Radiology

ORIGINAL RESEARCH • THORACIC IMAGING

DeepCOVID-XR: An Artificial Intelligence Algorithm to Detect COVID-19 on Chest Radiographs Trained and Tested on a Large U.S. Clinical Data Set

Ramsey M. Wehbe, MD • Jiayue Sheng, BS • Shinjan Dutta, BS • Siyuan Chai • Amil Dravid • Semih Barutcu, MS • Yunan Wu, MS • Donald R. Cantrell, MD, PhD • Nicholas Xiao, MD • Bradley D. Allen, MD, MS • Gregory A. MacNealy, MD • Hatice Savas, MD • Rishi Agrawal, MD • Nishant Parekh, MD • Aggelos K. Katsaggelos, PhD

From the Division of Cardiology, Department of Medicine and Bluhm Cardiovascular Institute (R.M.W.), Division of Neurointerventional Radiology (D.R.C.), Division of Interventional Radiology (N.X.), and Division of Thoracic Imaging (B.D.A., G.A.M., H.S., R.A., N.P.), Department of Radiology, Northwestern Memorial Hospital, 676 N St Clair St, Chicago, IL 60611; and Department of Electrical and Computer Engineering, McCormick School of Engineering, Northwestern University, Evanston, Ill (J.S., S.D., S.C., A.D., S.B., Y.W., A.K.K.). Received August 21, 2020; revision requested September 30; revision received October 26; accepted October 30. **Address correspondence to** R.M.W. (e-mail: *ramsey.wehbe@northwestern.edu*).

Conflicts of interest are listed at the end of this article.

See also the editorial by van Ginneken in this issue.

Radiology 2021; 299:E167–E176 • https://doi.org/10.1148/radiol.2020203511 • Content code: CH

Background: There are characteristic findings of coronavirus disease 2019 (COVID-19) on chest images. An artificial intelligence (AI) algorithm to detect COVID-19 on chest radiographs might be useful for triage or infection control within a hospital setting, but prior reports have been limited by small data sets, poor data quality, or both.

Purpose: To present DeepCOVID-XR, a deep learning AI algorithm to detect COVID-19 on chest radiographs, that was trained and tested on a large clinical data set.

Materials and Methods: DeepCOVID-XR is an ensemble of convolutional neural networks developed to detect COVID-19 on frontal chest radiographs, with reverse-transcription polymerase chain reaction test results as the reference standard. The algorithm was trained and validated on 14 788 images (4253 positive for COVID-19) from sites across the Northwestern Memorial Health Care System from February 2020 to April 2020 and was then tested on 2214 images (1192 positive for COVID-19) from a single hold-out institution. Performance of the algorithm was compared with interpretations from five experienced thoracic radiologists on 300 random test images using the McNemar test for sensitivity and specificity and the DeLong test for the area under the receiver operating characteristic curve (AUC).

Results: A total of 5853 patients (mean age, 58 years ± 19 [standard deviation]; 3101 women) were evaluated across data sets. For the entire test set, accuracy of DeepCOVID-XR was 83%, with an AUC of 0.90. For 300 random test images (134 positive for COVID-19), accuracy of DeepCOVID-XR was 82%, compared with that of individual radiologists (range, 76%–81%) and the consensus of all five radiologists (81%). DeepCOVID-XR had a significantly higher sensitivity (71%) than one radiologist (60%, $P < .001$) and significantly higher specificity (92%) than two radiologists (75%, $P < .001$; 84%, $P = .009$). AUC of DeepCOVID-XR was 0.88 compared with the consensus AUC of 0.85 ($P = .13$ for comparison). With consensus interpretation as the reference standard, the AUC of DeepCOVID-XR was 0.95 (95% CI: 0.92, 0.98).

Conclusion: DeepCOVID-XR, an artificial intelligence algorithm, detected coronavirus disease 2019 on chest radiographs with a performance similar to that of experienced thoracic radiologists in consensus.

Supplemental material is available for this article.

Coronavirus disease 2019 (COVID-19) is responsible for over 40 million cases and over 1.1 million deaths worldwide as of October 22, 2020 (1) and has strained critical health care resources. Although the reference standard for diagnosis of COVID-19 is a reverse-transcription polymerase chain reaction (RT-PCR) assay for severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) viral RNA, there are characteristic findings of COVID-19 on chest CT images or chest radiographs. This has inspired multiple efforts at developing an artificial intelligence (AI) algorithm for automated diagnosis of COVID-19 on chest images. However, imaging findings are neither sensitive nor specific enough to be used as a diagnostic tool for COVID-19, and studies that suggest otherwise are limited by selection bias (2–4). Instead, a potential application of AI for chest imaging analysis is for triage or infection control programs in a hospital or emergency department setting to provide early identification of patients with suspicious findings on chest images for further testing and isolation.

Although there have been promising results using AI for detection of COVID-19 on CT images (5–7), the use of CT for this purpose is limited by concerns regarding cost, time, radiation exposure, and decontamination procedures for equipment (8). In contrast, chest radiography can be

Abbreviations

AI = artificial intelligence, AUC = area under the receiver operating characteristic curve, CNN = convolutional neural network, COVID-19 = coronavirus disease 2019, RT-PCR = reverse-transcription polymerase chain reaction, SARS-CoV-2 = severe acute respiratory syndrome coronavirus 2

Summary

DeepCOVID-XR, an artificial intelligence algorithm for detecting coronavirus disease 2019 on chest radiographs, demonstrated performance similar to that of the consensus of experienced thoracic radiologists.

Key Results

- DeepCOVID-XR classified 2214 test images (1194 positive for coronavirus disease 2019 [COVID-19]) with an accuracy of 83% and an area under the receiver operating characteristic curve (AUC) of 0.90 compared with the reference standard of reverse-transcription polymerase chain reaction.
- On 300 random test images (134 positive for COVID-19), accuracy of DeepCOVID-XR was 82% (AUC, 0.88), as compared with five individual thoracic radiologists (accuracy, 76%–81%) and the consensus of all five radiologists (accuracy, 81%; AUC, 0.85).
- When using consensus interpretation of the radiologists as the reference standard, AUC of DeepCOVID-XR was 0.95.

performed rapidly and at the bedside, involves trivial radiation exposure, and poses a lower risk for viral spread (9,10). However, previously reported AI algorithms for identification of COVID-19 using chest radiography have been limited by small data sets and the use of publicly available images of variable quality and questionable validity (11–16). Here, we present DeepCOVID-XR, a deep learning algorithm to detect chest radiographs suspicious for COVID-19. DeepCOVID-XR was trained and tested on a large data set of clinical images from a major U.S. health care system (to our knowledge, the largest clinical data set of chest radiographs from the COVID-19 era used to train a published AI platform to date). In this study, we compare the performance of the DeepCOVID-XR algorithm with interpretations by experienced thoracic radiologists.

Materials and Methods

Patients

This retrospective study was approved by the Northwestern institutional review board (STU00212323) and was granted a waiver of Health Insurance Portability and Accountability Act authorization and a waiver of written informed consent. Our study sample included consecutive patients from over 20 sites (including hospitals, stand-alone emergency departments, and urgent care facilities) (Table E1 [online]) across the Northwestern Memorial Health Care System who were tested for COVID-19 from February 2020 to April 2020. Patients included adults aged at least 18 years with a documented RT-PCR test result for SARS-CoV-2 (whether positive or negative); a diagnosis of COVID-19 as defined by the *International Classification of Diseases, Tenth Revision* (ICD-10) code; or a COVID-19 definitive positive flag in the electronic health record. COVID-19 positivity was defined as any one positive RT-PCR result for SARS-CoV-2 during the associated clinical encounter (eg, a patient with multiple RT-PCR tests and only one positive result would be considered positive); a diagnosis of COVID-19 as defined by the ICD-10 code; or a COVID-19 definitive positive flag in the electronic health record (most patients with a diagnosis indicated only by the ICD-10 code or a COVID-19 definitive positive flag in the electronic health record had a prior documented positive RT-PCR test for SARS-CoV-2 at an institution outside the Northwestern Memorial Health Care System). Patients with only negative documented RT-PCR test results for SARS-CoV-2 during their clinical encounter were labeled as COVID-19 negative.

Image Labeling and Data Set Partitioning

Every chest radiograph obtained during the study period in patients who met inclusion criteria was included, regardless of quality. All chest radiographs acquired during a given clinical encounter were labeled as positive or negative for COVID-19 based on the previously mentioned patient-level criteria, regardless of the timing of chest radiographic findings compared with RT-PCR results. Images were filtered to include only frontal projections (ie, bedside anteroposterior images and only posteroanterior images from posteroanterior or lateral acquisitions).

Images from Northwestern Memorial Health Care's major academic teaching hospital, Northwestern Memorial Hospital, were combined with those from other sites, with the exception of images from a single community hospital, Lake Forest Hospital, which were held out as a test set that the algorithm was never exposed to during training or validation. Images from Northwestern Memorial Hospital and the other sites were then split into training and validation sets in an 80%-20% fashion (while ensuring no crossover of patients among groups). Figure 1 shows the breakdown of training, validation, and test data sets.

DeepCOVID-XR: An Ensemble of Deep Neural Networks for COVID-19 Prediction

Details regarding image preprocessing; the architecture of the deep convolutional neural network (CNN) ensemble model; algorithm training, validation, and testing; and saliency heatmap generation are provided in Appendix E1 (online). Briefly, DeepCOVID-XR is a weighted ensemble of deep neural networks (Fig 2). Every image in the data set is first preprocessed to produce four separate images (cropped and uncropped images at resolutions of 224 × 224 pixels and 331 × 331 pixels). Each image is then fed into six previously validated CNN architectures—DenseNet-121 (17), ResNet-50 (18), InceptionV3 (19), Inception-ResNetV2 (20), Xception (21), and EfficientNet-B2 (22)—for a total of 24 individually trained CNNs that served as members of the deep learning model ensemble. The CNNs in this ensemble were pretrained on a large publicly available data set of over 100 000 chest radiographs from the National Institutes of Health (23) and were then fine-tuned on our clinical training set of chest radiographs from the COVID-19 era using transfer learning. The validation data set was used to optimize hyperparameters. The final binary prediction of the neural network architecture was a weighted average of the predictions of these individual CNNs, classifying images as either COVID-19 positive or COVID-19 negative using an out-

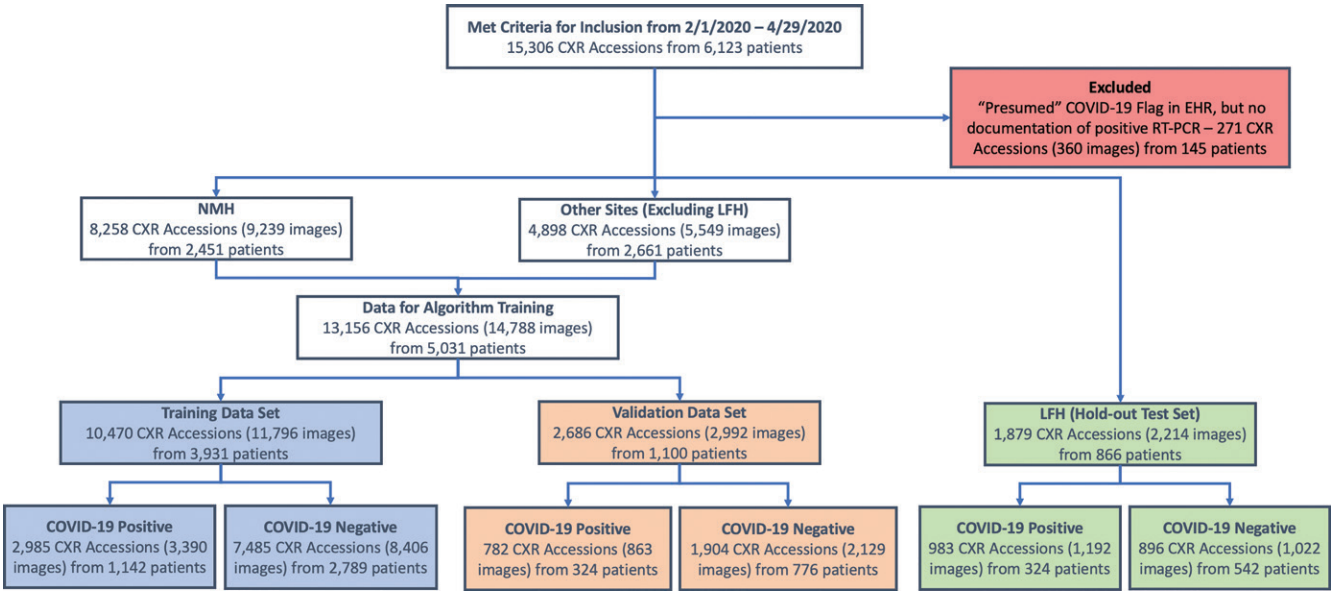


Figure 1: Flowchart for patient inclusion in the study and breakdown of training, validation, and hold-out test data sets. COVID-19 = coronavirus disease 2019, CXR = chest radiography, EHR = electronic health record, LFH = Lake Forest Hospital, NMH = Northwestern Memorial Hospital, RT-PCR = reverse-transcription polymerase chain reaction

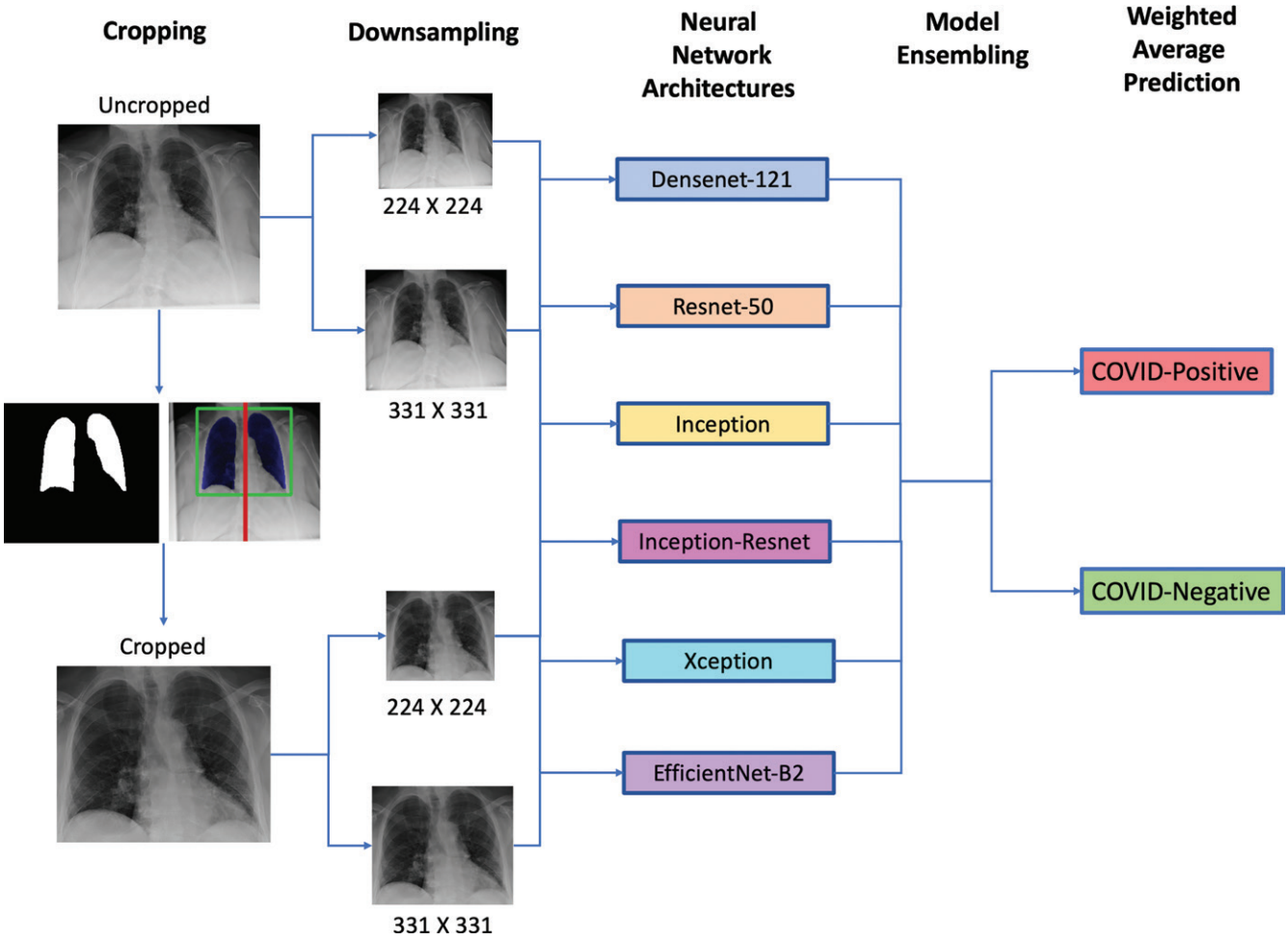


Figure 2: Schematic shows the general architecture of the DeepCOVID-XR deep learning ensemble model. Images are initially preprocessed to crop a square centered on the lungs, and then both uncropped and cropped images are downsampled to two different resolutions (224 × 224 pixels and 331 × 331 pixels) before being fed into each of six different previously validated convolutional neural network architectures (four images × six architectures = 24 models total). The predictions from each individual model are then ensembled using a weighted average to produce a single prediction of coronavirus disease 2019 (COVID-19) positivity or COVID-19 negativity for each image.

Table 1: Baseline Characteristics of Patients

Characteristic	Overall	Training			Validation			Testing		
		Total	COVID-19–pos	COVID-19–neg	Total	COVID-19–pos	COVID-19–neg	Total	COVID-19–pos	COVID-19–neg
Patient-level data	5853	3931	1142	2789	1100	324	776	866	324	542
Age	58 ± 19	58 ± 18	57 ± 18	58 ± 19	59 ± 18	56 ± 16	60 ± 19	57 ± 18	56 ± 17	58 ± 19
Female sex	3101 (53)	2081 (53)	577 (51)	1504 (54)	580 (53)	149 (46)	431 (56)	440 (54)	161 (51)	279 (55)
Inpatient	3629 (62)	2413 (61)	719 (63)	1694 (61)	672 (61)	216 (67)	456 (59)	544 (66)	237 (75)	307 (61)
Chest radiographs per patient	1 (1–3)	1 (1–3)	2 (1–3)	1 (1–3)	1 (1–2)	1 (1–3)	1 (1–2)	1 (1–2)	1 (1–2)	1 (1–2)
Chest radiography–level data	17 002	11 796	3390	8406	2992	863	2129	2214	1192	1022
AP, PA, and not listed	14 843 (87), 1105 (6), 1054 (6)	10 200 (86), 781 (7), 815 (7)	3024 (89), 111 (3), 255 (8)	7176 (85), 670 (8), 560 (7)	2502 (84), 264 (9), 226 (8)	770 (89), 32 (4), 61 (7)	1732 (81), 232 (11), 165 (8)	2141 (97), 60 (3), 13 (0)	1183 (99), 8 (1), 1 (0)	958 (94), 52 (5), 12 (1)
Chest radiograph prior to first positive RT-PCR result	NA	NA	1161 (34)	NA	NA	241 (28)	NA	NA	263 (22)	NA
Hours from chest radiography to first positive RT-PCR result*	NA	NA	34 (6–273)	NA	NA	21 (3–76)	NA	NA	10 (1–40)	NA

Note.—For categorical variables, values are presented as number of patients, with the percentage in parentheses. For continuous variables, values are presented as mean ± standard deviation for normally distributed data and as the median and interquartile range for nonnormally distributed data. AP = anteroposterior, COVID-19 =coronavirus disease 2019, NA = not applicable, neg = negative, PA = posteroanterior, pos = positive, RT-PCR = reverse-transcription polymerase chain reaction.

* Among images that were acquired prior to first positive RT-PCR result.

put threshold of greater than 0.5 (on a scale from 0 to 1). Gradient class activation mapping (24) was used to produce heatmaps to visualize feature importance for arriving at a prediction of COVID-19 positivity. Our code base, including trained weights for each of the 24 individual neural network architectures and their respective model weights for the weighted ensemble, is provided freely on GitHub at *https://github.com/IVPLatNU/deepcovidxr*.

Comparison with Experienced Thoracic Radiologist Interpretations

Three hundred images were selected at random from the hold-out test data set (ensuring only one image per patient and no patient overlap with training or validation sets) for expert interpretation. Expert interpretations were independently provided by five radiologists: four board-certified thoracic radiologists (R.A., N.P., H.S., and B.A.) with 8 years, 6 years, 6 years, and 1 year of posttraining experience, respectively, and one board-certified diagnostic radiologist (G.M.) with 38 years of posttraining experience. Radiologists were blinded to any identifying information or clinical characteristics and had access to the full radiologic study in our picture archiving and communication system (ie, radiologists were able to review lateral images for posteroanterior or lateral studies). Radiologists provided an overall interpretation of positive for COVID-19 or negative for COVID-19 (chest radiographs with abnormalities that were not deemed consistent with COVID-19 were graded as negative for COVID-19) and an associated confidence level with this assessment (graded as low, medium, or high). In this way, we derived a six-point scoring system ranging from −3 (high confidence in COVID-19–negative interpretation) to +3 (high confidence in COVID-19–positive interpretation). In addition to the overall interpretation, this six-point scoring system was used to calculate five separate decision thresholds for each radiologist for comparison to the algorithm. Finally, a consensus interpretation for the five radiologists was determined by taking the majority vote (mode) of the individual interpretations, and receiver operating characteristic curves for the consensus interpretation were produced by calculating an average of the six-point scores for all radiologists on each image.

Statistical Analysis

For comparison of DeepCOVID-XR and radiologist interpretations, 95% CIs were produced for sensitivity, specificity, and area

under the receiver operating characteristic curve (AUC) (using 2000 bootstrap samples). Sensitivity and specificity were compared using the McNemar test for paired samples (25), and AUCs were compared using the DeLong test (26). A two-tailed *P* value of .05 was considered to indicate statistical significance. Statistical analyses were performed using the packages DTComPair and pROC in R, version 3.6 (R Core Team; R Foundation for Statistical Computing).

Results

Patient Characteristics

A total of 5853 patients (mean age, 58 years ± 19 [standard deviation]; 3101 women, 1782 with positive results for COVID-19) were evaluated across data sets (Table 1). The rate of positivity for COVID-19 among chest radiographs in the hold-out test set (1192 of 2214 [54%]) was higher than in the training (3390 of 11 786 [29%]) and validation (863 of 2992 [29%]) sets. The proportion of patients with positive COVID-19 findings who underwent inpatient treatment (237 of 324 [75%]) was higher in the test set than in the training (719 of 1142; 63%) or validation (216 of 324; 67%) sets. Additionally, there was a higher proportion of anteroposterior images (97% [2141 of 2214]) in the test set than in the training (86% [10 200 of 11,786]) and validation (84% [2502 of 2992]) sets. In the test set, 263 of 1192 (22%) COVID-19–positive imaging findings were acquired prior to positive RT-PCR results.

Performance of DeepCOVID-XR

A performance comparison of individual model architectures and ensemble models is provided in the supplemental materials (Tables E2, E3 [online]). In the hold-out test set of 2214 images, the overall accuracy of DeepCOVID-XR for predicting COVID-19 was 83% (1846 of 2214), with a sensitivity of 75% (898 of 1192), a specificity of 93% (948 of 1022), and an AUC of 0.90 (confusion matrix in Fig 3a, receiver operating characteristic curve in Fig 3b). Notably, 156 of 1192 (13%) of COVID-19–positive imaging findings were acquired prior to RT-PCR results and were accurately labeled by the algorithm as suspicious for COVID-19. As approximately 5% (44 patients contributing 151 images) of patients in the hold-out test set also underwent chest radiography at one of the institutions in our training or validation sets during the study period, we performed a sensitivity analysis in which these images were dropped from the test set and the results were unchanged (Table E4 [online]). The most representative images from each class of DeepCOVID-XR predictions (true-positive, true-negative, false-positive, and false-negative predictions) are provided in Figure 4. Gradient class activation mapping heatmaps of feature importance for individual chest radiographs are provided in Figure 5. Heatmaps for COVID-19–positive imaging findings highlighted features in the lungs that identified areas of abnormalities (Fig 5a–5c), in contrast to the heatmaps for COVID-19–negative imaging findings (Fig 5d).

Comparison with Expert Thoracic Radiologists

A comparison of the performance of DeepCOVID-XR with expert chest radiologist interpretations of 300 patients' chest radiographs (134 with positive COVID-19 results) randomly selected from the hold-out test set is provided in Table 2. The overall accuracy of DeepCOVID-XR on this test set was 82% (247 of 300) compared with the reference standard of RT-PCR, whereas the accuracy of individual radiologists ranged from 76% (227 of 300) to 81% (242 of 300) and the accuracy of the consensus interpretation of all five radiologists was 81% (242 of 300). DeepCOVID-XR had a significantly higher specificity (92%, 152 of 166) than two of the radiologists (75% [125 of 166], $P < .001$; 84% [139 of 166], $P = .009$) and significantly higher sensitivity (71% [95 of 134]) than one radiologist (60% [81 of 134]; $P < .001$). A comparison of the receiver operating characteristic curve for DeepCOVID-XR with overall individual radiologist interpretations is provided in Figure 6a. A comparison of DeepCOVID-XR with individual radiologists on each of the five decision thresholds derived from the six-point scoring system is provided in Table E5 (online) and Figure E1 (online).

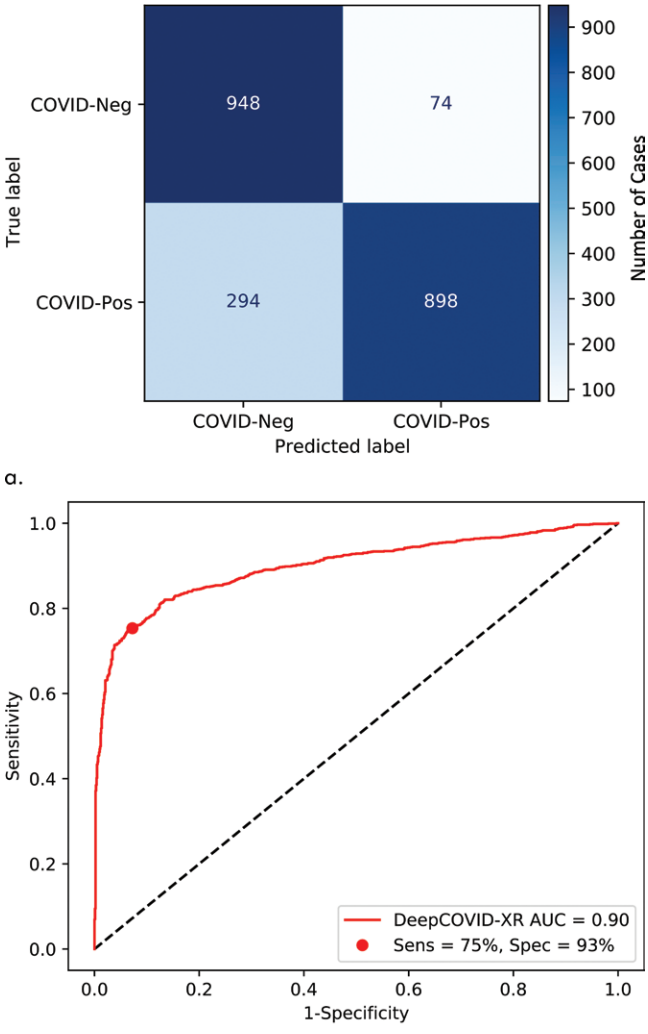


Figure 3: Performance of DeepCOVID-XR on hold-out test set of 2214 images. **(a)** Confusion matrix of algorithm predictions and **(b)** receiver operating characteristic (ROC) curve (red line) show discriminative performance of the algorithm, with an area under the ROC curve (AUC) of 0.90 and a prediction threshold for coronavirus disease 2019 (COVID-19) positivity (red point) with a sensitivity (Sens) of 75% (898 of 1192) and a specificity (Spec) of 93% (948 of 1022). Neg = negative, pos = positive.

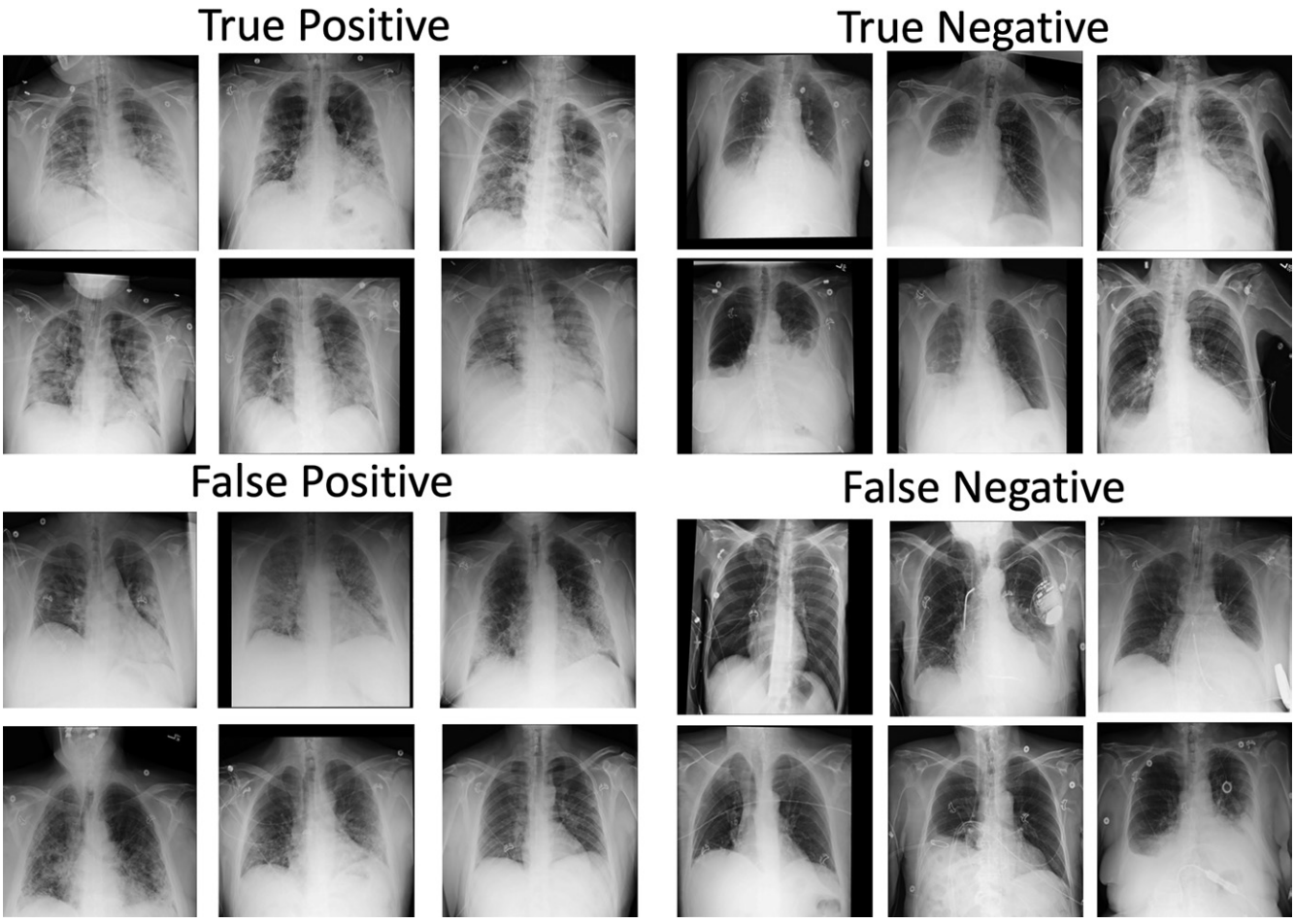


Figure 4: Sample of most representative images from different classes of DeepCOVID-XR predictions relative to the reference standard. Images classified as positive for coronavirus disease 2019 (COVID-19) by the algorithm (whether true-positive or false-positive findings) tend to have the typical features of COVID-19 pneumonia described in the literature, including patchy bilateral consolidations and ground-glass opacities with peripheral and lower-lung predominance. Images classified as negative for COVID-19 by the algorithm tend to show clear lungs, concomitant pleural effusions, or both, which are reported to be rare in COVID-19 pneumonia.

The AUC for DeepCOVID-XR was 0.88 (range, 0.84–0.92) compared with 0.85 (range, 0.80–0.89; P = .13 for comparison) for the consensus interpretation of all five radiologists (Fig 6b). When using the consensus interpretation rather than the RT-PCR assay as the reference standard, the AUC for DeepCOVID-XR was 0.95 (range, 0.92–0.98; Fig 6c). The time to analyze this subset of 300 images with DeepCOVID-XR on one NVIDIA Titan V graphics processing unit was approximately 18 minutes compared with approximately 2.5–3.5 hours for each expert radiologist.

Discussion

In this study, we present DeepCOVID-XR, an ensemble deep learning artificial intelligence (AI) algorithm to detect coronavirus disease 2019 (COVID-19) on chest radiographs. DeepCOVID-XR was trained and tested on, to our knowledge, the largest clinical data set of chest radiographs from the COVID-19 era of any other published AI platform to date, including images from multiple institutions across a large U.S. health care system (17 002 images from 5853 patients total). Of note, study patients were representative of a real-world population of patients presenting for emergency or inpatient care in the COVID-19 era; a proportion of patients with negative COVID-19 test results likely had a spectrum of non–COVID-19–related abnormalities on chest radiographs (including confounding variables like non–COVID-19 viral pneumonia) that one would expect in this patient population. On a hold-out test data set of 2214 images (1192 positive for COVID-19) from a single institution that the algorithm was not exposed to during model development, DeepCOVID-XR detected COVID-19 with an overall accuracy of 83% (sensitivity of 75%, specificity of 93%) and an area under the receiver operating characteristic curve (AUC) of 0.90. Additionally, on a random sample of 300 test images, the accuracy of DeepCOVID-XR was 82% compared with 76%–81% for individual experienced thoracic radiologists and 81% for the consensus interpretation of all five radiologists. Finally, the AUC for DeepCOVID-XR was 0.88 compared with 0.85 for the consensus interpretation (P = .13). When using the consensus radiologist interpretation rather than real-time polymerase chain reaction as the reference standard, the AUC for DeepCOVID-XR was 0.95, suggesting a discriminative ability of our algorithm similar to that of a consensus of experts.

Errors made by the algorithm were explainable. Images categorized as demonstrating positivity (whether true or false positivity) often depicted characteristic features of COVID-19 viral pneumonia that have previously been reported in the literature, including

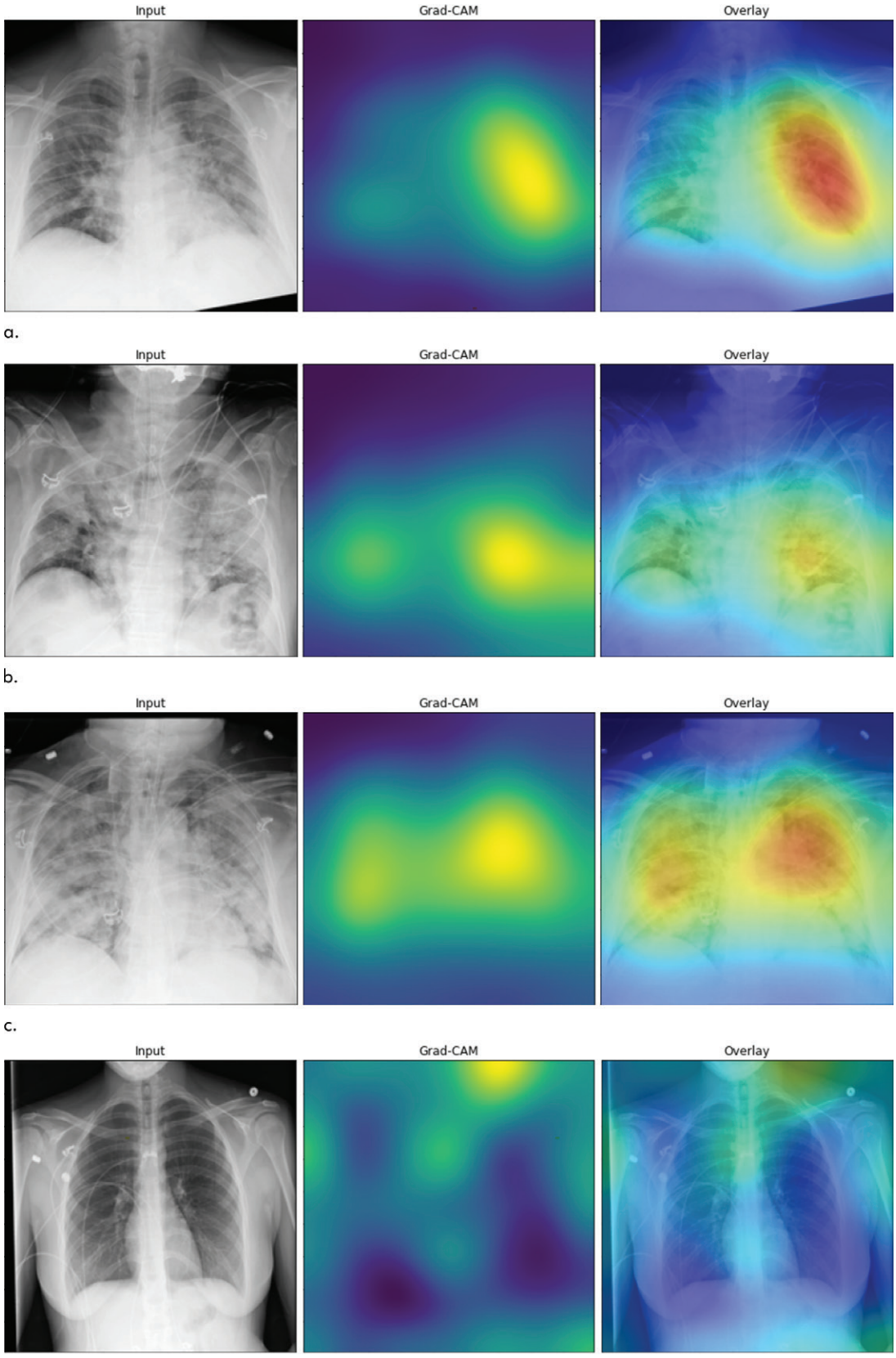


Figure 5: Gradient class activation mapping (Grad-CAM) heatmaps of feature importance for prediction of a positive coronavirus disease 2019 (COVID-19) finding. Generated heatmaps appropriately highlight abnormalities in the lungs in **(a–c)** those images accurately labeled as depicting positive COVID-19 findings, in contrast to **(d)** images that were accurately labeled as depicting negative COVID-19 findings. The intensity of colors on the heatmap corresponds to features of the image that are important for prediction of COVID-19 positivity.

Table 2: Performance of DeepCOVID-XR on Random Sample of 300 Images from Test Set Compared with Expert Thoracic Radiologists' Interpretations and Consensus Radiologist Interpretation

	Deep COVID-XR	Radiologist 1		Radiologist 2		Radiologist 3		Radiologist 4		Radiologist 5		Consensus	
Metric	Performance	Performance	*P* Value*	Performance	*P* Value*	Performance	*P* Value*	Performance	*P* Value*	Performance	*P* Value*	Performance	*P* Value*
Accuracy (%)	82	79	NA	81	NA	76	NA	76	NA	79	NA	81	NA
No. of TP findings	95	87	NA	92	NA	81	NA	102	NA	99	NA	94	NA
No. of TN findings	152	151	NA	150	NA	148	NA	125	NA	139	NA	148	NA
No. of FP findings	14	15	NA	16	NA	18	NA	41	NA	27	NA	18	NA
No. of FN findings	39	47	NA	42	NA	53	NA	32	NA	35	NA	40	NA
Sensitivity (%)	71 (63, 79)	65 (57, 73)	.09	69 (61, 77)	.47	60 (52, 69)	<.001	76 (69, 83)	.09	74 (66, 81)	.37	70 (62, 78)	.78
Specificity (%)	92 (87, 96)	91 (87, 95)	.8	90 (86, 95)	.62	89 (84, 94)	.32	75 (69, 82)	<.001	84 (78, 89)	.009	89 (84, 94)	.29
AUC (%)	0.88 (0.84, 0.92)	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	0.85 (0.80, 0.89)	.13

Note.—Data in parentheses are 95% CIs. AUC = area under the receiver operating characteristic curve, FN = false-negative, FP = false-positive, NA = not applicable, TP = true-positive, TN = true-negative.

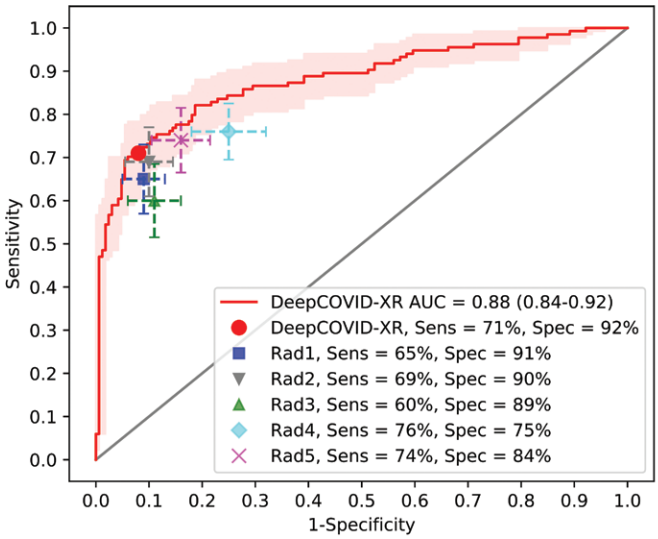
* *P* value for comparison with DeepCOVID-XR algorithm performance.

bilateral consolidations and ground-glass opacities with lower lung and peripheral predominance (27,28). In contrast, images categorized as demonstrating negativity by the algorithm (whether true or false negativity) often depicted clear lungs, concomitant pleural effusions, or both; interestingly, pleural effusions have been previously found to be quite rare (3%) in COVID-19–related pneumonia (27). Visualization of feature importance using gradient class activation mapping heatmaps revealed abnormalities in the lungs to be highly predictive of COVID-19 on chest radiographs, as expected. This serves as an important sanity check to reinforce confidence in algorithm predictions.

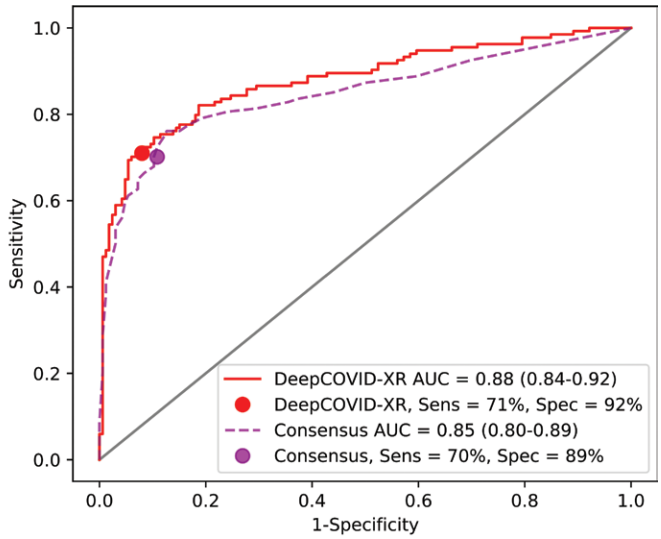
The explainable errors in algorithm predictions likely represent limitations of chest imaging in the radiologic diagnosis of COVID-19 rather than limitations of the algorithm itself. Prior clinical studies showed COVID-19 pneumonia produces characteristic features on chest images, but up to 56% of symptomatic patients can have normal findings on chest images, especially early in their disease course (9,27,29–31). Thus, imaging is inappropriate to rule out disease. In addition, many of the COVID-19 imaging findings are nonspecific and may overlap with findings characteristic of other conditions, particularly other viral pneumonias (32). Chest imaging should therefore not be used as a diagnostic tool for COVID-19 but could play an important role in earlier identification of patients likely to have the disease to aid in triage and infection control. Interestingly, Wong et al (27) found that approximately 9% of patients had abnormal imaging findings before they received a positive RT-PCR result, a proportion similar to the 13% (156 of 1192) of COVID-19–positive images in our study obtained prior to the patient's positive RT-PCR result and flagged as depicting COVID-19 positivity by our algorithm.

A number of groups from industry and academic sectors have published studies and non–peer-reviewed preprint reports with claims of extremely high sensitivity and specificity of AI algorithms in the detection of COVID-19 on chest radiographs (11–14). However, most of these studies have been limited by small sample sizes or have relied on images from publicly available data sets containing a mix of images from research articles and clinical reports of variable quality and questionable image-label accuracy (15). These data sets are subject to significant bias (33) and are simply not sufficient to train an algorithm ultimately intended for clinical use.

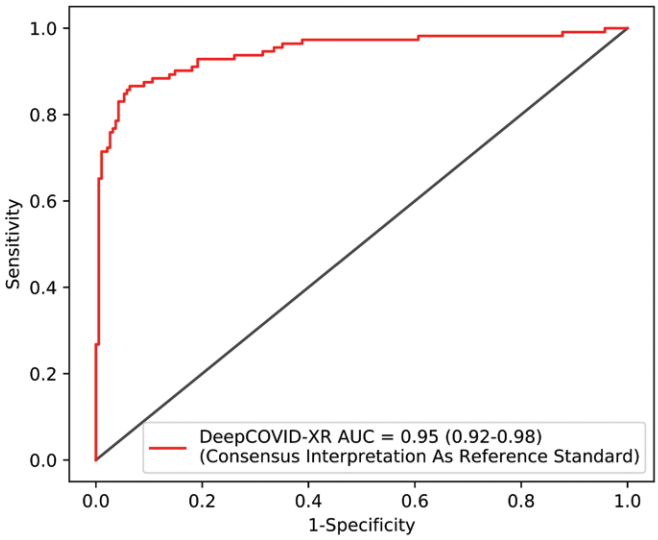
Murphy et al (16) presented a deep learning algorithm for detection of COVID-19 on chest radiographs that included both pre–COVID-19 era images for model pretraining and a data set composed of 606 clinical images for training and 468 clinical images for testing from patients at two Dutch hospitals during the COVID-19 era. The authors used a commercial patch-based CNN called CAD4COVID–x-ray, which had an AUC of 0.81 for predicting COVID-19 on a hold-out test set from one institution. By contrast, our model was trained and tested on more



a.



b.



c.

Figure 6: Comparison with expert radiologist interpretations. **(a)** Comparison of the performance of DeepCOVID-XR with individual expert radiologist interpretations on random sample of 300 images from the test set. For DeepCOVID-XR, the receiver operating characteristic (ROC) curve (red line) and decision threshold for overall interpretation of coronavirus disease 2019 (COVID-19) positivity or negativity (red point) is plotted with a 95% CI (red shaded area). For each radiologist, the overall interpretation sensitivity (Sens) and specificity (Spec) is plotted with 95% CIs (dashed lines). Radiologist 1 (Rad1) = blue square, radiologist 2 (Rad2) = gray downward arrow, radiologist 3 (Rad3) = green upward arrow, radiologist 4 (Rad4) = cyan diamond, radiologist 5 (Rad5) = magenta *X*. **(b)** Comparison of DeepCOVID-XR with the consensus interpretation of all five radiologists. ROC curves (lines) and decision thresholds (points) for DeepCOVID-XR (red) and the consensus interpretation (purple) (area under the ROC curve [AUC], 0.88 vs 0.85; *P* = .13). **(c)** ROC curve shows performance of DeepCOVID-XR (red line) using the consensus interpretation of all five radiologists rather than real-time polymerase chain reaction (RT-PCR) as the radiologic reference standard.

than 15 times the number of clinical images from the COVID-19 era. The discriminative performance of DeepCOVID-XR on an independent hold-out test set was superior to that reported for CAD4COVID–x-ray (AUC, 0.90 vs 0.81). Although differences in patient populations may partly account for this difference in performance, the AUC of our algorithm (ie, 0.95) when compared with a consensus of radiologists was far higher than that reported by Murphy et al (AUC range, 0.81–0.86), suggesting that DeepCOVID-XR more reliably produces predictions in line with the ground-truth radiologic diagnosis as determined by a consensus of experts. Finally, although the authors made their software available for use through a cloud-based interface, no details regarding algorithm development or code were made available, given the proprietary nature of their platform. We are freely providing all our code and pretrained neural network and model ensemble weights for open-source use in an effort to move toward a democratized approach to model development (*https://github.com/IVPLatNU/deepcovidxr*).

Our study had some limitations. First, the algorithm was evaluated on only those patients who were tested for COVID-19; thus, there was likely some degree of selection bias. Second, the performance of our algorithm was compared with that of RT-PCR assays as a reference standard, which has somewhat limited sensitivity itself because of sampling error or viral mutation (34). Finally, it is unclear how well the algorithm performs when COVID-19 is not the dominant viral pneumonia, as the study was performed at a time of considerable case load in our health care system.

In conclusion, DeepCOVID-XR is a deep learning artificial intelligence (AI) algorithm that detects coronavirus disease 2019 (COVID-19) on chest radiographs. The algorithm was trained and tested on a large U.S. clinical data set with performance similar to that of the consensus interpretation of experienced thoracic radiologists. We feel that this algorithm has the potential to benefit health care systems in mitigating unnecessary exposure to the virus by serving as an automated tool to rapidly flag patients with suspicious chest imaging findings for isolation and further testing. Planned future studies include a prospective evaluation of the algorithm (including in those patients not suspected of having COVID-19), a necessity for any AI algorithm prior to clinical implementation. In addition, we plan to incorporate other clinical

data (eg, demographics, vital signs, laboratory data) into the algorithm to further boost the performance and adapt the algorithm for risk prediction of clinically meaningful outcomes in patients with confirmed COVID-19. By providing the DeepCOVID-XR algorithm code base as an open-source resource, we hope investigators around the world will further improve, fine-tune, and test the algorithm using clinical images from their own institutions.

Acknowledgments: The authors acknowledge Scott M. Leonard for his assistance with transferring deidentified studies to our picture archiving and communications system for review by radiologists.

Author contributions: Guarantors of integrity of entire study, R.M.W., Y.W., D.R.C., N.P., A.K.K.; study concepts/study design or data acquisition or 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; agrees to ensure any questions related to the work are appropriately resolved, all authors; literature research, R.M.W., J.S., S.D., S.C., A.D., S.B., Y.W., N.X., H.S., N.P., A.K.K.; clinical studies, R.M.W., S.C., Y.W., D.R.C., N.X., B.D.A., H.S., N.P.; statistical analysis, R.M.W., J.S., S.D., S.C., A.D., S.B., Y.W., D.R.C., A.K.K.; and manuscript editing, R.M.W., J.S., S.D., S.C., A.D., S.B., Y.W., D.R.C., N.X., H.S., N.P., A.K.K.

Disclosures of Conflicts of Interest: **R.M.W.** disclosed no relevant relationships. **J.S.** disclosed no relevant relationships. **S.D.** disclosed no relevant relationships. **S.C.** disclosed no relevant relationships. **A.D.** disclosed no relevant relationships. **S.B.** disclosed no relevant relationships. **Y.W.** disclosed no relevant relationships. **D.R.C.** disclosed no relevant relationships. **N.X.** disclosed no relevant relationships. **B.D.A**. Activities related to the present article: disclosed no relevant relationships. Activities not related to the present article: received personal fees from Tempus Labs. Other relationships: disclosed no relevant relationships. **G.A.M.** disclosed no relevant relationships. **H.S.** disclosed no relevant relationships. **R.A.** Activities related to the present article: disclosed no relevant relationships. Activities not related to the present article: is a speaker for Boehringer Ingelheim. Other relationships: disclosed no relevant relationships. **N.P.** Activities related to the present article: disclosed no relevant relationships. Activities not related to the present article: is a speaker for Boehringer Ingelheim. Other relationships: disclosed no relevant relationships. **A.K.K.** disclosed no relevant relationships.

References

1. Gardner L, Ensheng D. Johns Hopkins Center for Systems Science and Engineering (CSSE) COVID-19 dashboard. Johns Hopkins University Website. https://coronavirus.jhu.edu/map.html. Published 2020. Accessed August 11, 2020.
2. Weinstock MB, Echenique A, Russell JW, et al. Chest x-ray findings in 636 ambulatory patients with COVID-19 presenting to an urgent care center: a normal chest x-ray is no guarantee. J Urgent Care Med 2020;14(7):13–18.
3. Hope MD, Raptis CA, Shah A, Hammer MM, Henry TS; Six Signatories. A role for CT in COVID-19? What data really tell us so far. Lancet 2020;395(10231):1189–1190.
4. Wynants L, Van Calster B, Collins GS, et al. Prediction models for diagnosis and prognosis of COVID-19 infection: systematic review and critical appraisal. BMJ 2020;369:m1328 [Published correction appears in BMJ 2020;369:m2204.].
5. Shi F, Wang J, Shi J, et al. Review of artificial intelligence techniques in imaging data acquisition, segmentation and diagnosis for COVID-19. IEEE Rev Biomed Eng 2020. 10.1109/RBME.2020.2987975. Published online April 16, 2020. Accessed July 14, 2020.
6. Li L, Qin L, Xu Z, et al. Using artificial intelligence to detect COVID-19 and community-acquired pneumonia based on pulmonary CT: evaluation of the diagnostic accuracy. Radiology 2020;296(2):E65–E71.
7. Huang L, Han R, Ai T, et al. Serial quantitative chest CT assessment of COVID-19: deep-learning approach. Radiol Cardiothorac Imaging 2020;2(2):e200075.
8. ACR Recommendations for the use of chest radiography and computed tomography (CT) for suspected COVID-19 infection. American College of Radiology Web site. https://www.acr.org/Advocacy-and-Economics/ACR-Position-Statements/Recommendations-for-Chest-Radiography-and-CT-for-Suspected-COVID19-Infection. Updated March 22, 2020. Accessed August 11, 2020.
9. Cleverley J, Piper J, Jones MM. The role of chest radiography in confirming COVID-19 pneumonia. BMJ 2020;370:m2426.
10. Jacobi A, Chung M, Bernheim A, Eber C. Portable chest x-ray in coronavirus disease-19 (COVID-19): a pictorial review. Clin Imaging 2020;64:35–42.
11. Yi PH, Kim TK, Lin CT. Generalizability of deep learning tuberculosis classifier to COVID-19 chest radiographs: new tricks for an old algorithm? J Thorac Imaging 2020;35(4):W102–W104.
12. Oh Y, Park S, Ye JC. Deep learning COVID-19 features on CXR using limited training data sets. IEEE Trans Med Imaging 2020;39(8):2688–2700.
13. Castiglioni I, Ippolito D, Interlenghi M, et al. Artificial intelligence applied on chest x-ray can aid in the diagnosis of COVID-19 infection: a first experience from Lombardy, Italy. medRxiv 2020.
14. Wang L, Zhong QL, Wong A. COVID-Net: a tailored deep convolutional neural network design for detection of COVID-19 cases from chest x-ray images. ArXiv 2003.09871 [preprint] https://arxiv.org/abs/2003.09871. Posted March 22, 2020. Accessed July 29, 2020.
15. Cohen JP, Morrison P, Dao L. COVID-19 image data collection. ArXiv 2003.11597 [preprint] https://arxiv.org/abs/2003.11597. Posted March 25, 2020. Accessed July 19, 2020.
16. Murphy K, Smits H, Knoops AJG, et al. COVID-19 on chest radiographs: a multireader evaluation of an artificial intelligence system. Radiology 2020;296(3):E166–E172.
17. Huang G, Liu Z, Van Der Maaten L, Weinberger KQ. Densely connected convolutional networks. In: Proceedings of the 30th IEEE conference on computer vision and pattern recognition. Piscataway, NJ: Institute of Electrical and Electronics Engineers, 2017; 2261–2269 .
18. He K, Zhang X, Ren S, Sun J. Deep residual learning for image recognition. In: Proceedings of the 29th IEEE conference on computer vision and pattern recognition. Piscataway, NJ: Institute of Electrical and Electronics Engineers, 2016; 770–778 https://doi.org/10.1109/CVPR.2016.90.
19. Szegedy C, Vanhoucke V, Ioffe S, Shlens J, Wojna Z. Rethinking the inception architecture for computer vision. In: Proceedings of the 29th IEEE conference on computer vision and pattern recognition. Piscataway, NJ: Institute of Electrical and Electronics Engineers, 2016; 2818–2826 https://doi.org/10.1109/CVPR.2016.308.
20. Szegedy C, Ioffe S, Vanhoucke V, Alemi AA. Inception-v4, inception-ResNet and the impact of residual connections on learning. In: Proceedings of the thirty-first AAAI conference on artificial intelligence. Menlo Park, CA: Association for the Advancement of Artificial Intelligence, 2017; 4278–4284.
21. Chollet F. Xception: deep learning with depthwise separable convolutions. In: Proceedings of the 30th IEEE conference on computer vision and pattern recognition. Piscataway, NJ: Institute of Electrical and Electronics Engineers, 2017; 1800–1807.
22. Tan M, Le QV. EfficientNet: rethinking model scaling for convolutional neural networks. In: Proceedings of the 36th international conference on machine learning. Cambridge, MA: PMLR, 2019;97:6105–6114.
23. Wang X, Peng Y, Lu L, Lu Z, Bagheri M, Summers RM. ChestX-ray8: hospital-scale chest x-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases. In: Proceedings of the 30th IEEE Conference on Computer Vision and Pattern Recognition. Piscataway, NJ: Institute of Electrical and Electronics Engineers, 2017; 3462–3471 https://doi.org/10.1109/CVPR.2017.369.
24. Selvaraju RR, Cogswell M, Das A, Vedantam R, Parikh D, Batra D. Grad-CAM: visual explanations from deep networks via gradient-based localization. In: Proceedings of the 2017 IEEE international conference on computer vision. Piscataway, NJ: Institute of Electrical and Electronics Engineers, 2017; 618–626 https://doi.org/10.1109/ICCV.2017.74.
25. McNemar Q. Note on the sampling error of the difference between correlated proportions or percentages. Psychometrika 1947;12(2):153–157.
26. DeLong ER, DeLong DM, Clarke-Pearson DL. Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach. Biometrics 1988;44(3):837–845.
27. Wong HYF, Lam HYS, Fong AHT, et al. Frequency and distribution of chest radiographic findings in patients positive for COVID-19. Radiology 2020;296(2):E72–E78.
28. Ng MY, Lee EY, Yang J, et al. Imaging profile of the COVID-19 infection: radiologic findings and literature review. Radiol Cardiothorac Imaging 2020;2(1):e200034.
29. Vancheri SG, Savietto G, Ballati F, et al. Radiographic findings in 240 patients with COVID-19 pneumonia: time-dependence after the onset of symptoms. Eur Radiol 2020;30(11):6161–6169.
30. Salehi S, Abedi A, Balakrishnan S, Gholamrezanezhad A. Coronavirus disease 2019 (COVID-19): a systematic review of imaging findings in 919 patients. AJR Am J Roentgenol 2020;215(1):87–93.
31. Bernheim A, Mei X, Huang M, et al. Chest CT findings in coronavirus disease-19 (COVID-19): relationship to duration of infection. Radiology 2020;295(3):200463.
32. Bai HX, Hsieh B, Xiong Z, et al. Performance of radiologists in differentiating COVID-19 from non-COVID-19 viral pneumonia at chest CT. Radiology 2020;296(2):E46–E54.
33. DeGrave AJ, Janizek JD, Lee SI. AI for radiographic COVID-19 detection selects shortcuts over signal. MedRxiv 2020.09.13.20193565 2020 [preprint] https://doi.org/10.1101/2020.09.13.20193565. Posted October 8, 2020. Accessed September 20, 2020.
34. Watson J, Whiting PF, Brush JE. Interpreting a COVID-19 test result. BMJ 2020;369(May):m1808.