JAMA Network | Open™

Original Investigation | Imaging

Association of Artificial Intelligence–Aided Chest Radiograph Interpretation With Reader Performance and Efficiency

Jong Seok Ahn, MD; Shadi Ebrahimian, MD; Shaunagh McDermott, MD; Sanghyup Lee, MD; Laura Naccarato, MD; John F. Di Capua, MD; Markus Y. Wu, MD; Eric W. Zhang, MD; Victorine Muse, MD; Benjamin Miller, BS; Farid Sabzalipour, BS; Bernardo C. Bizzo, MD, PhD; Keith J. Dreyer, DO, PhD; Parisa Kaviani, MD; Subba R. Digumarthy, MD; Mannudeep K. Kalra, MD

Abstract

IMPORTANCE The efficient and accurate interpretation of radiologic images is paramount.

OBJECTIVE To evaluate whether a deep learning–based artificial intelligence (AI) engine used concurrently can improve reader performance and efficiency in interpreting chest radiograph abnormalities.

DESIGN, SETTING, AND PARTICIPANTS This multicenter cohort study was conducted from April to November 2021 and involved radiologists, including attending radiologists, thoracic radiology fellows, and residents, who independently participated in 2 observer performance test sessions. The sessions included a reading session with AI and a session without AI, in a randomized crossover manner with a 4-week washout period in between. The AI produced a heat map and the image-level probability of the presence of the referrable lesion. The data used were collected at 2 quaternary academic hospitals in Boston, Massachusetts: Beth Israel Deaconess Medical Center (The Medical Information Mart for Intensive Care Chest X-Ray [MIMIC-CXR]) and Massachusetts General Hospital (MGH).

MAIN OUTCOMES AND MEASURES The ground truths for the labels were created via consensual reading by 2 thoracic radiologists. Each reader documented their findings in a customized report template, in which the 4 target chest radiograph findings and the reader confidence of the presence of each finding was recorded. The time taken for reporting each chest radiograph was also recorded. Sensitivity, specificity, and area under the receiver operating characteristic curve (AUROC) were calculated for each target finding.

RESULTS A total of 6 radiologists (2 attending radiologists, 2 thoracic radiology fellows, and 2 residents) participated in the study. The study involved a total of 497 frontal chest radiographs—247 from the MIMIC-CXR data set (demographic data for patients were not available) and 250 chest radiographs from MGH (mean [SD] age, 63 [16] years; 133 men [53.2%])—from adult patients with and without 4 target findings (pneumonia, nodule, pneumothorax, and pleural effusion). The target findings were found in 351 of 497 chest radiographs. The AI was associated with higher sensitivity for all findings compared with the readers (nodule, 0.816 [95% CI, 0.732-0.882] vs 0.567 [95% CI, 0.524-0.611]; pneumonia, 0.887 [95% CI, 0.834-0.928] vs 0.673 [95% CI, 0.632-0.714]; pleural effusion, 0.872 [95% CI, 0.808-0.921] vs 0.889 [95% CI, 0.862-0.917]; pneumothorax, 0.988 [95% CI, 0.932-1.000] vs 0.792 [95% CI, 0.756-0.827]). AI-aided interpretation was associated with significantly improved reader sensitivities for all target findings, without negative impacts on the specificity. Overall, the AUROCs of readers improved for all 4 target findings, with significant improvements in detection of pneumothorax and nodule. The reporting time with AI was 10% lower than without AI (40.8 vs 36.9 seconds; difference, 3.9 seconds; 95% CI, 2.9-5.2 seconds; *P* < .001).

(continued)

Key Points

Question Can an artificial intelligence (AI) engine used concurrently improve reader performance when reporting chest radiographs?

Findings In this cohort study involving 497 chest radiographs and 6 readers, the area under the receiver operating characteristic curve and reporting times improved when readers used AI during the interpretation. The improvement was significant for pneumothorax and nodule detection, and sensitivities of readers improved significantly for all findings when using AI.

Meaning These findings suggest that readers' performance in evaluating chest radiographs is improved by AI use.

+ Supplemental content

Author affiliations and article information are listed at the end of this article.

Abstract (continued)

CONCLUSIONS AND RELEVANCE These findings suggest that AI-aided interpretation was associated with improved reader performance and efficiency for identifying major thoracic findings on a chest radiograph.

JAMA Network Open. 2022;5(8):e2229289. doi:10.1001/jamanetworkopen.2022.29289

Introduction

Chest radiography is the most common imaging modality in the world for its portability, low cost, and accessibility.[1,2] It provides valuable information in detecting thoracic diseases and aiding with clinical decisions in managing them. Despite the abundance of test numbers, chest radiograph interpretation and reporting is an inherently difficult and subjective task, with previous research showing low to moderate interreader agreement in the final radiology report.[3-9] Furthermore, timely reporting of chest radiographs has been an issue in both developing and developed countries because of the shortage of qualified readers.[10,11] Accurate and efficient reading of chest radiographs is an important clinical target, especially in detecting clinically critical or urgent findings such as pneumothorax.[12]

Thus, there has been an increasing interest, with the rise of deep learning and artificial intelligence (AI) applications in medical imaging, to create chest radiograph AI algorithms that can help clinicians to accurately and efficiently detect key radiographic findings.[13,14] Research shows that AI algorithms can improve the performance of readers when used in a concurrent manner.[15-19] However, there are concerns about what the impact of AI would be in the real world, given that most research was conducted in a simulated setting without an observer performance tool that mimics the real-world workflow.

There is also a lack of evidence on the impact of AI in the reader efficiency, especially in terms of time taken for readers to complete their reports.[20] With previous computer-aided detection technology, such as in mammography, prior studies[21,22] reported workflow impediment due to the low specificity, which resulted in a high number of false positives. The concerns about AI reducing the reader efficiency are especially high for chest radiographs because of the sheer volume of chest radiographs in hospital settings, low reimbursement, and short reporting times for the experienced radiologists.[23,24] Therefore, for the wide adoption of AI algorithms in chest radiograph interpretation and reporting, it is crucial to show that there is no impedance in terms of accuracies and time taken to complete reporting with AI-assisted interpretation.

In this study, we explored the impact of AI on reader performance, both in terms of accuracy and efficiency. We have reflected the real-world environment by creating a custom version of the observer performance test platform that incorporates report templates and measures the time taken for readers to complete the interpretation and reporting task.

Methods

Data Sources and Approvals

The study used data from 2 sources. The first is the publicly available Medical Information Mart for Intensive Care–Chest X-Ray (MIMIC-CXR) database version 2.0.0, which is a large data set of chest radiographs from Beth Israel Deaconess Medical Center, Boston, Massachusetts. The second source of chest radiographs was another quaternary hospital (Massachusetts General Hospital [MGH], Boston, Massachusetts). This retrospective cohort study was approved by the institutional review board of MGH, which waived the need for informed consent because of the retrospective nature of the data collection and the use of anonymized images. The study met the requirements of the Health Insurance Portability and Accountability Act guidelines. This manuscript follows the Strengthening the Reporting of Observational Studies in Epidemiology (STROBE) reporting guidelines for observational studies.

Data Collection

The inclusion criteria for chest radiographs were adult patients regardless of gender, availability of frontal chest radiograph, and the presence of 1 or more of the 4 radiographic findings: pneumonia, lung nodule, pneumothorax, and pleural effusion. Because of high interobserver variation and subjectivity in evaluation of pulmonary nodules from chest radiographs,[25] only nodules larger than 6 mm in mean dimension on chest radiographs and with persistence on either follow-up or prior chest radiographs or a recent chest computed tomography image (within 3 months) were included in the study. We included only 1 chest radiograph per patient and excluded repeated or follow-up chest radiographs from the same patient. Repeated or follow-up chest radiographs from the same patient were excluded. Chest radiographs with incomplete inclusion of the entire chest or the presence of artifacts were also excluded. To avoid bias per STROBE guidelines, we chose consecutive chest radiographs with and without the 4 target findings. Both upright and portable chest radiographs were included in the cohort to ensure generalizability across radiographic techniques.

To identify the eligible cases, we reviewed the radiology reports of MIMIC-CXR data sets. For the MGH data sets, we used a proprietary radiology reports search engine, Render, with the keywords of nodule, pneumothorax, pleural effusion, and pneumonia. We also included 105 chest radiographs without any findings (normal radiographs) over the same time frame as those chest radiographs with the target findings.

AI Algorithm

A commercially available AI algorithm (Lunit INSIGHT CXR, version 3.1.2.0; Lunit Inc) was used to process the chest radiograph images. See the eAppendix in the Supplement for more details.

Ground Truth Creation

Two fellowship-trained thoracic radiologists (S.R.D. with 16 years of experience and M.K.K. with 14 years of experience) reviewed all chest radiographs and independently documented the presence of the radiographic findings. The target findings included pulmonary nodule, pneumothorax, pleural effusion, and pneumonia (including both the classic pattern of focal or multifocal consolidation and atypical pneumonia). In addition, because the presence of distracting or nontarget findings can influence the performance of both human readers and AI algorithms, we included nontarget findings such as enlarged cardiac silhouette, bone fractures, pleural thickening, atelectasis, and pericardial calcifications. The ground truths were created for both groups of findings to reflect the clinical reporting as much as possible. The reporting template used in the reading session has been created to match this (eFigure 1 in the Supplement).

The 2 radiologists specified the locations of each finding in different lung zones (upper, middle, and lower zones) in each lung. Any differences between the 2 ground-truthers were resolved in a joint review session to arrive at consensus.

Observer Performance Test Tool

We used a customized version of an online observer performance test tool (BestImage, version 6.0.0; IRM) for the reader study. The tool mimics the real-world picture archiving and communication system (PACS) and has several basic PACS viewer functionalities, such as window width and level adjustments, rotation, zooming, panning, and measurement features. In addition, the tool enables recording of radiographic findings and reporting time.

The tool enables users to report each chest radiograph using a multiple-choice question report template (eFigure 1 in the Supplement). When pneumothorax and pleural effusions were present, each reader selected the laterality of these findings (right, left, or bilateral). For lung nodules and pneumonia, each reader selected 1 or more lung zones for location of these findings: right upper, right mid, right lower, left upper, left mid, and left lower zone. For each of the 4 target radiographic findings (pleural effusion, pneumonia, pneumothorax, and nodule), each reader also recorded a confidence score ranging from 0% to 100%, on a 6-point Likert scale (0%, 1%-20%, 21%-40%, 41%-60%, 61%-80%, or 81%-100%). This represented the confidence of readers regarding the presence or the absence of lesions. In addition, the readers were also asked to comment on the other (nontarget findings) commonly found chest radiograph findings. This was done to ensure that the reading session reflected the real-life reporting as much as possible, such that performance and efficiency could be measured in a clinically relevant manner. Once a report was submitted, the readers could not make any further changes, thus reflecting the real-life reporting.

The tool also enabled worklist prioritization based on the AI score, at the user's discretion, which was designed to mimic the real-world PACS and worklist management. The physical server to run the tool was situated in the US to minimize the delay on the web interface. The screenshots of the reader study tool are presented in eFigure 2 in the Supplement. The reporting time was defined as the time from loading the chest radiographs to clicking the submit button on the report form.

Observer Performance Test

Six radiologists, including 2 thoracic radiologists (V.M. with 25 years of experience and S.M. with 15 years of experience), 2 thoracic imaging fellows (E.W.Z. and M.Y.W. with 6-8 months training as thoracic imaging fellows), and 2 second-year radiology residents (L.N. and J.F.D.), from MGH participated in the study as independent and blinded test readers. No radiologists involved in the ground truth creation participated as readers. The radiologists were blinded to the information pertaining to ground truthing, case selection, AI vendor, and specifics of study hypothesis. Before each review session, all radiologists reviewed 10 separate chest radiographs, which were not part of the analyzed data, as the training set, to enable user familiarization with the multiple-choice question report form and the observer performance tool.

The reader study was conducted between April and November 2021, with half of the readers interpreting chest radiographs with AI outputs and the other half reporting the chest radiographs without AI aid. **Figure 1** illustrates the different display modes of AI output. To avoid complications associated with mixed interpretation of alternate chest radiographs with and without AI outputs, we did not randomize the chest radiographs with and without AI outputs to any of the 6 readers. Each interpretation session, however, had a consecutive set of chest radiographs without AI output and a separate set of chest radiographs with AI output. Once the first session was complete, there was at least a 4-week washout before starting the second session. For the AI-aided session, each reader was able to review the original, unannotated chest radiograph and toggle the chest radiograph to view the AI output overlay.

Statistical Analysis

All statistical analyses were performed with R statistical software version 3.6.2 (R Project for Statistical Computing). We estimated the area under the receiver operating characteristic curve (AUROC) of AI stand-alone performance and the reader performance with and without AI outputs. The AUROCs were compared using the DeLong test. Since the purpose of our study was to note the finding level performance with AI, all analyses were performed on individual findings (ie, data measures included >1 data point for some chest radiographs with multiple findings of similar or different types). Sensitivity and specificity comparison between AI-aided and unaided interpretation was performed with generalized estimated equations. For the observer performance test, the sensitivity and the specificity of readers between 2 sessions were compared using McNemar test. To estimate sensitivity and specificity, we considered any finding with greater than 0% score as a predicted finding. The Kolmogorov-Smirnov test was used to test the normal distribution of data on the reporting times. For nonnormal distribution, we estimated median and IQR of values.

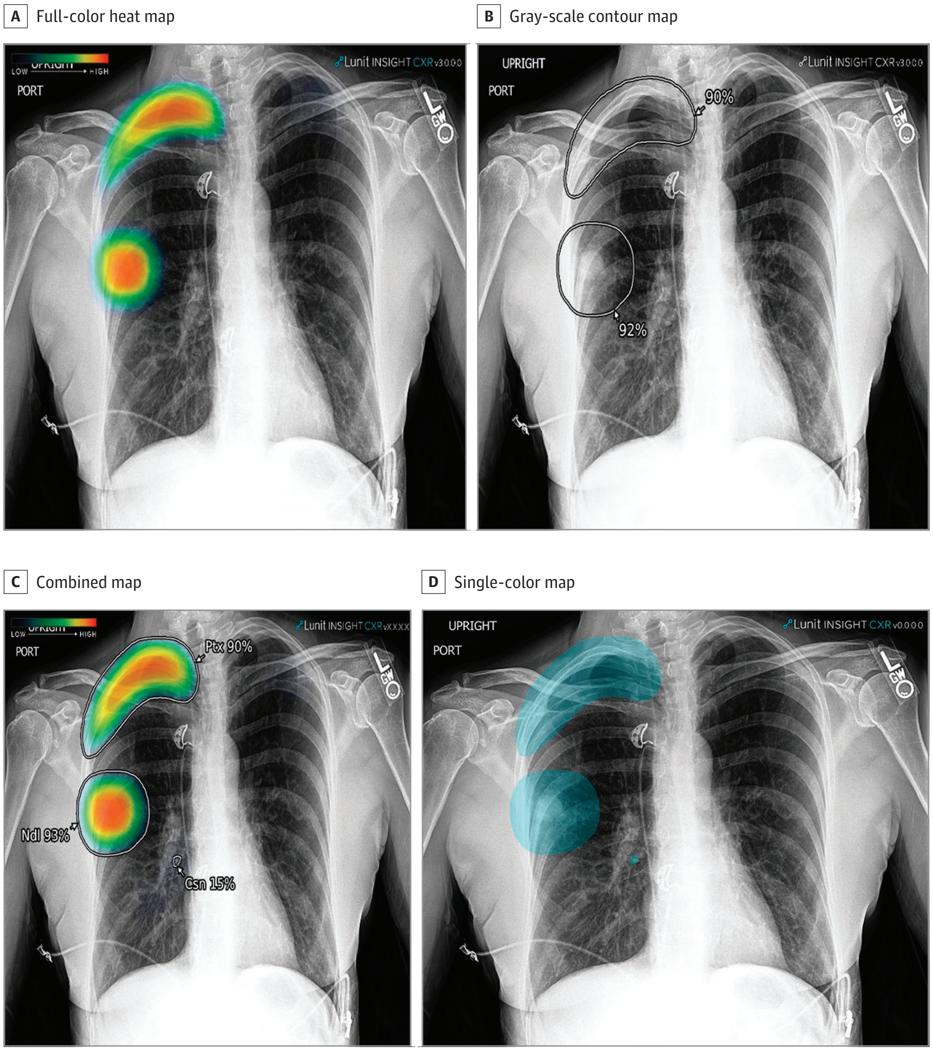
Reporting times for each chest radiograph were compared between 2 sessions using the paired *t* test. We excluded interpretation times for cases requiring more than 3 minutes (2-fold higher than the median times). These resulted from errors related to failure on part of the readers to click the submit button or failure to log off from the interpretation tool at the end of an interpretation session. These errors were realized upon statistical analysis of the data. For all tests, 2-sided $P < .05$ was considered as significant.

Results

The study included a total of 497 frontal chest radiographs (both portable anteroposterior and erect posteroanterior projections). The first 247 chest radiographs were randomly selected from the MIMIC-CXR data set[26] (patients' demographic data were unavailable), and the other 250 chest radiographs were from MGH (mean [SD] age, 63 [16] years; 133 men [53.2%] and 117 women [46.8%]). To simulate reporting volume in our hospital, we chose the number of chest radiographs to represent the approximate number of chest radiographs that individual radiologists report over 2 full days.

The final distribution of each of the target finding in 351 abnormal chest radiographs was as follows: 114 lung nodules, 195 pneumonia, 149 pleural effusion, and 80 pneumothorax. A total of 146

Figure 1. Different Display Modes Available for the Artificial Intelligence Output



A Full-color heat map

B Gray-scale contour map

C Combined map

D Single-color map

Shown are the color heat map (A), grayscale contour map (B), combined map (C), and single-color map (D).

chest radiographs had no abnormal target radiographic findings. The full distribution of the findings can be found in **Table 1**.

AI Stand-alone Performance

From the stand-alone performance perspective, AI identified the 4 target chest radiograph findings with sensitivities of 0.816 to 0.988 and specificities of 0.728 to 0.986. The highest sensitivity and specificity were calculated for pneumothorax detection (sensitivity, 0.988; specificity, 0.986; AUROC, 0.999 [95% CI, 0.997-1.00]). **Figure 2** summarizes the AUROC of a deep-learning algorithm for each target findings and comparison against the reader performance. The lowest diagnostic accuracy was calculated for lung nodule detection (sensitivity, 0.816; specificity, 0.731; AUROC, 0.858 [95% CI, 0.819-0.897]). AI stand-alone performance for detecting the 4 target findings is summarized in **Table 2**.

Compared with the ground-truth, on a stand-alone basis, AI was associated with detection rates of 82.5% for lung nodules (94 of 114 findings), 88.7% for pneumonia (173 of 195 findings), 87.2% for pleural effusions (130 of 149 findings), and 100.0% for pneumothoraces (80 of 80 findings). There was no significant overall AUROC difference in the AI stand-alone performance in chest radiographs with and without nontarget findings (0.955 [95% CI, 0.933-0.976] vs 0.898 [95% CI, 0.838-0.958]; *P* = .08) (**Table 3**).

Readers' Performance

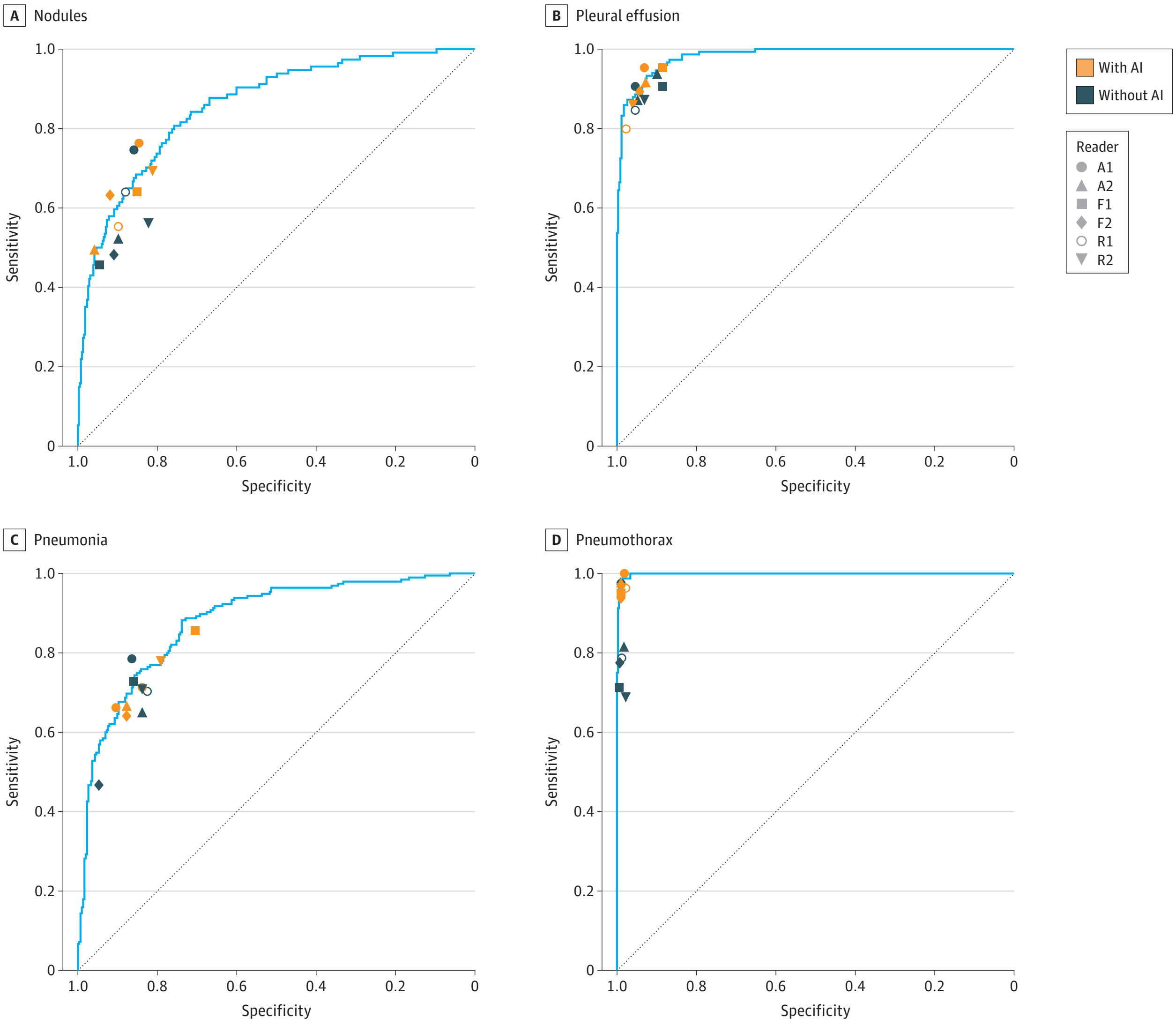
The Cohen κ statistic for the 2 ground truth radiologists before the consensus reading was 0.794 for the target findings. For all 4 target radiographic findings, AI-assisted interpretation was associated with a significant improvement in the sensitivities compared with unassisted reporting (without vs with AI: nodule, 0.567 [95% CI, 0.524-0.611] vs 0.629 [95% CI, 0.586-0.671]; pneumonia, 0.673 [95% CI, 0.632-0.714] vs 0.719 [95% CI, 0.679-0.758]; pleural effusion, 0.889 [95% CI, 0.862-0.917]

Table 1. Distribution of the Findings

Findings	Total, No. (%)
Origin of chest radiographs (n = 497)	
Massachusetts General Hospital	250 (50.3)
The Medical Information Mart for Intensive Care	247 (49.7)
Target findings (n = 538)	
Nodule	114 (21.2)
Pleural effusion	149 (27.7)
Pneumonia	195 (36.2)
Pneumothorax	80 (14.9)
Nontarget findings (n = 193)	
Extra pleural	7 (3.6)
Extra thoracic	181 (93.8)
Bone fractures	5 (2.6)
Target findings on chest radiographs (n = 497)	
0	146 (29.4)
1	198 (39.8)
2	120 (24.2)
3	32 (6.4)
4	1 (0.2)
Chest radiographs with findings (target and nontarget) (n = 497)	
0	105 (21.1)
1	174 (35.1)
2	108 (21.7)
≥3	110 (22.1)

vs 0.895 [95% CI, 0.868-0.922]; pneumothorax, 0.792 [95% CI, 0.756-0.827] vs 0.965 [95% CI, 0.949-0.981]; *P* < .001). However, there was no change in specificities of individual test radiologists between AI-assisted and unassisted interpretation (without vs with AI: nodule, 0.885 [95% CI, 0.858-0.913] vs 0.881 [95% CI, 0.852-0.929]; pneumonia, 0.862 [95% CI, 0.832-0.892] vs 0.832 [95% CI, 0.799-0.865]; pleural effusion, 0.928 [95% CI, 0.906-0.951] vs 0.937 [95% CI, 0.916-0.959]; pneumothorax, 0.988 [95% CI, 0.978-0.997] vs 0.986 [95% CI, 0.976-0.996]; *P* = .07). Compared with the AI stand-alone performance, human readers had lower sensitivity both with and without AI assistance (AI stand-alone vs human without AI vs human with AI: nodule, 0.816 [95% CI, 0.732-0.882] vs 0.567 [95% CI, 0.524-0.611] vs 0.629 [95% CI, 0.586-0.671]; pneumonia, 0.887 [95% CI, 0.834-0.928] vs 0.673 [95% CI, 0.632-0.714] vs 0.719 [95% CI, 0.679-0.758]; pleural effusion, 0.872 [95% CI, 0.808-0.921] vs 0.889 [95% CI, 0.862-0.917] vs 0.895 [95% CI, 0.868-0.922]; pneumothorax, 0.988 [95% CI, 0.932-1.000] vs 0.792 [95% CI, 0.756-0.827] vs 0.965 [95% CI, 0.949-0.981]). Table 2 summarizes the mean performance changes for all readers.

Figure 2. Receiver Operating Characteristic Curves of a Deep-Learning Artificial Intelligence (AI) Algorithm for the Target Findings and Comparison Against the Reader Performance



Graphs show data for nodules (A), pleural effusions (B), pneumonia (C), and pneumothorax (D). Diagonal lines denote lines of regression. A indicates attending radiologist; F, fellow; R, resident.

The individual reader performance in the cohort study is summarized in Table 2. The sensitivity was higher for all readers with different levels of experience for detection of pneumothorax with AI-assisted interpretation compared with unaided readout (without vs with AI: attending radiologist 1, 0.975 [95% CI, 0.913-0.997] vs 1.000 [95% CI, 0.955-1.000]; attending radiologist 2, 0.812 [95% CI, 0.71-0.891] vs 0.975 [95% CI, 0.913-0.997]; fellow 1, 0.713 [95% CI, 0.600-0.808] vs 0.950 [95% CI, 0.877-0.986]; fellow 2, 0.775 [95% CI, 0.668-0.861] vs 0.983 [95% CI, 0.860-0.979];

Table 2. Sensitivity, Specificity, and AUROC of Individual Readers

Findings and readers	Sensitivity (95% CI)		Specificity (95% CI)		AUROC (95% CI)	
	Without AI	With AI	Without AI	With AI	Without AI	With AI
Nodules						
AI	NA	0.816 (0.732-0.882)	NA	0.731 (0.684-0.775)	NA	0.858 (0.819-0.897)
Attending radiologist 1	0.746 (0.656-0.823)	0.765 (0.674-0.838)	0.859 (0.820-0.892)	0.846 (0.806-0.881)	0.799 (0.748-0.850)	0.801 (0.751-0.851)
Attending radiologist 2	0.518 (0.422-0.612)	0.486 (0.396-0.587)	0.898 (0.863-0.927)	0.958 (0.933-0.976)[a]	0.706 (0.645-0.766)	0.723 (0.662-0.784)
Fellow 1	0.456 (0.363-0.552)	0.640 (0.545-0.728)[a]	0.945 (0.917-0.966)	0.851 (0.812-0.885)[a]	0.699 (0.637-0.760)	0.773 (0.687-0.799)[a]
Fellow 2	0.482 (0.388-0.578)	0.632 (0.536-0.72)[a]	0.909 (0.875-0.936)	0.919 (0.887-0.944)	0.639 (0.632-0.755)	0.773 (0.716-0.829)
Resident 1	0.640 (0.545-0.728)	0.553 (0.457-0.646)	0.880 (0.843-0.911)	0.898 (0.863-0.927)	0.757 (0.701-0.814)	0.723 (0.664-0.782)
Resident 2	0.561 (0.465-0.654)	0.693 (0.600-0.776)[a]	0.822 (0.780-0.859)	0.812 (0.769-0.85)	0.689 (0.630-0.749)	0.749 (0.695-0.804)[a]
Mean[b]	0.567 (0.524-0.611)	0.629 (0.586-0.671)[a]	0.885 (0.858-0.913)	0.881 (0.852-0.909)[a]	0.724 (0.700-0.748)	0.752 (0.729-0.775)
Pneumonia						
AI	NA	0.887 (0.834-0.928)	NA	0.728 (0.675-0.778)	NA	0.880 (0.849-0.911)
Attending radiologist 1	0.785 (0.720-0.84)	0.662 (0.590-0.728)[a]	0.864 (0.820-0.901)	0.904 (0.865-0.935)	0.825 (0.784-0.765)	0.783 (0.738-0.828)[a]
Attending radiologist 2	0.646 (0.575-0.713)	0.662 (0.590-0.728)	0.838 (0.791-0.877)	0.877 (0.835-0.912)	0.742 (0.696-0.789)	0.770 (0.724-0.815)
Fellow 1	0.728 (0.660-0.789)	0.856 (0.799-0.902)[a]	0.861 (0.817-0.898)	0.705 (0.650-0.756)[a]	0.795 (0.752-0.938)	0.783 (0.741-0.825)
Fellow 2	0.467 (0.395-0.539)	0.641 (0.569-0.708)[a]	0.947 (0.915-0.969)	0.877 (0.835-0.912)[a]	0.707 (0.657-0.757)	0.759 (0.713-0.805)[a]
Resident 1	0.703 (0.633-0.766)	0.713 (0.644-0.775)	0.825 (0.777-0.866)	0.838 (0.791-0.877)	0.764 (0.719-0.809)	0.776 (0.731-0.820)
Resident 2	0.708 (0.638-0.770)	0.779 (0.715-0.836)	0.838 (0.791-0.877)	0.791 (0.741-0.836)	0.773 (0.728-0.817)	0.786 (0.743-0.829)
Mean[b]	0.673 (0.632-0.714)	0.719 (0.679-0.758)[a]	0.862 (0.832-0.892)	0.832 (0.799-0.865)[a]	0.768 (0.749-0.786)	0.776 (0.758-0.794)
Pleural effusion						
AI	NA	0.872 (0.808-0.921)	NA	0.960 (0.933-0.978)	NA	0.983 (0.974-0.992)
Attending radiologist 1	0.906 (0.847-0.948)	0.953 (0.906-0.981)	0.954 (0.926-0.973)	0.931 (0.899-0.955)	0.930 (0.900-0.960)	0.942 (0.917-0.967)
Attending radiologist 2	0.933 (0.880-0.967)	0.913 (0.855-0.953)	0.899 (0.863-0.929)	0.928 (0.896-0.953)	0.916 (0.887-0.946)	0.921 (0.890-0.951)
Fellow 1	0.906 (0.847-0.948)	0.953 (0.906-0.981)	0.885 (0.847-0.917)	0.885 (0.847-0.917)	0.896 (0.862-0.929)	0.916 (0.887-0.945)
Fellow 2	0.872 (0.808-0.921)	0.893 (0.831-0.937)	0.948 (0.919-0.969)	0.943 (0.913-0.965)	0.910 (0.876-0.944)	0.918 (0.886-0.950)
Resident 1	0.846 (0.777-0.900)	0.799 (0.725-0.860)	0.954 (0.926-0.973)	0.977 (0.955-0.990)	0.900 (0.864-0.936)	0.888 (0.848-0.927)
Resident 2	0.872 (0.808-0.921)	0.859 (0.793-0.911)	0.931 (0.899-0.955)	0.960 (0.933-0.978)	0.903 (0.869-0.938)	0.909 (0.875-0.944)
Mean[b]	0.889 (0.862-0.917)	0.895 (0.868-0.922)[a]	0.928 (0.906-0.951)	0.937 (0.916-0.959)[a]	0.909 (0.896-0.923)	0.916 (0.902-0.929)
Pneumothorax						
AI	NA	0.988 (0.932-1.000)	NA	0.986 (0.969-0.995)	NA	0.999 (0.997-1.000)
Attending radiologist 1	0.975 (0.913-0.997)	1.000 (0.955-1.000)[a]	0.990 (0.976-0.997)	0.981 (0.963-0.992)	0.977 (0.952-1.000)	0.984 (0.968-1.000)
Attending radiologist 2	0.812 (0.710-0.891)	0.975 (0.913-0.997)[a]	0.983 (0.966-0.993)	0.990 (0.976-0.997)	0.893 (0.841-0.945)	0.977 (0.952-1.000)[a]
Fellow 1	0.713 (0.600-0.808)	0.950 (0.877-0.986)[a]	0.995 (0.983-0.999)	0.990 (0.976-0.997)	0.849 (0.788-0.910)	0.958 (0.924-0.992)[a]
Fellow 2	0.775 (0.668-0.861)	0.938 (0.860-0.979)[a]	0.993 (0.979-0.999)	0.990 (0.976-0.997)	0.879 (0.823-0.935)	0.958 (0.924-0.992)[a]
Resident 1	0.787 (0.682-0.871)	0.963 (0.894-0.992)[a]	0.988 (0.972-0.996)	0.978 (0.959-0.99)	0.883 (0.828-0.937)	0.965 (0.936-0.993)[a]
Resident 2	0.688 (0.574-0.787)	0.963 (0.894-0.992)[a]	0.978 (0.959-0.99)	0.988 (0.972-0.996)	0.829 (0.766-0.891)	0.969 (0.940-0.997)[a]
Mean[b]	0.792 (0.756-0.827)	0.965 (0.949-0.981)[a]	0.988 (0.978-0.997)	0.986 (0.976-0.996)[a]	0.885 (0.863-0.907)	0.969 (0.957-0.980)[a]

Abbreviations: AI, artificial intelligence; AUROC, area under the receiver operating characteristic curve; NA, not applicable.

[a] Denotes metrics with statistically significant differences between the AUROCs, sensitivities, and specificities with and without AI (*P* < .05).

[b] The mean values represent mean reader performance and do not include stand-alone AI performance.

resident 1, 0.787 [95% CI, 0.682-0.871] vs 0.963 [95% CI, 0.894-0.992]; resident 2, 0.688 [95% CI, 0.574-0.787] vs 0.963 [95% CI, 0.894-0.992]; *P* = .009). The 2 thoracic imaging fellows reported pneumonia with significantly higher sensitivity with AI assistance (without vs with AI: fellow 1, 0.728 [95% CI, 0.660-0.789] vs 0.856 [95% CI, 0.799-0.902]; fellow 2, 0.467 [95% CI, 0.395-0.539] vs 0.641 [95% CI, 0.569-0.708]; *P* < .001). For lung nodule detection, 2 fellows and 1 resident witnessed higher sensitivity with AI compared with interpretation without AI (without vs with AI: fellow 1, 0.456 [95% CI, 0.363-0.552] vs 0.640 [95% CI, 0.545-0.728]; fellow 2, 0.482 [95% CI, 0.388-0.578] vs 0.632 [95% CI, 0.536-0.72]; resident 2, 0.561 [95% CI, 0.465-0.654] vs 0.693 [95% CI, 0.60-0.720]; *P* = .02) (Table 2).

The association of target and nontarget findings with reader performance was variable, as seen in eTable 1 in the Supplement. The sensitivity was higher for pneumonia (without vs with nontarget findings: without AI, 0.727 [95% CI, 0.676-0.778] vs 0.609 [95% CI, 0.542-0.677]; with AI, 0.775 [95% CI, 0.727-0.822] vs 0.654 [95% CI, 0.588-0.720]; *P* < .001), nodules (without vs with nontarget findings: without AI, 0.598 [95% CI, 0.542-0.653] vs 0.519 [95% CI, 0.450-0.588]; with AI, 0.674 [95% CI, 0.621-0.727] vs 0.557 [95% CI, 0.488-0.626]; *P* < .001), and pneumothorax (without vs with nontarget findings: without AI, 0.820 [95% CI, 0.777-0.864] vs 0.766 [95% CI, 0.707-0.825]; with AI, 0.965 [95% CI, 0.944-0.986] vs 0.964 [95% CI, 0.938-0.990]; *P* < .001) detection by the readers, when there were no nontarget findings present. Similarly, the specificity was higher for pneumonia (without vs with nontarget findings: without AI, 0.922 [95% CI, 0.892-0.952] vs 0.757 [95% CI, 0.698-0.817]; with AI, 0.893 [95% CI, 0.858-0.928] vs 0.726 [95% CI, 0.664-0.788]; *P* < .001), pleural effusion (without vs with nontarget findings: without AI, 0.953 [95% CI, 0.929-0.977] vs 0.870 [95% CI, 0.823-0.916]; with AI, 0.962 [95% CI, 0.940-0.984] vs 0.878 [95% CI, 0.832-0.923]; *P* < .001), and pneumothorax (without vs with nontarget findings: without AI, 0.994 [95% CI, 0.984-1.000] vs 0.979 [95% CI, 0.959-0.999]; with AI, 0.994 [95% CI, 0.984-1.000] vs 0.974 [95% CI, 0.956-0.991]; *P* < .001), when there were nontarget findings present. These findings can be found in eTable 1 in the Supplement.

Compared with the ground truth, without AI aid, radiologists detected 45.6% (52 of 114 findings) to 73.7% (84 of 114 findings) of lung nodules, 46.7% (91 of 195 findings) to 78.5% (153 of

Table 3. Summary of Artificial Intelligence Stand-alone Performance for Detection of the 4 Target Findings in Chest Radiographs With and Without Nontarget Findings

Target findings	Detected findings, No.	AUROC (95% CI)	Sensitivity (95% CI)	Specificity (95% CI)
Nodule				
Without extra findings	70	0.870 (0.823-0.918)	0.829 (0.720-0.908)	0.740 (0.678-0.796)
With extra findings	44	0.842 (0.778-0.906)	0.795 (0.647-0.902)	0.718 (0.640-0.787)
P value	NA	.49	.66	.63
Pneumonia				
Without extra findings	105	0.931 (0.899-0.963)	0.895 (0.820-0.947)	0.854 (0.796-0.901)
With extra findings	90	0.777 (0.712-0.841)	0.878 (0.792-0.937)	0.509 (0.412-0.606)
P value	NA	<.001	.70	<.001
Pleural effusion				
Without extra findings	50	0.985 (0.971-0.999)	0.860 (0.733-0.942)	0.980 (0.953-0.993)
With extra findings	99	0.975 (0.958-0.991)	0.879 (0.798-0.936)	0.911 (0.838-0.958)
P value	NA	.34	.75	.003
Pneumothorax				
Without extra findings	38	0.999 (0.997-1.000)	0.974 (0.862-0.999)	0.996 (0.979-1.000)
With extra findings	42	0.999 (0.997-1.000)	1.000 (0.916-1.000)	0.968 (0.928-0.990)
P value	NA	.93	.29	.02
Total				
Without extra findings	297	0.955 (0.933-0.976)	0.917 (0.868-0.952)	0.857 (0.775-0.918)
With extra findings	200	0.898 (0.838-0.958)	0.969 (0.928-0.99)	0.585 (0.421-0.737)
P value	NA	.08	.04	<.001

Abbreviations: AUROC, area under the receiver operating characteristic curve; NA, not applicable.

195 findings) of cases of pneumonia, 84.6% (126 of 149 findings) to 93.3% (139 of 149 findings) of pleural effusions, and 68.8% (55 of 80 findings) to 98.8% (79 of 80 findings) of pneumothoraces. With the AI aid, radiologists detected 49.1% (56 of 114 findings) to 77.2% (88 of 114 findings) of lung nodules, 64.1% (125 of 195 findings) to 85.6% (167 of 195 findings) of cases of pneumonia, 79.9% (119 of 149 findings) to 95.3% (142 of 149 findings) of pleural effusions, and 95.0% (76 of 80 findings) to 100.0% (80 of 80 findings) of pneumothoraces.

The range (minimum to maximum number of findings for individual readers) of additional or distracting findings beyond the 4 target chest radiograph findings detected by the readers included atelectasis (96-348 findings), pulmonary edema (42-116 findings), fibrosis (0-12 findings), pleural calcification (0-23 findings), pleural thickening (2-63 findings), and pleural nodules (0-9 findings). As noted from the wide ranges, there were considerable variations in the numbers of individual target findings detected by different radiologists.

Interpretation Time of Radiograph

eTable 2 in the Supplement summarizes interpretation times for individual test readers with and without AI-assisted interpretation. We excluded outlier data related to interpretation times (>3 minutes) for 81 unaided and 75 AI-aided interpretations. There was a small but significant reduction in interpretation time associated with AI-assisted interpretation compared with unaided interpretation (median [IQR], 36.9 [23.5-53.7] seconds vs 40.8 [27.5-58.2] seconds; difference, 3.9 seconds; 95% CI, 2.9-5.2 seconds; $P < .001$). There was a significant reduction in reporting times for the trainees (resident 1, resident 2, and fellow 2). On the entire data set of interpretation times including those with outliers (interpretation time >3 minutes), AI-assisted interpretation was significantly faster than unaided interpretation (median [IQR], 37.9 [24.1-56.1] seconds vs 42.0 [28.2-60.2] seconds; difference, 4.1 seconds; 95% CI, 3.0-5.4 seconds; $P < .001$).

Discussion

In this cohort study, the use of an AI algorithm was associated with sensitivity gains for all 4 target chest radiograph findings across all readers regardless of their experience and training status. Such improvement in detection of findings with AI has been reported in several prior studies.[15,16,19,27,28] Prior studies[27,29,30] with our AI as well as other research and commercial AI algorithms have reported comparable or lower diagnostic performance for detection of pneumothorax, pleural effusion, pneumonia, and pulmonary nodules. Although the stand-alone performance of our AI algorithm was comparable to that of the radiologists and trainees in terms of sensitivity, it had substantially lower specificity. One of the reasons for the lower specificity of the AI stand-alone performance in pneumonia may be that the AI engine output is designed to label both airspace and interstitial opacities as pneumonia. In contrast, the ground-truth and test radiologists used the classic airspace pattern on the chest radiographs to label a finding as pneumonia. This difference in labeling likely contributed to the lower specificity for the AI stand-alone performance and is further evident from higher AI specificity on chest radiographs without any nontarget findings.

Despite the higher frequency of false-positive findings with AI, the improved sensitivity with AI-aided interpretation did not come at the cost of a significant change in their specificities. In other words, all readers were able to reject AI-detected false-positive findings while benefiting from acceptance of true-positive findings detected and marked up by the AI algorithm. Examples of such cases can be found in eFigure 3 in the Supplement.

Most studies[27,31,32] either do not include distracting findings (nontarget AI findings) or do not evaluate the impact on interpretation efficiency assessed in our study. Given the tremendous volume of chest radiographs, the accuracy of detection is as important as the efficiency of reporting the chest radiographs in a timely manner within 12 to 24 hours of their acquisition.

In this context, our study found a small but measurably significant improvement in time to report chest radiographs with AI-aided interpretation compared with interpretation without AI.

Because the AI was limited for aiding with detection of 4 findings only, this could be a meaningful reduction in reporting time. We included 11 nontarget AI findings and incorporated them into our reporting template. This was done to reflect the real clinical practice and the thoracic abnormalities that can be detected, other than the 4 target findings. Our study highlights the method and need for future research and clinical adoption of AI algorithms with simultaneous evaluation of diagnostic accuracy and workflow efficiency.

The chief implication of our study is the improved accuracy in detecting 4 target chest radiograph findings (pneumonia, lung nodule, pleural effusion, and pneumothorax) with AI-aided interpretation while improving overall efficiency in reporting target and nontarget findings. Specifically, improved interpretation both in terms of finding detection and interpretation time was most notable for 3 of the 4 residents and thoracic imaging fellows. Although neither attending radiologist improved their reading efficiency with AI, there was an improvement in detection of target findings with use of AI for both radiologists. Demonstration of noninferiority of interpretation time with AI vs non-AI interpretation will become more critical as AI algorithms expand their target findings beyond a handful to a comprehensive, multifinding detection.[33]

Another implication of our study pertains to the use of a structured, form-based report format instead of the conventional, free-text, dictation-based reporting system used in our department. Although a similar structured, form-based reporting system is often used for screening mammography,[34] its use is limited in chest radiograph reporting. Yet, several studies[35] have highlighted the need to generate structured reporting templates and formats to improve consistency, reduce errors, and enhance readability between radiologists. Apart from the AI algorithm used in our study, other commercial AI vendors[36,37] also provide a structured list of AI-detected and annotated findings. In addition to providing measurable data for research, quality control, and audits of chest radiograph findings, such checklist-based reporting can also help as a gatekeeper of AI findings that get archived and transferred into PACS and/or electronic medical records (true positive or true negative) vs those that get deleted (false-positive outputs of AI) before the AI output becomes archived. Such information can be used for monitoring AI performance in a continuous, clinical use, since converting free-text reporting format is tedious, error prone, and inconsistent. Another implication of our study pertains to the likelihood that the reader improvement with AI assistance was influenced by some factors such as the size, extent, and/or number of findings. A true impact statement regarding AI performance would require separate assessment of these factors, which were not assessed in our study.[38-41]

Limitations

There are limitations in our study. First, the ground truths were obtained from frontal chest radiographs only, which could have resulted in some inaccuracies. However, most prior studies[27,28] have used multireader ground truths from chest radiographs alone. Second, given the difficulty in integrating research software into a clinical reporting interface, it was not possible to assess the real-world, clinical chest radiograph interpretation workflow. The reader study tool used in our study simulated the structured reporting format in our practice. Although the checklist type of reporting format used in our study does not conform with the field-based, free-text, structured reporting template in clinical workflow, it was not feasible to convert free-text reports into measurable data for statistical analysis. This limitation can, however, restrict the application of our research for assessing true reporting efficiency with and without AI. Third, we excluded data pertaining to interpretation times greater than 3 minutes, which represented less than 10% of the overall chest radiograph interpretations. As reported in the Results section, such exclusion did not change the data on reporting efficiency either toward or against the AI-assisted interpretation of chest radiographs. Fourth, we did not perform a power analysis to determine the adequacy of our sample size or the number of test readers. Fifth, although the AI algorithm used can detect more than 4 target findings, we focused on 4 findings and, therefore, cannot comment on reader performance or reporting efficiency when other AI findings are also included. It is possible that with a larger number of AI findings and mark-up, a greater number of false-positive or additional true-positive annotations can slow down the readers. The latter, however, will result in improved reader performance from detection of additional findings. Sixth, because of the small number of diverse nontarget findings (eg, lines and tubes, cardiac silhouette enlargement, mediastinal widening, and bony abnormalities), we did not assess the effect of AI-aided or unaided interpretation on detection of the nontarget findings. Although loss of performance in detection of nontarget findings from the use of any AI algorithm will be detrimental, none of the AI algorithms cleared by the US Food and Drug Administration can detect or triage all possible findings on chest radiographs.

Conclusions

In conclusion, the use of an AI algorithm was associated with an improved sensitivity for detection of 4 target chest radiograph findings (pneumonia, lung nodules, pleural effusion, and pneumothorax) for radiologists, thoracic imaging fellows as well as radiology residents, while maintaining the specificity. These findings suggest that an AI algorithm can improve the reader performance and efficiency in interpreting chest radiograph abnormalities.

ARTICLE INFORMATION

Accepted for Publication: July 15, 2022.

Published: August 31, 2022. doi:10.1001/jamanetworkopen.2022.29289

Corresponding Author: Mannudeep K. Kalra, MD, Division of Thoracic Imaging, Department of Radiology, Massachusetts General Hospital and Harvard Medical School, 75 Blossom Ct, Boston, MA 02114 (mkalra@mgh.harvard.edu).

Author Affiliations: Lunit, Inc, Seoul, South Korea (Ahn, Lee); Division of Thoracic Imaging, Department of Radiology, Massachusetts General Hospital and Harvard Medical School, Boston, Massachusetts (Ebrahimian, McDermott, Naccarato, Di Capua, Wu, Zhang, Muse, Miller, Sabzalipour, Bizzo, Dreyer, Kaviani, Digumarthy, Kalra); Internal Medicine, Icahn School of Medicine at Mount Sinai, Elmhurst Hospital Center, Elmhurst, New York (Ebrahimian); Data Science Office, Mass General Brigham, Boston, Massachusetts (Miller, Sabzalipour, Bizzo, Dreyer, Kalra).

Author Contributions: Dr Kalra had full access to all of the data in the study and takes responsibility for the integrity of the data and the accuracy of the data analysis. Drs Ahn and Ebrahimian contributed equally as coauthors.

Concept and design: Ahn, Lee, Miller, Bizzo, Dreyer, Kalra.

Acquisition, analysis, or interpretation of data: Ahn, Ebrahimian, McDermott, Lee, Naccarato, Di Capua, Wu, Zhang, Muse, Sabzalipour, Kaviani, Digumarthy, Kalra.

Drafting of the manuscript: Ahn, Ebrahimian, Di Capua, Miller, Sabzalipour, Digumarthy, Kalra.

Critical revision of the manuscript for important intellectual content: Ahn, Ebrahimian, McDermott, Lee, Naccarato, Di Capua, Wu, Zhang, Muse, Bizzo, Dreyer, Kaviani, Kalra.

Statistical analysis: Ahn, Ebrahimian, Di Capua, Wu, Kaviani.

Administrative, technical, or material support: Ahn, Ebrahimian, McDermott, Lee, Muse, Miller, Sabzalipour, Bizzo, Dreyer.

Supervision: Lee, Di Capua, Bizzo, Dreyer, Digumarthy, Kalra.

Conflict of Interest Disclosures: Dr Naccarato reported receiving personal fees from Massachusetts General Hospital for time interpreting chest radiographs during the conduct of the study. Dr Digumarthy reported receiving personal fees from Siemens Healthineers and grants from Lunit, GE, Vuno, and QureAI outside the submitted work; Dr Digumarthy also provides independent image analysis for hospital-contracted clinical research trials programs for Merck, Pfizer, Bristol Myers Squibb, Novartis, Roche, Polaris, Cascadian, Abbvie, Gradalis, Bayer, Zai Laboratories, Shanghai Bioscience, Biengen, Resonance, Riverain, and Analise. Dr Kalra reported receiving grants from Lunit, Inc, for an unrelated study outside the submitted work. No other disclosures were reported.

Funding/Support: This work was supported by a grant from Lunit, Inc.

Role of the Funder/Sponsor: The funder 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.

REFERENCES

1. McComb BL, Chung JH, Crabtree TD, et al; Expert Panel on Thoracic Imaging. ACR Appropriateness Criteria® routine chest radiography. *J Thorac Imaging*. 2016;31(2):W13-W15. doi:10.1097/RTI.0000000000000200

2. Mettler FA Jr, Mahesh M, Bhargavan-Chatfield M, et al. Patient exposure from radiologic and nuclear medicine procedures in the United States: procedure volume and effective dose for the period 2006-2016. *Radiology*. 2020;295(2):418-427. doi:10.1148/radiol.2020192256

3. de Groot PM, Carter BW, Abbott GF, Wu CC. Pitfalls in chest radiographic interpretation: blind spots. *Semin Roentgenol*. 2015;50(3):197-209. doi:10.1053/j.ro.2015.01.008

4. Austin JH, Romney BM, Goldsmith LS. Missed bronchogenic carcinoma: radiographic findings in 27 patients with a potentially resectable lesion evident in retrospect. *Radiology*. 1992;182(1):115-122. doi:10.1148/radiology.182.1.1727272

5. Johnson J, Kline JA. Intraobserver and interobserver agreement of the interpretation of pediatric chest radiographs. *Emerg Radiol*. 2010;17(4):285-290. doi:10.1007/s10140-009-0854-2

6. Moncada DC, Rueda ZV, Macías A, Suárez T, Ortega H, Vélez LA. Reading and interpretation of chest X-ray in adults with community-acquired pneumonia. *Braz J Infect Dis*. 2011;15(6):540-546. doi:10.1016/S1413-8670(11)70248-3

7. Albaum MN, Hill LC, Murphy M, et al; PORT Investigators. Interobserver reliability of the chest radiograph in community-acquired pneumonia. *Chest*. 1996;110(2):343-350. doi:10.1378/chest.110.2.343

8. Melbye H, Dale K. Interobserver variability in the radiographic diagnosis of adult outpatient pneumonia. *Acta Radiol*. 1992;33(1):79-81.

9. Campbell SG, Murray DD, Hawass A, Urquhart D, Ackroyd-Stolarz S, Maxwell D. Agreement between emergency physician diagnosis and radiologist reports in patients discharged from an emergency department with community-acquired pneumonia. *Emerg Radiol*. 2005;11(4):242-246. doi:10.1007/s10140-005-0413-4

10. Rimmer A. Radiologist shortage leaves patient care at risk, warns royal college. *BMJ*. 2017;359:j4683. doi:10.1136/bmj.j4683

11. Nakajima Y, Yamada K, Imamura K, Kobayashi K. Radiologist supply and workload: international comparison—Working Group of Japanese College of Radiology. *Radiat Med*. 2008;26(8):455-465. doi:10.1007/s11604-008-0259-2

12. Yarmus L, Feller-Kopman D. Pneumothorax in the critically ill patient. *Chest*. 2012;141(4):1098-1105. doi:10.1378/chest.11-1691

13. Lakhani P, Sundaram B. Deep learning at chest radiography: automated classification of pulmonary tuberculosis by using convolutional neural networks. *Radiology*. 2017;284(2):574-582. doi:10.1148/radiol.2017162326

14. Topol EJ. High-performance medicine: the convergence of human and artificial intelligence. *Nat Med*. 2019;25(1):44-56. doi:10.1038/s41591-018-0300-7

15. Hwang EJ, Park S, Jin K-N, et al; DLAD Development and Evaluation Group. 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. Park S, Lee SM, Lee KH, et al. Deep learning-based detection system for multiclass lesions on chest radiographs: comparison with observer readings. *Eur Radiol*. 2020;30(3):1359-1368. doi:10.1007/s00330-019-06532-x

17. Hwang EJ, Hong JH, Lee KH, et al. Deep learning algorithm for surveillance of pneumothorax after lung biopsy: a multicenter diagnostic cohort study. *Eur Radiol*. 2020;30(7):3660-3671. doi:10.1007/s00330-020-06771-3

18. Park S, Lee SM, Kim N, et al. Application of deep learning-based computer-aided detection system: detecting pneumothorax on chest radiograph after biopsy. *Eur Radiol*. 2019;29(10):5341-5348. doi:10.1007/s00330-019-06130-x

19. 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

20. Nam JG, Kim M, Park J, et al. Development and validation of a deep learning algorithm detecting 10 common abnormalities on chest radiographs. *Eur Respir J*. 2021;57(5):57. doi:10.1183/13993003.03061-2020

21. Thurfjell E, Thurfjell MG, Egge E, Bjurstam N. Sensitivity and specificity of computer-assisted breast cancer detection in mammography screening. *Acta Radiol*. 1998;39(4):384-388. doi:10.1080/02841859809172450

22. Fenton JJ, Abraham L, Taplin SH, et al. Effectiveness of computer-aided detection in community mammography practice. *J Natl Cancer Inst*. 2011;105(15):1152-1161. doi:10.1093/jnci/djr206

23. Meyl TP, de Bucourt M, Berghöfer A, et al. Subspecialization in radiology: effects on the diagnostic spectrum of radiologists and report turnaround time in a Swiss university hospital. *Radiol Med*. 2019;124(9):860-869. doi:10.1007/s11547-019-01039-3

24. Eng J, Mysko WK, Weller GER, et al. Interpretation of emergency department radiographs: a comparison of emergency medicine physicians with radiologists, residents with faculty, and film with digital display. *AJR Am J Roentgenol*. 2000;175(5):1233-1238. doi:10.2214/ajr.175.5.1751233

25. Singh SP, Gierada DS, Pinsky P, et al. Reader variability in identifying pulmonary nodules on chest radiographs from the national lung screening trial. *J Thorac Imaging*. 2012;27(4):249-254. doi:10.1097/RTI.0b013e318256951e

26. MIT Laboratory for Computational Physiology. Medical Information Mart for Intensive Care. Accessed April 2021. https://mimic-cxr.mit.edu

27. Homayounieh F, Digumarthy S, Ebrahimian S, et al. An artificial intelligence–based chest x-ray model on human nodule detection accuracy from a multicenter study. *JAMA Netw Open*. 2021;4(12):e2141096. doi:10.1001/jamanetworkopen.2021.41096

28. Ueda D, Yamamoto A, Shimazaki A, et al. Artificial intelligence-supported lung cancer detection by multi-institutional readers with multi-vendor chest radiographs: a retrospective clinical validation study. *BMC Cancer*. 2021;21(1):1120. doi:10.1186/s12885-021-08847-9

29. Hong W, Hwang EJ, Lee JH, Park J, Goo JM, Park CM. Deep learning for detecting pneumothorax on chest radiographs after needle biopsy: clinical implementation. *Radiology*. 2022;303(2):433-441. doi:10.1148/radiol.211706

30. Zhou L, Yin X, Zhang T, et al. Detection and semiquantitative analysis of cardiomegaly, pneumothorax, and pleural effusion on chest radiographs. *Radiol Artif Intell*. 2021;3(4):e200172. doi:10.1148/ryai.2021200172

31. Ebrahimian S, Homayounieh F, Rockenbach MABC, et al. Artificial intelligence matches subjective severity assessment of pneumonia for prediction of patient outcome and need for mechanical ventilation: a cohort study. *Sci Rep*. 2021;11(1):858. doi:10.1038/s41598-020-79470-0

32. Homayounieh F, Digumarthy SR, Febbo JA, et al. Comparison of baseline, bone-subtracted, and enhanced chest radiographs for detection of pneumothorax. *Can Assoc Radiol J*. 2021;72(3):519-524. doi:10.1177/0846537120908852

33. 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

34. Häberle L, Wagner F, Fasching PA, et al. Characterizing mammographic images by using generic texture features. *Breast Cancer Res*. 2012;14(2):R59. doi:10.1186/bcr3163

35. Qure.ai Technologies. qXR: AI for chest x-rays. Accessed February 4, 2022. https://qure.ai/product/qxr/

36. Annalise-AI. Comprehensive medical imaging AI solutions. Accessed February 4, 2022. https://annalise.ai/

37. Singh R, Kalra MK, Nitiwarangkul C, et al. Deep learning in chest radiography: detection of findings and presence of change. *PLoS One*. 2018;13(10):e0204155. doi:10.1371/journal.pone.0204155

38. He K, Zhang X, Ren S, Sun J. Deep residual learning for image recognition. In: 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). IEEE; 2016:770-778. Accessed August 1, 2022. https://www.computer.org/csdl/proceedings-article/cvpr/2016/8851a770/12OmNxvwoXv

39. Cubuk ED, Zoph B, Mane D, Vasudevan V, Le QV. AutoAugment: learning augmentation policies from data. Computer Vision Foundation. 2018. Accessed August 1, 2022. https://openaccess.thecvf.com/content_CVPR_2019/papers/Cubuk_AutoAugment_Learning_Augmentation_Strategies_From_Data_CVPR_2019_paper.pdf

40. Kim M, Park J, Na S, Park CM, Yoo D. Learning visual context by comparison. arXiv. Posted online July 15, 2020. Accessed August 1, 2022. https://arxiv.org/abs/2007.07506

41. Caruana R. Multitask learning. *Machine Learning*. 1997;28:41-75. doi:10.1023/A:1007379606734

SUPPLEMENT.

eAppendix. AI Algorithm
eFigure 1. Radiology Report Template
eFigure 2. Screenshots of Structured Radiology Report Tool With the AI-Aided (Top Image) and Unaided CXR Display
eTable 1. Summary of Reader Performance for Detection of the Four Target Findings in CXRs With and Without Nontarget Findings
eTable 2. Interpretation Times of Individual Readers Without and With AI-Aided Interpretation of CXRs
eFigure 3. Frontal Chest Radiographs Belonging to Four Separate Patients