



Multi-population generalizability of a deep learning-based chest radiograph severity score for COVID-19

Matthew D. Li, MD^a , Nishanth T. Arun, BTEch^a, Mehak Aggarwal^a, Sharut Gupta^a, Praveer Singh, PhD^a, Brent P. Little, MD^b, Dexter P. Mendoza, MD^b, Gustavo C.A. Corradi, MD^c, Marcelo S. Takahashi, MD^c, Suely F. Ferracioli MD^c, Marc D. Succi, MD^e, Min Lang, MD^b, Bernardo C. Bizzo, MD^{a,f}, Ittai Dayan, MD^f, Felipe C. Kitamura, MD^{c,d,*}, Jayashree Kalpathy-Cramer, PhD^{a,f,*} 

Abstract

To tune and test the generalizability of a deep learning-based model for assessment of COVID-19 lung disease severity on chest radiographs (CXRs) from different patient populations.

A published convolutional Siamese neural network-based model previously trained on hospitalized patients with COVID-19 was tuned using 250 outpatient CXRs. This model produces a quantitative measure of COVID-19 lung disease severity (pulmonary x-ray severity (PXS) score). The model was evaluated on CXRs from 4 test sets, including 3 from the United States (patients hospitalized at an academic medical center (N = 154), patients hospitalized at a community hospital (N = 113), and outpatients (N = 108)) and 1 from Brazil (patients at an academic medical center emergency department (N = 303)). Radiologists from both countries independently assigned reference standard CXR severity scores, which were correlated with the PXS scores as a measure of model performance (Pearson *R*). The Uniform Manifold Approximation and Projection (UMAP) technique was used to visualize the neural network results.

Tuning the deep learning model with outpatient data showed high model performance in 2 United States hospitalized patient datasets (*R* = 0.88 and *R* = 0.90, compared to baseline *R* = 0.86). Model performance was similar, though slightly lower, when tested on the United States outpatient and Brazil emergency department datasets (*R* = 0.86 and *R* = 0.85, respectively). UMAP showed that the model learned disease severity information that generalized across test sets.

A deep learning model that extracts a COVID-19 severity score on CXRs showed generalizable performance across multiple populations from 2 continents, including outpatients and hospitalized patients.

Abbreviations: AP = anterior-posterior, CXR = chest radiograph, COVID-19 = coronavirus disease 2019, mRALE = modified radiographic assessment of lung edema, PXS = pulmonary X-ray severity, PA = posterior-anterior, UMAP = uniform manifold approximation and projection.

keywords: COVID-19; chest radiograph; deep learning; artificial intelligence; generalizability

Conflicts of interest and sources of funding: This study was supported by sundry funds to J.K. 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. M.D.L., B.C.B., I.D., and J.K. report collaborating with Bayer Radiology on addressing regulatory requirements for potential clinical application of this technology (no funding provided for the work in this manuscript). M.D.L. reports funding from an RSNA R&E Fund Research Resident/Fellow Grant, outside of the submitted work. BPL is a textbook associate editor and author for Elsevier, Inc. and receives royalties. F.C.K. reports consulting for MD.ai. J.K. reports grants from GE Healthcare, nonfinancial support from AWS, and grants from Genentech Foundation, outside the submitted work. For the remaining authors none were declared.

The datasets generated during and/or analyzed during the current study are not publicly available, but are available from the corresponding author on reasonable request.

^a Athinoula A. Martinos Center for Biomedical Imaging, Department of Radiology, Massachusetts General Hospital, Harvard Medical School, Boston, MA, USA, ^b Division of Thoracic Imaging and Intervention, Department of Radiology, Massachusetts General Hospital, Harvard Medical School, Boston, MA, USA, ^c Diagnósticos da América SA (DASA), São Paulo, Brazil, ^d Department of Diagnostic Imaging, Universidade Federal de São Paulo, São Paulo, Brazil, ^e Division of Emergency Radiology, Department of Radiology, Massachusetts General Hospital, Harvard Medical School, Boston, MA, USA,

^f MGH and BWH Center for Clinical Data Science, Mass General Brigham, Boston, MA, USA.

*Correspondence: Felipe C. Kitamura, Diagnósticos da América SA (DASA) and Universidade Federal de São Paulo, Rua Gilberto Sabino 215, Pinheiros, São Paulo - SP CEP: 05425-020, Brazil (e-mail: kitamura.felipe@gmail.com); Jayashree Kalpathy-Cramer, Athinoula A. Martinos Center for Biomedical Imaging, 149 13th Street, Charlestown, MA 02129 (e-mail: kalpathy@nmr.mgh.harvard.edu).

Copyright © 2022 the Author(s). Published by Wolters Kluwer Health, Inc. This is an open-access article distributed under the terms of the Creative Commons Attribution-Non Commercial License 4.0 (CCBY-NC), where it is permissible to download, share, remix, transform, and buildup the work provided it is properly cited. The work cannot be used commercially without permission from the journal.

How to cite this article: Li MD, Arun NT, Aggarwal M, Gupta S, Singh P, Little BP, Mendoza DP, Corradi GC, Takahashi MS, Ferracioli SF, Succi MD, Lang M, Bizzo BC, Dayan I, Kitamura FC, Kalpathy-Cramer J. Multi-population generalizability of a deep learning-based chest radiograph severity score for COVID-19. *Medicine* 2022;101:29(e29587).

Received: 19 October 2020 / Received in final form: 21 April 2022 / Accepted: 28 April 2022

Publication history: This manuscript was previously posted to medRxiv: doi: <https://doi.org/10.1101/2020.09.15.20195453> <http://dx.doi.org/10.1097/MD.00000000000029587>

1. Introduction

Chest radiographs (CXRs) are routinely obtained in symptomatic patients with suspected or confirmed coronavirus disease 2019 (COVID-19) infection. While CXRs have limited sensitivity for the diagnosis of COVID-19,^[1-3] the severity of radiographic lung findings has been associated with worse clinical outcomes.^[4-6] Deep learning-based techniques have been used to automate the extraction of measures of lung disease severity from CXR image data, which correlate with manual scores of disease severity by radiologists and can be potentially used for patient risk stratification.^[7-12] These techniques are promising; however, the performance of CXR deep learning models are known to show variable generalization on external data.^[13] Thus, validation on data from different sources and patient populations is essential before such models can be deployed in clinical practice.

In this study, we aimed to test the generalizability of a previously published deep learning-based model for automated assessment of COVID-19 pulmonary disease severity, the Pulmonary X-Ray Severity (PXS) score model.^[7] A limitation of the original model was that it was trained and tested on CXRs from patients hospitalized with COVID-19, who tend to have more severe disease compared to the general population infected by COVID-19. In addition, portable anterior-posterior (AP) CXRs are overrepresented compared to standard posterior-anterior (PA) CXRs, which may be more common in outpatient settings. In this work, we tuned the PXS score model by training with primarily outpatient CXRs. We tested the hypothesis that this the PXS score model would generalize to different patient populations model by assessing performance in comparison to manual radiologist annotations for lung disease severity in 4 different test sets with different technical and patient characteristics, including CXRs acquired from patients in 2 countries (United States and Brazil).

2. Methods

This retrospective study was reviewed and exempted by the Institutional Review Board of Massachusetts General Brigham (Boston, USA), with waiver of informed consent. The parts of the

study involving data from Hospital Santa Paula were approved by the Institutional Review Board of the Universidade Federal de São Paulo (São Paulo, Brazil). The hospitals involved in this study include Massachusetts General Hospital (Hospital 1) (Boston, USA), Hospital Santa Paula (Hospital 2) (São Paulo, Brazil), and Newton Wellesley Hospital (Hospital 3) (Newton, USA), Hospitals 1 and 2 are large academic medical centers, while Hospital 3 is a community hospital in the Boston metropolitan area.

2.1. PXS score base model

For the base model for this study, we used a previously published convolutional Siamese neural network-based model that can extract a continuous measure of lung disease severity from CXRs in patients with COVID-19, the PXS score model.^[7] In brief, a Siamese neural network is composed of twinned subnetworks with identical weights; paired images can be passed as inputs, each image passing to a subnetwork.^[14] The Euclidean distance between the last fully connected layers of the subnetworks can serve as a continuous measure of disease severity similarity between the 2 input images.^[15] In the original PXS score model, a Siamese neural network composed of twinned DenseNet121 networks^[16] was pretrained using ~160,000 anterior-posterior (AP) chest radiographs from the publicly available CheXpert dataset.^[17] The model was then trained using 314 admission CXRs from hospitalized patients with COVID-19 at Hospital 1 annotated by radiologists using a manual scoring system for lung disease severity.^[7] During model inference, the image-of-interest was compared to a pool of normal CXRs from CheXpert, and the median of the Euclidean distances between the image-of-interest and each normal CXR served as the PXS score. Please refer to the previously published work for the technical details of this implementation.^[7] See Figure 1 for a study design schematic.

2.2. Chest radiograph data

We assembled 2 new CXR DICOM datasets for this study:

(1) *Hospital 1 Outpatient Dataset*. This dataset was composed of 358 CXRs from 349 unique patients who presented

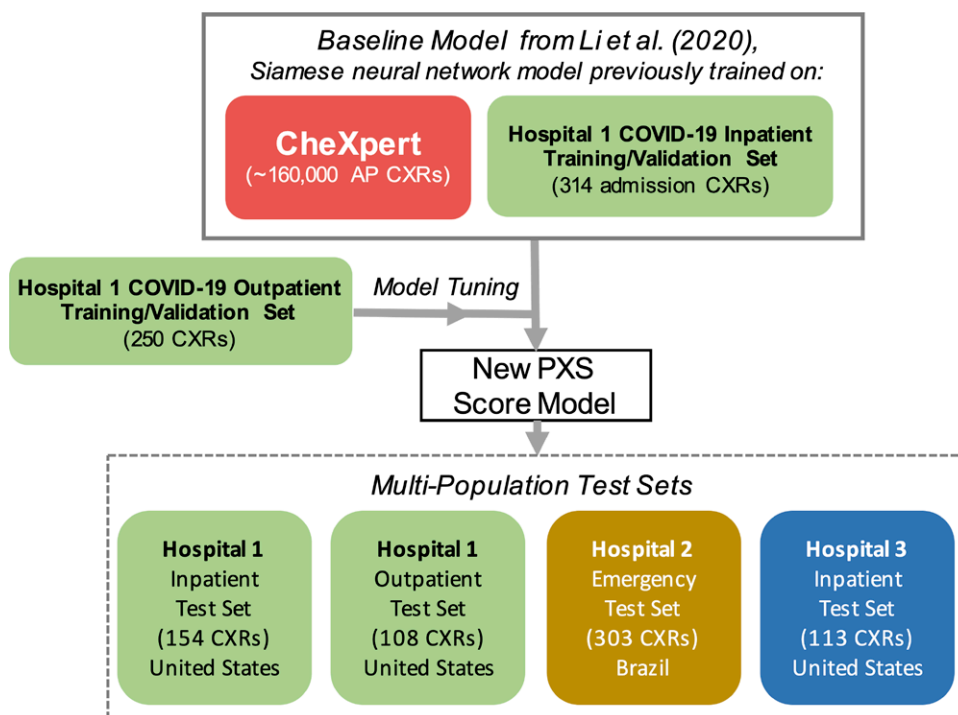


Figure 1. Schematic of study design. Previously published Siamese neural network-based model for extracting lung disease severity from CXRs^[7] was tuned using new CXR data and evaluated in 4 test sets.

for outpatient imaging at urgent care or respiratory illness clinics associated with Hospital 1 and tested positive for COVID-19 by nasopharyngeal swab RT-PCR obtained at their outpatient visit from March 15, 2020 to April 15, 2020. Raw DICOM data for the frontal view CXRs was extracted and anonymized directly from the institutional PACS. This dataset was composed of mostly CXRs acquired in the posterior-anterior (PA) position (342, 96%), with the remainder acquired in the anterior-posterior (AP) position (16, 4%). Some radiographs in this data set overlapped with the original PXS score model training data set (22, 6%) because the outpatient CXR was used as the admission CXR in some patients.^[7] Thus, in partitioning this outpatient CXR dataset, we included the overlapping radiographs in the planned training/validation partition and then randomly allocated the remaining CXRs up to a 70:30 distribution (250 for training/validation and 108 for testing). The training/validation partition was then randomly partitioned 90:10 (225 for training, 25 for validation). Associated age and sex data were extracted from the electronic health record.

(2) *Hospital 2 Emergency Test Set.* This dataset was composed of 303 CXRs from 242 unique patients who presented to the emergency department with suspected COVID-19 at Hospital 2. These CXRs were sampled from patients from February 1, 2020 to May 30, 2020 with at least 1 COVID-19 RT-PCR result within ± 3 days of the CXR. Sampling was stratified on RT-PCR test results, so that 70% of CXRs in the dataset would have at least 1 positive associated test and 30% would have all negative tests. In addition, a subset of patients was permitted to have multiple CXRs in the dataset (49 with 2 CXRs, 6 with 3 CXRs). Raw DICOM data for the frontal view CXRs was extracted and anonymized from the institutional PACS. The AP vs PA view position was not available in the DICOM metadata for this data. Age and sex data were extracted from the electronic health record.

In addition to these 2 data sets that were created for this study, we also used previously published data sets for model testing, including 154 admission CXRs from 154 unique patients hospitalized for COVID-19 at Hospital 1 (*Hospital 1 Inpatient Test Set*) and 113 admission CXRs from 113 unique patients hospitalized for COVID-19 at Hospital 3 (*Hospital 3 Inpatient Test Set*).^[7] X-ray equipment manufacturer information from all 4 test sets were extracted from the DICOM metadata tags.

The datasets generated during and/or analyzed during the current study are not publicly available as the datasets are not entirely anonymized of protected health information, but are available from the corresponding authors on reasonable request.

Raw pixel data from the CXR DICOMs used in training, validation, and testing were preprocessed using the same steps as used in the baseline PXS score model,^[7] including conversion to 8-bit, correction of photometric inversion, histogram equalization, and conversion to a JPEG file.

2.3. Radiologist annotations for lung disease severity

We used a manual scoring system for COVID-19 lung disease severity on CXRs previously used for training of the PXS score model, which is a modified version of the Radiographic Assessment of Lung Edema scoring system (mRALE).^[7,18] In brief, from the frontal view of the CXR, each lung is assigned a score from 0 to 4 for extent of consolidation or ground glass/hazy opacities (up to 0%, 25%, 50%, 75%, 100%) and a score from 1 to 3 for overall density (hazy, moderate, dense). The sum of the products of the extent and density scores for each lung is the mRALE score (range from 0 to 24). Higher mRALE scores have been associated with worse clinical outcomes in COVID-19.^[5] Two diagnostic radiologists with thoracic subspecialty expertise (B.P.L., D.P.M.) from Hospital 1 independently annotated the 358 CXRs from the Hospital 1 Outpatient Dataset for mRALE, viewing the images on a diagnostic PACS viewer. Three

diagnostic radiologists with nonthoracic subspecialty training (G.C.A.C., M.S.T., S.F.F.) from Hospital 2 independently annotated the 303 CXRs from the Hospital 2 Emergency Test Set for mRALE, viewing the images using the MD.ai annotation platform (New York, United States). The average of the rater mRALE scores served as the mRALE score for each CXR. All raters had previously rated 10 CXRs using mRALE with feedback on their scores, though the Hospital 1 raters had more experience, previously rating ~ 300 studies independently.

To assess the correlation between the radiologists from both hospitals in applying the mRALE score, the 2 thoracic radiologists from Hospital 1 rated a subset of 69 studies from the Hospital 2 dataset in PACS viewers. This subset was composed of studies with mRALE ≥ 3.0 assigned by the Hospital 3 raters, in order to focus reassessment on abnormal lungs, rather than normal/ near-normal lungs.

2.4. PXS score model retraining

The base PXS score Siamese neural network model was retrained (“tuned”) using the 250 CXR training/validation partition of the Hospital 1 Outpatient Dataset, using the same training strategy with mean square error (MSE) loss as previously reported.^[7] In brief, random CXR image pairs were fed to the Siamese neural network. The difference between the Euclidean distance between the final fully connected layers of the network and the absolute difference in mRALE scores between the 2 input images served as the “error” for the MSE loss function. During model training and validation, 1600 and 200 input image pairs were randomly sampled per epoch, respectively. For training, input images were randomly rotated $\pm 5^\circ$ and then randomly cropped to a scale of 0.8-1 and resized to 320 x 320 pixels. For validation, input images were resized to 336 x 336 pixels and center cropped to 320 x 320 pixels. The model training was implemented in Python (version 3.6.9) with the Pytorch package (version 1.5.0), using the Adam optimizer^[19] (initial learning rate = 0.00002, $\beta_1 = 0.9$, $\beta_2 = 0.999$). Training/validation batch sizes of 8 and early stopping at 7 epochs without improvement in validation loss were set. The lowest validation loss model was saved for evaluation. The code used for model training is available at <https://github.com/QTIM-Lab/PXS-score>.

2.5. PXS score model inference

The PXS score for an image-of-interest is the median of Euclidean distances calculated from paired image inputs passed through the Siamese neural network, where each paired image input consists of the image-of-interest and an image from a pool of N normal CXRs. In this study, we created a set of 15 manually curated normal chest x-rays with varying body habitus and field-of-view from CXRs from Hospital 1 to serve the pool of normal CXRs (age range 18–72 years, 7 women and 8 men).

In some CXR images, primarily in the Hospital 3 dataset, large black borders may surround the actual CXR. Immediately before the histogram normalization step described in the pre-processing step described above, a Python script for automated rectangular cropping for black borders was applied to the image (i.e. border pixels with normalized values < 2 were cropped). Code used for model inference and this cropping step is also available at the GitHub link above.

2.6. Statistics/data visualization

To evaluate differences in sex between the datasets, we used the Chi-square test. To evaluate differences in age and mRALE scores (treated as a continuous variable from 0 to 24), we used the Kruskal-Wallis test and post hoc Mann-Whitney tests (2-sided). Interrater correlations for mRALE labeling and correlations

Table 1.
Summary of dataset characteristics and radiologist mRALE scores.

	Hospital 1 Outpatient Dataset (United States) Patients presenting for outpatient imaging who tested positive by COVID-19 RT-PCR				Hospital 2 Emergency Test Set (Brazil) Patients presenting to emergency department with suspected COVID-19			
	All	Training/validation set	Outpatient test set	P-value*	All	RT-PCR positive	RT-PCR negative	P-value†
CXRs, N	358	250	108		303	203	100	
Unique Patients, N	349	248	106		242	167	75	
Age (years), median (Q1–Q3)	53 (41–64)	52 (41–65)	53 (41–63)	0.9	41 (33–52)	40 (33–50)	44 (33–52)	0.2
Sex, N women (%)	186 (52%)	132 (53%)	54 (50%)	0.7	175 (58%)	113 (56%)	62 (62%)	0.4
mRALE, median (Q1–Q3)	1.0 (0–3.5)	1.0 (0–3.0)	1.0 (0–4.5)	0.2	0.3 (0–2.7)	0.3 (0–2.8)	0.3 (0–1.8)	0.6
mRALE, N (%)								
mRALE = 0	123 (34%)	88 (35%)	35 (32%)		122 (40%)	84 (41%)	38 (38%)	
0 < mRALE ≤ 4	164 (46%)	122 (49%)	42 (39%)		126 (42%)	78 (38%)	48 (48%)	
4 < mRALE ≤ 10	58 (16%)	30 (12%)	28 (26%)		29 (10%)	22 (11%)	7 (7%)	
mRALE > 10	13 (4%)	10 (4%)	3 (3%)		26 (9%)	19 (9%)	7 (7%)	

*P-value for comparison of internal test set with training/validation set;

†p-value for comparison of patients who tested positive vs negative by COVID-19 RT-PCR.

mRALE, Modified Radiographic Assessment of Lung Edema, N, Number, Q1–Q3, Quartile 1 to Quartile 3 (i.e. interquartile range).

between PXS score and mRALE were assessed using Pearson correlations (*R*). Statistical tests were performed using the *scipy* Python package (version 1.1.0), with an a priori threshold for statistical significance set at *P* < 0.05.

The *Seaborn* Python package (version 0.10.0) was used for scatterplot data visualizations. To perform dimensionality reduction for visualizing the neural network results, we used the Python implementation of UMAP (Uniform Manifold Approximation and Projection) (version 0.4.2) (number of neighbors = 20, minimum distance = 0.6, metric = correlation).^{120,211} Each test set image was passed through a single subnetwork of the Siamese neural network and the last fully connected layer (1000 nodes in DenseNet121) from each image was used as an input for UMAP.

3. Results

3.1. Chest radiograph dataset characteristics

The Hospital 1 Outpatient Dataset (including training/validation and test partitions) and Hospital 2 Emergency Test Set characteristics are summarized in Table 1. The Hospital 1 Inpatient Test Set and Hospital 3 Inpatient Test Set characteristics were previously published.¹⁷ There were significantly different age distributions between the test sets (*P* = 0.003) (Fig. 2A).

The Hospital 1 Outpatient Test Set patient ages were significantly lower compared to the Hospital 1 and 3 Inpatient Test Sets (median 53 vs 59 years, *P* = 0.003, and median 53 vs 74 years, *P* < 0.001, respectively). The Hospital 2 Emergency Test Set ages were significantly lower than the Hospital 1 Outpatient Test Set ages (median 41 vs 53 years, *P* < 0.001). There was a significantly higher proportion of CXRs from women in the dataset from Brazil compared to the combined datasets from the United States (58% vs 45%, *P* = 0.001). The X-ray equipment used to obtain these CXRs came from a variety of manufacturers that differed by dataset (Table 2).

3.2. Radiologist annotations of COVID-19 lung disease severity

The correlation between the 2 raters at Hospital 1 for assigning mRALE scores to the 358-CXR Hospital 1 Outpatient Dataset was high (*R* = 0.89, *P* < 0.001). The correlation between the 3 raters at Hospital 2 for assigning mRALE scores to the 303-CXR Hospital 2 Emergency Test Set was lower (*R*=0.85, 0.81, and 0.84, for each pairwise comparison; *P*<0.001 in all comparisons). In the 69-CXR subset of the Hospital 2 Emergency Test Set that the Hospital 1 raters also evaluated, the correlation between the average Hospital 1 and average Hospital 2 rater mRALE scores was 0.86 (*P* < 0.001). However, the individual Hospital 2 rater mRALE

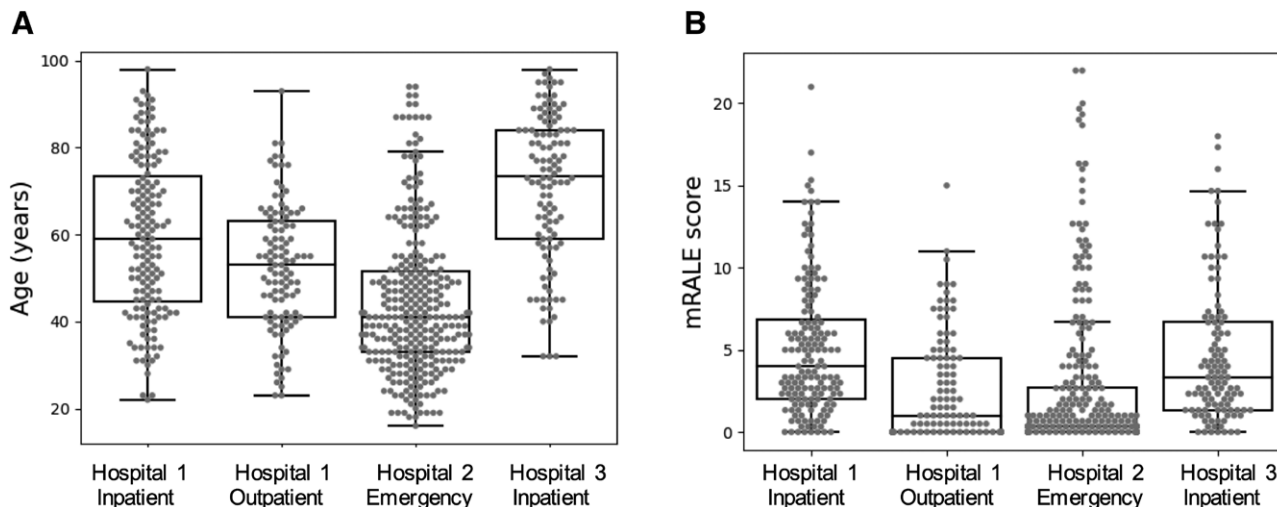


Figure 2. Boxplots show variable distributions in patient age (A) and lung disease severity by mRALE score (B) in the different CXR test sets. Boxplots show the median and interquartile range (IQR), where the whiskers extend up to 1.5 x IQR.

Table 2.

Summary of x-ray equipment manufacturers extracted from DICOM metadata.

Dataset	Manufacturer (headquarters)	Number of CXRs
Hospital 1 inpatient test set (United States)	Agfa (Mortsel, Belgium)	136
	GE Healthcare (Chicago, USA)	1
	Varian (Palo Alto, USA)	4
	Not available	13
Hospital 1 outpatient test set (United States)	Agfa (Mortsel, Belgium)	108
Hospital 2 emergency test set (Brazil)	Fujifilm Corporation (Tokyo, Japan)	303
Hospital 3 inpatient test set (United States)	Agfa (Mortsel, Belgium)	33
	Carestream (Rochester, USA)	72
	Kodak (Rochester, USA)	2
	Phillips (Amsterdam, Netherlands)	2
	Siemens (Munich, Germany)	2

scores showed variable correlation with the average Hospital 1 raters ($R = 0.65, 0.75, 0.86$).

There were significantly different mRALE distributions between the test sets ($P=0.011$) (Fig. 2B). The mRALE scores were significantly lower in the Hospital 1 Outpatient Dataset compared to the Hospital 1 and 3 Inpatient Test Sets (median 1.0 vs 4.0, $P<0.001$, and median 1.0 vs 3.3, $P<0.001$). The mRALE scores in the Hospital 2 Emergency Test Set were significantly lower compared to each of the other test sets (all $P<0.001$).

3.3. Deep learning model performance and testing of generalizability

The PXS score model tuned using the Hospital 1 Outpatient Training/Validation Set showed similar correlation between the model output (PXS score) and radiologist-determined mRALE scores in the Hospital 1 Inpatient and Hospital 3 Inpatient Test Sets ($R = 0.88$ and $R = 0.90$, respectively, $P < 0.001$; compared to $R = 0.86$ and $R = 0.86$ using the baseline model) (Fig. 3A, 3D)

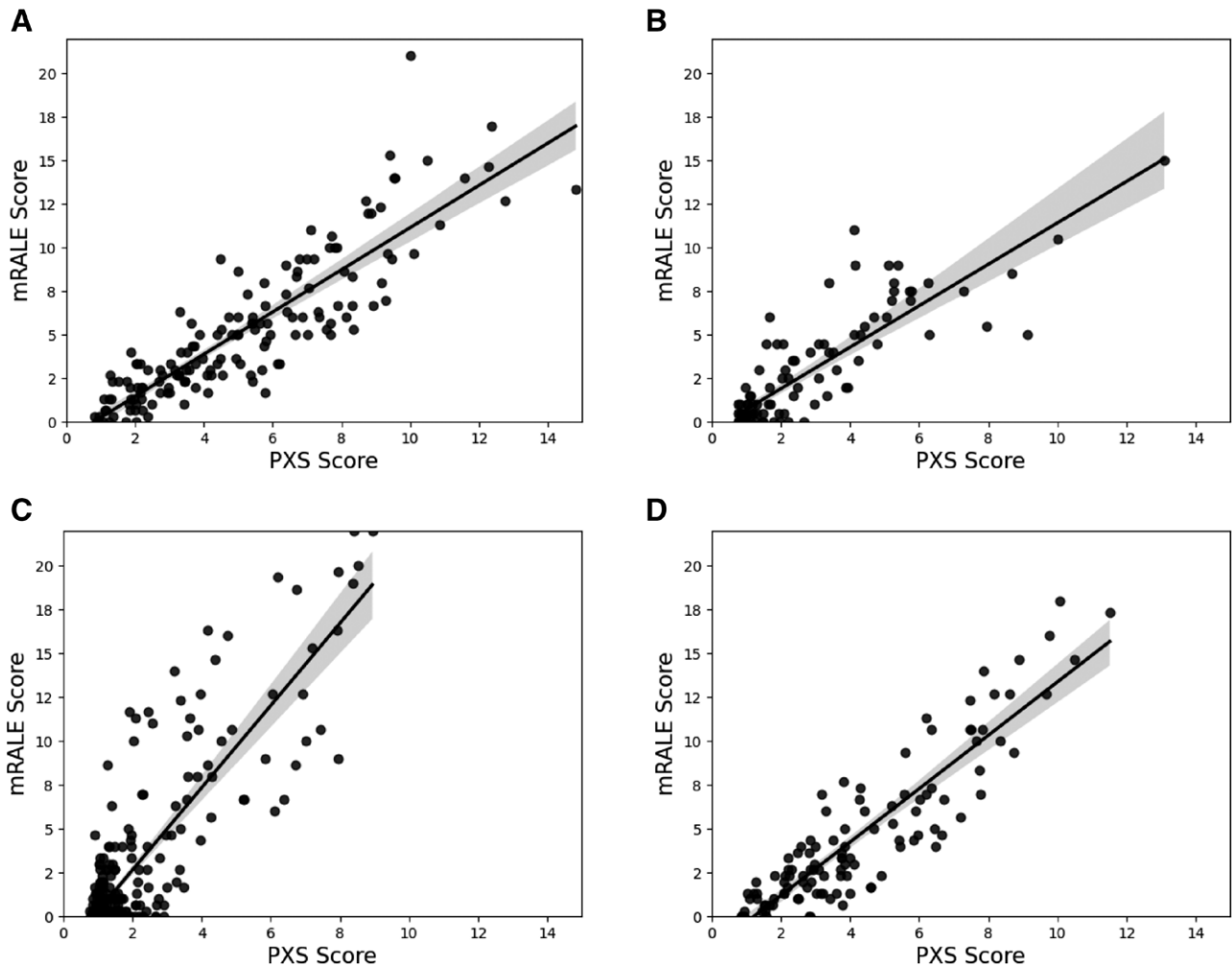


Figure 3. Scatterplots show the correlation between radiologist-determined mRALE score and the deep learning-based PXS score in the Hospital 1 Inpatient Test Set ($R = 0.88$) (A), Hospital 1 Outpatient Test Set ($R = 0.86$) (B), Hospital 2 Emergency Test Set ($R = 0.85$) (C), and Hospital 3 Inpatient Test Set ($R = 0.90$) (D). Linear regression 95% confidence intervals are shown in each scatterplot.

We further tested this tuned PXS score model on the Hospital 1 Outpatient and Hospital 2 Emergency Test Sets, which showed that the model could generalize to these additional datasets ($R = 0.86$ and $R = 0.85$ respectively, $P < 0.001$) (Fig. 3B, C). However, there was a steeper slope for the regression on the Hospital 2 Emergency Test Set data (slope = 2.3) vs the slope of the other test sets (in aggregate, slope = 0.6). While the model learned a measure of disease severity as evidenced by the significant correlation between mRALE and PXS scores, for this specific test set, the relationship between mRALE and PXS was scaled differently.

3.4. Visualizing test set relationships using dimensionality reduction

When CXRs from all 4 test sets were analyzed in aggregate (total $N = 678$), UMAP showed that the CXRs appear to cluster principally in relation to similar disease severity (PXS and mRALE scores) (Fig. 4A, B). Contrastingly, there was substantial overlap between the CXRs from different test sets (Fig. 4C). These findings support the finding that the PXS score model learned a generalizable representation of lung disease severity. However, the normal or near-normal severity CXRs appear to have a larger cluster in the Hospital 2 Emergency Test Set compared to the other test sets (Fig. 4C). We visually inspected these images and did not find a systematic perceptible difference in view position, body habitus, heart size, or x-ray exposure.

4. Discussion

We demonstrated the generalizability of a deep learning-based PXS score model for assessment of a quantitative measure of COVID-19 lung disease severity on CXRs on 4 test sets reflective of different populations from the United States and Brazil. The PXS model was originally trained using admission CXRs from hospitalized COVID-19 patients.^[7] In this study, we found that tuning the deep learning model using CXR data from outpatients showed similar performance on the test sets from hospitalized patients. Based on the correlation of the model results with manual radiologist annotations for lung disease severity in multiple test sets, the PXS score model does appear to generalize to different patient cohorts. This may be because the model was able to learn from different distributions of data, including both inpatients as in the original model and outpatients. Further supporting this conclusion, a dimensionality reduction technique showed that the CXRs from different test sets cluster primarily by lung disease severity, as opposed to by test set source.

While the correlation between the radiologist-determined severity score (mRALE) and the deep learning-based PXS score was generalizable between test sets, there was a difference in the mRALE/PXS slope between the test sets from the United States and Brazil. Thus, differences in calibration between mRALE and PXS scores may occur for CXRs coming from different sources and this needs to be considered before the use of such a model clinically. This phenomenon could be due to systematic differences in x-ray equipment manufacturers and acquisition technique (including parameters like x-ray tube voltage and current), which can alter the properties of tissue contrast in the image. Subjectively, our radiologist raters found a perceptible difference in exposure/contrast in the images from Brazil vs the United States. The PXS score model attempts to address this issue using histogram normalization, but this transformation may not eliminate all systematic differences. Model training on data from an increased variety of vendors could help address this calibration issue and is a direction of future research.

Despite the calibration issue, the finding that the PXS score model was able to correlate with manual radiologist annotations at multiple test sites has potential clinical application for the reproducible assessment of COVID-19 lung disease severity at different sites. This reproducible assessment is important

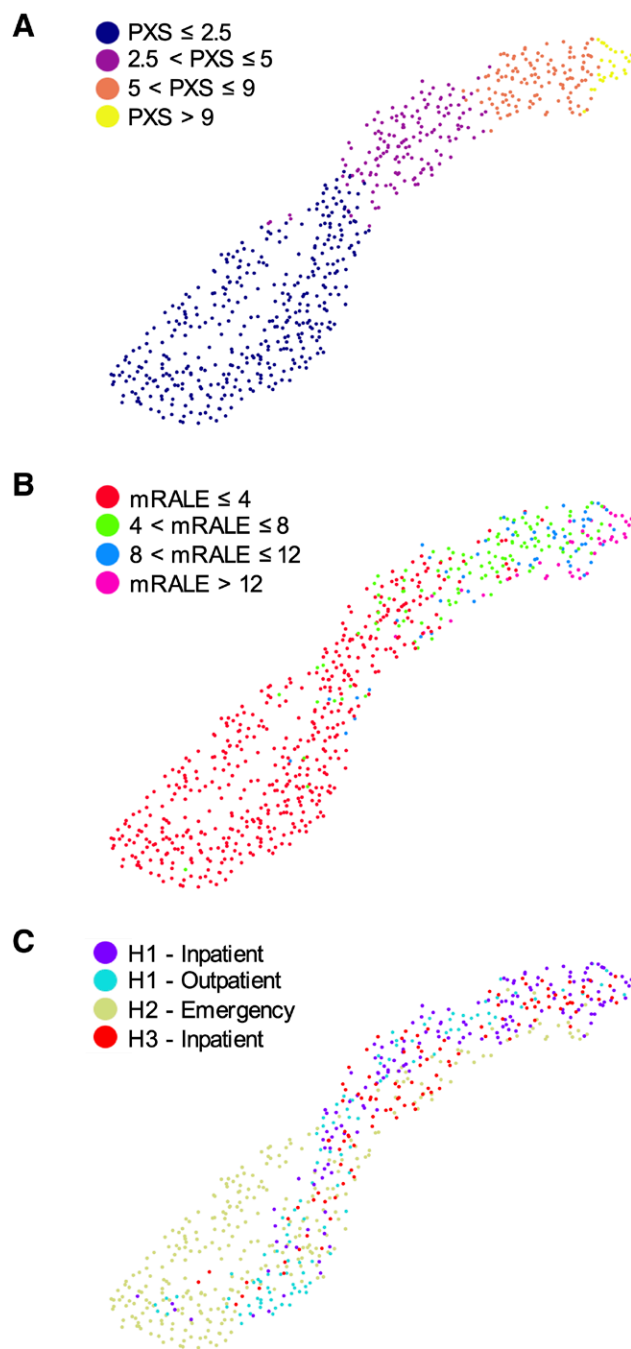


Figure 4. Dimensionality reduction using UMAP shows the relationships between CXR data passed through the deep learning-based PXS score model from all 4 test sets (total $N = 678$), color coded for PXS score (A), mRALE score (B), and test set (C). For the legend in (C), *H* indicates Hospital. Across the different test sets, a representation of lung disease severity is learned by the PXS score model.

because CXR findings have been associated with worse clinical outcomes in patients with COVID-19,^[4-6] which may be useful for clinical risk stratification, and there is interrater variation between radiologist assessments (which will be more pronounced in the “real world” where radiologists are not uniformly trained on the use of a scoring system). Another possible application is for radiologist worklist prioritization, which could help expedite identification of the sickest patients.^[22]

Previous work on developing deep-learning based models to assess COVID-19 lung disease severity on CXRs have shown correlations between various systems of manual radiologist

assessments and deep learning outputs, though often without external testing. For example, Cohen et al split their 94 PA CXR dataset 50:50 for training and testing,^[12] Zhu et al split their 131 portable CXR dataset 80:20 for training and testing,^[8] and Blain et al reported performance on a 65 CXR dataset using 5-fold cross validation.^[10] On the other hand, work from Amer et al and Signoroni et al^[9,11] does include external testing of their deep learning models on CXRs from the Cohen et al dataset,^[12] and Barbosa et al also perform external testing on an 86 CXR dataset.^[23] Future work in this field should continue to include assessment of model performance across multiple sites, to characterize generalizability for different x-ray acquisition techniques and patient populations before these artificial intelligence-based tools can be deployed for possible clinical use. In particular, training and testing using heterogeneous patient populations will be critical for ensuring that such models benefit diverse patients and avoid the worsening of health disparities.^[24]

There are limitations to this study. First, the reference standard label used for disease severity assessment on CXRs is determined by radiologists, which has inherent variability. Furthermore, the United States CXRs were annotated by thoracic subspecialist radiologists, while the Brazilian CXRs were annotated by nonthoracic subspecialty training, leading to a potential systematic difference in the reference standard between these datasets. We used the average of multiple radiologist raters for the reference standard to decrease the variability in this study. However, other reference standards such as CT-derived scores may be promising, as has been found using digitally reconstructed radiographs from CT.^[23] Second, while studying the technical properties of deep learning-based models like PXS score is necessary, making such CXR-based severity scores clinically useful in addressing the COVID-19 pandemic is a different avenue of important research. Future work into how radiologists and other clinicians can use the PXS score (and other developed lung disease severity scores) to guide patient management or workflows will be essential to deliver value.

In this work, we show that the performance of a deep learning model that extracts a COVID-19 severity score on CXRs can generalize across multiple populations from 2 continents, including outpatients and hospitalized patients.

Acknowledgment

We thank members of the QTIM Lab at Massachusetts General Hospital, and the MGH and BWH Center for Clinical Data Science for their support in this work.

Author contributions

Study concepts/study design, M.D.L., F.C.K., J.K.C.; data acquisition and data analysis/interpretation, all authors; manuscript drafting or manuscript revision for important intellectual content, all authors; approval of final version of submitted manuscript, all authors.

References

- [1] Wong HYF, Lam HYS, Fong AH-T, et al. Frequency and Distribution of Chest Radiographic Findings in Patients Positive for COVID-19. *Radiology*. 2019;296:E72–8.
- [2] Smith DL, Grenier J-P, Batte C, et al. A characteristic chest radiographic pattern in the setting of COVID-19 pandemic. *Radiol Cardiothorac Imaging*. 2020;2:e200280.
- [3] Cozzi D, Albanesi M, Cavigli E, et al. Chest X-ray in new coronavirus disease 2019 (COVID-19) infection: findings and correlation with clinical outcome. *Radiol Medica*. 2020;125:730–7.
- [4] Toussie D, Voutsinas N, Finkelstein M, et al. Clinical and chest radiography features determine patient outcomes in young and middle-aged adults with COVID-19. *Radiology*. 2020;297:E197–206.
- [5] Joseph NP, Reid NJ, Som A, et al. Racial and ethnic disparities in disease severity on admission chest radiographs among patients admitted with confirmed coronavirus disease 2019: a retrospective cohort study. *Radiology*. 2020;297:E303–12.
- [6] Kim HW, Capaccione KM, Li G, et al. The role of initial chest X-ray in triaging patients with suspected COVID-19 during the pandemic. *Emerg Radiol*. 2020;27:617–21.
- [7] 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.
- [8] Zhu J, Shen B, Abbasi A, et al. Deep transfer learning artificial intelligence accurately stages COVID-19 lung disease severity on portable chest radiographs. *PLoS One*. 2020;15:e0236621.
- [9] Signoroni A, Savardi M, Benini S, et al. End-to-end learning for semi-quantitative rating of COVID-19 severity on chest X-rays. 2020. Available at: <http://arxiv.org/abs/2006.04603> [accessed September 8, 2020].
- [10] Blain M, Kassim MT, Varble N, et al. Determination of disease severity in COVID-19 patients using deep learning in chest X-ray images. *Diagn Interv Radiol*. 2020;27:20–7.
- [11] Amer R, Frid-Adar M, Gozes O, et al. COVID-19 in CXR: from detection and severity scoring to patient disease monitoring. 2020. Available at: <http://arxiv.org/abs/2008.02150> [accessed September 11, 2020].
- [12] Cohen JP, Dao L, Roth K, et al. Predicting COVID-19 Pneumonia Severity on Chest X-ray with deep learning. *Cureus*. 2020;12:e9448.
- [13] 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:e1002683.
- [14] Bromley J, Bentz JW, Bottou L, et al. Signature verification using a “Siamese” time delay neural network. *Int J Pattern Recognit Artif Intell*. 1993;07:669–88.
- [15] 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;3:1–9.
- [16] Huang G, Liu Z, van der Maaten L, et al. Densely connected convolutional networks. *Proc - 30th IEEE Conf Comput Vis Pattern Recognition, CVPR 2017*. 2016:2261–2269. Available at: <http://arxiv.org/abs/1608.06993> [accessed March 29, 2020].
- [17] 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 January 4, 2020].
- [18] 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;73:840–6.
- [19] Kingma DP, Ba J. Adam: a method for stochastic optimization. 2014. Available at: <http://arxiv.org/abs/1412.6980> [accessed June 16, 2019].
- [20] McInnes L, Healy J, Saul N, et al. UMAP: uniform manifold approximation and projection. *J Open Source Softw*. 2018;3:861.
- [21] McInnes L, Healy J, Melville J. UMAP: uniform manifold approximation and projection for dimension reduction. 2018. Available at: <http://arxiv.org/abs/1802.03426> [accessed June 11, 2019].
- [22] Richardson ML, Garwood ER, Lee Y, et al. Noninterpretive uses of artificial intelligence in radiology. *Acad Radiol*. 2021;28:1225–35.
- [23] Barbosa EM, Geftter WB, Yang R, et al. Automated detection and quantification of COVID-19 airspace disease on chest radiographs: a novel approach achieving radiologist-level performance using a CNN trained on digital reconstructed radiographs (DRRs) from CT-based ground-truth. 2020. Available at: <http://arxiv.org/abs/2008.06330> [accessed September 11, 2020].
- [24] Zou J, Schiebinger L. Ensuring that biomedical AI benefits diverse populations. *EBioMedicine*. 2021;6:7.