Performance of a Chest Radiograph AI Diagnostic Tool for COVID-19: A Prospective Observational Study

Ju Sun, PhD • Le Peng, BS • Tainhui Li, MS • Dyah Adila, BS • Zach Zaiman, BS • Genevieve B. Melton-Meaux, MD, PhD • Nicholas E. Ingraham, MD • Eric Murray • Daniel Boley, PhD • Sean Switzer, DO • John L. Burns, MS • Kun Huang, PhD • Tadashi Allen, MD • Scott D. Steenburg, MD • Judy Wawira Gichoya, MD, MS • Erich Kummerfeld, PhD* • Christopher J. Tignanelli, MD, MS*

Purpose: To conduct a prospective observational study across 12 U.S. hospitals to evaluate real-time performance of an interpretable artificial intelligence (AI) model to detect COVID-19 on chest radiographs.

Materials and Methods: A total of 95,363 chest radiographs were included in model training, external validation, and real-time validation. The model was deployed as a clinical decision support system, and performance was prospectively evaluated. There were 5335 total real-time predictions and a COVID-19 prevalence of 4.8% (258 of 5335). Model performance was assessed with use of receiver operating characteristic analysis, precision-recall curves, and F1 score. Logistic regression was used to evaluate the association of race and sex with AI model diagnostic accuracy. To compare model accuracy with the performance of board-certified radiologists, a third dataset of 1638 images was read independently by two radiologists.

Results: Participants positive for COVID-19 had higher COVID-19 diagnostic scores than participants negative for COVID-19 (median, 0.1 [IQR, 0.0–0.8] vs 0.0 [IQR, 0.0–0.1], respectively; P < .001). Real-time model performance was unchanged over 19 weeks of implementation (area under the receiver operating characteristic curve, 0.70; 95% CI: 0.66, 0.73). Model sensitivity was higher in men than women (P = .01), whereas model specificity was higher in women (P = .001). Sensitivity was higher for Asian (P = .002) and Black (P = .046) participants compared with White participants. The COVID-19 AI diagnostic system had worse accuracy (63.5% correct) compared with radiologist predictions (radiologist 1 = 67.8% correct, radiologist 2 = 68.6% correct; McNemar P = .046) participants compared with White participants. The COVID-19 AI diagnostic system had worse accuracy (63.5% correct) compared with radiologist predictions (radiologist 1 = 67.8% correct, radiologist 2 = 68.6% correct; McNemar P = .001 for both).

Conclusion: AI-based tools have not yet reached full diagnostic potential for COVID-19 and underperform compared with radiologist prediction.

Supplemental material is available for this article.

©RSNA, 2022

COVID-19 has overwhelmed health care systems worldwide, resulting in critical equipment and supply shortages (1). The absence of curative treatment early in the pandemic gave rise to rapid identification and supportive treatment of infected individuals as a key tool in curtailing COVID-19.

The mainstay of COVID-19 diagnosis is nucleic acid testing of upper or lower respiratory tract swab specimens with use of reverse-transcription polymerase chain reaction (RT-PCR) (2). Early in the pandemic, RT-PCR testing remained a bottleneck and delay for COVID-19 diagnosis, with studies reporting clinical sensitivity of approximately 70% (3,4). Attempts have been made to develop artificial intelligence (AI) diagnostic models of COVID-19. A recent review by Roberts et al identified 62 AI diagnostic and prognostic models for COVID-19 using biomedical imaging (5). This review identified serious limitations in AI models published to date, including lack of external validation (6), lack of reporting participant demographics, inadequate number of images, and the use of so-called unrealistic training datasets (7) that do not represent the environment in which the model will ultimately be deployed (5,8). It is important to note that Roberts et al (5) reported that in none of the 45 diagnostic models published to date was a prospective real-time validation of model performance conducted.
Abbreviations

AI = artificial intelligence, AUC = area under the receiver operating characteristic curve, AUPRC = area under the precision-recall curve, ED = emergency department, IU = Indiana University, NPV = negative predictive value, OR = odds ratio, PCR = polymerase chain reaction, PPV = positive predictive value, RT-PCR = reverse-transcription PCR

Summary

This 12-site prospective study characterizes the real-time performance of an artificial intelligence–based diagnostic tool for COVID-19, which may serve as an adjunct to, but not as a replacement for, clinical decision-making in the diagnosis of COVID-19.

Key Points

- The COVID-19 artificial intelligence (AI) diagnostic tool achieved an area under the receiver operating characteristic curve of 0.70 on real-time validation.
- At equity and subgroup analysis, the AI tool demonstrated improved diagnostic capabilities in patients with more severe disease and in non-White participants, improved sensitivity in men, and improved specificity in women during real-time and external validations.
- The COVID-19 AI diagnostic system had significantly lower accuracy (63.5%) compared with radiologists (radiologist 1 = 67.8% correct, radiologist 2 = 68.6% correct; McNemar $P < .001$).

Keywords

Diagnosis, Classification, Application Domain, Infection, Lung

In November 2020, researchers at the University of Minnesota investigated the real-time performance of an interpretable AI diagnostic model for COVID-19 implemented as a clinical decision support system across 12 hospitals in the state of Minnesota. A benefit of interpretable (vs so-called black box) AI is the use of visual aids (ie, heatmaps) to display the basis for which models are used to make predictions. This study represents a preplanned prospective observational study to investigate real-time performance, model equity, and model drift of an image-based diagnostic AI system for COVID-19 over a 19-week period after implementation, which has not previously been well established in the literature (9,10).

Materials and Methods

This study was supported in part by an AI for Health COVID-19 grant (Microsoft). This grant provided graphical processing unit support for this project. No additional support was provided by that grantor, and the authors had control of the data and information submitted for publication. This study was approved by the University of Minnesota institutional review board, and the requirement for written informed consent was waived (STUDY 00011158). External validation at Indiana University (IU) was deemed exempt by the institutional review board because all secondary data were fully de-identified and remained within IU (STUDY 2010169012). External validation of the model at Emory University was approved by the institutional review board (STUDY 00000506).

Model Development

M Health Fairview model development (training) dataset.— We obtained 2220 chest radiographs from 837 adult participants (mean age, 60 years ± 16 [SD]; 51% women) who presented to the emergency department (ED) with polymerase chain reaction (PCR)–confirmed COVID-19 (radiography performed either 2 weeks before COVID-19 diagnosis or during a COVID-19–associated hospitalization) and 36,288 COVID-19–negative chest radiographs from 19,908 participants (mean age, 59 years ± 19; 51% women) from M Health Fairview for model training and optimization. All the COVID-19–positive chest radiographs, hereinafter referred to as cases, were obtained between March 2, 2020, and June 30, 2020. Negative controls, hereinafter referred to as controls, were chosen to mimic the real-world environment in which the AI model would be deployed. Because the model would be deployed on every adult ED chest radiograph, controls included any adult ED chest radiograph obtained between October 26, 2016, and March 3, 2020 (prior to the confirmed onset of COVID-19 within our organization). All chest radiographs were obtained in Minnesota at an M Health Fairview clinic or hospital. Participant demographics for the training dataset are provided in Table 1.

Publicly available COVID-19 datasets.— COVID-19 chest radiographs were collected from two open-source COVID-19 databases: the Valencian Region Medical ImageBank, or BIMCV, COVID-19+ database (11) and the COVID Chest X-ray GitHub database (12) (Table 1). The BIMCV COVID-19+ database contains 2261 chest radiographs collected from 11 hospitals from the Valencian Region, Spain; the cases were collected between February 26, 2020, and April 18, 2020. From the BIMCV COVID-19+ and COVID Chest X-Ray GitHub databases, we included all frontal (ie, images with at least one of the following “view” column attribute values: “posteroanterior,” “anteroposterior,” “anteroposterior supine,” or “anteroposterior semi-erect” in the GitHub metadata) chest radiographs with “COVID-19,” “COVID-19, acute respiratory distress syndrome,” or “severe acute respiratory syndrome” labels. In total, we collected 504 images from this database. The COVID-19 Medical Imaging and Data Resource Center was not used because our model was already developed and temporally validated by August 2020 (13).

Publicly available non–COVID-19 datasets.— For the COVID-19–negative dataset, we collected frontal images combined from two databases: 2011–2016 Medical Information Mart for Intensive Care Chest X-ray, or MIMIC-CXR, (14) (random sample of 23,611 images) and Open-i 2013 Indiana University Chest Radiograph Collection (15) (random sample of 3814 images) (Table 1). Images in the MIMIC-CXR and Open-i sets are dated earlier than December 2019, resulting in 27,425 images of participants with no particular medical status except the absence of COVID-19. Participant demographic information in publicly available datasets was not available because it was removed by the originating institutions to facilitate participant de-identification.
For model development, 38 508 M Health Fairview chest radiographs (2220 cases and 36 288 controls) were used for training. Model training was supplemented to maximize model generalizability using publicly available images of participants with and without COVID-19 (9592 total images; positive-to-negative ratio, 1:16). In the training set, 444 positive and 7257 negative images were held out for tuning the hyperparameters of the deep learning models, and the remainder were used to train the models. Our main model pipeline (Fig 1) consisted of lung segmentation (Fig 1a) and subsequent detection of COVID-19 (Fig 1b). The table below provides patient demographics for training and validation datasets:

| Dataset and COVID-19 Status | No. of Radiographs | Age (y) | Men | Race |
|-----------------------------|--------------------|---------|-----|------|
| MHFV training               |                    |         |     |      |
| Positive                    | 2220               | 60 ± 16 | 49  | NA   |
| Negative                    | 36 288             | 59 ± 19 | 49  | NA   |
| Public                      |                    |         |     |      |
| Positive                    | 2765               | NA      | NA  | NA   |
| Negative                    | 27 425             | NA      | NA  | NA   |
| MHFV temporal validation    |                    |         |     |      |
| Positive                    | 1777               | 62 ± 16 | 69  | NA   |
| Negative                    | 5228               | 58 ± 19 | 31  | NA   |
| MHFV real-time week 1 (early-implementation pilot) validation | | | | |
| Positive                    | 139                | NA      | NA  | NA   |
| Negative                    | 544                | NA      | NA  | NA   |
| MHFV real-time week 8 to week 19 validation | | | | |
| Positive                    | 258                | 63 ± 19 | 56  | Asian: 7.4 Black: 14.7 Other: 10.1 White: 67.8 |
| Negative                    | 5077               | 59 ± 19 | 48  | Asian: 3.2 Black: 9.9 Other: 6.3 White: 80.6 |
| Indiana University          |                    |         |     |      |
| Positive                    | 3001               | 63 ± 16 | 57  | NA   |
| Negative                    | 7001               | 59 ± 19 | 49  | NA   |
| Emory University            |                    |         |     |      |
| Positive                    | 1001               | 62 ± 17 | 52  | Black: 68.5 Other: 10.8 White: 20.7 |
| Negative                    | 1001               | 60 ± 19 | 48  | Black: 50.4 Other: 6.7 White: 42.9 |

Note.—Training dataset discussed in the “M Health Fairview Model development (training) dataset” subsection of the Materials and Methods; public dataset discussed in “Publicly available COVID-19 datasets” and “Publicly available non–COVID-19 datasets” subsections of the Materials and Methods; MHFV temporal validation discussed in the “M Health Fairview temporal validation dataset” subsection of the Materials and Methods; MHFV real-time week 1 (pilot) validation discussed in the “M Health Fairview real-time validation dataset (week 1 [early-implementation pilot])” subsection of the Materials and Methods; MHFV real-time week 8 to week 19 validation discussed in the “Week 8 to week 19 real-time validation” subsection of the Materials and Methods; Indiana University and Emory University datasets discussed in the “External validation datasets” subsection of the Materials and Methods. MHFV = M Health Fairview, NA = not available.

* Data are means ± SDs.
† Data are percentages.
‡ Pilot study: interim goal for overall performance evaluation.
§ Other race represents Alaska Native, American Indian, Native Hawaiian, or Pacific Islander participants. In the case of Emory University, other race also represents Asian participants.
external validations. To simulate real-time performance, temporal validation included all adult chest radiographs within the M Health Fairview system obtained between July 1, 2020, and July 30, 2020. To investigate model performance under differing COVID-19 prevalence, varying ratios of case imbalance were evaluated using a ratio of 1:1 (50% positive-to-negative) to 1:20 (4.8%). The area under the precision-recall curve (AUPRC) was calculated for each ratio. During this prospect-
tive period, 5228 chest radiographs were obtained from 5228 participants (mean age, 58 years ± 19; 69% women) who tested negative for COVID-19, and 1777 chest radiographs were obtained from 1777 participants (mean age, 62 years ± 16; 69% men) with PCR-confirmed COVID-19 (prevalence, 25%). Participant demographics for the temporal validation dataset are provided in Table 1. A subanalysis was conducted to evaluate model performance for participants with severe disease, which was defined as participants who required intensive care unit admission (as compared with participants without COVID-19 infection) and moderate disease, which was defined as participants who required hospital (but not intensive care unit) admission (as compared with participants without COVID-19 infection). Intensive care unit admission criteria for participants with COVID-19 infection were defined as the need for mechanical ventilation, greater than 60% fraction of inspired oxygen, or use of humidified high-flow nasal cannula with greater than 30 L per minute of support.

External validation datasets.— The first external validation dataset included 10 002 ED chest radiographs obtained from participants aged 18 years and older within the 15-hospital IU system, of which 7001 participants (mean age, 59 years ± 19; 49% men) were negative for COVID-19 (radiographs obtained between February 1, 2019, and July 15, 2019) and 3001 participants (mean age, 63 years ± 16; 57% men) were PCR-confirmed positive for COVID-19 (radiographs obtained between March 13, 2020, and November 7, 2020) (prevalence rate, 30%). The second external validation dataset included 2002 chest radiographs obtained between March 1, 2020, and July 30, 2020, from participants aged 18 years and older within the Emory University hospital system. COVID-19–positive and –negative chest radiographs were equally distributed. Participant demographics for each dataset are provided in Table 1. The Youden index was used to identify the optimal threshold for classification prediction for reporting sensitivity, specificity, negative predictive value (NPV), and positive predictive value (PPV) (16).

Model Implementation

In collaboration with Epic System’s Cognitive Computing, the AI model was integrated into the M Health Fairview production instance of Epic on November 10, 2020 (Fig E1 [supplement]). Additional details regarding implementation are provided in Appendix E1 (supplement).

Prospective Observational Study

M Health Fairview real-time validation dataset (week 1 [pilot]).— The chest radiograph AI model was implemented into the M Health Fairview Epic electronic health record system for investigative purposes on November 10, 2020. The AI model evaluated all ED and inpatient chest radiographs in adults aged 18 years and older obtained between November 11, 2020, and November 16, 2020 (week 1) at all M Health Fairview hospitals (n = 12) in real time with a COVID-19 status of unknown or negative. A total of 683 images met these criteria (Table 1). During this period, chest radiographs were obtained from 544 participants negative for COVID-19 infection and 139 participants with confirmed PCR-positive COVID-19 infection (prevalence, 20.4%).

Week 8 to week 19 real-time validation.— An investigation for model drift was conducted for all chest radiographs that triggered the clinical decision support system in participants with unknown or negative COVID-19 status between January 1, 2021, and March 18, 2021. A total of 5335 images met the criteria. During this period, chest radiographs were obtained from 5077 participants negative for COVID-19 infection (mean age, 59 years ± 19; 52% women) and from 258 participants with confirmed PCR-positive COVID-19 infection (mean age, 63 years ± 19; 56% men) (prevalence, 5%). Participant demographics are provided in Table 1. Heatmaps were generated to facilitate interpretable AI by comparing the gradient of the output class score with respect to the input image pixel values.

Model Performance Compared with Board-certified Radiologist Performance

Model performance was evaluated against an additional third external dataset previously generated at four Emory University hospitals to assess board-certified radiologist performance in predicting COVID-19 status from chest radiographs. Ten board-certified radiologists (1–13 years of experience after completion of training) with primary practice in the ED were recruited from four academic institutions (IU, Emory University, Wake Forest University, and University of Florida). Readers were only provided the participants’ chest radiograph and were blinded to the RT-PCR results (ground truth) and any participant clinical information. Each image was reviewed by a random sample of two of the 10 radiologists. An image was labeled as correct if a radiologist predicted that the participant had COVID-19 and the participant was RT-PCR–positive for COVID-19.

Statistical Analysis

We set a minimum threshold to achieve 80% power or higher with a predetermined minimum sample size of 5000 AI predictions. The sample size needed for adequate power varies based on the prevalence (17). Assuming a prevalence of 5%, 4860 AI predictions would be needed for investigation with over 80% power. Thus, with 5335 AI predictions and a prevalence of 5%, our study achieved our minimum threshold of 80% power.

Model performance was evaluated in real time across the 12-hospital M Health Fairview system. The COVID-19 diagnostic score from the model ranged from 0 to 1, indicating the likelihood of COVID-19. The Wilcoxon rank sum test, which is used to compare differences between two groups, was used to evaluate differences in the COVID-19 diagnostic score and COVID-19 positivity. The Kruskal-Wallis test, which is used to compare differences between three groups, was used to evaluate differences in the COVID-19 diagnostic AI score and
disease severity for the following groups: COVID-19–negative, mild to moderate COVID-19 disease, and severe COVID-19 disease. The McNemar test was used to compare the diagnostic accuracy of radiologist versus AI predictions (18). The model was evaluated during two time periods—1 week early-implementation pilot and between 8 weeks and 19 weeks after implementation—to assess for model performance, model drift, and model equity. To evaluate model drift, images were split into four time quartiles, three with 1334 images and one with 1333 images: quartile 1, January 1, 2021, through January 11, 2021; quartile 2, January 12, 2021, through January 20, 2021; quartile 3, January 21, 2021, through February 1, 2021; and quartile 4, February 2, 2021, through March 18, 2021. Areas under the receiver operating characteristic curve (AUCs), 95% CIs, precision, recall, and F1 scores were calculated during each period. To assess equity, model performance was evaluated across self-reported race and sex (when available) in the electronic health record with use of logistic regression. For equity analysis, two binary measures (dependent variables of the logistic regression model) of model accuracy were assessed: (a) sensitivity, or the ability to classify a patient correctly if they have COVID-19 and (b) specificity, or the ability to classify a patient as not having COVID-19 if they do not have COVID-19. Self-reported race and sex were included in each regression model as independent variables to assess the association of patient demographics with model accuracy. P < .05 was considered to indicate a statistically significant difference. All analyses were conducted using Stata/MP (version 16). The Checklist for Artificial Intelligence in Medical Imaging was used for reporting in this study.

Model Availability

The model is not publicly available; however, it can be made available upon request for investigative purposes.

Results

Preimplementation Validation

Temporal and publicly available dataset validation.— To investigate how prevalence and disease severity may affect real-time model performance, we investigated model performance before implementation with use of M Health Fairview images collected between July 1, 2020, and July 30, 2020. The mean AUC was 0.80 during temporal validation. The AUPRC varied with prevalence, with values of 0.84 (prevalence, 50%) to 0.41 (prevalence, 5%) (Table E1 [supplement]).

A subanalysis was conducted to evaluate model performance for participants with severe (AUC, 0.83) and moderate disease (AUC, 0.70) compared with participants negative for COVID-19, as shown in Table E1 (supplement). Similarly, we assessed the effect of prevalence on performance among participants with severe or moderate disease compared with participants negative for COVID-19. At a disease prevalence of 50%, the AUPRC was 0.48 and 0.22 for participants with severe and moderate disease, respectively. Figure E2 (supplement) displays curves for each AUC (labeled ROC [receiver operating characteristic] in the figure) and AUPRC (labeled PRC [precision-recall curve] in the figure), where each random subsampling result is represented by a line; mean AUCs and AUPRCs are displayed with each figure. The distribution of COVID-19 diagnostic scores for cases and controls during July 2020 is shown in Figure E3 (supplement).

For the publicly available COVID-19 dataset, the AUC was 0.96 and the AUPRC ranged from 0.98 (50% prevalence) to 0.93 (5% prevalence) (Table E1 [supplement]).

External validation.— Models were externally validated at IU and Emory University (Table E2 [supplement]). All sites were noted to have median diagnostic AI scores near 0 for controls (Figs 3, E4 [supplement]). Variation was noted in the median diagnostic AI scores for cases across each site. Thus, to tailor model classification for each site, institution-specific diagnostic thresholds were defined to maximize the Youden index. The AUC was 0.72 (95% CI: 0.71, 0.73) for external validation on images from IU (prevalence, 30%). With the chosen thresholds, specificity was 77%, sensitivity was 67%, PPV was 85%, and NPV was 56% (Table E2 [supplement]). The AUC was 0.66 (95% CI: 0.64, 0.68) for external validation on images from Emory University (prevalence, 50%). With the chosen thresholds, specificity was 86%, sensitivity was 46%, PPV was 63%, and NPV was 75% (Table E2 [supplement]). Box and whisker plots of COVID-19 diagnostic scores for cases and controls at IU and Emory University are shown in Figure E4 (supplement).

Real-time Evaluation and Assessment for Model Drift

In the evaluation of AI model performance, participants positive for COVID-19 had significantly higher scores than participants negative for COVID-19 (median, 0.1 [IQR, 0.0–0.8] vs median, 0.0 [IQR, 0.0–0.1], respectively; P < .001) (Fig 3). Heatmaps were generated to display the basis for model decisions (Fig 4). Participants with severe COVID-19 had higher scores (median, 0.2 [IQR, 0.0–1.0]) than participants with mild or moderate COVID-19 disease (median, 0.1 [IQR, 0.0–0.8]), and participants without COVID-19 infection had the lowest scores (median, 0.0 [IQR, 0.0–0.1]; P < .001). Real-time performance during early-implementation pilot achieved an AUC of 0.70 (95% CI: 0.65, 0.75) and for the time period 8 weeks to 19 weeks achieved an AUC of 0.70 (95% CI: 0.66, 0.73) (Table E3 [supplement]). The F1 score during the time period 8 weeks to 19 weeks was 0.62 (Table 2). AI model performance was similar across time quartiles (quartile 1: AUC = 0.69 [95% CI: 0.63, 0.75], F1 score = 0.61; quartile 2: AUC = 0.66 [95% CI: 0.60, 0.74], F1 score = 0.56; quartile 3: AUC = 0.70 [95% CI: 0.62, 0.78], F1 score = 0.62; and quartile 4: AUC = 0.74 [95% CI: 0.65, 0.83], F1 score = 0.70) (Table 2). Model performance peaked during the final quartile (February 2, 2021–March 18, 2021) (AUC = 0.74, precision = 98%, recall = 55%, F1 score = 0.70) (Table 2).
Subgroup Analysis by Race and Sex

Data on race or ethnicity and sex were available for controls and cases for both external validation at Emory University and real-time validation at M Health Fairview. Model performance was evaluated by subgroup analysis as shown in Table 3. Logistic regression was used to evaluate the association between race and sex with model accuracy. Among participants with COVID-19 infection, the odds ratio (OR) for correct classification of men was 2.3 (95% CI: 1.2, 4.47; \( P = .01 \)) relative to women. Among participants with COVID-19 infection, Black participants (OR, 2.44 [95% CI: 1.02, 5.86]; \( P = .046 \)), Asian participants (OR, 26.9 [95% CI: 3.3, 219]; \( P = .002 \)), and Alaska Native, American Indian, Native Hawaiian, or Pacific Islander participants (OR, 4.66 [95% CI: 1.51, 14.4]; \( P = .008 \)) were more likely to be correctly classified compared with White participants. Among COVID-19–negative participants, the OR for correct classification of women was 1.62 (95% CI: 1.2, 2.1; \( P = .001 \)) relative to men.

Figure 3: COVID-19 diagnostic artificial intelligence (AI) scores for participants positive and negative for COVID-19. Box and whisker plots show (A) COVID-19 diagnostic scores (y-axis) for non–COVID-19 versus polymerase chain reaction–confirmed COVID-19 from real-time implementation at M Health Fairview and (B) initial COVID-19 diagnostic scores (y-axis) for non–COVID-19 versus mild or moderate (mild/mod) COVID-19 versus severe COVID-19. Boxes represent the IQR [25%–75%], with the median denoted by the horizontal line within each box. The Wilcoxon rank sum test, which is used to compare differences between two groups, was used to evaluate differences in COVID-19 diagnostic AI score and COVID-19 positivity. The Kruskal-Wallis test, which is used to compare differences between three groups, was used to evaluate differences in COVID-19 diagnostic AI score and disease severity. * = \( P < .001 \) compared with COVID-19–negative disease. Circles represent outlier points in the box distribution plots.

Figure 4: Interpretable artificial intelligence (AI) heatmaps from real-time implementation indicating features used for AI model prediction. COVID-19–negative and –positive chest radiographs (grayscale) and representative heatmaps (in color) from real-time implementation output are shown. Features of importance for model predictions are represented by the red end of the heatmap spectrum, and blue represents features of least importance. Representative images with both high and low diagnostic scores are provided. Black boxes were added to obscure potentially patient-identifying information.
(95% CI: 1.22, 2.15; \( P = .001 \)) relative to men. We found no evidence of a difference between White versus Black participants being more correctly classified if they did not have COVID-19 (OR, 1.5 [95% CI: 0.998, 2.34]; \( P = .051 \)).

**AI Model and Radiologist Performance**

Radiologists demonstrated better diagnostic accuracy (radiologist 1 = 67.8%, radiologist 2 = 68.6%) than the AI model (63.5%) (McNemar \( P < .001 \)). The AI model (AUC, 0.75 [95% CI: 0.71, 0.78]) did not outperform independent radiologists (radiologist 1: AUC = 0.78 [95% CI: 0.76, 0.80]; radiologist 2: AUC = 0.76 [95% CI: 0.73, 0.78]).

**Discussion**

To our knowledge, no other prospective observational studies to investigate the real-world performance of an AI model for COVID-19 diagnosis based on chest radiograph findings alone have been performed. Several key observations can be made based on our findings. First, COVID-19 chest radiograph diagnostic models perform well for participants with severe COVID-19; however, they do not differentiate participants with mild COVID-19 who may present with minimal chest radiograph findings. Second, for real-time performance of the AI model in participants with unknown or previously negative COVID-19 status, the AUC was 0.70. Third, there was no significant model drift over a period of 2.5 months. Fourth, validation using publicly available datasets provides unrealistic performance estimates. Finally, the diagnostic accuracy of the AI model was inferior to predictions made by board-certified radiologists.

At the beginning of the COVID-19 pandemic, we and others sought to generate AI models to successfully predict COVID-19 from biomedical imaging (19,20). Unfortunately, 2 years into the pandemic, no such generalizable model exists, and few models have been investigated in real time (9,21). There may be several reasons for the lack of any successfully deployed diagnostic model. First, it may be impossible to develop a model solely based on chest radiograph findings to differentiate between participants with COVID-19 and non–COVID-19 diagnoses. This was an early hypothesis by our clinical content experts, as the radiographic appearance of COVID-19 is heterogeneous. The radiographic appearance of COVID-19 may range from no or minimal observable pathologic abnormality to severe acute respiratory distress syndrome and can progress and defervesce depending on the time of exposure and stage of disease. However, the possibility that AI could differentiate these diseases based on features not observed by the naked eye promulgated efforts to test this hypothesis.

Second, it is possible that adequate training data have not yet been collected to train such a generalizable model. Our model, which used approximately 50,000 images both locally and internationally, achieved an AUC of 0.70 on real-world validation.
A recent review of 62 AI models for COVID-19 from biomedical imaging identified substantial limitations in AI models published to date and identified a majority of articles as being at high risk of bias (5). These biases include lack of external validation, lack of equity analysis by race and sex, lack of reporting patient demographics, inadequate number of images, and the use of unrealistic training datasets. The review also noted that of the 45 COVID-19 diagnostic AI models published, none evaluated real-time performance, as done in this study (5). We encourage model developers to implement and accurately evaluate real-world performance before publishing overly optimistic findings. We also encourage exercising maximal discretion when interpreting or using reported performance with publicly available datasets. Our model showed unrealistic performance (AUCs > 0.96) with use of such publicly available datasets.

Finally, it is possible that false-positive results with the AI model occur in patients with COVID-19 who had a negative PCR test despite being positive for COVID-19. The current sensitivity of rapid PCR testing varies greatly based on the viral load of a patient. Patients with viral load cycle threshold levels less than 25 have a sensitivity of 90%; in patients with lower viral loads (higher cycle threshold), however, PCR sensitivity drops to 76% (22).

Our findings suggest that AI analysis of chest radiographs alone is not adequate to diagnose COVID-19. However, an AI-enabled clinical decision support system may add additional information, which ED providers can integrate into clinical decision-making when developing a differential diagnosis and determining whether confirmatory testing and isolation for COVID-19 are necessary. We propose the integration of structured and unstructured (23) data from electronic medical records into model training in the future.

A source of bias in most models is the lack of adequate analysis ensuring it performs similarly across different populations, specifically sex and racial or ethnic groups. We and others have reported that COVID-19 has disproportionately burdened minority populations (24,25). To ensure the model performed equitably, we tested the model across race and sex. Notably, the model had significantly higher sensitivity in men and minority groups—populations that have been found to be at higher risk for severe disease (24–26). In fact, one study found disease severity at imaging to be higher across patients in minority populations compared with White patients (27). This may explain the improved performance we noted in non-White participants; our preimplementation model performance was superior for participants with severe versus moderate COVID-19 (Table E1 [supplement]). Importantly, the model performs equitably across races, and there is limited risk that it would further widen the disparate COVID-19 outcomes being experienced by minority populations.

This study is not without limitations. First, our controls were not selected from a target population of patients with suspected COVID-19. We included all radiographs during training to model a real-world environment to optimize realistic performance; however, this limits the potential usefulness of the model outside of the ED and early inpatient setting. Second, our model ran only on radiographs in participants with unknown or negative COVID-19 status. Given the high PPV of COVID-19 PCR testing, it is unnecessary to deploy an AI model when the diagnosis is already confirmed. Third, although lung segmentation may minimize AI shortcuts and focus the AI model specifically to the lung window, it may exclude the lung areas of the left lower lobe that are hidden by the heart, as well as areas of both lung bases that are hidden by the opacity of the abdomen below the diaphragm. This may limit the AI model’s ability to identify lung opacities in these regions. Fourth, these models were trained and validated on fixed data; it is anticipated that the models will evolve as new data arrive. It is possible to modify the models to allow for gradual improvement, thus leveraging advances in machine learning. Fifth, although all sites had low diagnostic AI scores for controls, variation in median diagnostic AI scores was noted for cases across all three institutions. Thus, to tailor performance to each institution, institution-specific diagnostic thresholds were defined.

In conclusion, AI-based diagnostic tools may serve as an adjunct to, but not a replacement for, clinical decision-making concerning COVID-19 diagnosis, which largely hinges on exposure history, signs, and symptoms. Although AI-based tools have not yet reached full diagnostic potential in COVID-19, they may still offer valuable information to clinicians when taken into consideration along with clinical signs and symptoms.

Acknowledgment: The authors acknowledge the Minnesota Supercomputing Institute (http://www.msi.umn.edu) at the University of Minnesota for providing resources that contributed to the research results reported in this article.

Author contributions: Guarantors of integrity of entire study, TL., D.A., Z.Z., C.J.T.; 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, J.S., L.P., TL., D.A., G.B.M.M., N.E.I., E.M., S.S., K.H., TA., C.J.T.; clinical studies, G.B.M.M., N.E.I., S.S., TA., S.D.S., C.J.T.; experimental studies, J.S., L.P., TL., D.A., Z.Z., E.M., K.H., TA., J.W.G.; statistical analysis, J.S., L.P., TL., D.A., Z.Z., D.B., J.L.B., K.H., J.W.G., E.K., C.J.T.; and manuscript editing. J.S., L.P., TL., D.A., G.B.M.M., N.E.I., E.M., D.B., S.S., J.L.B., K.H., TA., S.D.S., J.W.G., E.K., C.J.T.

Disclosures of conflicts of interest: J.S. No relevant relationships. L.P. No relevant relationships. TL. No relevant relationships. D.A. No relevant relationships. Z.Z. No relevant relationships. G.B.M.M. Payments to institution: Fairview, University of Minnesota, NIH, AHQR; payment or honoraria for lectures, presentations, or speakers’ bureaus from AMIA, ACM, Washington University Informatics, and NIH; support for attending meetings and/or travel from AMIA, ACM, Washington University Informatics, and NIH; patent planned, issued, or pending for application 17/302,573; participation on a data safety monitoring board or advisory board not related to this article; leadership or fiduciary role at AMIA and ACMI. N.E.I. No relevant relationships. E.M. No relevant relationships. D.B. No relevant relationships. S.S. No relevant relationships. J.L.B. No relevant relationships. K.H. Internal funding from Indiana University Precision Health Initiative; author received research grants from NIH; author was principal investigator (PI) for a research contract with Merck (not relevant to this manuscript); author is PI for a research contract with Eli Lilly (not relevant to this manuscript). TA. Speakers’ bureau for Genentech and Boehringer Ingelheim; gives nonbranded lectures on interstitial lung disease unrelated to this manuscript. S.D.S. No relevant relationships. J.W.G. Medical Imaging and Data Resource Center grant from NIH; Nightingale Open Science grant; NSF Future of Work grant; Kheiron Breast AI Validation grant; BDC Data Anonymization grant; Radiology: Artificial Intelligence trained editorial board lead. E.K. Grant UL1TR002349; this technology submitted for a U.S. patent. C.J.T. Microsoft AI for Health provided GPU resources via a COVID-19 Research Grant (no payments, other support, or any role in the actual conduct of the research was had by Microsoft); technology was submitted for a U.S. patent.

References

1. White DB, Lo B. A Framework for rationing ventilators and critical care beds during the COVID-19 pandemic. JAMA 2020;323(18):1773–1774.
2. Centers for Disease Control and Prevention. Interim Guidelines for Collecting, Handling, and Testing Clinical Specimens from Persons for Coronavirus Disease 2019 (COVID-19). https://www.cdc.gov/coronavirus/2019-ncov/lab/guidelines-clinical-specimens.html. Published 2020. Accessed April 26, 2020.

3. Woloshin S, Patel N, Kesselheim AS. False negative tests for SARS-CoV-2 infection — challenges and implications. N Engl J Med 2020;383(6):e38.

4. Kortela E, Kirjavainen V, Ahava MJ, et al. Real-life clinical sensitivity of SARS-CoV-2 RT-PCR test in symptomatic patients. PLoS One 2021;16(5):e0251661.

5. Roberts M, Driggs D, Thorpe M, et al. Common pitfalls and recommendations for using machine learning to detect and prognosticate for COVID-19 using chest radiographs and CT scans. Nat Mach Intell 2021;3(3):199–217.

6. DeGrave AJ, Janizek JD, Lee SI. AI for radiographic COVID-19 detection selects shortcuts over signal. medRxiv. 2020.09.13.20193565 [preprint] https://www.medrxiv.org/content/10.1101/2020.09.13.20193565v2... Posted October 8, 2020. Accessed November 2, 2020.

7. Banerjee I, Sinha P, Purkayastha S, et al. Was there COVID-19 back in 2012? Challenge for AI in Diagnosis with Similar Indications. arXiv 2006.13262 [preprint] https://arxiv.org/abs/2006.13262. Posted June 23, 2020. Accessed October 7, 2020.

8. Shamsoddin E. Can medical practitioners rely on prediction models for COVID-19? A systematic review. Evid Based Dent 2020;21(3):84–86.

9. Wang B, Jin S, Yan Q, et al. AI-assisted CT imaging analysis for COVID-19 screening: building and deploying a medical AI system. Appl Soft Comput 2021;98:106897.

10. Carlile M, Hurt B, Hsiao A, Hogarth M, Longhurst CA, Dameff C. Deployment of artificial intelligence for radiographic diagnosis of COVID-19 pneumonia in the emergency department. J Am Coll Emerg Physicians Open 2020;1(6):1459–1464.

11. da Iglésia Vayá M, Saborit JM, Montell JA, et al. BIMCV COVID-19+ : a large annotated dataset of RX and CT images from COVID-19 patients. arXiv 2020.01174 [preprint] https://arxiv.org/abs/2006.01174. Posted June 1, 2020. Accessed August 20, 2020.

12. Covid-19 ChestXray Dataset. https://github.com/ieee8023/covid-chestxray-dataset. Accessed October 18, 2020.

13. Medical Imaging and Data Resource Center (MIDRC) website. https://www.midrc.org/midrc-data. Accessed May 20, 2021.