European Radiology (2021) 31:9664–9674
https://doi.org/10.1007/s00330-021-08074-7

AI-based improvement in lung cancer detection on chest radiographs: results of a multi-reader study in NLST dataset

Hyunsuk Yoo[1] · Sang Hyup Lee[1] · Chiara Daniela Arru[2,3] · Ruhani Doda Khera[2,3] · Ramandeep Singh[2,3] · Sean Siebert[2,3] · Dohoon Kim[4] · Yuna Lee[4] · Ju Hyun Park[5] · Hye Joung Eom[6] · Subba R. Digumarthy[2,3] · Mannudeep K. Kalra[2,3]

Received: 2 February 2021 / Revised: 20 March 2021 / Accepted: 17 May 2021 / Published online: 4 June 2021

Abstract

Objective Assess if deep learning–based artificial intelligence (AI) algorithm improves reader performance for lung cancer detection on chest X-rays (CXRs).

Methods This reader study included 173 images from cancer-positive patients (n = 98) and 346 images from cancer-negative patients (n = 196) selected from National Lung Screening Trial (NLST). Eight readers, including three radiology residents, and five board-certified radiologists, participated in the observer performance test. AI algorithm provided image-level probability of pulmonary nodule or mass on CXRs and a heatmap of detected lesions. Reader performance was compared with AUC, sensitivity, specificity, false-positives per image (FPPI), and rates of chest CT recommendations.

Results With AI, the average sensitivity of readers for the detection of visible lung cancer increased for residents, but was similar for radiologists compared to that without AI (0.61 [95% CI, 0.55–0.67] vs. 0.72 [95% CI, 0.66–0.77], $p = 0.016$ for residents, and 0.76 [95% CI, 0.72–0.81] vs. 0.76 [95% CI, 0.72–0.81, $p = 1.00$ for radiologists), while false-positive findings per image (FPPI) was similar for residents, but decreased for radiologists (0.15 [95% CI, 0.11–0.18] vs. 0.12 [95% CI, 0.09–0.16], $p = 0.13$ for residents, and 0.24 [95% CI, 0.20–0.29] vs. 0.17 [95% CI, 0.13–0.20], $p < 0.001$ for radiologists). With AI, the average rate of chest CT recommendation in patients positive for visible cancer increased for residents, but was similar for radiologists (54.7% [95% CI, 48.2–61.2%] vs. 70.2% [95% CI, 64.2–76.2%], $p < 0.001$ for residents and 72.5% [95% CI, 68.0–77.1%] vs. 73.9% [95% CI, 69.4–78.3%], $p = 0.68$ for radiologists), while that in cancer-negative patients was similar for residents, but decreased for radiologists (11.2% [95% CI, 9.6–13.1%] vs. 9.8% [95% CI, 8.0–11.6%], $p = 0.32$ for residents and 16.4% [95% CI, 14.7–18.2%] vs. 11.7% [95% CI, 10.2–13.3%], $p < 0.001$ for radiologists).

Conclusions AI algorithm can enhance the performance of readers for the detection of lung cancers on chest radiographs when used as second reader.

Key Points

- *Reader study in the NLST dataset shows that AI algorithm had sensitivity benefit for residents and specificity benefit for radiologists for the detection of visible lung cancer.*
- *With AI, radiology residents were able to recommend more chest CT examinations (54.7% vs 70.2%, p < 0.001) for patients with visible lung cancer.*
- *With AI, radiologists recommended significantly less proportion of unnecessary chest CT examinations (16.4% vs. 11.7%, p < 0.001) in cancer-negative patients.*

✉ Mannudeep K. Kalra
mkalra@mgh.harvard.edu

[1] Lunit, Seoul, Korea

[2] Division of Thoracic Imaging, Department of Radiology, Massachusetts General Hospital, 75 Blossom Court, Boston, MA 02114, USA

[3] Harvard Medical School, Boston, MA, USA

[4] Department of Radiology, Seoul National University College of Medicine, Seoul, Korea

[5] Suwon Total Healthcare Center, Kangbuk Samsung Hospital, Sungkyunkwan University School of Medicine, Youngin-si, Gyeongi-do 16954, Korea

[6] Cheju Halla General Hospital, 65 Doryeong-ro, Yeon-dong, Jeju-si, Jeju-do, Korea

Keywords Radiography · Lung cancer · Mass screening · Artificial intelligence · Computer-assisted radiographic image interpretation

Abbreviations

ACRIN	American College of Radiology Imaging Network
AI	Artificial intelligence
AUC	Area under the ROC curve
CAD	Computer-aided diagnosis
CR	Computed radiograph
CXR	Chest X-ray
DICOM	Digital imaging and communications in medicine
DR	Digital radiograph
FPPI	False-positives per image
GEE	Generalized estimating equations
LDCT	Low-dose CT
NLST	National Lung Screening Trial
ROC	Receiver operating characteristic

Introduction

Lower cost, ease of acquisition, portability, and wider accessibility make chest radiography the most commonly used imaging test for initial workup of thoracic diseases [1]. However, the projectional nature of chest X-rays (CXRs) makes the detection of pulmonary nodules a difficult task as nodules may be obscured by anatomical structures [2], causing lesions located in the blind spots to be missed by radiologists [3, 4]. Results from the National Lung Screening Trial (NLST) show that screening with CXRs does not lower the mortality of patients due to low sensitivity of CXR for nodule detection as compared to low-dose CT (LDCT) [5, 6]. Nonetheless, the sheer prevalence of CXRs in modern medicine implies that with improved sensitivity, CXRs can play an important role in detection of lung cancers presenting as incidental pulmonary nodules [7].

Studies have demonstrated that artificial intelligence (AI) algorithms improve the performance of radiologists for the detection of lung cancer in CXRs [8–11]. Jang et al reported that AI helps observers detect overlooked lung cancers that were either missed or detected with misinterpretation on prior CXRs [11]. The authors suggested that AI may help observers reduce the number of overlooked cancer [11]. However, these algorithms are yet being adapted to clinical practice due to concerns that, like conventional computer-aided diagnosis (CAD) systems, AI CAD systems may decrease the specificity of the readers and lead to an increase in detection of false-positive and/or benign nodules which trigger further workup with CT and/or invasive tissue biopsies [12, 13]. Therefore, for clinical implementation of AI algorithms, it is important to demonstrate that these systems improve malignant nodule detection without increasing the number of false-positive findings.

In our previous study, we validated the performance of an AI algorithm for the detection of malignant pulmonary nodules in the NLST data set [9]. Our previous study suggested that AI can help improve lung cancer detection on CXRs, but did not assess the performance improvement of blinded readers with AI-aided interpretation. In this study, we present the results of an observer study of eight readers, including three radiology residents, and five radiologists, in the NLST data set. The goal of this study was to assess whether an AI algorithm improves the reader performance for lung cancer detection without increasing unnecessary false-positive findings on CXRs.

Methods

Ethics review and approval were obtained from the institutional review board (IRB) of Massachusetts General Hospital. IRB approval was a required step to sign the data use agreement for access and use of NLST data. The need for informed consent was waived because our retrospective reader study used previously acquired data from other clinical trial.

Study population

A total of 519 screening CXRs from 294 patients were retrospectively selected from NLST, a multicenter randomized clinical trial comparing low-dose CT (LDCT) with CXRs for screening high-risk population for lung cancer [6, 14]. The trial enrolled 53,454 participants at 33 screening centers in the USA from August 2002 through April 2004. Participants were randomized to three annual screens (at T0, T1, and T2) with either LDCT or CXRs [14]. Among patients enrolled through American College of Radiology Imaging Network (ACRIN), 5491 participants were within 83% random sample and had available screening CXRs. One hundred seventy-three CXRs from 98 participants with diagnosis of lung cancer during screening examinations or within 1 year of the final screening examination were selected as cancer-positive subgroup. Three hundred forty-six CXRs from 196 cancer-negative patients were then consecutively sampled as cancer-negative subgroup based on the following criteria: the distribution of patients with just one CXR (T0), those with two CXRs (T0, T1), and those with all three CXRs (T0, T1, T2) in the cancer-negative subgroup was identical to the cancer-positive subgroup (Fig. 1).

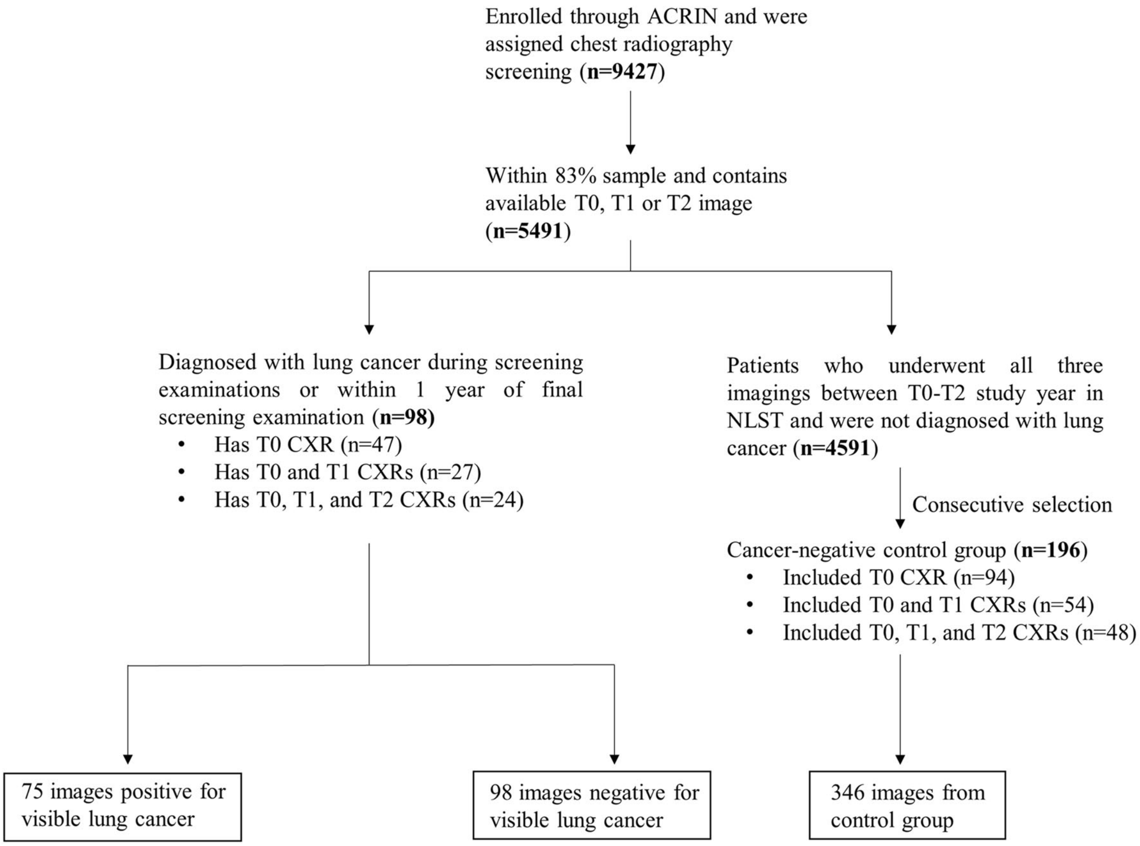
Enrolled through ACRIN and were assigned chest radiography screening (n=9427)
Within 83% sample and contains available T0, T1 or T2 image (n=5491)
Diagnosed with lung cancer during screening examinations or within 1 year of final screening examination (n=98)
• Has T0 CXR (n=47)
• Has T0 and T1 CXRs (n=27)
• Has T0, T1, and T2 CXRs (n=24)
Patients who underwent all three imagings between T0-T2 study year in NLST and were not diagnosed with lung cancer (n=4591)
Consecutive selection
Cancer-negative control group (n=196)
• Included T0 CXR (n=94)
• Included T0 and T1 CXRs (n=54)
• Included T0, T1, and T2 CXRs (n=48)
75 images positive for visible lung cancer
98 images negative for visible lung cancer
346 images from control group

Fig. 1 Flow chart summarized selection of CXRs for observer performance testing. There were 75 CXRs from 68 patients with visible lung cancers. Three hundred forty-six CXRs from 196 cancer-negative patients were used as the control group

AI algorithm

We used a commercially available AI algorithm (Lunit INSIGHT for Chest Radiography; version 2.4.11.0; Lunit Inc.), trained with 12408 abnormal CXRs with lung nodules or masses, annotated by at least 1 of 15 board-certified radiologists, and 72704 normal CXRs [9]. None of the NLST data was used in the training of the AI algorithm. The algorithm is a ResNet34-based deep convolutional neural network with a self-attention mechanism to generate a more distinguishable image representation [15]. The model takes Digital Imaging and Communications in Medicine (DICOM) file as input and produces a probability map and an abnormality score, ranging between 0 and 100, for 10 common abnormalities in CXRs: atelectasis, calcification, cardiomegaly, consolidation, fibrosis, mediastinal widening, nodule or mass, pleural effusion, pneumoperitoneum, and pneumothorax. Please refer to examples of lung nodules and/or masses with different scores in Supplemental Figure S1. The model does not produce output map when the abnormality score is below 15.0, the operating point chosen using Youden criteria in the internal validation set. Although we only used output map and score corresponding to pulmonary nodules in this study, the model was still trained using multi-task learning scheme, in which binary cross-entropy loss computed for each of the 10 abnormalities updated the parameters during model training to improve generalizability [9, 16]. None of the NLST data was included in the training set. A more detailed description of the development of the AI algorithm can be found in previous studies [9, 17].

Establishing the ground truth

Two senior radiologists (S.R.D. and M.K.K., with 16 and 13 years of experience in thoracic radiology, respectively) independently annotated all 519 CXRs included in the test set for the presence of lesion(s) suspicious for lung cancer as the ground truth radiologists. For CXRs with suspicious for lung cancers, the ground truth radiologists drew a closed contour around the suspicious lesions. The suspicious lesions were defined as pulmonary or pleural nodules, masses, opacities, and hilar lymphadenopathy, which were concerning for lung cancer either on the CXRs. Characteristics such as size, irregular or spiculated margins, and growth over time on serial CXRs were considered suspicious features for cancer. Lesions with typical benign features (dense or popcorn calcification) and/or stability over serial CXRs were considered benign. To improve their accuracy, these radiologists referred to each patient's cancer characteristics as well as available sequential CXRs during the radiologic evaluation. Disagreements over annotations were resolved with consensual review of CXRs.

A CXR was considered positive for visible lung cancer if the location of the suspicious lesion marked by the ground truth radiologists matched the location described in the histopathology report. CXR lesions not corresponding to the location of lung cancer described in the pathology report were deemed as negative for visible lung cancer. Patients were deemed to have visible lung cancer on prior CXRs if their radiograph taken 1 year before the final screening CXR was positive for visible lung cancer. For patients with multifocal lung cancer (n = 5 patients), all suspicious lesions (n = 2 per patient) were annotated.

Design of observer performance test

Three 2nd year radiology residents (one US resident and two Korean residents) and five residency-trained radiologists (two US radiologists and three Korean radiologists, respective experience of 5, 4, 5, 14, and 9 years) participated as independent and blinded test radiologists. First year residents were not included since they undergo chest radiography rotations at different timepoints of the year. Second year residents were selected to ensure that they all had completed one supervised clinical rotation in chest radiography and had similar length of experience/exposure in interpretation of chest radiographs. Availability of the third and fourth year residents was limited at the time of ongoing pandemic when several radiology residents were called to serve in overflowing inpatient services. The two ground truth chest radiologists (S.R.D. and M.K.K.) did not participate as test radiologists.

Prior to the observer performance test, each test radiologist was instructed to mark lesions (per CXR) for any lesions suspicious for cancer, not limited to pulmonary nodules or masses (such as hilar lymphadenopathy, and pleural nodule or thickening), while ignoring benign findings such as calcification, subsegmental atelectasis, and linear scars. For CXRs with multiple concerning findings, we instructed each reader to annotate and score the two most suspicious lesions for lung cancer. Each reader annotated ten training CXRs to enhance their understanding of the study objectives before beginning the observer performance test. Test radiologists could refer to previous CXRs when available for comparison. The serial CXRs were presented in chronologic order, and the readers could not refer to future radiographs during the annotation. On CXRs with suspicious lung cancer, each independent reader drew a separate closed contour around the lesion and specified a confidence rating from 1 (confidence level 0–20%) to 5 (confidence level 80–100%) for up to two suspicious lesions per CXR. All contours were considered when estimating FPPIs. The normalized confidence rating for the readers was calculated by dividing the readers' confidence rating by the highest possible confidence rating. The readers also specified need for a chest CT examination for workup of suspicious lesions.

Each reader reviewed each CXR twice: first, without AI and then with AI with at least 4 weeks of wash-out within the two reviews. When reviewing images with AI, the readers were able to toggle between original CXR and that with the AI heatmap. All the readers who participated in the observer performance test evaluated the whole test set. The reader study was conducted between January 1, 2020, through March 31, 2020. A screen-capture of the web-based observer performance test tool is shown in Supplemental Figure S2.

Statistical analysis

Differences in the selected characteristics of the study population were compared between cancer-positive and cancer-negative patients using Student's t test for continuous variables (age, follow-up period, and mortality) and chi-square test for categorical variables (sex, race, and ethnicity). The confidence rating of the readers was compared with Student's t test. To evaluate the performance of readers with and without AI for lung cancer detection, receiver operating characteristic (ROC) analyses were performed. Comparison of individual observer-level and average area under the ROC curve (AUC) was made using DeLong's method and Hillis' method, respectively [18, 19]. Sensitivity and specificity of readers with and without AI were calculated using threshold confidence level of 15% as the operating point, and were compared with generalized estimating equations (GEEs) [11, 20]. False-positive markings per image (FPPI) were defined as total number of false-positive markings divided by the total number of CXRs, and was compared by Poisson regression [10]. Rates of CT recommendations and detection rates of missed lung cancer visible on prior CXRs were assessed with GEE [20]. The inter-reader agreements for lung cancer detection were assessed with weighted κ using linear weighting [21]. Kappa results were interpreted as follows: values ≤ 0 as indicating no agreement and 0.01–0.20 as none to slight, 0.21–0.40 as fair, 0.41–0.60 as moderate, 0.61–0.80 as substantial, and 0.81–1.00 as almost perfect agreement. For all tests, $p < 0.05$ was considered statistically significant. All statistical analyses were conducted using R software, version 3.6.1 (R Foundation for Statistical Computing).

Results

Patient characteristics

The differences in mean age, gender, race distribution, ethnicity, and smoking status distribution between the cancer-positive and cancer-negative groups were statistically non-significant. Compared to that of the cancer-positive patients, the

Table 1 Baseline demographic characteristics of cancer-positive and cancer-negative patients for the observer performance test. Data are presented as no./tot no. (%) of patients, unless otherwise indicated.

Characteristics	Total	Cancer-positive	Cancer-negative	*p* value
N	294 (100.0)	98 (33.3)	196 (66.7)	
Number of screening examinations				
One (T0)	141 (48.0)	47 (48.0)	94 (48.0)	
Two (T0, T1)	81 (27.6)	27 (27.6)	54 (27.6)	
Three (T0, T1, and T2)	72 (24.5)	24 (24.5)	48 (24.5)	
Age, mean (SD)	62.6 (5.1)	62.8 (5.0)	62.4 (5.2)	0.49
Sex				
Male	164 (55.8)	59 (60.2)	105 (53.6)	0.28
Female	130 (44.2)	39 (39.8)	91 (46.4)	
Race				
White	276 (93.9)	93 (94.9)	183 (93.4)	0.51
Black or African American	11 (3.7)	5 (5.1)	6 (3.1)	
Asian	1 (0.3)	0 (0.0)	1 (0.5)	
American Indian or Alaskan Native	1 (0.3)	0 (0.0)	1 (0.5)	
Native Hawaiian or other Pacific Islander	0 (0.0)	0 (0.0)	0 (0.0)	
> 1 race	4 (1.4)	0 (0.0)	4 (2.0)	
Unavailable	1 (0.3)	0 (0.0)	1 (0.5)	
Ethnicity				
Hispanic or Latino	3 (1.0)	1 (1.0)	2 (1.0)	1.00
Not Hispanic or Latino	291 (99.0)	97 (99.0)	194 (99.0)	
Unavailable	0 (0.0)	0 (0.0)	0 (0.0)	
Smoking status				
Former	140 (47.6)	44 (44.9)	96 (49.0)	0.51
Current	154 (52.4)	54 (55.1)	100 (51.0)	
Outcomes				
Follow-up, median (IQR), year	6.3 (5.5–7.2)	3.7 (0.0–7.9)	6.5 (5.9–7.1)	< 0.001
Mortality	61 (20.7)	59 (60.2)	2 (1.0)	< 0.001

median time from T0 screen to last follow-up date was greater for cancer-negative patients (3.7 (0.0–7.9) vs. 6.5 (5.9–7.1), $p < 0.001$). Patients in the cancer-positive group had significantly higher mortality compared to patients in the cancer-negative group (60.2% vs. 1.0%, $p < 0.001$). The demographic characteristics are summarized in Table 1.

Observer performance assessment for visible lung cancer detection

Among 98 CXRs selected from cancer-positive patients, 23 CXRs in which there were no visible lesions suggestive of lung cancer were excluded in our primary analysis, and 75 CXRs labeled by the ground truth radiologists as positive for visible lung cancer were selected as case group (17.8%, 74/421). Three hundred forty-six CXRs from cancer-negative patients were used as the control group (82.2%, 346/421). The distribution of the confidence rating and the total and per CXR positive markings for each reader is shown in Supplemental Table S1, and Supplemental Table S2, respectively.

The performance of readers for the detection of visible lung cancer detection is summarized in Table 2. Compared to that without AI, the average AUC for the detection of visible lung cancer increased significantly for radiology residents with AI (0.76 [95% CI, 0.67–0.86] vs. 0.82 [95% CI, 0.75–0.89], $p = 0.003$), but for radiologists, the average AUC was similar (0.82 [95% CI, 0.74–0.91] vs. 0.84 [95% CI, 0.79–0.89], $p = 0.24$). Compared to that without AI, the average sensitivity increased significantly for radiology residents (0.61 [95% CI, 0.55–0.67] vs. 0.72 [95% CI, 0.66–0.77], $p = 0.016$), but specificity was similar with AI (0.88 [95% CI, 0.86–0.90] vs. 0.88 [95% CI, 0.86–0.90], $p = 0.89$). For radiologists, average sensitivity (0.76 [95% CI, 0.72–0.81] vs. 0.76 [95% CI, 0.72–0.81], $p = 1.00$) was similar, but specificity increased with AI (0.79 [95% CI, 0.77–0.81] vs. 0.86 [95% CI, 0.84–0.87], $p < 0.001$). Average FPPI without and with AI was similar for radiology residents (0.15 [95% CI, 0.11–0.18] vs. 0.12 [95% CI, 0.09–0.16], $p = 0.13$), but was significantly lower with AI for radiologists (0.24 [95% CI, 0.20–0.29] vs. 0.17 [95% CI, 0.13–0.20], $p < 0.001$). The performance of the readers for the detection of all lung cancer, when no exclusion is applied, is presented in Supplemental Table S3.

Table 2 The performance of readers for the detection of visible lung cancer in 421 CXRs. CXRs of cancer-positive patients without visible lesions are excluded. *FPPI* false-positives per image

Group	AUC			Sensitivity			Specificity			FPPI		
	Without AI	With AI	*p* value	Without AI	With AI	*p* value	Without AI	With AI	*p* value	Without AI	With AI	*p* value
Radiology residents												
1	0.75 (0.69–0.81)	0.85 (0.79–0.90)	< 0.001	0.56 (0.45–0.67)	0.73 (0.63–0.83)	0.024	0.90 (0.87–0.94)	0.92 (0.89–0.95)	0.42	0.09 (0.06–0.12)	0.08 (0.05–0.11)	0.56
2	0.74 (0.67–0.80)	0.79 (0.73–0.85)	0.06	0.57 (0.46–0.69)	0.67 (0.56–0.77)	0.24	0.86 (0.83–0.90)	0.86 (0.83–0.90)	1.00	0.18 (0.14–0.22)	0.15 (0.11–0.19)	0.31
3	0.81 (0.75–0.87)	0.83 (0.78–0.89)	0.38	0.69 (0.59–0.80)	0.75 (0.65–0.85)	0.47	0.86 (0.82–0.90)	0.85 (0.81–0.89)	0.67	0.17 (0.13–0.21)	0.14 (0.10–0.18)	0.33
Average	0.76 (0.67–0.86)	0.82 (0.75–0.89)	0.003	0.61 (0.55–0.67)	0.72 (0.66–0.77)	0.016	0.88 (0.86–0.90)	0.88 (0.86–0.90)	0.89	0.15 (0.11–0.18)	0.12 (0.09–0.16)	0.13
Radiologists												
1	0.90 (0.85–0.94)	0.89 (0.85–0.94)	0.79	0.89 (0.82–0.96)	0.84 (0.76–0.92)	0.34	0.74 (0.70–0.79)	0.85 (0.81–0.88)	<0.001	0.27 (0.23–0.33)	0.18 (0.14–0.22)	0.003
2	0.72 (0.65–0.78)	0.78 (0.72–0.84)	0.02	0.67 (0.56–0.77)	0.71 (0.60–0.81)	0.6	0.66 (0.61–0.71)	0.82 (0.78–0.86)	<0.001	0.46 (0.39–0.52)	0.22 (0.18–0.27)	< 0.001
3	0.82 (0.77–0.88)	0.82 (0.76–0.88)	0.91	0.79 (0.69–0.88)	0.80 (0.71–0.89)	0.84	0.73 (0.69–0.78)	0.73 (0.68–0.77)	0.80	0.29 (0.23–0.34)	0.31 (0.26–0.36)	0.53
4	0.82 (0.77–0.88)	0.86 (0.81–0.91)	0.09	0.68 (0.57–0.79)	0.75 (0.65–0.85)	0.37	0.94 (0.91–0.96)	0.96 (0.94–0.98)	0.24	0.06 (0.04–0.09)	0.06 (0.04–0.08)	0.89
5	0.86 (0.81–0.91)	0.84 (0.79–0.89)	0.48	0.79 (0.69–0.88)	0.72 (0.62–0.82)	0.34	0.87 (0.84–0.91)	0.93 (0.90–0.96)	0.010	0.14 (0.11–0.18)	0.06 (0.04–0.09)	< 0.001
Average	0.82 (0.74–0.91)	0.84 (0.79–0.89)	0.24	0.76 (0.72–0.81)	0.76 (0.72–0.81)	1.00	0.79 (0.77–0.81)	0.86 (0.84–0.87)	<0.001	0.24 (0.20–0.29)	0.17 (0.13–0.20)	< 0.001

The inter-reader agreements for lung cancer detection without and with AI are presented in Supplemental Table S4. There was fair to moderate interobserver agreement between three residents without AI, and consistently moderate agreement with AI-assisted interpretation. Likewise, radiologist improved from fair or moderate (5/10 radiologist pairwise comparison in each) interobserver agreement for CXR interpretation without AI to moderate (9/10 radiologists pairwise comparison) or good (1/10) agreement with AI-assisted interpretation.

Table 3 summarizes the percentages of chest CT recommendation for patients with and without visible lung cancer. For patients with visible lung cancer on CXR, the average chest CT recommendation rate increased significantly for residents, but was similar for radiologists without and with AI (54.7% [95% CI, 48.2–61.2%] vs. 70.2% [95% CI, 64.2–76.2%], $p < 0.001$ for residents and 72.5% [95% CI, 68.0–77.1%] vs. 73.9% [95% CI, 69.4–78.3%], $p = 0.68$ for radiologists). Conversely, in patients without visible lung cancer, the average chest CT recommendation rate was similar without and with AI for residents, but decreased for radiologists (11.2% [95% CI, 9.6–13.1%] vs. 9.8% [95% CI, 8.0–11.6%], $p = 0.32$ for residents and 16.4% [95% CI, 14.7–18.2%] vs. 11.7% [95% CI, 10.2–13.3%], $p < 0.001$ for radiologists).

Detection rate for lung cancer visible in previous chest radiographs

Among 98 cancer-positive patients, 51 patients had two or more CXRs, and 7 patients had visible lung cancers on prior CXRs. In this study, these 7 CXRs were regarded as having missed lung cancers. Of these 7 missed lung cancers, average residents detected significantly more lung cancer with AI than without AI (39% [2.7 of 7] vs. 71% [5.0 of 7], $p = 0.021$), but such gain was not seen in radiologists (57% [4.0 of 7] vs. 51% [3.6 of 7], $p = 0.63$). Similarly, average residents recommended significantly more chest CT examination for these CXRs with AI than without AI (33% [2.3 of 7] vs. 71% [5.0 of 7], $p = 0.008$), but the recommendation rate was similar for radiologists (51% [3.6 of 7] vs. 49% [3.4 of 7], $p = 0.81$). The performance of individual readers for the detection of missed lung cancers is shown in Table 4.

Discussion

This study assessed how an AI algorithm benefits reader for detecting visible lung cancer in CXRs. When AI was used as a second reader, residents detected more visible lung cancer (0.61 vs. 0.72, $p = 0.016$), and were able to detect more missed lung cancer present in prior CXRs (39% vs. 71%, $p = 0.021$). In comparison, radiologists had higher specificity (0.79 vs. 0.86, $p < 0.001$) and lower FPPI (0.24 vs. 0.17, $p < 0.001$)

Table 3 The percentages of chest CT recommendation in 75 CXRs (n = 68) positive for visible lung cancer and 346 CXRs (n = 196) selected from the cancer-negative control group

Group	75 CXRs from patients positive for visible lung cancer (n = 68)			346 CXRs from cancer-negative patients (n = 196)		
	Without AI	With AI	*p* value	Without AI	With AI	*p* value
Radiology residents						
1	45.3 (34.1–56.6)	62.7 (51.7–73.6)	0.031	8.7 (5.7–11.6)	5.5 (3.1–7.9)	0.10
2	57.3 (46.1–68.5)	74.7 (64.8–84.5)	0.022	16.5 (12.6–20.4)	14.2 (10.5–17.8)	0.40
3	61.3 (50.3–72.4)	73.3 (63.3–83.3)	0.11	8.4 (5.5–11.3)	9.8 (6.7–13.0)	0.51
Average	54.7 (48.2–61.2)	70.2 (64.2–76.2)	< 0.001	11.2 (9.3–13.1)	9.8 (8.0–11.6)	0.32
Radiologists						
1	82.7 (74.1–91.2)	81.3 (72.5–90.2)	0.83	16.8 (12.8–20.7)	8.4 (5.5–11.3)	< 0.001
2	65.3 (54.6–76.1)	69.3 (58.9–79.8)	0.60	33.5 (28.6–38.5)	17.9 (13.9–22.0)	< 0.001
3	68.0 (57.4–78.6)	73.3 (63.3–83.3)	0.47	13.9 (10.2–17.5)	21.1 (16.8–25.4)	0.012
4	68.0 (57.4–78.6)	74.7 (64.8–84.5)	0.37	6.4 (3.8–8.9)	4.3 (2.2–6.5)	0.24
5	78.7 (69.4–87.9)	70.7 (60.4–81.0)	0.26	11.6 (8.2–14.9)	6.9 (4.3–9.6)	0.04
Average	72.5 (68.0–77.1)	73.9 (69.4–78.3)	0.68	16.4 (14.7–18.2)	11.7 (10.2–13.3)	< 0.001

with AI. Improved inter-reader agreement for both residents and radiologists on AI-assisted interpretation was likely related to improved reader confidence in “calling” lesions as present or absent with AI assistance rather than without AI on an otherwise highly subjective interpretation of projection radiographs.

Our results suggest that AI algorithm benefits less-experienced readers in terms of sensitivity, and more-experienced readers in terms of specificity. Previous studies suggested that less-experienced readers are prone to missing lung cancers, especially for lesions that have low visibility score, and those that are overlapping with anatomical structures [11]. In this study, the AI algorithm likely reduced such errors by locating such subtle lesions for the residents [8, 11]. As shown in other studies, AI helped improve the performance of the residents up to the level that was on par with radiologists, which led to a significant increase in the chest CT recommendation rate for patients with visible lung cancer [8, 11, 17]. In comparison, such benefit may have not been realized for the more-experienced readers in our study who detected visible lung cancers without AI.

For the radiologists in our study, AI improved specificity and reduced false-positive nodules without decreasing their sensitivity. As a result, radiologists recommended fewer chest CT examinations for patients without visible lung cancer, while maintaining a similar chest CT recommendation rate

Table 4 The detection rate and chest CT recommendation rate for lung cancers visible in previous chest radiographs

Group	Detection rate (image-level)			Detection rate (lesion-level)			Chest CT recommendation rate		
	Without AI	With AI	*p* value	Without AI	With AI	*p* value	Without AI	With AI	*p* value
Radiology residents									
1	57 [4/7]	71 [5/7]	0.57	57 [4/7]	71 [5/7]	0.57	43 [3/7]	57 [4/7]	0.59
2	29 [2/7]	71 [5/7]	0.08	29 [2/7]	71 [5/7]	0.08	29 [2/7]	86 [6/7]	0.008
3	29 [2/7]	71 [5/7]	0.08	29 [2/7]	71 [5/7]	0.08	29 [2/7]	71 [5/7]	0.076
Average	39 [2.7/7]	71 [5.0/7]	0.021	39 [2.7/7]	71 [5.0/7]	0.021	33 [2.3/7]	71 [5/7]	0.008
Radiologists									
1	100 [7/7]	57 [4/7]	0.025	100 [7/7]	57 [4/7]	0.025	71 [5/7]	57 [4/7]	0.57
2	29 [2/7]	57 [4/7]	0.26	29 [2/7]	43 [3/7]	0.57	29 [2/7]	57 [4/7]	0.26
3	57 [4/7]	57 [4/7]	1.00	57 [4/7]	57 [4/7]	1.00	57 [4/7]	43 [3/7]	0.59
4	29 [2/7]	43 [3/7]	0.57	29 [2/7]	43 [3/7]	0.57	29 [2/7]	43 [3/7]	0.57
5	71 [5/7]	43 [3/7]	0.26	71 [5/7]	43 [3/7]	0.26	71 [5/7]	43 [3/7]	0.26
Average	57 [4.0/7]	51 [3.6/7]	0.63	57 [4.0/7]	49 [3.4/7]	0.47	51[3.6/7]	49 [3.4/7]	0.81

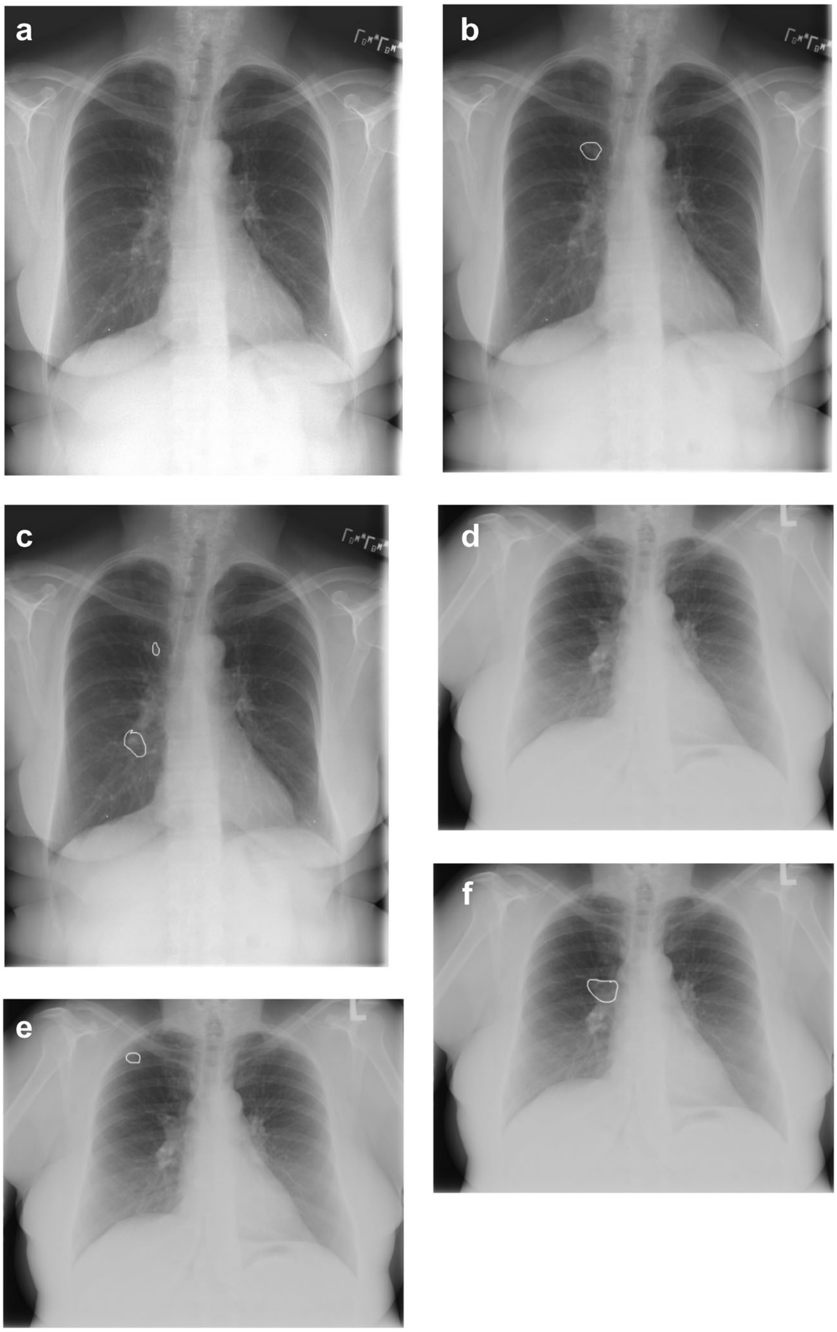

Fig. 2 CXRs of patients without visible lung cancer for which the AI helped reduce the false-positive annotation of the radiologists. **a** CXRs of a woman in her 50s, and (**d**) a man in his 50s who were negative for visible lung cancer. **b**, **c**, **e**, **f** Examples where radiologists initially had false-positive annotation without AI (drawn with a white circle), but correctly dismissed the annotations with AI. Because the abnormality score of the AI was less than the operating point, the AI did not display any heatmap

for patients positive for visible lung cancer. Without AI, radiologists tended to overcall ambiguous findings that were in fact benign findings, which led to unnecessary chest CT recommendations. However, with AI, radiologists were able to rule out such benign lesions while still ruling in positive findings (Fig. 2).

We believe our results support the use of AI CAD systems for the detection of lung cancer in CXRs. Prior studies with conventional CAD systems, which rely on hand-crafted features, document good sensitivity for lung nodule detection, but their application is limited by high false-positive outputs [12, 13, 22, 23]. High false-positive rates can trigger unnecessary chest CT examinations and patient anxiety [23]. Our AI CAD system can help avoid such issues while assisting readers in identifying subtle lesions that may otherwise be missed. This finding is consistent with other studies on AI CAD systems with markedly decreased false-positive rate, high specificity, and preserved excellent nodule detection performance [8, 10].

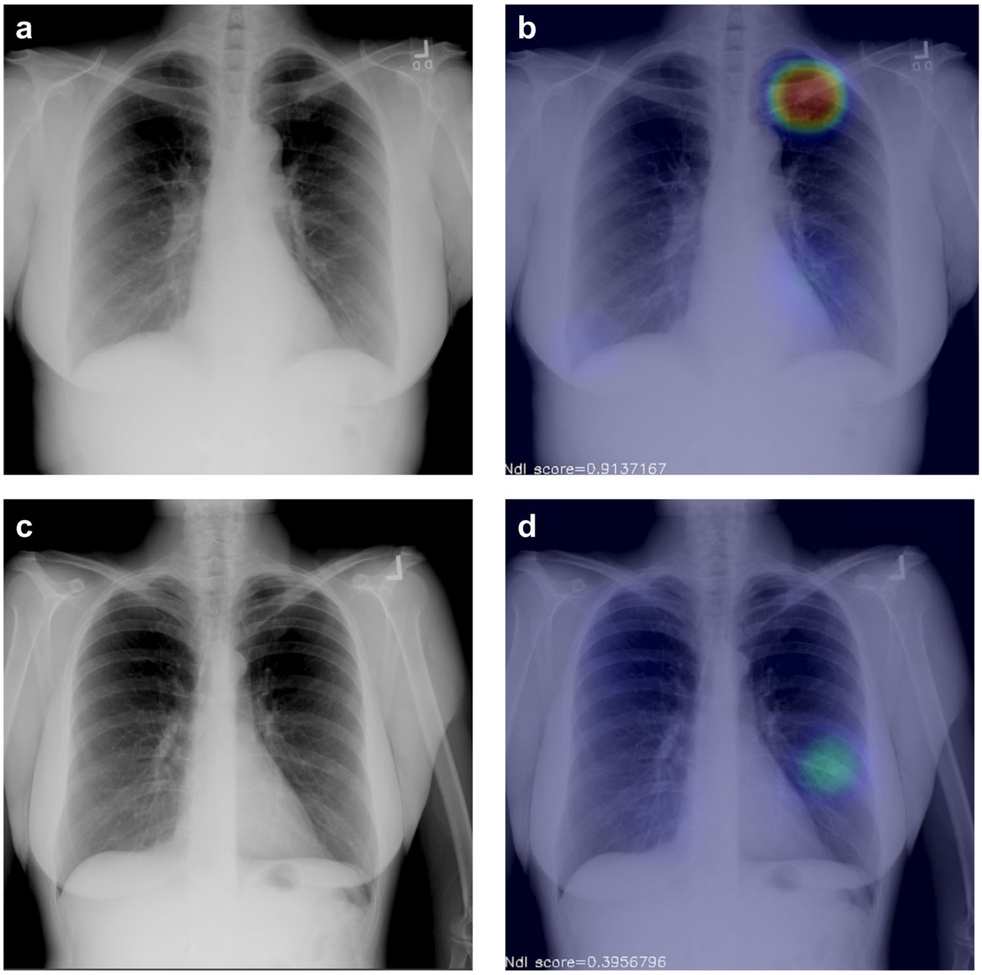

Fig. 3 Missed lung cancers on prior CXRs of two patients that were detected by the AI algorithm. The location of the lesion is marked with a red circle. **a** CXR of a woman in her 60s who was diagnosed with lung cancer 454 days after her baseline imaging. The AI algorithm (**b**) detected lung nodule in the left upper lung overlapping with the left clavicle. **c** CXR of a woman in her 50s who was diagnosed with lung cancer 747 days after her baseline CXR. **d** The AI algorithm detected a subtle lung nodule in the left lower lung that is overlapping with the rib

The relative advantage of AI CADs, especially those trained using deep learning–based algorithms, over conventional CAD systems in terms of improved specificity may be derived from large training datasets and "experiential learning" approach of AI CAD. Such learning enables AI CAD systems to map input image into a latent feature space which can help differentiate concerning findings from other structures and benign lesions [24]. During training, loss function penalizes false-positive as well as false-negative predictions, forcing the AI algorithm to learn representations of nodules that distinguishes them from those of non-target findings such as spurious lesions related to calcifications and vessels [25].

The increased detection of missed lung cancer in prior CXRs can enable earlier detection of lung cancer. AI detected five of the seven missed lung cancers (Fig. 3), but failed to detect the remaining two of the seven cancers (Fig. 4). These two cancers missed by the AI were present on conventional radiographs (CRs), the type of images for which the AI has been shown to have decreased performance as compared to digital radiographs (DRs) [9]. Since the five cancers detected with AI-aided interpretation were difficult cases that were mostly missed by residents without AI, AI increased the detection rate for radiology residents. Conversely, AI did not increase the detection rate of these lesions for the experienced radiologists who detected these cases without AI. In fact, for radiologists 1 and 5, AI led to false-negative interpretation of lung cancers that were reported as present without AI. Such result suggests that readers should be cautious about interpreting AI results, especially when AI is implemented on a setting with characteristics vastly different from the training set.

A strength of our study is evaluation of readers with AI in a data set selected from the NLST, a multicenter randomized clinical trial in which patients had a wide spectrum of abnormalities that may be encountered in the clinical practice [14]. In contrast to the previous studies that used normal CXRs as the control group, to simulate real-world practice and provide sufficient challenge to our AI algorithm, we did not

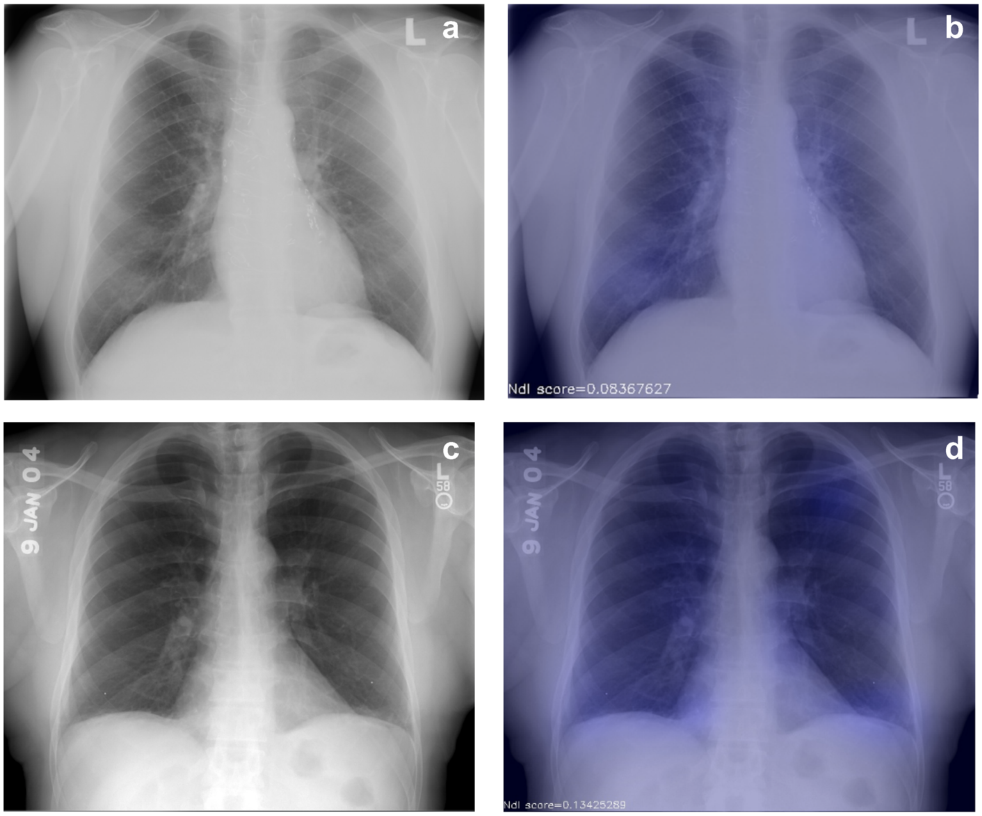

Fig. 4 Baseline CXRs of patients with missed lung cancer which were also missed by the AI algorithm. The location of the lesions is marked with a red circle. Both of these images were conventional radiographs (CR). **a** CXR of a male in his 60s who was diagnosed with lung cancer 412 days after his baseline imaging. **b** The AI algorithm missed the abnormality in the right lower lung. **c** CXR of a male in his 60s who was diagnosed with lung cancer 372 days after his baseline imaging. **d** The AI algorithm missed the subtle nodule in the left upper lung that is overlapping with the left clavicle

intentionally exclude any patients with other pathologies, such as calcified granuloma, consolidation, emphysema, and other thoracic diseases [8–10]. Also, all radiologists and residents in our study had access to prior screening CXRs during their evaluation, reflecting the actual interpretation workflow. Thus, our study has greater value in terms of applicability to the real-word settings and further establishes the generalizability of the results [8, 10].

There are several limitations to this study. First, although NLST is a community cohort of participants at high risk of lung cancer, the prevalence of lung cancer was low, and only 98 patients with lung cancer were included in our analysis. Of these 98 patients, 75 patients had visible cancer, and seven patients had missed lung cancer present in prior CXRs. Because of the small number of lung cancer patients, it was hard to achieve statistical significance, especially for assessing detection of missed lung cancer present in prior CXRs. Second, since we were unable to conduct observer performance test on entire NLST data set, we conducted an observer study data set consisting of 519 CXRs. In clinical practice, the prevalence of lung cancer encountered may be lower than that encountered in this reader study. Third, as pointed out in the previous study [9], AI had lower performance in the CR images, which led to the underperformance of the AI algorithm. The added value of AI may be greater when applied to CXRs acquired with modern equipment. Fourth, although two ground truth radiologists referred to all available clinical and pathology information during the annotation, the ground truth visibility labels were generated without paired CT images, so they may have been inaccurate.

In conclusion, the AI algorithm improved sensitivity and reduced false-positives for lung cancer detection for residents and radiologists, respectively. AI can help enhance the value of CXRs for detecting lung cancer by improving the quality of reading for various reader groups.

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/s00330-021-08074-7.

Funding The authors state that this work has not received any funding.

Declarations

Guarantor The scientific guarantor of this publication is Mannudeep K. Kalra.

Conflict of interest Mannudeep K. Kalra has received research grants from Siemens Healthineers and Riverain Technologies LLC for unrelated works. Subba R. Digumarthy reported receiving grants from Lunit during the conduct of the study and providing independent image analysis for hospital-contracted clinical trials for Abbvie, Bristol-Myers Squibb, Cascadian Therapeutics, Clinical Bay Laboratories, Gradalis, Merck, Novartis, Pfizer, Roche, Polaris Pharmaceuticals, and Zai Laboratories, and receiving honoraria from Siemens outside the submitted work. Hyunsuk Yoo and Sanghyup Lee are employees of Lunit, South Korea. No funding was received for this project.

Statistics and biometry Soo Young Kwak kindly provided statistical advice for this manuscript. Hyunsuk Yoo carried out actual statistical analysis of the data.

Informed consent Written informed consent was waived by the Institutional Review Board.

Ethical approval Ethics review and approval were obtained from the institutional review board of Massachusetts General Hospital.

Study subjects or cohorts overlap This study was carried out using randomly sampled data of patients enrolled through ACRIN arm of the public National Lung Screening Trial (NLST) dataset, which has been used in prior publications by the NLST investigators and others who have used the public dataset for various research purposes. We are the first to carry out a reader study comparing the performance of readers with and without AI using NLST dataset for the detection of visible lung cancer.

Methodology
- retrospective
- diagnostic or prognostic study
- multicenter study

References

1. McComb BL, Chung JH, Crabtree TD et al (2016) ACR Appropriateness Criteria® routine chest radiography. J Thorac Imaging 31:W13–W15. https://doi.org/10.1097/RTI.0000000000000200
2. de Groot PM, Carter BW, Abbott GF, Wu CC (2015) Pitfalls in chest radiographic interpretation: blind spots. Semin Roentgenol 50:197–209. https://doi.org/10.1053/j.ro.2015.01.008
3. Gavelli G, Giampalma E (2000) Sensitivity and specificity of chest X-ray screening for lung cancer. Cancer 89:2453–2456. https://doi.org/10.1038/sj.bjc.6604351
4. Austin JH, Romney BM, Goldsmith LS (1992) Missed bronchogenic carcinoma: radiographic findings in 27 patients with a potentially resectable lesion evident in retrospect. Radiology 182:115–122. https://doi.org/10.1148/radiology.182.1.1727272
5. Aberle DR, DeMello S, Berg CD et al (2013) Results of the two incidence screenings in the National Lung Screening Trial. N Engl J Med 369:920–931. https://doi.org/10.1056/NEJMoa1208962
6. Aberle DR, Adams AM, Berg CD et al (2011) Reduced lung-cancer mortality with low-dose computed tomographic screening. N Engl J Med 365:395–409. https://doi.org/10.1056/NEJMoa1102873
7. Quadrelli, S, Lyons G, Colt H, Chimondeguy D, Buero A (2015) Clinical characteristics and prognosis of incidentally detected lung cancers. Int J Surg Oncol 2015:287604. https://doi.org/10.1155/2015/287604
8. Nam JG, Park S, Hwang EJ et al (2019) Development and validation of deep learning-based automatic detection algorithm for malignant pulmonary nodules on chest radiographs. Radiology 290:218–228. https://doi.org/10.1148/radiol.2018180237
9. Yoo H, Kim KH, Singh R, Digumarthy SR, Kalra MK (2020) Validation of a deep learning algorithm for the detection of malignant pulmonary nodules in chest radiographs. JAMA Netw Open 3:e2017135–e2017135. https://doi.org/10.1001/jamanetworkopen.2020.17135
10. Sim Y, Chung MJ, Kotter E et al (2019) Deep convolutional neural network–based software improves radiologist detection of malignant lung nodules on chest radiographs. Radiology 294:199–209. https://doi.org/10.1148/radiol.2019182465
11. Jang S, Song H, Shin YJ et al (2020) Deep learning–based automatic detection algorithm for reducing overlooked lung cancers on chest radiographs. Radiology 296:652–661. https://doi.org/10.1148/radiol.2020200165
12. Li F, Engelmann R, Metz CE, Doi K, MacMahon H (2008) Lung cancers missed on chest radiographs: results obtained with a commercial computer-aided detection program. Radiology 246:273–280. https://doi.org/10.1148/radiol.2461061848
13. Schalekamp S, van Ginneken B, Koedam E et al (2014) Computer-aided detection improves detection of pulmonary nodules in chest radiographs beyond the support by bone-suppressed images. Radiology 272:252–261. https://doi.org/10.1148/radiol.14131315
14. Aberle DR, Berg CD, Black WC et al (2011) The National Lung Screening Trial: overview and study design. Radiology 258:243–253. https://doi.org/10.1148/radiol.10091808
15. Kim M, Park J, Na S, Park CM, Yoo D (2020) Learning visual context by comparison. In: Vedaldi A, Bischof H, Brox T, Frahm JM (eds) Computer vision – ECCV 2020. ECCV 2020. Lecture Notes in Computer Science, vol 12350. Springer, Cham. https://doi.org/10.1007/978-3-030-58558-7_34
16. Caruana R (1997) Multitask learning. Mach Learn 28:41–75. https://doi.org/10.1023/A:1007379606734
17. Hwang EJ, Park S, Jin K-N et al (2019) Development and validation of a deep learning–based automated detection algorithm for major thoracic diseases on chest radiographs. JAMA Netw Open 2:e191095–e191095. https://doi.org/10.1001/jamanetworkopen.2019.1095
18. DeLong ER, DeLong DM, Clarke-Pearson DL (1988) Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach. Biometrics 44:837–845. https://doi.org/10.2307/2531595
19. Hillis SL (2007) A comparison of denominator degrees of freedom methods for multiple observer ROC analysis. Stat Med 26:596–619. https://doi.org/10.1002/sim.2532
20. Genders TSS, Spronk S, Stijnen T, Steyerberg EW, Lesaffre E, Hunink MGM (2012) Methods for calculating sensitivity and specificity of clustered data: a tutorial. Radiology 265:910–916. https://doi.org/10.1148/radiol.12120509
21. McHugh ML (2012) Interrater reliability: the kappa statistic. Biochem Med 22:276–282
22. Meziane M, Mazzone P, Novak E et al (2012) A comparison of four versions of a computer-aided detection system for pulmonary nodules on chest radiographs. J Thorac Imaging 27:58–64. https://doi.org/10.1097/RTI.0b013e3181f240bc
23. Lee KH, Goo JM, Park CM, Lee HJ, Kwang Jin KN (2012) Computer-aided detection of malignant lung nodules on chest radiographs: effect on observers' performance. Korean J Radiol 13:564–571. https://doi.org/10.3348/kjr.2012.13.5.564
24. O'Mahony N, Campbell S, Carvalho A et al (2020) Deep learning vs. traditional computer vision. In: Arai K, Kapoor S (eds) Advances in computer vision. CVC 2019. Advances in Intelligent Systems and Computing, vol 943. Springer, Cham. https://doi.org/10.1007/978-3-030-17795-9_10
25. Bengio Y, Courville A, Vincent P (2013) Representation learning: a review and new perspectives. IEEE Trans Pattern Anal Mach Intell. https://doi.org/10.1109/TPAMI.2013.50

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.