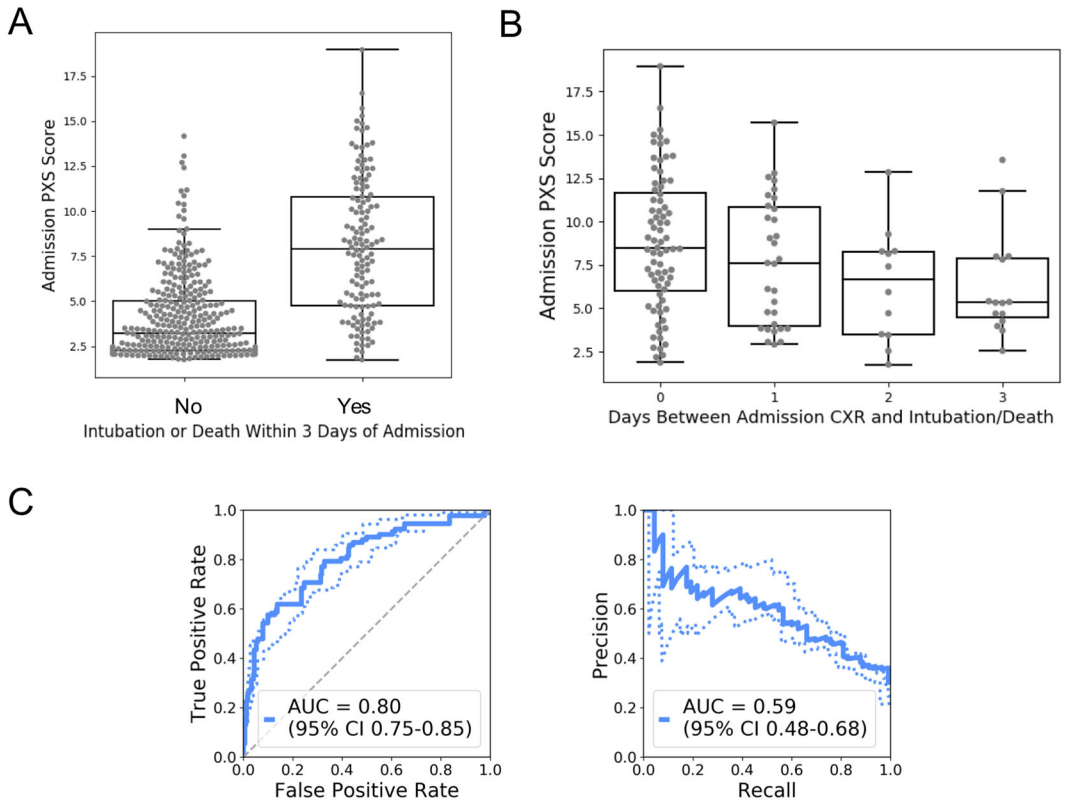


Figure 4: Siamese neural network-based Pulmonary X-Ray Severity (PXS) score is associated with intubation in patients hospitalized with COVID-19. A, Boxplot shows the PXS score is significantly higher in patients intubated within three days of hospital admission (P<0.001). B, Boxplot shows that a higher PXS score is associated with a shorter time interval before

intubation (ρ=0.25, P=0.004), C, Receiver operating characteristic and precision recall curves show the performance of the PXS score for predicting subsequent intubation within three days of hospital admission, in patients without an endotracheal tube on their admission chest radiograph (AUC, area under the curve; dashed lines indicate bootstrap 95% confidence intervals).

SUPPLEMENTAL MATERIALS

Table of Contents

Supplemental Methods

Chest Radiograph Image Pre-processing

Full size CXR images in JPEG format from CheXpert were all resized to 320 x 320 pixels, which is within the resolution range of optimal performance for CXR binary classification tasks (27). DICOM files from the COVID-19 CXRs were all pre-processed in the same manner as in ChexPert, with image pixel array extraction using pydicom (28), followed by normalization to [0, 255], conversion to 8-bit, correction of photometric inversion, histogram equalization in OpenCV (29), and conversion to a JPEG file. These DICOMs were anonymized at the time of study export from the PACS. In the external test set CXR images, some images included a large black border around the actual radiograph, which was mostly removed using an automatic cropping algorithm in Python (border pixels with a 0 pixel value were removed).

mRALE annotator instructions and CXR image viewing conditions

All annotators were instructed on use of mRALE and practiced on ~10 cases before annotating the complete datasets independently. There were instructed that the goal of mRALE is to grade pulmonary opacity, regardless of cause (e.g. fibrosis or pulmonary edema still presents with a lung opacity, and should be graded as such). In the overall density score, the term ‘moderate’ is used which is from the original RALE paper. We have interpreted it to mean anything in between hazy opacities and dense consolidation. The lung may have different densities in different parts (e.g. ~50% of the left lung shows opacities, but some is ‘moderate’ density and some is ‘dense.’ The rater decides on the predominant density to assign the score. Pleural effusions are not included in the scoring system, though concurrent “basal opacities” which may be due to atelectasis does contribute to the mRALE score.

For the training set, CXR images were viewed by annotators using JPEG images pre-processed from the DICOMs on personal computers (due to convenience during the COVID-19 pandemic). For the internal and external test sets, CXR images were viewed by annotators

using PACS stations routinely used for clinical work in the hospital, in standard diagnostic conditions, so as to simulate the real-world radiologist work environment.

For the longitudinal image pair annotations for change, CXR images were displayed as side-by-side pre-processed JPEG images to allow for convenience of comparison.

Convolutional Siamese Neural Network Training

We used a two-step training strategy, that involves pre-training with weak labels on the large CheXpert data set followed by transfer learning to the relatively small COVID-19 training set, as follows:

Step 1. To pre-train the Siamese neural network on CheXpert data, the contrastive loss function is used to train the network parameters, as defined by the equation ($L = (1 - Y){D_w}^2 + (Y)\{max(0, m - D_w)\}^2$; Y = 0 if same class (i.e. no change) and Y = 1 if different class (i.e. change), D_w = Euclidean distance, and m = margin) (9). The contrastive loss function minimizes when there is a small Euclidean distance for no change and large Euclidean distance for change in class. The margin hyperparameter is empirically set to 50, which gives the maximum D_w for which dissimilar image input pairs will not contribute further to the loss, helping to stabilized training. As the goal of this algorithm is to generate a measure of disease severity, we trained the convolutional Siamese neural network to maximize Euclidean distance when the input images showed a difference in labels that identify lung parenchymal abnormalities. In the CheXpert data set, there are annotations that represent pulmonary parenchymal findings, including “lung opacity,” “lung lesion,” “consolidation,” “pneumonia,” “atelectasis,” and “edema.” If an image had any one of these labels (marked positive or uncertain), it was assigned an abnormal lung label. If an image did not have any one of those labels, it was assigned a normal lung label. In training the network, paired CXR images were sampled from the training data and passed to the subnetworks separately, where the contrastive loss function label Y = 0 if the CXRs both have the ChexPert label “No Finding” label (i.e. no difference between normal lungs)

and Y = 1 if one CXR has an abnormal lung label and the other has a “No Finding” label (i.e. difference between lungs). The paired input CXR images were randomly sampled in a manner so that an equal number of Y = 0 and Y = 1 labels were assigned, for both training and validation. We empirically set the number of CXR image pairs sampled per epoch of training and validation at 6400 and 200 image pairs respectively. For both training and validation, each input image is resized to 336 x 336 pixels followed by a center crop to 320 x 320 pixels. This algorithm was implemented in Python with the PyTorch package, using the Adam optimizer (30) (initial learning rate = 0.00002, β_1 = 0.9, β_2 = 0.999). Batch sizes were fixed at 8 for training and validation. Early stopping of training occurred when the validation loss showed no further improvement after 3 training epochs. The model with the lowest validation loss was saved for further training.

Step 2. After pre-training on ChexPert data using weak labels, we train the Siamese neural network on the 314 image COVID-19 training set using mean square error (MSE) loss. Each image pair fed to the Siamese neural network results in an output of the Euclidean distance between the final fully connected layers. This Euclidean distance is an abstraction of difference in pulmonary disease severity between the two input CXRs. The “error” of the MSE loss is the difference between the Euclidean distance and the absolute difference in the labeled mRALE scores between the two input images. The input image pairs are randomly sampled during training and validation, with 1600 and 200 image pairs sampled per epoch, respectively. For training, each input image is resized to 336 x 336 pixels followed augmentation with random rotations of ±5° and random crop of 320 x 320 pixels. For validation, each input image is resized to 336 x 336 pixels followed by a center crop to 320 x 320 pixels. This training step was also implemented using the Adam optimizer, with the same hyperparameters as the previous step, and batch sizes of 8. Early stopping was set at 7 epochs without improvement in validation loss. The model with the lowest validation loss was saved for testing evaluation.

The rationale for using contrastive loss in pre-training (Step 1), but MSE loss instead in training (Step 2), is related to the available data labels. The pre-training dataset has weak binary labels (lung opacity versus no lung opacity), which can be used to create same versus different labels for paired inputs for the contrastive loss function. MSE loss cannot be used in this case, as the label is binary. Contrastingly, the training dataset has labels that provide a more granular severity grade (mRALE), which can be used as inputs for MSE loss. By using MSE loss during training, the model can learn that the differences between varying magnitudes of differences in mRALE scores (e.g. mRALE 1 versus 4 is different from 5 vs 16, but 3 versus 6 should have the same difference in mRALE between two inputs as 1 versus 4). This allows the PXS score model to learn a linear representation of lung disease severity.

Occlusion sensitivity maps for visualizing Siamese neural network outputs

To generate an occlusion map, patches of 32 x 32 pixels in the paired input images are occluded (patch area pixel intensities equal the mean of the patch) iteratively across the entire image (stride length 16 pixels). For each iteration, both occluded images are passed through the Siamese neural network and a Euclidean distance is calculated. An increased difference between this Euclidean distance and the non-occluded baseline Euclidean distance indicates that part of the image is important to the network and can be represented as a heat map. When evaluating disease severity in a single image, as in this case, an occlusion sensitivity map is generated for each comparison of the image-of-interest to each image in the pool of ChexPert “No Finding” images. The median of these occlusion sensitivity maps is used for visualization.

Statistics

To evaluate differences in patient gender and age in the COVID-19 data sets, we used the Chi-square test and Mann-Whitney test (two-sided), respectively. To evaluate differences in mRALE score between the COVID-19 data sets, treated mRALE as a continuous variable (scale 0-24)

and used the Mann-Whitney test (two-sided). The associations between variables including the PXS score, mRALE score, and inter-rater mRALE scores were calculated using the Pearson correlation (r) (and also Spearman rank correlation (ρ) for comparison of correlations with and without pretraining). For evaluating inter-rater reliability of longitudinal CXR change labels, linear Cohen's kappa (κ) was used. For comparing radiologist agreement with the algorithm on longitudinal change assessment, we used the Spearman rank correlation. For comparison of the PXS score between patients with and without intubation/death, the Mann-Whitney test was used (two-sided). For evaluating the correlation between the time interval between admission and intubation/death, we used the Spearman rank correlation. For correlation coefficients and area under the receiver operating characteristic curve analysis, bootstrap 95% confidence intervals were calculated. Odds ratios and p-values for the association of PXS score thresholds with clinical outcomes were calculated using Fisher's exact test (unconditional maximum likelihood estimate). These calculations were all performed using the *scipy* and *sklearn* Python packages. The threshold for statistical significance was considered *a priori* to be $P<0.05$. Data visualizations were performed using the *Seaborn* Python package.

Supplemental Results

Inter-rater correlation in assigning mRALE scores

In the 314-patient COVID-19 training set, the correlation between the assigned mRALE score of the two raters was good ($r=0.87$, $P<0.001$). In the 154-patient COVID-19 internal test set, the rank correlations of the assigned mRALE score between the three raters was similar ($r=0.85$, 0.87, 0.85, $P<0.001$ in all cases). In the 113-patient COVID-19 external test set, the rank correlations of the assigned mRALE score between the three raters was also similar (0.84, 0.86, 0.88, $P<0.001$ in all cases).

Impact of CheXpert pre-training on model performance

To evaluate the impact of pre-training, we also trained a Siamese neural network model without CheXpert pre-training. We also found that this pre-training resulted in improved model performance, as demonstrated by increased Pearson and Spearman correlations on the internal test set and increased Spearman correlation on the external test set (Table 3). The Pearson correlation on the external test set was essentially the same. A model trained using only abnormal versus normal lung labels derived from the CheXpert data set (weak supervision) had worse performance (Table 3).

Impact of image anchor pool size on model performance

We empirically set the size of the pool of normal studies for comparison to $N = 12$, which were used as image comparisons to calculate the PXS score as described in the Methods. During model development, we found that increasing the N improves model performance, particularly for smaller Euclidean distances (i.e. PXS scores), though with diminishing improvement with larger N (e.g. $N = 30$ resulted in the same performance as $N = 12$ in the internal test set). However, model inference time increases as N increases, due to the increased number of comparisons that are made using the Siamese neural network.

Supplemental Table 1. Distribution of abnormal lung labels in the CheXpert image dataset.

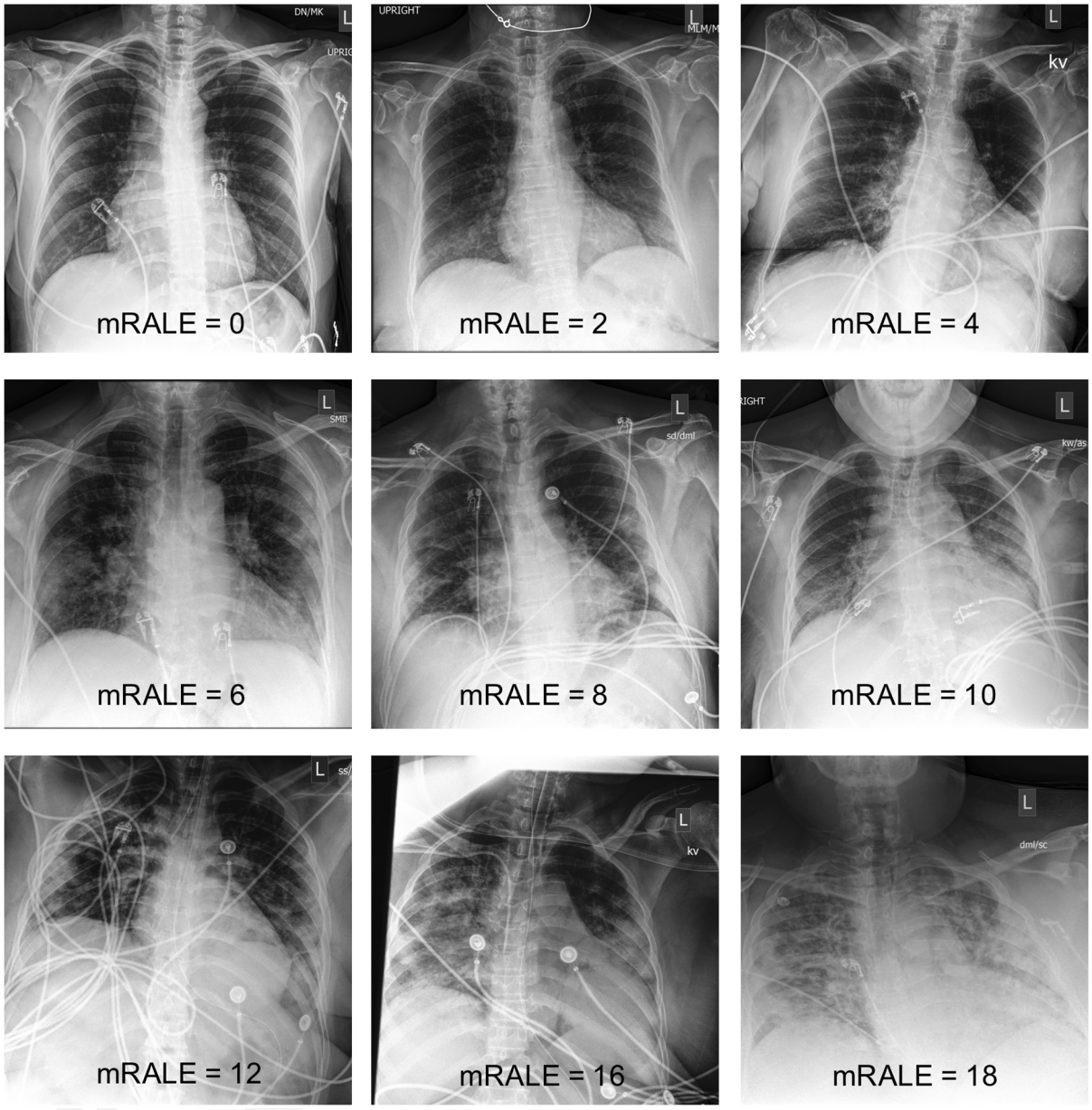
	Complete CheXpert dataset, *N* *(% total)*	**AP CheXpert images only, *N* *(% total)***
Abnormal lung label	Training: 166,291 (74%) Validation: 126 (54%)	Training: 130,934 (81%) Validation: 109 (64%)
Normal lung label	Training: 57,123 (26%) Validation: 108 (46%)	Training: 30,656 (19%) Validation: 60 (36%)
TOTAL	Training: 223,414 Validation: 234	Training: 161,590 Validation: 169

For any image with a CheXpert annotation (marked positive or uncertain) that represents pulmonary parenchymal findings, including “lung opacity,” “lung lesion,” “consolidation,” “pneumonia,” “atelectasis,” and “edema.”, it was assigned an abnormal lung label. If an image did not have any one of those labels, it was assigned a normal lung label. AP, anterior-posterior view.

Supplemental Table 2. Distribution of CXR view position in the COVID-19 image datasets.

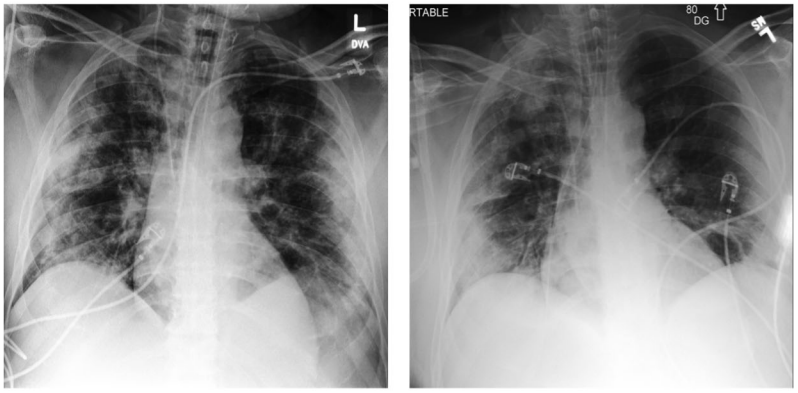
	CXRs, *N*	**AP view, *N (% set)***	**PA view, *N (% set)***
Training Set **Training Partition** **Validation Partition**	314 282 32	268 (85%) 239 (85%) 29 (91%)	46 (15%) 43 (15%) 3 (9%)
Internal Test Set	154	128 (83%)	26 (17%)
External Test Set	113	107 (95%)	6 (5%)
TOTAL	581	503 (87%)	78 (13%)

AP, anterior-posterior view; PA, posterior-anterior view.

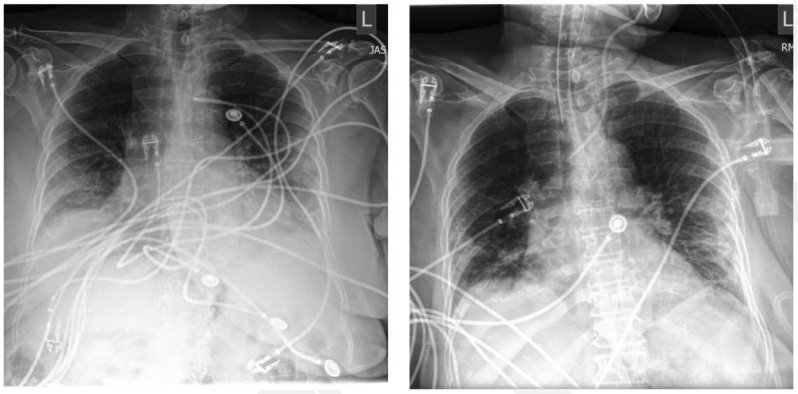


Supplemental Figure 1. Representative example images of mRALE scores in CXRs from patients with COVID-19 (average of scores assigned by multiple raters).

A

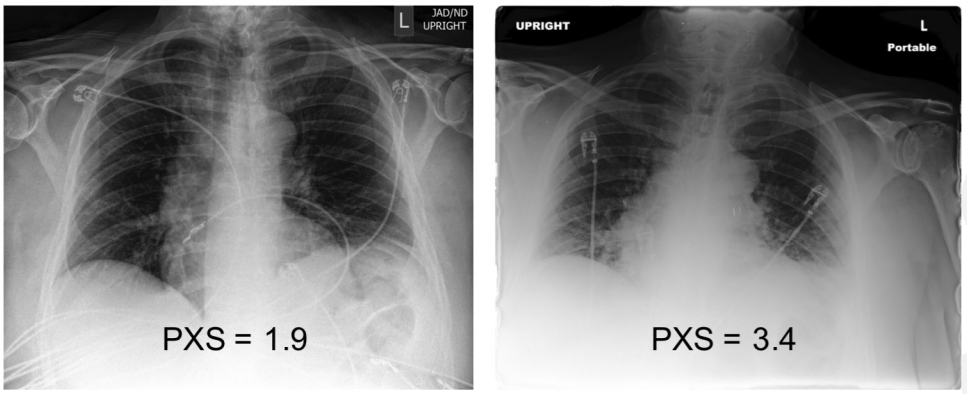


B

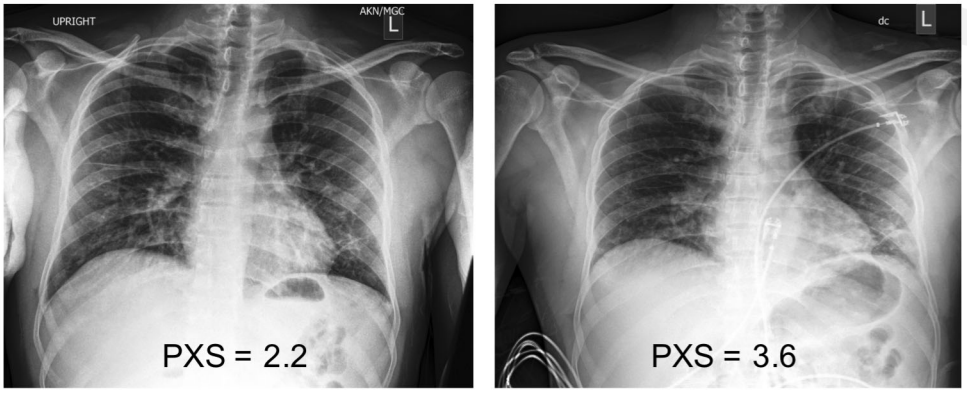


Supplemental Figure 2. Examples of longitudinally acquired CXRs in patients with COVID-19 where there was no majority vote for a change label by radiologist raters (i.e. one vote for no change, one vote for worse disease, and one for better disease). In A, the PXS score showed a change of -2.8 (10.4 to 7.6), suggesting decreased lung disease severity. In B, the PXS score showed a change of +1.2 (3.9 to 5.1), suggesting slightly increased lung disease severity.

A



B



Supplemental Figure 3. Illustrative examples of the potential impact of differences in inspiratory effort and positioning on PXS score. In A and B, the paired radiographs are from the same patient acquired at different time points (from the longitudinal analysis). In both cases, the CXR from the second time point has a higher PXS score, but this appears to be at least in part due to lower lung volumes with mild atelectasis in both cases. In case A, the patient positioning is also different. The majority vote of radiologist annotators in both of these paired cases was for no change in lung disease severity between the CXRs.