Radiology

ORIGINAL RESEARCH • THORACIC IMAGING

Performance of a Deep Learning Algorithm Compared with Radiologic Interpretation for Lung Cancer Detection on Chest Radiographs in a Health Screening Population

Jong Hyuk Lee, MD • Hye Young Sun, MD* • Sunggyun Park, PhD • Hyungjin Kim, MD, PhD • Eui Jin Hwang, MD • Jin Mo Goo, MD, PhD • Chang Min Park, MD, PhD*

From the Department of Radiology and Institute of Radiation Medicine, Seoul National University College of Medicine, 101 Daehak-ro, Jongno-gu, Seoul 03080, Korea (J.H.L., H.K., E.J.H., J.M.G., C.M.P.); Department of Radiology, Healthcare Research Institute, Healthcare System Gangnam Center, Seoul National University Hospital, Seoul, Korea (H.Y.S.); and Lunit Inc, Seoul, Korea (S.P.). Received March 31, 2020; revision requested May 19; revision received July 20; accepted July 30. **Address correspondence to** C.M.P. (e-mail: *cmpark.morphius@gmail.com*).

This study was supported by the Seoul National University Hospital Research Fund (grant no. 03-2019-0190).

* J.H.L. and H.Y.S. contributed equally to this work.

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

See also the editorial by Armato in this issue.

Radiology 2020; 00:1–11 • https://doi.org/10.1148/radiol.2020201240 • Content codes: CH OI

Background: The performance of a deep learning algorithm for lung cancer detection on chest radiographs in a health screening population is unknown.

Purpose: To validate a commercially available deep learning algorithm for lung cancer detection on chest radiographs in a health screening population.

Materials and Methods: Out-of-sample testing of a deep learning algorithm was retrospectively performed using chest radiographs from individuals undergoing a comprehensive medical check-up between July 2008 and December 2008 (validation test). To evaluate the algorithm performance for visible lung cancer detection, the area under the receiver operating characteristic curve (AUC) and diagnostic measures, including sensitivity and false-positive rate (FPR), were calculated. The algorithm performance was compared with that of radiologists using the McNemar test and the Moskowitz method. Additionally, the deep learning algorithm was applied to a screening cohort undergoing chest radiography between January 2008 and December 2012, and its performances were calculated.

Results: In a validation test comprising 10 285 radiographs from 10 202 individuals (mean age, 54 years ± 11 [standard deviation]; 5857 men) with 10 radiographs of visible lung cancers, the algorithm's AUC was 0.99 (95% confidence interval: 0.97, 1), and it showed comparable sensitivity (90% [nine of 10 radiographs]) to that of the radiologists (60% [six of 10 radiographs]; $P = .25$) with a higher FPR (3.1% [319 of 10 275 radiographs] vs 0.3% [26 of 10 275 radiographs]; $P < .001$). In the screening cohort of 100 525 chest radiographs from 50 070 individuals (mean age, 53 years ± 11; 28 090 men) with 47 radiographs of visible lung cancers, the algorithm's AUC was 0.97 (95% confidence interval: 0.95, 0.99), and its sensitivity and FPR were 83% (39 of 47 radiographs) and 3% (2999 of 100 478 radiographs), respectively.

Conclusion: A deep learning algorithm detected lung cancers on chest radiographs with a performance comparable to that of radiologists, which will be helpful for radiologists in healthy populations with a low prevalence of lung cancer.

Online supplemental material is available for this article.

Lung cancer is a leading cause of cancer-related deaths worldwide, accounting for up to one quarter of all cancer deaths (1). Because lung cancers are diagnosed in an advanced stage in most cases, screening of early-stage lung cancer has emerged as a strategy for reducing lung cancer mortality (2–4). In fact, the National Lung Screening Trial clearly demonstrated that low-dose CT using an average effective dose of 1.5 mSv could reduce lung cancer mortality (5–9). In contrast, the value of using chest radiography as a screening modality could not be proven for either early lung cancer detection or lung cancer mortality reduction (7). Thus, it is controversial whether lung cancer screening should be performed using chest radiography. Nonetheless, chest radiography is widely used as an initial screening tool for several important thoracic diseases, including lung cancer, in the general population thanks to its low cost, easy accessibility, negligible radiation dose, and reasonable diagnostic capability (10–12). However, the low sensitivity of lung cancer detection, substantial inter- and intrareader variability, and vulnerability to observer error remain persistent weaknesses of chest radiography as a screening tool (13–15).

Recently, deep learning algorithms have staked out a place in lung cancer detection on chest radiographs (16–18) and have demonstrated excellent diagnostic performance in disease-enriched settings (16–18). Nam et al (16) reported that a deep learning algorithm achieved a sensitivity of 71%–91%, a specificity of 93%–100%, and an area under the receiver operating characteristic curve (AUC) of 0.92–0.99 in their validation data sets with a lung cancer prevalence of approximately 60%–68%. Sim et al (18) reported that

Abbreviations

AUC = area under the receiver operating characteristic curve, CI = confidence interval, FPR = false-positive rate, NPV = negative predictive value, PPV = positive predictive value

Summary

A deep learning algorithm detected lung cancer nodules on chest radiographs with a performance comparable to that of radiologists, which will be helpful for radiologists in healthy populations with a low prevalence of lung cancer.

Key Results

- In a validation study composed of 10 285 chest radiographs from 10 202 individuals, a deep learning algorithm showed comparable sensitivity to that of radiologists in detecting visible lung cancers (90% vs 60%; $P = .25$).
- Of the 10 285 chest radiographs, 3.2% were classified as abnormal by the deep learning algorithm compared with 0.3% by the pooled radiologists' interpretations (positive predictive value, 2.7% vs 19%; $P < .001$).
- In a health screening cohort composed of 100 525 chest radiographs from 50 070 individuals, the deep learning algorithm detected 39 of 47 (83%) visible lung cancers on chest radiographs, and its area under the receiver operating characteristic curve was 0.97.

another deep learning algorithm had comparable diagnostic performance to that of radiologists in the detection of lung cancer in their study population composed of 75% lung cancer–containing chest radiographs. Also, they showed that assistance from their algorithm improved sensitivity (from 65% to 73%) and reduced the false-positive rate (FPR) (from 20% to 18%) (18). However, those previous studies (16,18) validated their deep learning algorithms with arbitrarily selected test data sets, instead of data sets reflecting real-world clinical practice. The purpose of this study was to evaluate the performance of a deep learning algorithm for lung cancer nodule detection on chest radiographs, in a healthy screening population with average risk for lung cancer, compared with radiologic interpretation.

Materials and Methods

Lunit (Seoul, Korea) provided technical support for analyzing chest radiographs with a deep learning algorithm, and one author (S.P.) is an employee of Lunit. However, Lunit did not have any role in the study design; in the collection, analysis, and interpretation of the data; in the writing of the report; or in the decision to submit the article for publication. The first author (J.H.L.) had full access to and controlled all data in the study without any conflict of interest.

This retrospective study was approved by the institutional review board of Seoul National University Hospital, and the requirement for written informed consent was waived. The population of this study has not been reported before.

Study Population and Data Collection

We collected all chest radiographs from individuals who participated in the medical check-up and screening program at Seoul National University Hospital Healthcare System Gangnam Center in Seoul, Korea, between January 2008 and December 2012. The center provides a comprehensive medical check-up and screening program for noncommunicable diseases such as malignancies (19), and chest radiography is a core test in this screening program to help detect any lung disease requiring further diagnostic tests or treatments (19). The participants in this study paid the screening costs at their own expense, and they were not assessed according to predefined lung cancer risk factors. In this regard, the study population was an average-risk general population, rather than a lung cancer screening population, that was carefully selected in terms of age and history of cigarette smoking. All individuals underwent chest radiography as part of a comprehensive medical check-up, not for an evaluation of specific symptoms or signs.

In this study, out-of-sample testing of a deep learning algorithm was performed in two steps. First, a validation test, including a reader study, was performed using screening radiographs obtained between July 2008 and December 2008 as part of a screening cohort. After the validation test, the deep learning software was applied to an entire screening cohort undergoing radiography between January 2008 and December 2012 (Fig 1).

Standard for the Determination of Lung Cancer

Two radiologists (H.Y.S. and H.K., with 12 and 10 years of experience in thoracic radiology, respectively) created a lung cancer registry by searching the electronic medical records of our institution between January 2008 and December 2018. The health examination center has a patient referral system through which individuals requiring further diagnostic or therapeutic steps are referred to Seoul National University Hospital. Therefore, this registry included all individuals with lung cancer who underwent chest radiography during the study period, and any exclusion criteria, such as emphysema or diffuse interstitial disease, were not applied to this registry.

For individuals diagnosed with lung cancer, we determined cancer-positive chest radiographs using the following criteria: *(a)* lung cancer shown on a chest CT scan obtained within 3 months of the chest radiograph, *(b)* if a chest CT scan was not available, a chest radiograph taken within 12 months before the patient was diagnosed with lung cancer. We included a buffer of 3 additional months to account for variability in patients' follow-up strategy. When the pathologic diagnosis was made 15 months or longer after the chest radiograph, the chest radiograph was excluded because we could not guarantee whether there was lung cancer on the chest radiograph.

For individuals in the lung cancer registry, chest radiographs were classified as cancer-negative if lung cancer was not present on a chest CT scan taken within 3 months of the chest radiograph. The chest radiographs of individuals who were not diagnosed with lung cancer were designated as cancer-negative according to the following criteria to ensure that those chest radiographs did not have any lung cancer: *(a)* if a chest CT scan taken within 3 months of the chest radiograph did not demonstrate any significant lung nodules (ie, noncalcified nodules ≥6 mm); *(b)* if significant lung nodules were present on a chest CT scan but were confirmed as benign lesions, either pathologically

Figure 1: Flowchart of **(a)** validation test cohort and **(b)** screening cohort.

or clinically (ie, they were stable during 2 years of follow-up); or *(c)* if a chest CT scan within 3 months of the chest radiograph was not available and a follow-up chest radiograph after 12 months or longer revealed cancer-negative results.

The following individuals were excluded from the study population: *(a)* individuals with a history of lung cancer, *(b)* those with lung lesions pathologically confirmed as preinvasive lesions of lung cancer and not definitive lung cancer, *(c)* those with significant lung nodules (≥6 mm) on a chest CT scan that were neither pathologically nor clinically confirmed, *(d)* individuals whose chest radiographs were obtained more than 15 months before the pathologic diagnosis, and *(e)* individuals who did not have lung cancer and had no available follow-up radiographs after 12 months.

Assessment of Lung Cancer Visibility on Chest Radiographs

For cancer-positive chest radiographs, two board-certified radiologists (J.H.L. and E.J.H., with 7 and 9 years of experience in thoracic radiology, respectively) independently assessed the visibility of lung cancer on each chest radiograph, referring to the available chest CT scans. In this assessment, the lung cancers on chest radiographs were dichotomized as visible or invisible. Finally, lung cancers on chest radiographs designated as visible by either radiologist were clas-

sified as visible lung cancers on chest radiographs, and chest radiographs concordantly judged as visible by both radiologists were categorized as clearly visible lung cancers on chest radiographs. Lung cancers determined as invisible by both radiologists were classified as invisible lung cancers on chest radiographs.

Deep Learning Algorithm

A commercially available deep learning algorithm (Lunit Insight for Chest Radiography, version 4.7.2; Lunit, Seoul, Korea) was used in this study. The algorithm was developed for the detection of major thoracic diseases. Further detailed information about its development and validation is presented in Appendix E1 (online) and has been reported in a previous article (17). The algorithm provides both an image-wise probability value of a chest radiograph being abnormal and a per-pixel localization map overlaid on the input chest radiograph identifying the location of abnormalities.

In this study, the algorithm was applied with a predefined threshold (0.16), with which 95% sensitivity was achieved in previous study (17). The threshold of 0.16 was selected in this study because the primary purpose of screening lies in sensitively detecting lung cancers. All localization maps of chest radiographs with positive results from the deep learning algorithm were checked to ensure that the algorithm adequately localized each lung cancer lesion on the chest radiographs.

Table 1: Baseline Clinical Characteristics of Individuals and Chest Radiographs in Validation Test and Screening Cohorts

Patient and Radiograph Characteristics	Validation Test Cohort	Screening Cohort
No. of individuals	10 206	50 098
No. of chest radiographs	10 289	100 576
Mean age ± SD (y)*	54 ± 11 (18–95)	53 ± 11 (18–99)
Sex		
M	5859	28 105
F	4347	21 993
Median no. of chest radiographs per individual*	1 (1–4)	1 (1–20)
No. of cancer-positive chest radiographs	14 (0.1)	98 (0.1)
No. of chest radiographs with visible lung cancers on chest radiographs†	10 (0.1)	47 (0.05)
No. of chest radiographs with clearly visible lung cancers‡	Not evaluated	28 (0.03)

Note.—Except where indicated, numbers in parentheses are percentages. SD = standard deviation.

* Numbers in parentheses are ranges.

† In the analysis of visible lung cancer, four and 51 chest radiographs of invisible lung cancers were excluded in validation test and screening cohort, respectively. Thus, data are for 10 285 radiographs in 10 202 individuals for the validation test cohort and 100 525 radiographs in 50 070 individuals in the screening cohort.

‡ In the analysis of clearly visible lung cancer, 70 chest radiographs were excluded from the screening cohort because these radiographs were judged as having invisible lung cancer by at least one of the two radiologists. Thus, data are for 100 506 radiographs in 50 057 individuals.

Table 2: Comparison between Diagnostic Performance of Deep Learning Algorithm and That of Three Board-certified Radiologists for Detection of Visible Lung Cancers on Chest Radiographs in Validation Test Cohort

Variable	Sensitivity (%)	*P* Value	Specificity (%)	*P* Value	FPR (%)	*P* Value	NPV (%)	*P* Value	PPV (%)	*P* Value	Accuracy (%)
Pooled performance of three radiologists	60 (6/10) [26, 88]	…	100 (10 249/10 275) [100, 100]	…	0.3 (26/10 275) [0.2, 0.4]	…	100 (10 249/10 253) [100, 100]	…	19 (6/32) [7, 36]	…	100 (10 255/10 285) [100, 100]
Deep learning algorithm*	90 (9/10) [55, 100]	.25	97 (9956/10 275), [97, 97]	<.001	3.1 (319/10 275) [2.8, 3.5]	<.001	100 (9956/9957) [100, 100]	.09	2.7 (9/328) [1.3, 5.1]	<.001	97 (9965/10 285) [97, 97]
Matched threshold, 0.847†	70 (7/10) [35, 93]	>.99	100 (10 249/10 275) [100, 100]	NA	0.3 (26/10 275) [0.2, 0.4]	NA	100 (10 249/10 252) [100, 100]	.56	21 (7/33) [9, 39]	.26	100 (10 256/10 285) [100, 100]

Note.—Visible lung cancers: 10 individuals (0.1% of 10 202 individuals) with 10 radiographs (0.1% of 10 285 radiographs). Numbers in parentheses are raw data. Numbers in brackets are 95% confidence intervals. *P* values are for comparisons with the pooled diagnostic performance of three board-certified radiologists. FPR = false-positive rate, NA = not applicable, NPV = negative predictive value, PPV = positive predictive value.

* A predefined threshold of 0.16 was used.

† Corresponding threshold, sensitivity, negative predictive value, and positive predictive value when the specificity of the algorithm matched that of the radiologists.

Reader Study in the Validation Test

All chest radiographs in the validation test were reviewed once by one of three board-certified radiologists (one with 7 years of experience in reading chest radiographs, and two with 6 years of experience in reading chest radiographs; their subspecialties were not thoracic radiology). The three radiologists were blinded to all clinical information and asked to determine whether each chest radiograph had suspicious abnormalities for lung cancer.

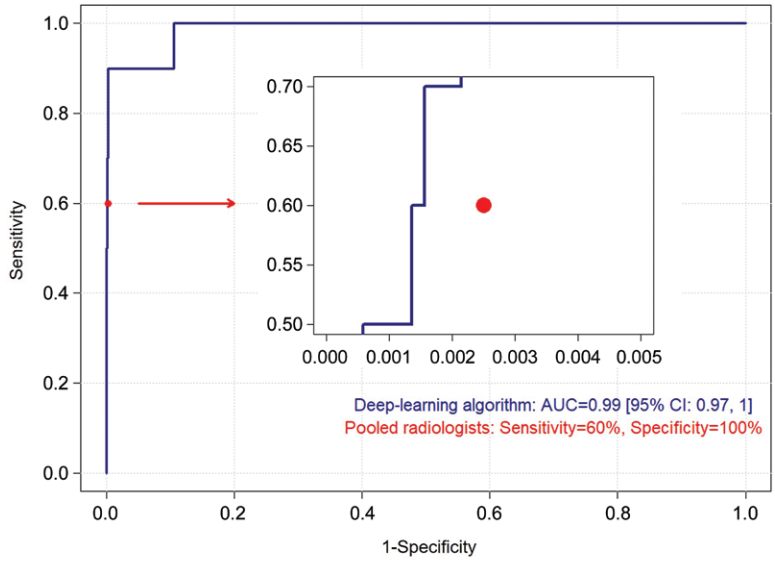
Statistical Analysis

In the validation test, the diagnostic performance of the deep learning algorithm was appraised through the following two tasks: *(a)* detection of visible lung cancers on chest radiographs and *(b)* discrimination between cancer-positive chest radiographs and cancer-negative chest radiographs. These two evaluations were performed independently.

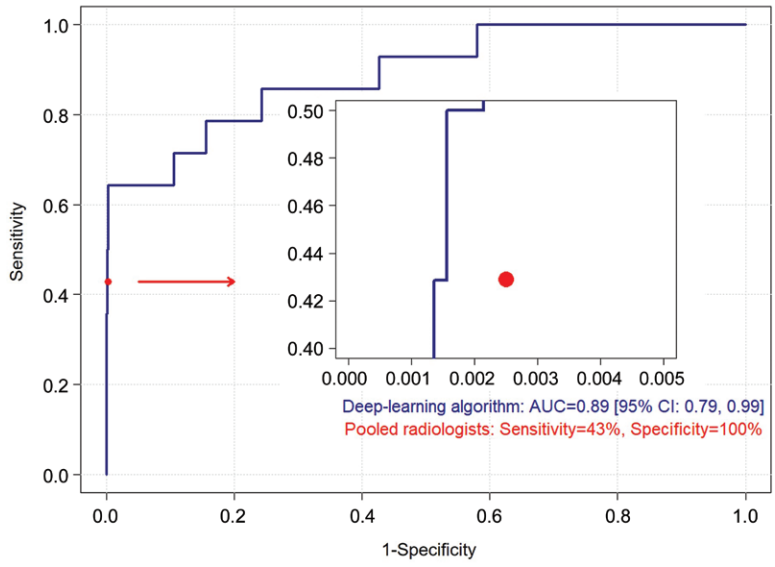
For each task, receiver operating characteristic curve analysis was performed and the AUC was used as the performance measure. Sensitivity, specificity, FPR, negative predictive value (NPV), positive predictive value (PPV), and accuracy were calculated. As diagnostic measures of the radiologists, the sensitivity, specificity, FPR, NPV, PPV, and accuracy of the pooled radiologists were calculated and compared with those of the deep learning algorithm using the McNemar tests for sensitivity, specificity, and FPR and the method described by Moskowitz and Pepe (20) for PPV and NPV. In addition, we calculated the threshold value where the specificity of the algorithm matched that of the pooled radiologists. The corresponding sensitivity, FPR, NPV, and PPV at this threshold were also calculated and compared with those of the radiologists.

For the entire screening cohort, in addition to the two tasks from the validation test, an additional task of detecting clearly visible lung cancers on chest radiographs was evaluated. These three evaluations were also performed independently.

Data were collected and saved as a spreadsheet, using Microsoft Excel 2016 (Microsoft, Redmond, Wash). All statistical analyses were performed using R software (version 4.0.0; The R Project for Statistical Computing, Vienna, Austria), and $P < .05$ was considered to indicate a statistically significant difference.



a.



b.

Figure 2: Receiver operating characteristic curves of deep learning algorithm for **(a)** detection of visible lung cancer on chest radiographs and **(b)** cancer-positive chest radiographs compared with board-certified radiologists in validation test. In validation test composed of 10 285 chest radiographs (**a**), including 10 chest radiographs with visible lung cancer, the algorithm had an area under the receiver operating characteristic curve (AUC) of 0.99 (95% confidence interval [CI]: 0.97, 1), and the radiologists showed a sensitivity of 60% and a specificity of 100%. In magnified illustration, red dot that represents radiologists' performance is below receiver operating characteristic curve of algorithm. In validation test composed of 10 289 chest radiographs (**b**), including 14 cancer-positive chest radiographs, the deep learning algorithm had an AUC of 0.89 (95% CI: 0.79, 0.99). In comparison, three board-certified radiologists showed a sensitivity of 43% and a specificity of 100% for this task. In magnified figure, red dot that represents radiologists' performance is below the receiver operating characteristic curve of the algorithm.

Results

Participant Characteristics

For the validation test, 13 640 individuals with 13 760 chest radiographs were initially included; 3434 individuals with 3471 radiographs were excluded due to the exclusion criteria. Finally, 10 206 individuals (5859 men and 4347 women; mean age, 54 years ± 11; age range, 18–95 years) with 10 289 chest radiographs were included in the task of detecting cancer-positive radiographs. For detecting visible lung cancer, 10 285 radiographs from 10 202 individuals (5857 men and 4345 women; mean age, 54 years ± 11) were included. Four individuals were excluded because their lung cancers were invisible on chest radiographs (Fig 1a, Table 1).

Table 3: Comparison between Diagnostic Performance of the Deep Learning Algorithm and That of Three Board-Certified Radiologists for Detection of Cancer-Positive Chest Radiographs in Validation Test

Variable	Sensitivity (%)	*P* Value	Specificity (%)	*P* Value	FPR (%)	*P* Value	NPV (%)	*P* Value	PPV (%)	*P* Value	Accuracy (%)
Pooled performance of three radiologists	43 (6/14) [18, 71]	…	100 (10 249/10 275) [100, 100]	…	0.3 (26/10 275) [0.2, 0.4]	…	100 (10 249/10 257) [100, 100]	…	19 (6/32) [7, 36]	…	100 (10 255/10 289) [100, 100]
Deep learning algorithm*	64 (9/14) [35, 87]	.25	97 (9956/10 275) [97, 97]	<.001	3.1 (319/10 275) [2.8, 3.4]	<.001	100 (9956/9961) [100, 100]	.11	3 (9/328) [1.3, 5.1]	<.001	97 (9965/10 289) [96, 97]
Matched threshold†, 0.808	64 (9/14) [35, 87]	.25	100 (10 249/10 275) [100, 100]	NA	0.2 (26/10 275) [0.2, 0.4]	NA	100 (10 249/10 254) [100, 100]	.08	26 (9/35) [12, 43]	.19	100 (10 258/10 289) [100, 100]

Note.—Cancer-positive chest radiographs: 14 individuals (0.1% of 10 206 individuals) with 14 chest radiographs (0.1% of 10 289 chest radiographs). Numbers in parentheses are raw data. Numbers in brackets are 95% confidence intervals. *P* values are for comparisons with the pooled diagnostic performance of three board-certified radiologists. FPR = false-positive rate, NA = not applicable, NPV = negative predictive value, PPV = positive predictive value.

* A predefined threshold of 0.16 was used.

† Corresponding threshold, sensitivity, negative predictive value, and positive predictive value when the specificity of the algorithm matched with that of the radiologists.

For the screening cohort, 71 951 individuals (37 938 men and 34 013 women; mean age, 50 years ± 12) with 132 207 chest radiographs were initially included, and 21 853 individuals with 31 631 radiographs were excluded according to the exclusion criteria. Thus, 50 098 individuals (28 105 men and 21 993 women; mean age, 53 years ± 11; age range, 18–99 years) with 100 576 radiographs were included in the task of detecting cancer-positive radiographs. For the analysis of visible lung cancer detection, 51 radiographs of invisible lung cancers from 37 individuals were excluded, and 50 070 individuals (28 090 men and 21 980 women; mean age, 53 years ± 11; age range, 18–99 years) with 100 525 radiographs were used. In the analysis of clearly visible lung cancer, 19 additional radiographs from 15 individuals were excluded because these radiographs were judged as having invisible lung cancer by one of the two radiologists (J.H.L. and E.J.H.) (Fig 1b, Table 1). The baseline characteristics of the study populations area described in Table 1.

Prevalence of Lung Cancer in Study Populations

In the validation test, 14 individuals (0.1% of 10 206 individuals) with 14 radiographs (0.1% of 10 289 chest radiographs) were confirmed to have lung cancers, and 10 192 individuals (99.9% of 10 206 individuals) with 10 275 radiographs (99.9% of 10 289 radiographs) were judged to have no lung cancers. Ten radiographs (0.1% of 10 285 radiographs) in 10 individuals (0.1% of 10 202 individuals) were judged to have visible lung cancers (Fig 1a, Table 1).

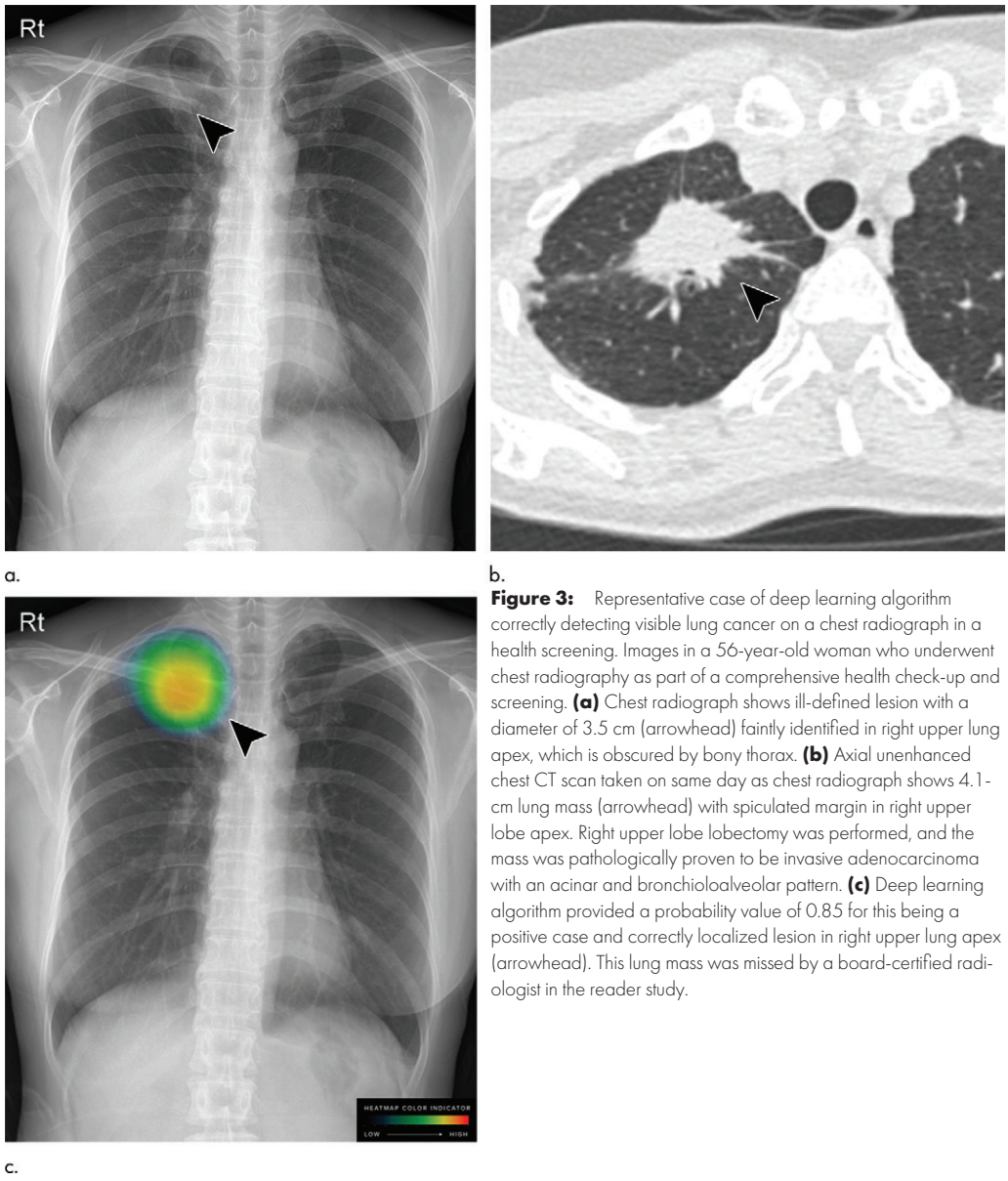
In the entire screening cohort, 77 individuals (0.2% of 50 098 individuals) with 98 radiographs (0.1% of 100 576 chest radiographs) were confirmed to have lung cancers, and 50 031 individuals (99.9% of 50 098 individuals) with 100 478 radiographs (99.9% of 100 576 radiographs) were judged to have no lung cancers. Ten individuals were included in both categories because their chest radiographs were initially cancer-negative, but they were later diagnosed with lung cancer and their chest radiographs were cancer-positive (demonstration on chest CT, *n* = 8; within 12 months of lung cancer diagnosis, *n* = 2). Detailed information about all lung cancers in this study population is presented in Table E1 (online). Among the 98 cancer-positive radiographs, 47 radiographs from 41 individuals were categorized as having visible lung cancers, and 28 chest radiographs from 27 individuals were determined as having clearly visible lung cancers (Fig 1b, Table 1). In one patient, the initial chest radiograph had invisible lung cancer, while the latter radiograph obtained 12 months after the initial radiograph had visible lung cancer.

Lung Cancer Detection Performance in the Validation Test

The detection performances of the deep learning algorithm and pooled radiologists for visible lung cancers on chest radiographs are shown in Table 2, and those of each radiologist are summarized in Table E2 (online). The algorithm's AUC was 0.99 (95% confidence interval [CI]: 0.97, 1), and it detected three more lung cancers (nine of 10 chest radiographs; sensitivity, 90%) than the radiologists (six of 10 chest radiographs; sensitivity, 60%) (Fig 2a). However, this difference was not statistically significant (*P* = .25). The algorithm had an NPV equivalent to that of the radiologists (100% vs 100%, *P* = .09), but it had a lower specificity and PPV and a higher FPR (specificity, 97% vs 100%, *P* < .001; PPV, 2.7% vs 19%, *P* < .001; FPR, 3.1% vs 0.3%, *P* < .001). At the threshold where the algorithm's specificity matched that of the radiologists (0.847), the algorithm's sensitivity, NPV, and PPV were 70%, 100%, and 21%, respectively, and all diagnostic measures of the algo-

rithm were comparable to those of the radiologists (sensitivity, $P >$ .99; NPV, P = .56; PPV, P = .26).

The classification of cancer-positive chest radiographs is presented in Table 3 and Figure 2b. The algorithm's AUC was 0.89 (95% CI: 0.79, 0.99). It detected three more lung cancers (nine of 14 chest radiographs; sensitivity, 64%) than the radiologists (six of 14 chest radiographs; sensitivity, 43%) (P = .25). However, it had a higher FPR (3.1% vs 0.3%; $P <$ .001) (Fig 3).



a. b.

c.

Figure 3: Representative case of deep learning algorithm correctly detecting visible lung cancer on a chest radiograph in a health screening. Images in a 56-year-old woman who underwent chest radiography as part of a comprehensive health check-up and screening. **(a)** Chest radiograph shows ill-defined lesion with a diameter of 3.5 cm (arrowhead) faintly identified in right upper lung apex, which is obscured by bony thorax. **(b)** Axial unenhanced chest CT scan taken on same day as chest radiograph shows 4.1-cm lung mass (arrowhead) with spiculated margin in right upper lobe apex. Right upper lobe lobectomy was performed, and the mass was pathologically proven to be invasive adenocarcinoma with an acinar and bronchioloalveolar pattern. **(c)** Deep learning algorithm provided a probability value of 0.85 for this being a positive case and correctly localized lesion in right upper lung apex (arrowhead). This lung mass was missed by a board-certified radiologist in the reader study.

Lung Cancer Detection Performance of the Deep Learning Algorithm in the Entire Screening Cohort

The performance metrics of the deep learning algorithm for lung cancer detection in the entire screening cohort are tabulated in Table 4. The algorithm classified 3% (3038 of 100 576 radiographs for cancer-positive radiographs, 3038 of 100 525 radiographs for visible lung cancers on chest radiographs, and 3027 of 100 506 radiographs for clearly visible lung cancers on chest radiographs) of chest radiographs as abnormal.

For the classification of cancer-positive radiographs, the AUC of the algorithm was 0.78 (95% CI: 0.73, 0.83) (Fig 4a). The algorithm correctly classified 39 of the 98 cancer-positive radiographs (sensitivity, 40%). The specificity, NPV, and PPV of the algorithm were 97%, 100%, and 1.3%, respectively.

In the detection of visible lung cancers on the chest radiograph, the algorithm had an AUC of 0.97 (95% CI: 0.95, 0.99) (Fig 4b). Visible lung cancers were correctly detected on 39 of 47 radiographs (sensitivity, 83%). The specificity, NPV, and PPV of the algorithm for detecting visible lung cancers were 97%, 100%, and 1.3%, respectively.

For the detection of clearly visible lung cancers on chest radiographs, the algorithm showed an AUC of 0.99 (95% CI: 0.99, 0.99) (Fig 4c). Clearly visible lung cancers were correctly detected on 28 of 28 chest radiographs (sensitivity, 100%). The specificity, NPV, and PPV of the algorithm were 97%, 100%, and 0.9%, respectively (Fig 5).

When the three receiver operating characteristic curves were compared with each other, the performance of the algorithm improved with increased visibility (cancer-positive chest radiographs [AUC = 0.78; 95% CI: 0.73, 0.83] vs visible lung cancers on chest radiographs [AUC = 0.97; 95% CI: 0.95, 0.99], $P <$ .001; cancer-positive chest radiographs vs clearly visible lung cancer on chest radiographs [AUC = 0.99; 95% CI: 0.99, 0.99], $P <$ .001; visible lung cancers on chest radiographs vs clearly visible lung cancers on chest radiographs, P = .01).

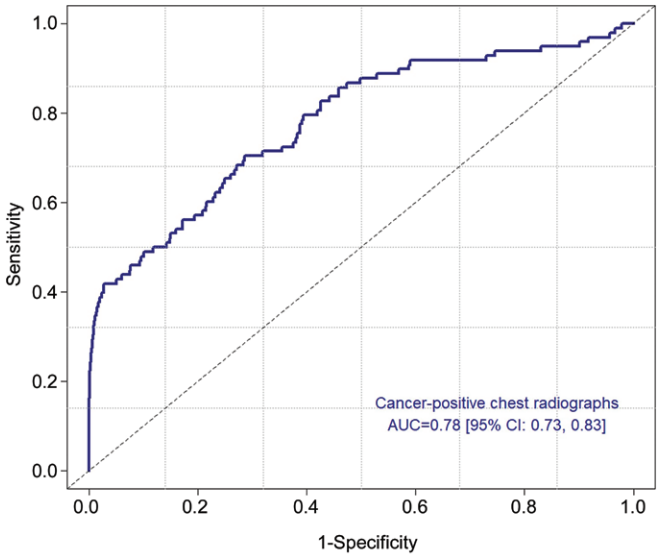
Discussion

To test whether a deep learning algorithm could evaluate chest radiographs for lung cancer in a large-scale health screening population, we applied a commercially available deep learning algorithm and compared its performance to that of radiolo-

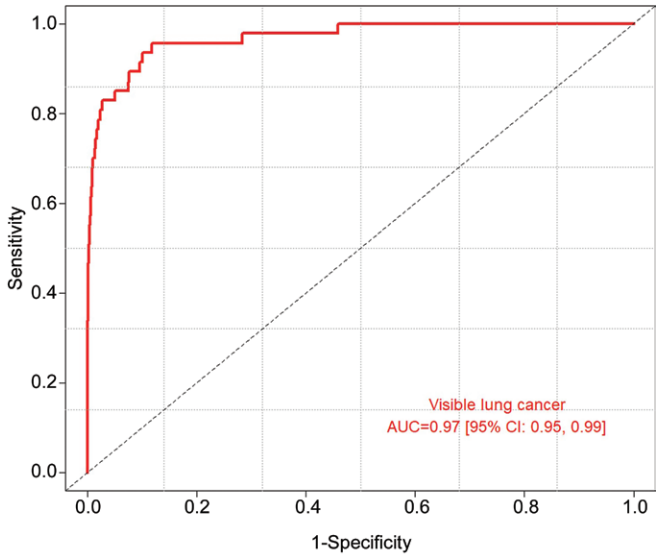
Table 4: Diagnostic Performance of Deep Learning Algorithm for Detection of Lung Cancers on Health Screening Cohort Chest Radiographs

Variable	Sensitivity (%)	Specificity (%)	FPR (%)	NPV (%)	PPV (%)	Accuracy (%)
Cancer-positive chest radiographs	40 (39/98) [30, 50]	97 (97 479/100 478) [97, 97]	3 (2999/100 478) [2.9, 3.1]	100 (97 479/97 538) [100, 100]	1.3 (39/3038) [0.9, 1.8]	97 (97 518/100 576) [97, 97]
Visible cancers on chest radiographs	83 (39/47) [69, 92]	97 (97 479/100 478) [97, 97]	3 (2999/100 478) [2.9, 3.1]	100 (97 479/97 487) [100, 100]	1.3 (39/3038) [0.9, 1.8]	97 (97 518/100 525) [97, 97]
Clearly visible cancers on chest radiographs	100 (28/28) [88, 100]	97 (97 479/100 478) [97, 97]	3 (2999/100 478) [2.9, 3.1]	100 (97 479/97 479) [100, 100]	0.9 (28/3027) [0.7, 1.3]	97 (97 507/100 506) [97, 97]

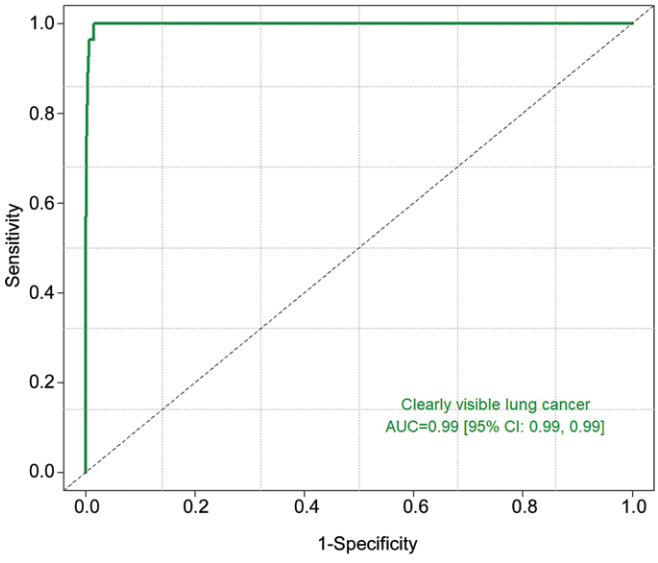
Note.—A predefined threshold of 0.16 was used. Cancer-positive chest radiographs: 77 individuals (0.2% of 50 098 individuals) with 98 radiographs (0.1% of 100 576 chest radiographs). Visible lung cancers: 41 individuals (0.1% of 50 070 individuals) with 47 radiographs (0.05% of 100 525 chest radiographs). Clearly visible lung cancers: 27 individuals (0.05% of 50 057 individuals) with 28 radiographs (0.03% of 100 506 chest radiographs). Numbers in parentheses are raw data. Numbers in brackets are 95% confidence intervals. FPR = false-positive rate, NPV = negative predictive value, PPV = positive predictive value.



a.



b.



c.

Figure 4: Receiver operating characteristic curves of deep learning algorithm for detection of lung cancer on chest radiographs in a health screening cohort. **(a)** Receiver operating characteristic curve of deep learning algorithm for classification of cancer-positive chest radiographs in a health screening. Area under the receiver operating characteristic curve (AUC) was 0.78 (95% confidence interval [CI]: 0.73, 0.83). **(b)** Receiver operating characteristic curve of deep learning algorithm for visible lung cancers on chest radiographs, with an AUC of 0.97 (95% CI: 0.95, 0.99). **(c)** Receiver operating characteristic curve of deep learning algorithm for detection of clearly visible lung cancers on chest radiographs. AUC of algorithm was 0.99 (95% CI: 0.99, 0.99).

gists. We found that the deep learning algorithm had an area under the receiver operating characteristic curve of 0.99 (95% confidence interval: 0.97, 1) and a comparable sensitivity (90% vs 60%, $P = .25$) to radiologists, with a higher false-positive rate (3.1% vs 0.3%, $P < .001$). At the threshold where the algorithm's specificity matched that of the radiologists (0.847), its sensitivity, negative predictive value, and positive predictive value were comparable to those of the radiologists.

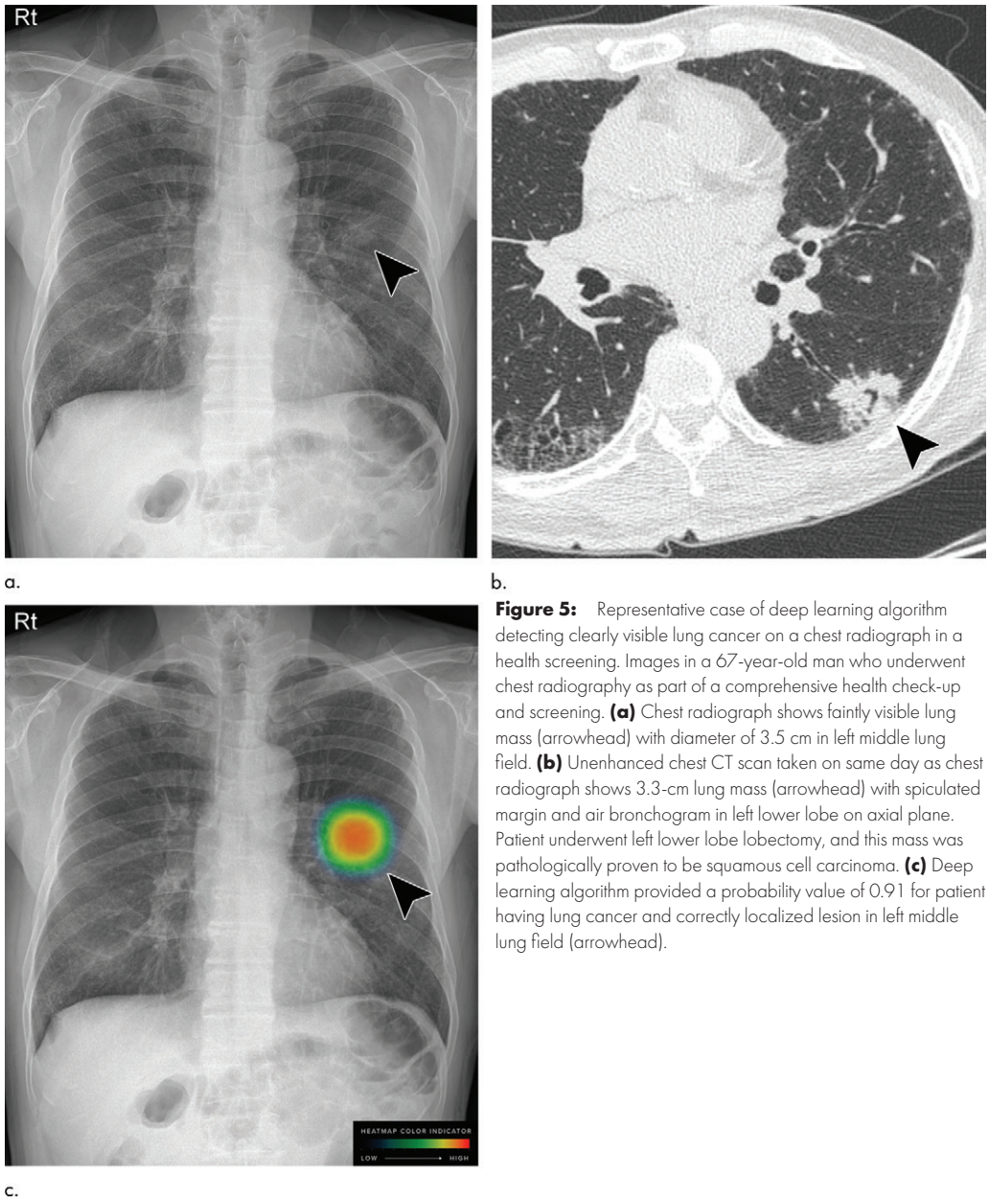


Figure 5: Representative case of deep learning algorithm detecting clearly visible lung cancer on a chest radiograph in a health screening. Images in a 67-year-old man who underwent chest radiography as part of a comprehensive health check-up and screening. **(a)** Chest radiograph shows faintly visible lung mass (arrowhead) with diameter of 3.5 cm in left middle lung field. **(b)** Unenhanced chest CT scan taken on same day as chest radiograph shows 3.3-cm lung mass (arrowhead) with spiculated margin and air bronchogram in left lower lobe on axial plane. Patient underwent left lower lobe lobectomy, and this mass was pathologically proven to be squamous cell carcinoma. **(c)** Deep learning algorithm provided a probability value of 0.91 for patient having lung cancer and correctly localized lesion in left middle lung field (arrowhead).

Previous studies of deep learning algorithms for the detection of lung cancers on chest radiographs had limitations in their applicability to real-world settings for the following reasons (21–23): *(a)* they were tested using disease-enriched data sets (prevalence, 16%–75%), which were clearly unrealistic (16–18); *(b)* their test data sets were arbitrarily selected in terms of the size, number, and location of the lung cancers (16–18); and *(c)* their test data sets comprised clearly dichotomized cases (chest radiographs with lung cancer vs normal chest radiographs), which intentionally excluded any indeterminate chest radiographs or radiographs with other pathologies (16–18). By contrast, we performed our study in a real-world setting, using a real-world health screening population (23). Therefore, we believe that the algorithm analyzed in our study could be reasonably applied for the detection of lung cancers on chest radiographs in a health screening population with an average risk of lung cancer.

The sensitivity for the detection of visible lung cancers on chest radiographs has been reported to be highly variable at 20%–92%, and radiologists' perceptual errors are reported to be the most common and preventable cause for failure to diagnose lung cancers or to detect lung cancers on chest radiographs (13–15,24). In this study, the deep learning algorithm consistently showed a much higher sensitivity (90% in the validation test and 83% in the health screening of this study) than previous studies for the detection of visible lung cancers on chest radiographs, classifying only 3% of chest radiographs as having a high probability of being abnormal (13–15,24). The deep learning algorithm can help reduce diagnostic errors caused by simple mistakes or perceptual errors due to radiologists' insufficient expertise, as this algorithm showed a consistent and high detection performance for lung cancers on chest radiographs and was not vulnerable to the perceptual errors of human readers (16,17).

In previous studies, the sensitivity of algorithms for lung cancer detection on chest radiographs increased when lesions were large and clearly demonstrated (12,16,18). Concordantly, our study also showed that the AUC increased when lung cancers were clearly visible on chest radiographs. The algorithm detected 100% of clearly visible lung cancers on radiographs in the health screening setting. By contrast, the sensitivity decreased in the

cases of lung cancers with lower visibility on radiographs. The decline in sensitivity was even more noticeable in the analysis of cancer-positive radiographs, which included all cancer-positive cases regardless of lesion visibility on radiographs. A clear explanation for this phenomenon is that the deep learning algorithm used in our study was trained with data sets of chest radiographs with visible lung cancers as determined by radiologists (16,17).

The most important limitation of this study is that not all patients had contemporaneous chest CT examinations at the time of their chest radiographs. This fact is problematic, as nodules could be missed by our method of relying on longitudinal follow-up as the reference standard. Very slowly growing lung cancers might not be found using our method for follow-up. Second, we did not evaluate the added value of the deep learning algorithm to the diagnostic performance of the radiologists. But given the results of previous studies, in which the diagnostic performance of radiologists improved with the aid of deep learning algorithms (16,18), we believe that this algorithm can improve radiologists' performance even in a health screening population. Third, no lateral radiographs were used in this study. Sometimes, pulmonary nodules are easier to find on lateral radiographs because of overlap of the nodule with normal structures on the frontal radiographs. Fourth, although the three radiologists participating in the validation test had 6–7 years of experience in reading chest radiographs, their performance results might not represent those of chest radiologists. Fifth, the diagnostic performances of other deep learning algorithms could differ from our results, and, thus, the deep learning algorithm used in our study cannot represent all other deep learning algorithms. Finally, this external validation study was performed at a single institution.

In conclusion, a deep learning algorithm detected lung cancer nodules on chest radiographs with a performance comparable to that of radiologists, which will be helpful for radiologists in healthy populations with a low prevalence of lung cancer. To generalize the clinical use of the deep learning algorithm in a screening program of the general population, further studies covering a variety of races and medical environments will be needed.

Acknowledgments: The reader study group included the following three members: Jae Hyun Kim, MD, Department of Radiology, Seoul National University Hospital, College of Medicine, Seoul, Korea; Hyoung In Choi, MD, Department of Radiology, Seoul National University Hospital, College of Medicine, Seoul, Korea; and Juil Park, MD, Department of Radiology, Seoul National University Hospital, College of Medicine, Seoul, Korea. We also thank Yeongho Choi, MD (Department of Emergency Medicine, Seoul National University Hospital, College of Medicine, Seoul, Korea) for consultation on the appropriate methods of data processing for this study and Lunit for providing technical support.

Author contributions: Guarantors of integrity of entire study, J.H.L., C.M.P.; study concepts/study design or data acquisition or data analysis/interpretation, all authors; manuscript drafting or manuscript revision for important intellectual content, all authors; approval of final version of submitted manuscript, all authors; agrees to ensure any questions related to the work are appropriately resolved, all authors; literature research, J.H.L., H.Y.S., H.K., E.J.H., C.M.P.; clinical studies, J.H.L., H.Y.S., E.J.H., J.M.G.; experimental studies, J.H.L.; statistical analysis, J.H.L., H.Y.S.; and manuscript editing, J.H.L., H.Y.S., S.P., H.K., J.M.G., C.M.P.

Disclosures of Conflicts of Interest: J.H.L. disclosed no relevant relationships. **H.Y.S.** disclosed no relevant relationships. **S.P.** Activities related to the present article: disclosed no relevant relationships. Activities not related to the present article: works for Lunit; holds stock/stock options in Lunit. Other relationships: disclosed no relevant relationships. **H.K.** Activities related to the present article: disclosed no relevant relationships. Activities not related to the present article: has grants/grants pending with Lunit. Other relationships: disclosed no relevant relationships. **E.J.H.** Activities related to the present article: institution received a grant from Lunit. Activities not related to the present article: disclosed no relevant relationships. Other relationships: disclosed no relevant relationships. **J.M.G.** Activities related to the present article: disclosed no relevant relationships. Activities not related to the present article: has grants/grants pending with Infinitt Healthcare and Dong Kook Lifescience. Other relationships: disclosed no relevant relationships. **C.M.P.** Activities related to the present article: disclosed no relevant relationships. Activities not related to the present article: receive a research grant from Lunit. Other relationships: disclosed no relevant relationships.

References

1. Siegel RL, Miller KD, Jemal A. Cancer statistics, 2020. CA Cancer J Clin 2020;70(1):7–30.
2. van Iersel CA, de Koning HJ, Draisma G, et al. Risk-based selection from the general population in a screening trial: selection criteria, recruitment and power for the Dutch-Belgian randomised lung cancer multi-slice CT screening trial (NELSON). Int J Cancer 2007;120(4):868–874.
3. Wille MM, Dirksen A, Ashraf H, et al. Results of the randomized Danish lung cancer screening trial with focus on high-risk profiling. Am J Respir Crit Care Med 2016;193(5):542–551.
4. Hocking WG, Hu P, Oken MM, et al. Lung cancer screening in the randomized Prostate, Lung, Colorectal, and Ovarian (PLCO) cancer screening trial. J Natl Cancer Inst 2010;102(10):722–731.
5. National Lung Screening Trial Research Team; Aberle DR, Adams AM, et al. Reduced lung-cancer mortality with low-dose computed tomographic screening. N Engl J Med 2011;365(5):395–409.
6. Oken MM, Hocking WG, Kvale PA, et al. Screening by chest radiograph and lung cancer mortality: the Prostate, Lung, Colorectal, and Ovarian (PLCO) randomized trial. JAMA 2011;306(17):1865–1873.
7. Dominioni L, Poli A, Mantovani W, et al. Assessment of lung cancer mortality reduction after chest X-ray screening in smokers: a population-based cohort study in Varese, Italy. Lung Cancer 2013;80(1):50–54.
8. Nakayama T, Baba T, Suzuki T, Sagawa M, Kaneko M. An evaluation of chest X-ray screening for lung cancer in gunma prefecture, Japan: a population-based case-control study. Eur J Cancer 2002;38(10):1380–1387.
9. Horeweg N, Scholten ET, de Jong PA, et al. Detection of lung cancer through low-dose CT screening (NELSON): a prespecified analysis of screening test performance and interval cancers. Lancet Oncol 2014;15(12):1342–1350.
10. Shankar A, Saini D, Dubey A, et al. Feasibility of lung cancer screening in developing countries: challenges, opportunities and way forward. Transl Lung Cancer Res 2019;8(Suppl 1):S106–S121.
11. Gossner J. Lung cancer screening-don't forget the chest radiograph. World J Radiol 2014;6(4):116–118.
12. Dominioni L, Rotolo N, Mantovani W, et al. A population-based cohort study of chest x-ray screening in smokers: lung cancer detection findings and follow-up. BMC Cancer 2012;12(1):18.
13. Quekel LG, Goei R, Kessels AG, van Engelshoven JM. Detection of lung cancer on the chest radiograph: impact of previous films, clinical information, double reading, and dual reading. J Clin Epidemiol 2001;54(11):1146–1150.
14. Berlin NI, Buncher CR, Fontana RS, Frost JK, Melamed MR. The National Cancer Institute Cooperative Early Lung Cancer Detection Program. Results of the initial screen (prevalence). Early lung cancer detection: Introduction. Am Rev Respir Dis 1984;130(4):545–549.
15. Gavelli G, Giampalma E. Sensitivity and specificity of chest X-ray screening for lung cancer: review article. Cancer 2000;89(11 Suppl):2453–2456.
16. Nam JG, Park S, Hwang EJ, et al. Development and validation of deep learning–based automatic detection algorithm for malignant pulmonary nodules on chest radiographs. Radiology 2019;290(1):218–228.
17. Hwang EJ, Park S, Jin KN, et al. Development and Validation of a Deep Learning-Based Automated Detection Algorithm for Major Thoracic Diseases on Chest Radiographs. JAMA Netw Open 2019;2(3):e191095.
18. Sim Y, Chung MJ, Kotter E, et al. Deep Convolutional Neural Network-based Software Improves Radiologist Detection of Malignant Lung Nodules on Chest Radiographs. Radiology 2020;294(1):199–209.
19. Lee C, Choe EK, Choi JM, et al. Health and Prevention Enhancement (H-PEACE): a retrospective, population-based cohort study conducted at the Seoul National University Hospital Gangnam Center, Korea. BMJ Open 2018;8(4):e019327.
20. Moskowitz CS, Pepe MS. Comparing the predictive values of diagnostic tests: sample size and analysis for paired study designs. Clin Trials 2006;3(3):272–279.
21. Ting DSW, Tan TE, Lim CCT. Development and Validation of a Deep Learning System for Detection of Active Pulmonary Tuberculosis on Chest Radiographs: Clinical and Technical Considerations. Clin Infect Dis 2019;69(5):748–750.
22. Ting DSW, Yi PH, Hui F. Clinical Applicability of Deep Learning System in Detecting Tuberculosis with Chest Radiography. Radiology 2018;286(2):729–731.
23. Park SH. Diagnostic Case-Control versus Diagnostic Cohort Studies for Clinical Validation of Artificial Intelligence Algorithm Performance. Radiology 2019;290(1):272–273.
24. Muhm JR, Miller WE, Fontana RS, Sanderson DR, Uhlenhopp MA. Lung cancer detected during a screening program using four-month chest radiographs. Radiology 1983;148(3):609–615.