An Artificial Intelligence–Based Chest X-ray Model on Human Nodule Detection Accuracy From a Multicenter Study

Fatemeh Homayounieh, MD; Subba Digumarthy, MD; Shadi Ebrahimian, MD; Johannes Rueckel, MD; Bej Friedrich Hoppe, MD; Bastian Oliver Sabel, MD; Saiseli Conjeti, PhD; Karsten Ridder, MD; Markus Sistermanns, MD; Lei Wang, PhD; Alexander Preuhs, PhD; Florin Ghesu, PhD; Awaais Mansoor, PhD; Mateen Moghbel, MD; Ariel Botwin, MD; Ramandeep Singh, MD; Samuel Cartmell, MD; John Patti, MD; Christian Huemmer, PhD; Andreas Fieselmann, PhD; Clemens Joerger, PhD; Negar Mirshahzadeh, MSc; Victorine Muse, MD; Mannudeep Kalra, MD

Abstract

IMPORTANCE Most early lung cancers present as pulmonary nodules on imaging, but these can be easily missed on chest radiographs.

OBJECTIVE To assess if a novel artificial intelligence (AI) algorithm can help detect pulmonary nodules on radiographs at different levels of detection difficulty.

DESIGN, SETTING, AND PARTICIPANTS This diagnostic study included 100 posteroanterior chest radiograph images taken between 2000 and 2010 of adult patients from an ambulatory health care center in Germany and a lung image database in the US. Included images were selected to represent nodules with different levels of detection difficulties (from easy to difficult), and comprised both normal and nonnormal control.

EXPOSURES All images were processed with a novel AI algorithm, the AI Rad Companion Chest X-ray. Two thoracic radiologists established the ground truth and 9 test radiologists from Germany and the US independently reviewed all images in 2 sessions (unaided and AI-aided mode) with at least a 1-month washout period.

MAIN OUTCOMES AND MEASURES Each test radiologist recorded the presence of 5 findings (pulmonary nodules, atelectasis, consolidation, pneumothorax, and pleural effusion) and their level of confidence for detecting the individual finding on a scale of 1 to 10 (1 representing lowest confidence; 10, highest confidence). The analyzed metrics for nodules included sensitivity, specificity, accuracy, and receiver operating characteristics curve area under the curve (AUC).

RESULTS Images from 100 patients were included, with a mean (SD) age of 55 (20) years and including 64 men and 36 women. Mean detection accuracy across the 9 radiologists improved by 6.4% (95% CI, 2.3% to 10.6%) with AI-aided interpretation compared with unaided interpretation. Partial AUCs within the effective interval range of 0 to 0.2 false positive rate improved by 5.6% (95% CI, −1.4% to 12.0%) with AI-aided interpretation. Junior radiologists saw greater improvement in sensitivity for nodule detection with AI-aided interpretation as compared with their senior counterparts (12%; 95% CI, 4% to 19% vs 9%; 95% CI, 1% to 17%) while senior radiologists experienced similar improvement in specificity (4%; 95% CI, −2% to 9%) as compared with junior radiologists (4%; 95% CI, −3% to 5%).

CONCLUSIONS AND RELEVANCE In this diagnostic study, an AI algorithm was associated with improved detection of pulmonary nodules on chest radiographs compared with unaided interpretation for different levels of detection difficulty and for readers with different experience.

Key Points

Question Can artificial intelligence (AI) improve detection of pulmonary nodules on chest radiographs at different levels of detection difficulty?

Findings In this diagnostic study, AI-aided interpretation was associated with significantly improved detection of pulmonary nodules on chest x-rays as compared with unaided interpretation of chest x-rays.

Meaning These results suggest that an AI algorithm may improve diagnostic performance of radiologists with different levels of experience for detecting pulmonary nodules on chest radiographs compared with unaided interpretation.
Introduction

Because of its accessibility, low cost, low radiation dose exposure, and reasonable diagnostic accuracy, chest radiographs are frequently used for early detection of pulmonary nodules despite their inferiority to low-dose computed tomography (CT). Pulmonary nodules are common initial radiologic manifestations of lung cancer, but can be easily missed when they are subtle, small, or in difficult areas.

Detection of pulmonary nodules on chest radiographs has been a focus of several computer-aided detection commercial and research software solutions for the past 2 decades; however, early solutions were limited due to their low sensitivity or high false positive rates. Although several recent studies have examined how AI may improve the performance of radiologists in detection of pulmonary nodules on CXRs, they are limited in their scope owing to the lack of multisite readers with different levels of clinical experience, detection difficulties, and settings. To address these limitations in the evaluation of a novel AI algorithm, we assessed if our AI algorithm can help detect pulmonary nodules on chest radiographs at different levels of detection difficulty with both normal and nonnormal control images in a multireader, multicenter setting.

Methods

The study included 100 posterior-anterior (PA) chest radiographs belonging to 100 adult patients (mean [SD] age, 55 [20] years; 64 men [64%], 36 women [36%]) from 2 different sources in 2019—an ambulatory health care center in Germany (site A) and the Lung Image Database Consortium in the US (site B). The images belonged to patients with ages binned by decade (from ages 20 years to 90 years). Furthermore, we included images acquired on radiography units from 9 different vendors (eTable 1 in the Supplement). This study followed the Standards for Reporting of Diagnostic Accuracy (STARD) reporting guideline.

Seventy-two images were sourced from site A, while the remaining 28 images came from site B (Figure 1). Inclusion criteria entailed adult patients, availability of PA chest radiographs, absence of any abnormality (25% of images), and presence of pulmonary nodules (50% of images). The remaining 25% of images were selected consecutively to represent non-nodular abnormalities (such as consolidation, atelectasis, and pleural effusion). The images with pulmonary nodules (the positive cohort) were not restricted to only nodules (ie, the presence of non-nodule pathologies was not an exclusion criterion). All included images had to meet additional predefined image quality criteria such as the Digital Imaging and Communications in Medicine (DICOM) image format standard, coverage of entire lungs inside the imaged chest radiograph, and absence of image artifacts. We excluded all non-PA chest radiographs (such as obliques or lateral projection images and portable images) and radiographs from patients with lung surgery (such as lobectomy and pneumonectomy). The initial selection cohort included 200 chest radiographs based on some exclusion criteria (such as age restriction and x-ray projection view). The selection criteria ensured only PA chest radiographs of adult patients were included, and that among these patients there was representation of both genders, different age groups, and radiographic equipment vendors. No images were excluded based on the image quality (such as noise or exposure). Images were retrospectively selected from the participating sites regardless of the clinical indication and status of the patient. We obtained approval from Massachusetts General Brigham’s human research committee for retrospective evaluation of deidentified radiograph images without need for informed consent.

AI Algorithm

All images were processed with the AI Rad Companion Chest X-Ray algorithm (Siemens Healthineers AG) (eAppendix in the Supplement). The distribution of cases used for training the algorithm are given in Table 1.
Establishing Ground Truth for the Study

To establish the ground truth for the study, 2 experienced thoracic radiologists (S.D. [with 16 years of experience] and M.K. [14 years of experience]) assessed all unprocessed images to establish the ground truth in a consensus reading fashion. The radiologists were provided with all available patients' clinical data including their corresponding CTs (available for all patients with pulmonary nodules) and lateral chest radiographs where available. Chest CTs were not available for patients without pulmonary nodules on their PA images. The ground truthers did not have access to the AI results or annotations to avoid any unintentional bias. They generated labels for absence and presence (scored as a 0:1 binary) of nodules, pleural effusion, pneumothorax, atelectasis, and consolidation with the one constraint that the detected findings should be visible on PA images. Findings not visible on PA radiographs but seen on non-PA projection images or CT were deemed as

Figure 1. Flowchart of Chest Radiograph Selection, Inclusion and Exclusion Criteria, Ground Truthing, and Multireader Study

Table 1. Distribution of Training and Validation Cases Used for the Development of the Artificial Intelligence Algorithm

| Characteristics   | Training cases | Validation cases |
|-------------------|----------------|------------------|
|                   | Total, No.     | Positive, No. (%)| Negative, No. (%)| Total, No. | Positive, No. (%)| Negative, No. (%)|
| Lesions           | 7776           | 5086 (65.4)      | 2690 (34.6)      | 444        | 138 (31.1)       | 306 (68.9)        |
| Consolidation     | 7732           | 2640 (34.1)      | 5092 (65.9)      | 333        | 67 (20.1)        | 266 (79.9)        |
| Atelectasis       | 7732           | 1120 (14.5)      | 6612 (85.5)      | 333        | 72 (21.6)        | 261 (78.4)        |
| Pleural effusion  | 7631           | 1724 (22.6)      | 5907 (77.4)      | 332        | 69 (20.8)        | 263 (79.2)        |
| Pneumothoraxa     | 75 067         | 3993 (5.3)       | 71 074 (94.7)    | 318        | 61 (19.2)        | 257 (80.8)        |

* The algorithm for detection and classification of pulmonary lesions, consolidation, pleural effusion, and atelectasis was trained jointly in a multiclass setting; the detection model for pneumothorax was trained separately.
absent. For each finding, the radiologists specified if the finding was easy, moderate, or challenging to detect on PA radiograph. To challenge the AI algorithm and assess relative effect of detection difficulty with and without AI-aided interpretation, we enriched the data with pulmonary nodules with different levels of detection difficulty by including 20 challenging, 7 moderately difficult, and 23 easy to detect lung nodules. The nodule size ranged between 4 to 28 mm (mean [SD] size, 14 [6] mm). One-half of images without pulmonary nodules (chosen in consecutive manner) were normal whereas the remaining 25 images had findings other than pulmonary nodules (such as but not limited to atelectasis, consolidation, pneumothorax, and pleural effusion).

**Multireader Study**

Each radiograph image was assigned a unique identification number, and then all were randomized across readers and reading sessions by generating pseudorandom permutation with the randperm function in MATLAB. To cover a wide spectrum of professional experience, we recruited 7 staff radiologists and 3 radiology residents as test readers. Staff radiologists belonged to 3 hospitals and had varied experience as thoracic radiologists (J.P., K.R., V.M. [with 35, 25, and 21 years of experience, respectively]) and as general radiologists (J.R., B.F.H., B.S., and M.S. [with 3.5, 2.5, 9, and 3 years of experience, respectively]). All 3 residents were in the first year of their radiology residency and had completed 1 clinical rotation in chest radiography (A.B., M.M., S.C.). Because of differences in residency training years for radiology at the 3 sites (3 or 5 years), we set a cutoff of 3 years for classifying participating radiologists into junior and senior groups. To ensure that there was no change in experience level of the junior radiologists caused by increased exposure or work hours in thoracic imaging, we ensured that none of the participating residents had a second rotation over the washout period or during the AI-aided interpretation phase. Prior to interpretation, each reader participated in a training session to understand study objectives and become familiar with the user interface.

All radiologists reviewed images in 2 sessions (without and with the aid of AI outputs) and recorded their findings on an electronic case record form (**Figure 2; eFigure 1 in the Supplement**). In the first, unaided reading session, each test reader independently read all original images without AI outputs and recorded the presence of 5 findings (pulmonary nodules, atelectasis, consolidation, pneumothorax, and pleural effusion), their level of confidence for detecting the individual finding on a scale from 1 to 10 (with 10 denoting the highest confidence). A 10-point scale was preferred to obtain a smoother distribution and allow for a smoother computation of nonparametric statistical testing. Furthermore, the location (upper [regions above the inferior aspect of the anterior second rib], middle [between ribs 2-4], and lower [below the inferior aspect of the anterior end of fourth rib]) and the size (the maximum single dimension in millimeters) were also recorded for each pulmonary nodule. After a 1-month washout period, we conducted the second reading session, in which readers reviewed both original DICOM images and AI-processed DICOM images simultaneously to record the findings in the same manner. All reading sessions were conducted on a PACS (Picture Archiving and Communication System) workstation or a computer with diagnostic radiology-grade monitor. In the training phase, we informed readers about data enrichment, they were not aware of the exact ratio of nodule positive or negative radiographs.

**Statistical Analysis**

Data were recorded in an editable case report form using Microsoft Excel 2019 (Microsoft Inc). All analyses and computations were performed in Python using statistical library SciPy. We conducted a sample size estimation described in multireader receiver operating characteristics curve (ROC) studies with the assumption that readers’ areas under the curve (AUC) would have a range of 0.15 and a 0.8 correlation between the unaided and AI-aided interpretation of images, for a fully paired study design. For 80% power at 5% 2-tailed type I error rate, the sample size estimation projected need for 50 images with and 50 images without pulmonary nodules, as well as 10 test readers. One test reader did not follow the exact instruction for the 2 reading sessions and thus was excluded from...
data analysis retrospectively. Confidence scores correlated with the detection accuracy. Therefore, to have readers’ binary detection of nodules, we applied a cutoff value of 5 to each reader’s confidence scores (confidence scores >5 were classified as positive and ≤5 classified as negative findings).

Figure 2. Deidentified Radiograph Images From 4 Adult Patients With True Positive Pulmonary Nodules

Panel A, artificial intelligence (AI) helped detect right upper lobe nodule (arrowheads) for 3 junior (J) radiologists as well as improved the confidence score for 1 senior (S) radiologist (score >5 implies positive finding; score ≤5 is negative). Likewise, AI-aided interpretation led to detection of missed right upper lobe nodule (arrowheads) for the radiograph in panel B for 2 junior radiologists and improved confidence of 2 junior radiologists. For panel C, AI helped 2 junior and 2 senior radiologists detect right lower lung nodule (arrowheads) they had missed on unaided interpretation. AI also helped either detect the right mid- and left-lower lung nodules (L1 and L2; arrowheads) (1 junior and 2 senior radiologists for each nodule) or improve confidence for detecting nodules (3 junior and 1 senior radiologists).
Results

The distribution of findings in the 100 images was 50% (50 of 100) pulmonary nodules, 12% (12 of 100) consolidation, 13% (13 of 100) atelectasis, and 10% (10 of 100) pleural effusion. A quarter of images (25 of 100) had no abnormal findings.

Sensitivity, Specificity, and Accuracy

The standalone performance of the AI for detection of pulmonary nodules included 46 true negative, 32 true positive, 18 false negative and 4 false positive identifications (Table 2). The mean performance of test radiologists improved with AI-aided interpretation with a decrease in both false negative and false positive identifications compared with unaided interpretation.

With the AI-aided interpretation, the mean sensitivity, specificity, and detection accuracy across the 9 radiologists improved by 10.4% (95% CI, 4%-16.6%), 2.4% (95% CI, 1.8%-6.5%), and 6.4% (95% CI, 2.3%-10.6%) compared with unaided interpretation (Figure 2 and eFigure 3 in the Supplement). Sensitivities and specificities of 5 of 9 radiologists improved with AI-aided interpretation while 4 of 9 radiologists had improvement of either sensitivity or specificity with AI-aided interpretation compared with unaided reading mode. The mean sensitivity, specificity, and accuracy for AI-aided interpretation increased by 11% (95% CI, 4% to 7%), 3% (95% CI, −2% to 5%), and 7% (95% CI, 2% to 11%) compared with unaided interpretation (Table 3).

For partial AUCs, the mean case level accuracies in unaided and AI-aided interpretation of images were 0.717 (95% CI, 0.640-0.798) and 0.772 (95% CI, 0.702-0.836), respectively. The mean sensitivities of junior radiologists were 0.400 (95% CI, 0.295-0.505) and 0.516 (95% CI, 0.439-0.624) for unaided and aided interpretive modes. The corresponding mean sensitivities for senior radiologists were 0.510 (95% CI, 0.423-0.602) and 0.600 (95% CI, 0.500-0.689). Mean specificities of junior radiologists were 0.948 (95% CI, 0.911-0.990) for unaided interpretation and 0.960 (95% CI, 0.921-0.994) for the AI-aided mode. The corresponding mean specificities of senior radiologists were 0.900 (95% CI, 0.869-0.939) and 0.940 (95% CI, 0.880-0.985).

AUC Analysis

All readers operated in a high level of specificity within the range of 0.8 to 1 (corresponding to a false positive rate range of 0 to 0.2, which we deemed as the effective interval) associated with different level of sensitivity. Partial AUC within the effective interval range of 0 to 0.2 false positive rate improved by 5.6% (95% CI, −1.4 to 12.0%) with AI-aided interpretation (mean AUC, 77.2%; 95% CI, 70.2%-83.6%) compared with unaided interpretation (mean AUC, 71.7%; 95% CI, 64.0%-79.8%) (eTable 2 in the Supplement). For summaries of mean partial AUCs with unaided and AI-aided detection of pulmonary nodules within the effective interval, see eFigure 3 in the Supplement.

Table 2. Distribution of Pulmonary Nodules With Unaided and AI-Aided Sessions for Nodule Detection

| Radiologists | Radiograph images, No. | Unaided interpretation | AI-aided interpretation |
|--------------|------------------------|------------------------|------------------------|
|              | TN | FN | TP | FP | TN | FN | TP | FP |
| J1           | 49 | 27 | 23 | 1  | 48 | 19 | 31 | 2  |
| S1           | 43 | 13 | 37 | 7  | 46 | 20 | 30 | 4  |
| J2           | 48 | 35 | 15 | 2  | 48 | 14 | 36 | 2  |
| S2           | 44 | 21 | 29 | 6  | 46 | 19 | 31 | 4  |
| J4           | 41 | 23 | 27 | 9  | 48 | 21 | 29 | 2  |
| S3           | 47 | 34 | 16 | 3  | 48 | 34 | 16 | 2  |
| J5           | 47 | 26 | 24 | 3  | 47 | 23 | 27 | 3  |
| J6           | 49 | 32 | 18 | 1  | 49 | 34 | 16 | 1  |
| S4           | 49 | 35 | 15 | 1  | 47 | 16 | 34 | 3  |
| Mean (SD) images | 45.5 (5.6) | 26.4 (8.3) | 23.6 (8.2) | 4.1 (3.3) | 47.5 (5.1) | 23.6 (7.7) | 26.4 (8.1) | 2.5 (1.7) |

Abbreviations: AI, artificial intelligence; FN, false negative; FP, false positive; J, junior radiologist; S, senior radiologist; TN, true negative; TP, true positive.
eTable 3 in the Supplement summarizes reader performance for easy and challenging nodules vs the control group of images.

**Junior vs Senior Readers**

There were significant differences in performance of senior and junior radiologists. Junior radiologists had a 12% (95% CI, 4% to 19%) improvement for sensitivity for nodule detection with AI-aided interpretation over unaided interpretation. The corresponding improvement for senior radiologists was 9% (95% CI, 0.5% to 17%) (eTable 4 in the Supplement). On the other hand, senior radiologists (for 3 of 4 senior radiologists) witnessed a larger improvement in specificity with AI-aided interpretation (mean improvement, 4%; 95% CI, −2% to 9%) compared with the junior group (for 1 of 5 junior radiologists, mean improvement, 1%; 95% CI, −3% to 5%). Both groups had similar improvements in accuracy (6%) and partial AUC (5%) with AI-aided interpretation. AI-aided interpretation improved sensitivities of 6 radiologists (5 of 5 junior radiologists and 1 of 4 senior radiologists) for detection of multiple nodules in 16 images (mean sensitivity of 64% vs 55% in unaided mode) with a minor change in specificity (2% increase, 93% vs 95% in unaided mode). One junior and 1 senior radiologist had disproportionate improvement upon AI-aided interpretation (eTable 2 in the Supplement). Despite exclusion of these 2 radiologists, we found a significant improvement in performance in AI-aided vs unaided interpretation modes (eFigure 4 in the Supplement).

**Discussion**

Our multicenter, multireader diagnostic study found a significant improvement in sensitivity, specificity, and accuracy for detecting of pulmonary nodules on chest radiographs. Compared with unaided interpretation, most test readers reported improved sensitivity (9% to 12%; 6 of 9 radiologists) with AI-aided reading, whereas specificity (1% to 4%) improved for 4 test readers with different levels of experience. These results are compatible with other similar studies with improved reader performance with use of AI algorithms for detecting pulmonary nodules on chest radiographs. Specifically, Yoo et al12 evaluated 5485 images both with and without malignant pulmonary nodules from National Lung Cancer Screening Trial data and reported a greater improvement in sensitivity (8% with AI-aided interpretation using a different algorithm than the one used in our study, as compared with unaided interpretation) and a much smaller gain in specificity (2%).12 Likewise, in another lung cancer screening data set, Jang et al13 reported an improved detection of malignant pulmonary nodules with another AI algorithm to an AUC of 76% from 67% AUC with unaided interpretation. The AUCs reported in our study are comparable with those indicated in Table 3. Case-Level Sensitivity, Specificity, and Accuracy of Unaided and AI-Aided Interpretation Modes for Pulmonary Nodule Detection

| Readers | Sensitivity, mean (SE), % | Specificity, mean (SE), % | Accuracy, mean (SE), % |
|---------|--------------------------|--------------------------|------------------------|
|         | Unaided | AI-aided | Change | Unaided | AI-aided | Change | Unaided | AI-aided | Change |
| J1      | 46 (7)  | 62 (7)  | 16 (7)  | 98 (2)  | 96 (3)  | −2 (2)  | 72 (4)  | 79 (4)  | 7 (4)  |
| S1      | 74 (6)  | 60 (7)  | −14 (8) | 86 (5)  | 92 (3)  | 6 (6)   | 80 (3)  | 76 (4)  | −4 (4) |
| J2      | 30 (7)  | 72 (6)  | 42 (7)  | 96 (3)  | 96 (3)  | 0 (4)   | 63 (5)  | 84 (4)  | 21 (4) |
| S2      | 58 (7)  | 62 (7)  | 4 (8)   | 88 (4)  | 92 (4)  | 4 (4)   | 73 (5)  | 77 (4)  | 4 (4)  |
| J4      | 54 (7)  | 58 (7)  | 4 (6)   | 82 (6)  | 96 (3)  | 14 (6)  | 68 (4)  | 77 (5)  | 9 (4)  |
| S3      | 30 (7)  | 30 (7)  | 0 (6)   | 94 (3)  | 98 (2)  | 4 (4)   | 62 (5)  | 64 (5)  | 2 (4)  |
| J5      | 46 (7)  | 54 (7)  | 8 (7)   | 94 (4)  | 94 (3)  | 0 (4)   | 70 (5)  | 74 (4)  | 4 (4)  |
| J6      | 36 (7)  | 32 (7)  | −4 (6)  | 98 (2)  | 98 (2)  | 0       | 67 (5)  | 65 (5)  | −2 (3) |
| S4      | 30 (7)  | 68 (7)  | 38 (6)  | 98 (2)  | 94 (3)  | −4 (3)  | 64 (4)  | 81 (4)  | 17 (4) |
| Total, mean (95% CI), % | 45 (38 to 53) | 55 (48 to 63) | 11 (4 to 17) | 93 (89 to 96) | 95 (91 to 99) | 3 (~2 to 5) | 69 (62 to 77) | 75 (70 to 81) | 2 (1 to 7) |

Abbreviations: AI, artificial intelligence; J, junior radiologist; S, senior radiologist.

* The cutoff for specificity was 0.5.

b Change values were estimated statistically with bootstrapping method and not computed numerically.
reported in prior studies\textsuperscript{13,14} but lower than those in other studies on chest radiographs in lung cancer screening settings (with AU\textsubscript{Cs} $\geq$ 90\%).\textsuperscript{15,16} The latter may be a result of high proportion of normal images or those with less complex findings in screening population as compared with our data set, in which the non-nodule group was enriched with images with other radiographic findings.

The novel AI algorithm used in this study was associated with improvements in both sensitivity and specificity of detecting pulmonary nodules on chest radiographs vs unaided interpretation, and demonstrated favorable diagnostic outcomes regardless of radiologists' experience. Given the high miss rate and importance of identifying pulmonary nodules on chest radiographs, AI algorithms such as the one assessed in our study and those tested in prior studies can play an important role in improving quality and diagnostic value of radiographs in patients with pulmonary nodules, not least in helping to reconcile differences in radiographic equipment from different vendors. Although our study did not assess the clinical relevance of pulmonary nodules with AI-aided interpretation as opposed to unaided interpretation, the importance of our findings can be inferred from a 2021 targeted study on improved detection of malignant pulmonary nodules with AI algorithms.\textsuperscript{17}

Another implication of our study is the improved sensitivity of junior radiologists with AI-aided interpretation, because such improvement can help radiologists detect (or finesse their search patterns for) pulmonary nodules on chest radiographs. Greater improvement in specificity for senior radiologists vs their junior colleagues implies that AI outputs can increase radiologists' confidence for ruling out pulmonary nodules and to avoid labeling non-nodular abnormalities and normal structures as pulmonary nodules. Standalone performance of the AI had higher sensitivity (32 true positive nodules) than all readers in both unaided and AI-aided interpretive modes. AI assistance was associated with an increase in readers' sensitivity. And although AI had almost similar specificity to interpretations performed without AI assistance, the specificity of readers increased from 93\% (unaided) to 95\% (aided). The lack of substantial difference in detection improvement for challenging and easy nodules with AI-aided interpretation compared with unaided mode may be related to the fact that challenging nodules are also hard for AI to detect. Among junior radiologists, AI-aided interpretation helps avoid missed pulmonary nodules caused by satisfaction of search-related bias.

**Limitations**

Our study had several limitations. First, the study sample size was limited to 100 images, which may not have been enough to assess how reader performance with and without AI varied based on different levels of detection difficulty for pulmonary nodules. To overcome the limitation of small sample size, we enriched the data with nodules of different detection difficulties as well as normal and nonnormal control images. Furthermore, with a total of 1800 reads (9 readers $\times$ 100 images $\times$ 2 reading sessions), this pilot study had adequate size with a focus only on 1 pathological finding (lesions) with an enriched data set to increase measurement reliability, both to test the study hypothesis and collect information about the effect size (such as magnitude of improvement with AI), interreader variability, and readers' mean AU\textsubscript{C} without AI. To avoid overtesting on a limited data set, we did not assess reader performance on detection of other pathologies. While use of enriched data sets is not representative of routine clinical workflow, prior studies have reported similar enrichment and distribution of normal and abnormal findings in chest radiographs.\textsuperscript{16-18} We further addressed the issue of enrichment with inclusion of images with nodules and non-nodular distractors to simulate routine clinical workflow and avoid bias.

Second, we did not assess the clinical significance of additional nodules detected or missed in the unaired or AI-aided interpretation of chest radiographs. Third, there were fewer junior, in-training radiologists as compared with the senior radiologists because the former were posted in inpatient service units (medical floors and intensive care units) during the time of the COVID-19 pandemic. Fourth, AI-aided evaluation was limited to pulmonary nodules even though both our and other commercial AI algorithms\textsuperscript{14-16} can detect other radiograph findings; this was necessary to avoid overanalysis and repetitive statistics for multiple radiographic findings in a limited data set of images.
Fifth, the establishment of ground truth was conducted with a small and even number of radiologists instead of multiple radiologists and majority opinion on the presence of radiographic abnormalities. However, findings of the 2 radiologists coincided with the observations reported or noted within the radiographs from both the German and US sites, and thus provided a robust ground truth (with availability of corresponding chest CT for all images with pulmonary nodules). Although lack of chest CT exams for nodule-negative radiographs creates a different ground truth standard than for chest radiographs with nodules, there is a much lower use of CT in patients with normal radiograph results or with other distracting findings (eg, pneumonia, pleural effusion, or pneumothorax), and review of CT in such patients could conceivably bias the radiologists generating the reference standard to mark-up nodules on images that are not apparent on standalone radiograph review. Sixth, despite explicit instructions and preinterpretation training sessions, radiologists did not use the entire scale of 1 to 10 for their confidence in detecting nodules; most confidence scores were clustered at the higher or lower end of detection confidence both with and without AI-aided interpretation. Seventh, we did not assess the extent of intraobserver variations in the detection of pulmonary nodules with unaided and AI-aided interpretation of images. This may have required interpretation of images multiple times over an extended duration, which could have resulted in a recall bias in the study. Finally, data from 1 radiologist had to be excluded since they did not follow instructions for interpreting images outlined in the training sessions.

**Conclusions**

This study found that a novel AI algorithm was associated with improved accuracy and AUCs for junior and senior radiologists for detecting pulmonary nodules. Four of 9 radiologists had a lower number of missed and false positive pulmonary nodules with help from AI-aided interpretation of chest radiographs.
Administrative, technical, or material support: Homayounieh, Rueckel, Sabel, Wang, Preuhs, Singh, Mirshahzadeh, Muse.

Supervision: Digumarthy, Ridder, Mirshahzadeh, Kalra.

Conflict of Interest Disclosures: Drs Conjeti, Wang, Preuhs, Ghesu, Mansoor, Huemmer, Fieselmann, Joerger, and Mirshahzadeh are full-time paid employees of Siemens Healthineers. Dr Digumarthy reported receiving an honorarium from Siemens Medical Solutions; he reported receiving grant funding from Lunit Inc; he reported consultation fees for independent image analysis from Pfizer, Merck, Novartis, Roche, Bayer, Abbvie, Zai Laboratories, Biengen, Gradalis, and Bristol Mayer Squibb outside the submitted work; he reported receiving grant funding from GE outside the submitted work. Dr Rueckel reported receiving compensation for conference presentations from Siemens Healthineers GmbH outside the submitted work. Dr Sabel reported receiving grants from Siemens Healthcare GmbH Department of Radiology and University Hospital of Ludwig Maximilian University of Munich; he reported receiving compensation for conference presentations from Siemens Healthcare GmbH during the conduct of the study. Dr Sistermanns reported grants from Siemens Healthineers outside the submitted work. Mrs Mirshahzadeh reported receiving personal fees from Siemens Healthcare GmbH during the conduct of the study. Dr Kalra reported receiving grants from Riverain Technologies outside the submitted work. No other disclosures were reported.

Funding/Support: Massachusetts General Hospital received research grants from GE Healthcare, Lunit Inc, Riverain Technologies Inc, and Siemens Healthineers AG.

Role of the Funder/Sponsor: Siemens Healthineers contributed to the design and conduct of the study. Collection, management, analysis, and interpretation were performed independently at non-Siemens sites without influence and interference from the sponsor. The decision to publish the study and manuscript preparation was made by non-Siemens coauthors. All study coauthors always had complete access to the study data and the manuscript. The other institutions providing grant support had no role in the design and conduct of the study; collection, management, analysis, and interpretation of the data; preparation, review, or approval of the manuscript; and decision to submit the manuscript for publication.

Additional Information: The full study protocol is available on request from the corresponding author of the study.

REFERENCES

1. Greene R, Francis H, Williams, MD: father of chest radiology in North America. Radiographics. 1991;11(2):325-332. doi:10.1148/radiographics.11.2.2028067
2. Yoo H, Lee SH, Arru CD, et al. AI-based improvement in lung cancer detection on chest radiographs: results of a multi-reader study in NLST dataset. Eur Radiol. 2021;31:9664-9674. doi:10.1007/s00330-021-08074-7
3. van Beek EJ, Mirdadrae S, Murchison JT. Lung cancer screening: computed tomography or chest radiographs? World J Radiol. 2015;7(8):189-193. doi:10.4329/wjr.v7.i8.189
4. Del Ciello A, Franchi P, Contegiacomo A, Cichetti G, Bonomo L, Larici AR. Missed lung cancer: when, where, and why? Diag Interv Radiol. 2017;23(2):118-126. doi:10.5152/dir.2016.16187
5. Tack D, Howarth N. Missed Lung Lesions: Side-by-Side Comparison of Chest Radiography with MOCT. In: Hodler J, Kubik-Huch RA, von Schulthess GK, eds. Diseases of the Chest, Breast, Heart and Vessels 2019-2022: Diagnostic and Interventional Imaging. Springer; 2019:17-26. doi:10.1007/978-3-030-11449-6_2
6. Elemraid MA, Muller M, Spencer DA, et al. Accuracy of the interpretation of chest radiographs for the diagnosis of paediatric pneumonia. PLoS One. 2014;9(8):e106051. doi:10.1371/journal.pone.0106051
7. Donald JJ, Barnard SA. Common patterns in 558 diagnostic radiology errors. J Med Imaging Radiat Oncol. 2012;56(2):173-178. doi:10.1111/j.1754-9485.2012.02348.x
8. Chen S, Han Y, Lin J, Zhao X, Kong P. Pulmonary nodule detection on chest radiographs using balanced convolutional neural network and classic candidate detection. Artif Intell Med. 2020;107:101881. doi:10.1016/j.artmed.2020.101881
9. Liang CH, Liu YC, Wu MT, Garcia-Castro F, Alberich-Bayarri A, Wu FZ. Identifying pulmonary nodules or masses on chest radiography using deep learning: external validation and strategies to improve clinical practice. Clin Radiol. 2020;75(1):38-45. doi:10.1016/j.crad.2019.08.005
10. MATLAB. Release R2021b. MathWorks database. Accessed April 12, 2021. https://www.mathworks.com/help/matlab/ref/randperm.html
11. Obuchowski NA. Nonparametric analysis of clustered ROC curve data. Biometrics. 1997;53(2):567-578. doi:10.2307/2533958
12. Yoo H, Kim KH, Singh R, Digumarthy SR, Kalra MK. Validation of a deep learning algorithm for the detection of malignant pulmonary nodules in chest radiographs. *JAMA Netw Open*. 2020;3(9):e2017135. doi:10.1001/jamanetworkopen.2020.17135

13. Jang S, Song H, Shin YJ, et al. Deep learning-based automatic detection algorithm for reducing overlooked lung cancers on chest radiographs. *Radiology*. 2020;296(3):652-661. doi:10.1148/radiol.2020200165

14. Rajpurkar P, Irvin J, Ball RL, et al. Deep learning for chest radiograph diagnosis: a retrospective comparison of the CheXNeXt algorithm to practicing radiologists. *PLoS Med*. 2018;15(11):e1002686. doi:10.1371/journal.pmed.1002686

15. Hwang EJ, Park S, Jin KN, et al. Development and validation of a deep learning-based automated detection algorithm for major thoracic diseases on chest radiographs. *JAMA Netw Open*. 2019;2(3):e191095. doi:10.1001/jamanetworkopen.2019.1095

16. Nam JG, Park S, Hwang EJ, et al. Development and validation of deep learning-based automatic detection algorithm for malignant pulmonary nodules on chest radiographs. *Radiology*. 2019;290(1):218-228. doi:10.1148/radiol.2018180237

17. Seah JCY, Tang CHM, Buchlak QD, et al. Effect of a comprehensive deep-learning model on the accuracy of chest x-ray interpretation by radiologists: a retrospective, multireader multicase study. *Lancet Digit Health*. 2021;3(8):e496-e506. doi:10.1016/S2589-7500(21)00106-0

18. Schalekamp S, van Ginneken B, Koedam E, et al. Computer-aided detection improves detection of pulmonary nodules in chest radiographs beyond the support by bone-suppressed images. *Radiology*. 2014;272(1):252-261. doi:10.1148/radiol.14131315

**SUPPLEMENT.**

**eAppendix.** AI Algorithm and Statistical Information

eTable 1. Vendor Distribution of Different Radiographic Equipment in Our Study

eTable 2. Partial AUC Values Under Effective Interval Partial AUCs Within FPR Interval 0-0.2 for Unaided and AI-aided Interpretation Modes

eTable 3. Summary of Stratified Readers Performance for Detection of Easy and Challenging Pulmonary Nodules Versus Control Cases Without Nodules

eTable 4. Sensitivity, Specificity, Accuracy and Partial AUC Within Different Reader Groups in Unaided and AI-aided Interpretation Modes

eFigure 1. Deidentified CXRs from 3 Adult Patients Without Pulmonary Nodules

eFigure 2. AI ROC Curve Specificity and Sensitivity

eFigure 3. ROC Curve Demonstrates Averaged Partial AUCs for Unaided and AI-aided Detection of Pulmonary Nodules on the Included CXRs Within the Effective Interval

eFigure 4. Histograms Depicting Distribution of Partial AUCs in Aided and Unaided Modes Over a 100 Times Sampling for Different Readers Groups