Validation of a Deep Learning Algorithm for the Detection of Malignant Pulmonary Nodules in Chest Radiographs

Hyunsuk Yoo, MD; Ki Hwan Kim, MD, PhD; Ramandeep Singh, MD; Subba R. Digumarthy, MD; Mannudeep K. Kalra, MD

Abstract

IMPORTANCE The improvement of pulmonary nodule detection, which is a challenging task when using chest radiographs, may help to elevate the role of chest radiographs for the diagnosis of lung cancer.

OBJECTIVE To assess the performance of a deep learning–based nodule detection algorithm for the detection of lung cancer on chest radiographs from participants in the National Lung Screening Trial (NLST).

DESIGN, SETTING, AND PARTICIPANTS This diagnostic study used data from participants in the NLST to assess the performance of a deep learning–based artificial intelligence (AI) algorithm for the detection of pulmonary nodules and lung cancer on chest radiographs using separate training (in-house) and validation (NLST) data sets. Baseline (T0) posteroanterior chest radiographs from 5485 participants (full T0 data set) were used to assess lung cancer detection performance, and a subset of 577 of these images (nodule data set) were used to assess nodule detection performance. Participants aged 55 to 74 years who currently or formerly (ie, quit within the past 15 years) smoked cigarettes for 30 pack-years or more were enrolled in the NLST at 23 US centers between August 2002 and April 2004. Information on lung cancer diagnoses was collected through December 31, 2009. Analyses were performed between August 20, 2019, and February 14, 2020.

EXPOSURES Abnormality scores produced by the AI algorithm.

MAIN OUTCOMES AND MEASURES The performance of an AI algorithm for the detection of lung nodules and lung cancer on radiographs, with lung cancer incidence and mortality as primary end points.

RESULTS A total of 5485 participants (mean [SD] age, 61.7 [5.0] years; 3030 men [55.2%]) were included, with a median follow-up duration of 6.5 years (interquartile range, 6.1-6.9 years). For the nodule data set, the sensitivity and specificity of the AI algorithm for the detection of pulmonary nodules were 86.2% (95% CI, 77.8%-94.6%) and 85.0% (95% CI, 81.9%-88.1%), respectively. For the detection of all cancers, the sensitivity was 75.0% (95% CI, 62.8%-87.2%), the specificity was 83.3% (95% CI, 82.3%-84.3%), the positive predictive value was 3.8% (95% CI, 2.6%-5.0%), and the negative predictive value was 99.8% (95% CI, 99.6%-99.9%). For the detection of malignant pulmonary nodules in all images of the full T0 data set, the sensitivity was 94.1% (95% CI, 86.2%-100.0%), the specificity was 83.3% (95% CI, 82.3%-84.3%), the positive predictive value was 3.4% (95% CI, 2.2%-4.5%), and the negative predictive value was 100.0% (95% CI, 99.9%-100.0%). In digital radiographs of the nodule data set, the AI algorithm had higher sensitivity (96.0% [95% CI, 88.3%-100.0%] vs 88.0% [95% CI, 75.3%-100.0%]; P = .32) and higher specificity (93.2% [95% CI, 89.9%-96.5%] vs 82.8% [95% CI, 77.8%-87.8%]; P = .001) for nodule detection compared with the NLST radiologists. For malignant pulmonary nodule detection on digital (continued)
radiographs of the full T0 data set, the sensitivity of the AI algorithm was higher (100.0% [95% CI, 100.0%-100.0%]) vs 94.1% [95% CI, 82.9%-100.0%]; P = .32) compared with the NLST radiologists, and the specificity (90.9% [95% CI, 89.6%-92.1%] vs 91.0% [95% CI, 89.7%-92.2%]; P = .91), positive predictive value (8.2% [95% CI, 4.4%-11.9%] vs 7.8% [95% CI, 4.1%-11.5%]; P = .65), and negative predictive value (100.0% [95% CI, 100.0%-100.0%] vs 99.9% [95% CI, 99.8%-100.0%]; P = .32) were similar to those of NLST radiologists.

CONCLUSIONS AND RELEVANCE In this study, the AI algorithm performed better than NLST radiologists for the detection of pulmonary nodules on digital radiographs. When used as a second reader, the AI algorithm may help to detect lung cancer.

JAMA Network Open. 2020;3(9):e2017135. doi:10.1001/jamanetworkopen.2020.17135

Introduction

Large randomized clinical trials investigating chest radiography and low-dose computed tomography (CT) as screening tools for lung cancer have reported that low-dose CT screening reduces lung cancer mortality in high-risk populations. Therefore, only low-dose CT is recommended for lung cancer screening among high-risk populations in most countries. However, a substantial cost is associated with low-dose CT; compared with chest radiography, CT is less accessible and more expensive, exposes patients to a higher dose of radiation, and produces a higher proportion of false-positive and incidental findings, which may lead to additional laboratory testing and increase patient anxiety.

Chest radiography avoids many of the problems associated with low-dose CT, but the survival benefits of chest radiography as a screening tool have not been fully examined. The detection of lung cancer on chest radiographs is challenging for radiologists because of the limited contrast resolution and 2-dimensional projectional nature of radiography, which can obscure findings owing to the superimposition of lesions by anatomical structures and produce high false-negative rates with low intraobserver and interobserver agreement. Previous studies have reported that, in retrospect, radiographic evidence of cancer was present in up to 90% of patients with peripheral cancer and in 65% to 70% of patients with centrally located cancer before the actual cancer diagnosis was made. Tumor characteristics, such as lesion size, conspicuity, and location, are all independent factors in detection error and can lead to missed lesions during the interpretation of chest radiographs.

For these reasons, low-dose CT is the recommended modality for lung cancer screening among high-risk populations in most countries. Yet, substantial proportions of incidentally detected lung cancer cases are detected as malignant pulmonary nodules on routine chest radiographs. Although improved detection of lung nodules could produce overdiagnosis of low-grade cancer without decreasing mortality, patients with incidental diagnoses have better prognoses, as they are likely to have earlier stages of cancer and smaller tumors. Therefore, if errors in the detection of malignant pulmonary nodules can be reduced, more lung cancer cases that are visible but overlooked by the observer on routine chest radiographs may be detected, with subsequent survival benefits for those with detected lung cancer.

Existing studies have examined various deep learning models and techniques for abnormal image classification on chest radiographs. In the last 5 years, several artificial intelligence (AI) algorithms have been tested in an effort to decrease radiologist errors and increase the detection rate of pulmonary nodules on chest radiographs. Compared with these previous studies, the control group in our study comprised participants without disease as well as those with various types of disease, including calcified granuloma, consolidation, emphysema, and other thoracic diseases. Thus, our control group was more reflective of the wide spectrum of disease encountered in clinical practice. The goal of this study was to evaluate whether an AI algorithm trained for pulmonary
nodule detection could be applied for lung cancer detection on chest radiographs from participants in the multicenter National Lung Screening Trial (NLST) to validate the generalizability and accuracy of our AI approach.22

Methods
This diagnostic study assessed the performance of a deep learning–based AI algorithm for the detection of pulmonary nodules and lung cancer on chest radiographs using separate training (in-house) and validation (NLST) data sets. Ethics review and approval were obtained from the institutional review board of Massachusetts General Hospital. The need for informed consent was waived because our retrospective study used previously acquired data from other clinical trials (NLST). This study followed the Standards for Reporting of Diagnostic Accuracy (STARD) reporting guideline.

Overview
Our retrospective analysis used data from participants in the NLST, a multicenter randomized clinical trial comparing low-dose CT with chest radiography for the screening of a high-risk population.22,23 From August 2002 through April 2004, eligible participants were enrolled in 33 US centers and randomized to receive either low-dose CT screening or chest radiography screening.22 Participants were eligible if they were aged 55 to 74 years with at least 30 pack-years of cigarette smoking; participants who formerly smoked were eligible if they had quit smoking within the previous 15 years. The participants were offered 3 annual chest radiography screenings (at T0, T1, and T2), and were followed up until December 31, 2009, for lung cancer incidence and mortality, which were the primary end points of the study. Of 9362 participants enrolled through the American College of Radiology Imaging Network and randomized to the chest radiography arm, 5485 participants (83% of whom were randomly selected) from 21 sites who had valid Digital Imaging and Communications in Medicine files (full T0 data set) were included in our analysis.

The chest radiographs comprised images from multiple vendors that were obtained using various types of chest radiography, including screen-film, computed, digital, and thoravision radiography. For each image, radiologists who participated in the NLST study annotated the type of the abnormality present, and commented on the presence of any finding suggestive of lung cancer (ie, nodule[s] ≥4 mm or enlarging nodule[s], mass[es], and other nonspecific abnormalities) that warranted further diagnostic evaluation.22

The radiologists who participated in the NLST had American Board of Radiology certification or its equivalent, radiologic training during residency, involvement in the supervision and interpretation of at least 300 chest CT acquisitions, performance of at least 200 chest radiography acquisitions per year, and participation in continuing medical education in accordance with American College of Radiology standards.22

Preparation and Annotation
To evaluate the performance of the AI algorithm for the detection of pulmonary nodules, a subset of 577 baseline (T0) images (nodule data set) were selected and reannotated for the presence of nodules with the assistance of clinical information or follow-up imaging examinations. First, we selected chest radiographs of 48 patients who received lung cancer diagnoses within 1 year of the T0 screening. Second, we selected chest radiographs of 50 patients who received all 3 years of screening radiographs and had noncalcified nodules (as annotated by NLST radiologists) but did not have lung cancer diagnoses. Third, we selected chest radiographs of 480 patients who did not meet the first 2 criteria. To reduce selection bias, chest radiographs were selected sequentially, in order of the coded patient identification numbers. One chest radiograph that did not include an image of the entire lung was removed. In the selection process, a maximum of 1 chest radiograph was selected per patient.
The nodule data set was labeled by 2 radiologists (K.H.K., with 6 years of experience, and M.K., with 21 years of experience). Each radiologist independently evaluated 10 chest radiographs for the presence of noncalcified nodules that were 4 mm or greater on the image level. To improve the accuracy of the radiologic evaluation, each patient’s cancer characteristics, all available sequential chest radiographs, and NLST radiologist labels for the chest radiographs were provided during annotation. A chest radiograph was labeled positive for noncalcified nodules if the 2 radiologists agreed on the presence of noncalcified nodules and negative if the 2 radiologists did not agree. The final labels generated in this study are publicly available.24

Lung cancer was considered to be present if cancer was diagnosed during the interval from 1 chest radiography screening to the next screening or 1 year after the last chest radiography screening, whichever occurred first.3,5 A chest radiograph was labeled positive for malignant pulmonary nodules if it had positive findings for the presence of both lung cancer and noncalcified nodules. The 2 radiologists (K.H.K. and M.K.) retrospectively reviewed cancer cases that were not labeled malignant pulmonary nodules and recorded whether any visible radiologic findings (other than pulmonary nodules) suggestive of lung cancer were present. An image was labeled positive for other visible radiologic findings if the 2 radiologists agreed on the presence of the findings.

Algorithm for Analysis
We used a commercially available AI algorithm (Lunit INSIGHT CXR) to analyze the initial chest radiographs. The AI algorithm we selected is a deep convolutional neural network that uses residual neural network 34 (ResNet-34) as its foundational architecture.25 In the preprocessing step, the raw pixel map of the Digital Imaging and Communications in Medicine file was normalized with windowing information, and the normalized pixel map was used as input for the AI model. The model then produced a coarse probability map with multiple channels, with an array value that defined the probability of the target lesions being present, and was generated using a weakly supervised object localization technique.26 Each channel defined the probability map for each target lesion, and the probability map was used to localize possible abnormal regions for the target lesion. The image-level probability score was then max pooled (a technique that reduces the dimensionality of images by reducing the number of pixels in the output from the previous convolutional layer) from the output probability map to generate a final abnormality score between 0 and 100. The commercial AI algorithm produced a probability map and score for 10 abnormalities; however, only output corresponding to noncalcified nodules was used in this study.

During training of the model, we used augmentation policies from AutoAugment (Google Brain), which defines information about which image processing operations to use, such as translation, rotation, or shearing, and the magnitude of the operations.27 When image-level labels were available, binary cross-entropy loss was computed using image-level probability scores and image-level labels; when ground truth annotation maps (ie, pixelwise lesion contour–label maps used for supervised training) were available, the binary cross-entropy loss was computed using both probability maps and annotation maps.

The training set included 12,408 abnormal images with lung nodules or masses that were obtained from patients with pathologically proven and/or radiologically confirmed diseases, which were verified by at least 1 of 15 board-certified radiologists (with 7-14 years of experience), and 72,704 normal images, all of which were obtained from multiple hospitals in South Korea.28 The training images included both digital and computed radiographs, and none of the NLST data were used during the training of the model. A more detailed description of the data collection process and the development of the AI algorithm can be found in a previous study.9

Assessment of Performance
Images analyzed by the AI algorithm were considered positive if the abnormality score produced by the AI algorithm for the entire image was higher than the operating point. The operating point was used to classify the presence of noncalcified nodules at the image level. In the nodule data set, both
malignant and benign noncalcified nodules were labeled positive, and detection of either type of nodule was regarded as true positive. In our analysis, the operating point of the AI model was set at 15.0, which was chosen using Youden criteria in the internal validation set. At this operating point, the AI model had sensitivity of 96.6% and specificity of 94.1% for the classification of abnormal images in the internal validation set. To assess the performance of NLST radiologists, we used 2 labels provided in the NLST data set: the nodule label, which recorded the presence of noncalcified nodules or masses, and the cancer label, which recorded nodule(s) of 4 mm or greater or enlarging nodule(s), mass(es), or other nonspecific abnormalities suggestive of lung cancer.

The change in the performance of the AI algorithm compared with the performance of the NLST radiologists for the detection of nodules, lung cancer, and malignant pulmonary nodules was analyzed on balanced test data sets. In the nodule data set, nonnodules were randomly selected to have 1:1, 1:2, and 1:3 nodule to nonnodule ratios. In the full TO data set, noncancer and nonmalignant pulmonary nodules were randomly selected to have 1:1, 1:2, and 1:3 cancer to noncancer ratios and malignant pulmonary nodule to nonmalignant pulmonary nodule ratios, respectively. Nodule labels were used to assess NLST radiologist performance for nodule detection; cancer labels were used for cancer and malignant nodule detection.

**Statistical Analysis**

Receiver operating characteristic (ROC) analyses were performed to evaluate the classification performance of the AI algorithm in the nodule data set. For comparison of the sensitivities and specificities of the AI algorithm vs the NLST radiologists, the McNemar test was used. For comparison of positive predictive values (PPVs) and negative predictive values (NPVs) of the AI algorithm vs NLST radiologists, the generalized score statistic was used.29 The performance of the AI algorithm vs the NLST radiologists was compared in all radiographs, digital radiographs, and computed radiographs. The agreement for noncalcified nodule annotation between each ground truth annotator and the NLST radiologists was assessed with weighted κ using linear weighting. We computed 95% CIs for the performance and agreement measures with 10,000 bootstrap replications. For all tests, \( P < .05 \) was considered statistically significant. All statistical analyses were conducted using R software, version 3.6.1 (R Foundation for Statistical Computing). Analyses were performed between August 20, 2019, and February 14, 2020.

**Results**

After excluding 1 participant with only a lateral chest radiograph and 5 participants with corrupted Digital Imaging and Communications in Medicine files, a total of 5485 patients (mean [SD] age, 61.7 [5.0] years; 3030 men [55.2%]; median follow-up duration, 6.5 years [interquartile range, 6.1-6.9 years]) were included. Of the 5485 participants with valid TO chest radiographs selected for further analysis, 48 patients (0.9%) received a diagnosis of lung cancer within 1 year of the baseline imaging. A subset of 577 participants was selected for the nodule data set. The demographic characteristics of participants with valid TO posteroanterior images and participants selected for the nodule data set are shown in Table 1.

Of the 48 participants who received a cancer diagnosis within 1 year of baseline imaging, 34 participants had visible malignant nodules present in their chest radiographs. Among the other 14 participants, 11 participants had no visible lesions present in their chest radiographs, and 3 participants had other radiologic manifestations that were suggestive of lung cancer, including atelectasis (n = 1), pleural thickening (n = 1), and hilar lymphadenopathy (n = 1).

The agreement between each ground truth annotator and the NLST radiologists for noncalcified nodule annotation was assessed. Moderate agreement was observed between each ground truth annotator (\( \kappa = 0.55; 95\% \text{ CI}, 0.47-0.64 \)) and the NLST radiologists (\( \kappa = 0.60; 95\% \text{ CI}, 0.52-0.69 \)).
Nodule Detection Performance
The performance of the AI algorithm vs the NLST radiologists for the detection of noncalcified nodules in the nodule data set was assessed. The area under the ROC curve (AUROC) of the AI algorithm was 0.93 (95% CI, 0.90-0.96) for all chest radiographs, 0.99 (95% CI, 0.97-1.00) for digital radiographs, and 0.86 (95% CI, 0.79-0.93) for computed radiographs (Figure 1). The sensitivity and specificity pairs of the NLST radiologists (indicated by Xs in Figure 1) were under the ROC curve for digital radiographs and above the ROC curve for computed radiographs.

The differences between the AI algorithm and the NLST radiologists in both sensitivity (86.2% [95% CI, 77.8%-94.6%] vs 87.7% [95% CI, 79.7%-95.7%], respectively;  with respect to specificity, 85.0% [95% CI, 81.9%-88.1%] vs 86.7% [83.8%-89.7%];  were statistically nonsignificant in all chest radiographs at the operating point chosen from the internal validation set. The nodule detection performance at other operating points is shown in Figure 1 in the Supplement.

The sensitivity and specificity of the AI algorithm were higher compared with those of the NLST radiologists for digital radiographs (for sensitivity, 96.0% [95% CI, 88.3%-100.0%] vs 88.0% [95% CI, 75.3%-100.0%], respectively;  for specificity, 93.2% [95% CI, 89.9%-96.5%] vs 82.8% [95% CI, 77.8%-87.8%];  but were lower compared with those of the NLST radiologists for computed radiographs (for sensitivity, 77.8% [95% CI, 64.2%-91.4%] vs 86.1% [95% CI, 74.8%-97.4%];  for specificity, 78.8% [95% CI, 73.9%-83.8%] vs 90.4% [95% CI, 86.8%-94.0%];  ). Of the 65 total noncalcified nodules or masses present in the nodule data

### Table 1. Participant Characteristics

| Characteristic                        | No. (%)          |         |         |
|--------------------------------------|------------------|---------|---------|
|                                      | Full T0 data set | Nodule data set |
| Total participants, No.              | 5485             | 577     |
| Participants with cancer             | 48 (0.9)         | 48 (8.3)|         |
| Age, mean (SD)                       | 61.7 (5.0)       | 62.2 (5.1)|       |
| Sex                                  |                  |         |         |
| Male                                 | 3030 (55.2)      | 322 (55.8)|       |
| Female                               | 2455 (44.8)      | 255 (44.2)|       |
| Race                                 |                  |         |         |
| White                                | 5145 (93.8)      | 539 (93.4)|       |
| Black or African American            | 222 (4.0)        | 22 (3.8)|         |
| Asian                                | 39 (0.7)         | 5 (0.9)|         |
| American Indian or Alaskan Native    | 17 (0.3)         | 3 (0.5)|         |
| Native Hawaiian or other Pacific Islander | 1 (0.02)    | 0       |         |
| >1 race                              | 36 (0.7)         | 6 (1.0)|         |
| Unavailable                          | 25 (0.5)         | 2 (0.4)|         |
| Ethnicity                            |                  |         |         |
| Hispanic or Latino                   | 77 (1.4)         | 5 (0.9)|         |
| Not Hispanic or Latino               | 5385 (98.2)      | 572 (99.1)|       |
| Unavailable                          | 23 (0.4)         | 0       |         |
| Smoking status                       |                  |         |         |
| Former                               | 2765 (50.4)      | 306 (53.0)|       |
| Current                              | 2720 (49.6)      | 271 (47.0)|       |
| Type of chest radiography used       |                  |         |         |
| Screen film                          | 47 (0.9)         | 4 (0.7)|         |
| Computed                             | 2861 (52.2)      | 296 (51.3)|       |
| Digital                              | 2108 (38.4)      | 246 (42.6)|       |
| Thoravision                          | 465 (8.5)        | 31 (5.4)|         |
| Unavailable                          | 4 (0.07)         | 0       |         |
| Outcomes                             |                  |         |         |
| Follow-up, median (IQR), y           | 6.5 (6.1-6.9)    | 6.5 (6.1-6.8)|       |
| Mortality                            | 380 (6.9)        | 53 (9.2)|         |

Abbreviations: IQR, interquartile range; T0, baseline.
set, 56 nodules or masses were detected by the AI algorithm (including 7 nodules or masses that were missed by NLST radiologists), 57 nodules or masses were detected by NLST radiologists (including 8 nodules or masses that were missed by the AI algorithm), 49 nodules or masses were detected by both, and 1 nodule or mass was missed by both.

**Cancer Detection Performance**

The performance of the AI algorithm compared with the NLST radiologists for the detection of all cancers and malignant pulmonary nodules in the nodule data set and the full T0 data set is shown in Table 2 and eTable 1 in the Supplement. The sensitivity, specificity, PPV, and NPV of the AI algorithm were 75.0% (95% CI, 62.8%-87.2%), 83.3% (95% CI, 82.3%-84.3%), 3.8% (95% CI, 2.6%-5.0%), and 99.8% (95% CI, 99.6%-99.9%) for the detection of all cancers in all chest radiographs of the full T0 data set. In digital radiographs of the full T0 data set, the AI algorithm and the NLST radiologists (as assessed by the cancer label) had similar sensitivity (76.0% [95% CI, 59.3%-92.7%] vs 80.0% [95% CI, 64.3%-95.7%], respectively; P = .65), similar specificity (90.0% [95% CI, 89.7%-92.2%] vs 91.1% [95% CI, 89.9%-92.3%]; P = .82), similar PPV (9.1% [95% CI, 5.2%-13.0%] vs 9.8% [95% CI, 5.7%-13.8%]; P = .62), and similar NPV (99.7% [95% CI, 99.4%-99.9%] vs 99.7% [95% CI, 99.5%-100.0%]; P = .65) for cancer detection. In computed radiographs of the full T0 data set, the AI algorithm had lower sensitivity (68.4% [95% CI, 47.5%-89.3%] vs 89.5% [95% CI, 75.7%-100.0%]; P = .10), lower specificity (76.7% [95% CI, 75.2%-78.3%] vs 91.4% [95% CI, 90.3%-92.4%]; P < .001), lower PPV (1.9% [95% CI, 0.9%-3.0%] vs 6.3% [95% CI, 3.5%-9.5%]; P < .001), and similar NPV (99.7% [95% CI, 99.5%-99.9%] vs 99.9% [95% CI, 99.8%-100.0%]; P = .07) compared with the NLST radiologists. Among all images of the 48 participants who received lung cancer diagnoses within 1 year of screening, 36 cases were detected by the AI algorithm, 41 cases were detected by both, and 4 cases were missed by both.

In all radiographs of the full T0 data set, the sensitivity, specificity, PPV, and NPV of the AI algorithm were 94.1% (95% CI, 86.2%-100.0%), 83.3% (95% CI, 82.3%-84.3%), 3.4% (95% CI, 2.2%-4.5%), and 100.0% (95% CI, 99.9%-100.0%), respectively, for the detection of malignant pulmonary nodules. In digital radiographs of the full T0 data set, the AI algorithm had higher sensitivity (100.0% [95% CI, 100.0%-100.0%] vs 94.1% [95% CI, 82.9%-100.0%]; P = .32), similar specificity (90.9% [95% CI, 89.6%-92.1%] vs 91.0% [95% CI, 89.7%-92.2%]; P = .91), similar PPV

![Figure 1. Receiver Operating Characteristic Curve of the Performance of the Artificial Intelligence Algorithm vs NLST Radiologists for the Detection of Noncalcified Nodules in the Nodule Data Set](image-url)
| Variable                      | All images | Digital radiographic images | Computed radiographic images | P value |
|------------------------------|------------|----------------------------|-----------------------------|---------|
|                              |            | AI vs NLST nodule | AI vs NLST cancer | AI vs NLST nodule | AI vs NLST cancer | AI vs NLST nodule | AI vs NLST cancer | AI vs NLST nodule | AI vs NLST cancer | AI vs NLST nodule | AI vs NLST cancer |
| Sensitivity                  |            |                       |                            |         |
| (all cancer detection)       |            |                       |                            |         |
| Nodule data set              | 75.0       | (62.8-87.2)            | 85.4                       | (75.4-95.4) | .78 | .13 |
|                              | 76.0       | (59.3-92.7)            | 80.0                       | (64.3-95.7) | .41 | .65 |
| Full T0 data set             | 75.0       | (62.8-87.2)            | 85.4                       | (75.4-95.4) | .78 | .13 |
|                              | 76.0       | (59.3-92.7)            | 80.0                       | (64.3-95.7) | .41 | .65 |
| Sensitivity                  |            |                       |                            |         |
| (malignant pulmonary nodule detection) |         |                        |                            |         |
| Nodule data set              | 94.1       | (86.2-100.0)           | 94.1                       | (86.2-100.0) | .65 | >.99 |
|                              | 100.0      | (100.0-100.0)          | 94.1                       | (82.9-100.0) | .16 | .32 |
| Full T0 data set             | 94.1       | (86.2-100.0)           | 94.1                       | (82.0-100.0) | .65 | >.99 |
|                              | 100.0      | (100.0-100.0)          | 94.1                       | (82.0-100.0) | .16 | .32 |
| Specificity                  |            |                       |                            |         |
| (malignant pulmonary nodule detection) |         |                        |                            |         |
| Nodule data set              | 81.4       | (78.4-85.0)            | 82.7                       | (79.5-85.9) | .56 | .56 |
|                              | 80.4       | (78.4-85.0)            | 81.2                       | (79.5-85.9) | .002 | .005 |
| Full T0 data set             | 83.3       | (82.3-84.3)            | 91.1                       | (90.3-91.8) | <.001 | <.001 |
|                              | 90.9       | (89.7-92.2)            | 91.0                       | (89.7-92.2) | <.001 | <.001 |

Abbreviations: AI, artificial intelligence; NLST, National Lung Screening Trial; NLST cancer, National Lung Screening Trial radiologists using cancer label; NLST nodule, National Lung Screening Trial radiologists using nodule label; T0, baseline.
Validation of Deep Learning Algorithm for Malignant Pulmonary Nodule Detection in Chest Radiographs

(8.2% [95% CI, 4.4%-11.9%] vs 7.8% [95% CI, 4.1%-11.5%]; \( P = .65 \)), and similar NPV (100.0% [95% CI, 100.0%-100.0%] vs 99.9% [95% CI, 99.8%-100.0%]; \( P = .32 \)) compared with the NLST radiologists (as assessed by the cancer label). In computed radiographs of the full TO data set, the AI algorithm had lower sensitivity (85.7% [95% CI, 67.4%-100.0%] vs 92.9% [95% CI, 79.4%-100.0%]; \( P = .56 \)), lower specificity (76.7% [95% CI, 75.2%-78.3%] vs 91.3% [95% CI, 90.2%-92.3%]; \( P < .001 \)), lower PPV (1.8% [95% CI, 0.8%-2.8%] vs 5.0% [95% CI, 2.3%-7.6%]; \( P < .001 \)), and similar NPV (99.9% [95% CI, 99.8%-100.0%] vs 100.0% [95% CI, 99.9%-100.0%]; \( P = .48 \)) compared with the NLST radiologists. In all images of the 34 patients with malignant pulmonary nodules who received lung cancer diagnoses within 1 year after imaging, 32 cases were detected by the AI algorithm, 32 cases were detected by the NLST radiologists, 30 cases were detected by both, and 0 cases were missed by both.

The performance of the AI algorithm (as measured by AUROC, sensitivity, and specificity) and the performance of the NLST radiologists (as measured by sensitivity and specificity) remained consistent at different ratios (1:1, 1:2, and 1:3) of control images for all tasks (eTable 2 in the

Figure 2. Frontal Chest Radiographs of Patients With Malignant Pulmonary Nodules Missed by NLST Radiologists But Detected by Artificial Intelligence Algorithm

A, Chest radiograph of woman in her 60s (without AI detection). The woman was diagnosed with lung cancer 86 days after baseline imaging. B, Chest radiograph of woman in her 60s (with AI detection). The AI algorithm detected the missed subtle abnormality (in green, with nodule score of 38%) in the left perihilar region. C, Chest radiograph of man in his 50s (without AI detection). The man was diagnosed with lung cancer 127 days after baseline imaging. D, Chest radiograph of man in his 50s (with AI detection). The AI algorithm detected the missed subcentimeter nodule (in green, with nodule score of 53%) in the right upper lung zone. AI indicates artificial intelligence; and NLST, National Lung Screening Trial.
Supplement). The small variations in the performance of the AI algorithm and the NLST radiologists at different control group ratios were all within the 95% CI of the performance in the nodule data set and the full TO data set for all tasks.

**Discussion**

In this study, we applied an AI algorithm, which was originally trained for pulmonary nodule detection, for the detection of lung cancer on chest radiographs. In the nodule data set, the sensitivity and specificity of the AI algorithm for nodule detection were 86.2% and 85.0%, respectively. When the same AI algorithm was applied for cancer detection, the sensitivity (76.0% vs 80.0%) and PPV (9.1% vs 9.8%) of the AI algorithm were similar to those of the NLST radiologists for the detection of all cancers. In digital radiographs of the full TO data set, the sensitivity (100.0% vs 94.1%) and PPV (8.2% vs 7.8%) were also similar to those of the NLST radiologists for the detection of malignant pulmonary nodules. We also observed that the performance of the AI algorithm remained consistent even with changes in the ratio of control images. Although the sensitivity of the AI algorithm for lung cancer detection in this study is lower than the sensitivity reported for low-dose CT (93.8%), it is similar to the reported sensitivity of radiologists for chest radiographs (73.5%).

The nodule data set used in this study included chest radiographs of both benign and malignant nodules as well as chest radiographs of other abnormalities, such as consolidation and emphysema. The data set was designed to reflect the distribution of chest radiographs that may be encountered in a cancer screening setting. Despite our nodule data set design, the AI algorithm retained high performance for nodule detection. In our study, the overall AUROC for nodule detection was 0.93 (95% CI, 0.90-0.96), which is similar to the AUROCs of 0.92 to 0.99 reported by Nam et al and the AUROC of 0.91 reported by Majkowska et al in a study of data from the National Institutes of Health Chest x-ray Data Set of 14 Common Thorax Disease Categories.

Notably, we also observed that the AI algorithm performed better in digital radiographs (AUROC, 0.99) than in computed radiographs (AUROC, 0.86). Although the data used to train the AI model included both digital and computed radiographs, the inferior quality of the older computed radiographs included in the nodule data set may have been unencountered during training, which would account for the decreased performance of the AI algorithm in computed radiographs. As a consequence, the AI model could not equal the performance of the NLST radiologists for the detection of noncalcified nodules in computed radiographs.

In this study, the AI algorithm had lower PPV and similar NPV compared with the NLST radiologists for the detection of all cancers and malignant pulmonary nodules. In older computed radiographs, the AI algorithm frequently provided false-positive readings, which produced low PPV. However, on the more contemporary digital radiographs, the AI algorithm performed better than the NLST radiologists with regard to sensitivity and had a PPV similar to that of the NLST radiologists for the detection of all cancers and malignant pulmonary nodules. This finding suggests that the AI results should be interpreted with caution when the algorithm is performed on a test set with characteristics different from those of the training set.

One of the most important contributions of our research is its assessment of the performance of an AI algorithm that was originally designed to detect pulmonary nodules for the detection of cancer using a data set from the NLST, the multicenter randomized clinical trial on which the selection criteria for current lung cancer screenings via imaging are based. Although the prevalence of lung cancer may be higher in the NLST data set compared with the general population, the test set included a spectrum of diseases that may be encountered in clinical practice.

There was a moderate decrease in the detection performance of the AI algorithm when it was applied for the detection of any lung cancer, but the AI algorithm had high performance for the detection of malignant pulmonary nodules. Retrospective radiologic assessment of all lung cancer cases in the full TO data set indicated that only 34 of 48 all-cancer cases presented as malignant nodules. Of the remaining 14 cases, 11 cases had no visible lesions, and 3 cases had other radiologic...
manifestations, such as hilar lymphadenopathy, pleural abnormalities, and sublobar atelectasis, as described in the literature. Because the AI algorithm used in this study is designed to specifically detect lung nodules or masses, the algorithm likely missed lung cancer cases that did not present as malignant pulmonary nodules.

In the digital radiographs of the full TO data set, the sensitivity of the AI algorithm for malignant pulmonary nodule detection was greater than that of the NLST radiologists, as assessed by both nodule and cancer labels. In addition, although the AI model had lower sensitivity compared with the NLST radiologists for the detection of all cancers, the sensitivity remained greater than that of the NLST radiologists, as assessed by the nodule label. Such results suggest that the AI algorithm has excellent sensitivity for detecting not only noncalcified nodules but also malignant pulmonary nodules, even performing better than radiologists.

In this study, modest agreement was observed between the NLST radiologists and the ground truth annotators. Given the interreader variability of pulmonary nodule detection in chest radiographs, we believe that the likelihood of mislabeling by radiologists will be high in instances when only a single reader interprets a chest radiograph. A recent study by Majkowska et al indicated that a substantial number of nonoverlapping true-positive findings exist for various types of lesions detected by an AI algorithm and radiologists. This finding is consistent with that of our study, which found that several noncalcified nodules and malignant pulmonary nodules that were missed by radiologists were detected only by the AI algorithm and vice versa. Of the 65 total noncalcified nodules or masses present in the nodule data set, 8 nodules or masses were missed by the NLST radiologists, 7 of which were also detected only by the AI algorithm. In addition, of the 34 total malignant pulmonary nodules, 2 nodules missed by the NLST radiologists were detected only by the AI algorithm (Figure 2). These findings underscore the value of an AI algorithm as a second reader during the interpretation of chest radiographs for lung nodules.

Limitations
Our study has several limitations. First, although NLST included a community cohort of participants with a high risk of lung cancer, the prevalence of lung cancer remained low, and only 48 participants with lung cancer were available for inclusion in the present analysis. Because of the small number of participants with cancer, it was difficult to achieve statistical significance for the differences in sensitivity between the AI algorithm and the NLST radiologists. Second, because NLST data were collected between 2002 and 2004, the quality of the chest radiographs might have been inferior to those obtained using modern equipment; this inferiority may have led to underperformance of the AI algorithm (particularly for computed radiographs), which was trained on chest radiographs that were all obtained after 2010. The decrease in performance was especially pronounced for computed radiographs; the 2 malignant nodules missed by the AI algorithm but detected by the NLST radiologists were both from computed radiographs (eFigure 2 in the Supplement).

Third, the term NLST radiologist was used to refer to a pool of radiologists who participated in the NLST, and the actual performance of individual radiologists may be different from the pooled performance of NLST radiologists. Fourth, ground truth labels of nodules were generated without paired CT images; thus, the labels may have been inaccurate. Fifth, we did not assess the incremental value of the AI algorithm as a second reader by conducting a prospective study that comprised separate sets of radiologists, who may have had different performance than the NLST radiologists. Sixth, chest radiographs are no longer recommended for lung cancer screening; therefore, the results of this study might not be implemented for lung cancer screening.

Conclusions
The AI algorithm performed better than the NLST radiologists for the detection of all noncalcified nodules and malignant pulmonary nodules on digital radiographs. The AI algorithm may help to detect lung cancer by detecting additional malignant pulmonary nodules.
ARTICLE INFORMATION
Accepted for Publication: July 8, 2020.
Published: September 24, 2020. doi: 10.1001/jamanetworkopen.2020.17135
Open Access: This is an open access article distributed under the terms of the CC-BY License. © 2020 Yoo H et al.
JAMA Network Open.

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

Author Affiliations: Lunit, Seoul, South Korea (Yoo, Kim); Division of Thoracic Imaging, Department of Radiology, Massachusetts General Hospital, Boston (Singh, Digumarthy, Kalra); Harvard Medical School, Boston, Massachusetts (Singh, Digumarthy, Kalra).

Author Contributions: Drs Yoo and Kalra had full access to all of the data in the study and take responsibility for the integrity of the data and the accuracy of the data analysis. Drs Yoo and Kim contributed equally to this work.

Concept and design: Yoo, Kim, Kalra.

Acquisition, analysis, or interpretation of data: Yoo, Singh, Digumarthy, Kalra.

Drafting of the manuscript: All authors.

Critical revision of the manuscript for important intellectual content: Yoo, Singh, Digumarthy, Kalra.

Statistical analysis: Yoo, Kim, Singh.

Administrative, technical, or material support: Yoo, Kim, Singh.

Supervision: Yoo, Kim, Digumarthy, Kalra.

Conflict of Interest Disclosures: Dr Yoo reported receiving personal fees from Lunit during the conduct of the study. Dr Kim reported receiving personal fees from Lunit during the conduct of the study. Dr 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; receiving grants from Lunit; and receiving honoraria from Siemens outside the submitted work. Dr Kalra reported receiving grants from Riverain Technologies and Siemens Healthineers outside the submitted work. No other disclosures were reported.

REFERENCES
1. de Koning HJ, van der Aalst CM, de Jong PA, et al. Reduced lung-cancer mortality with volume CT screening in a randomized trial. N Engl J Med. 2020;382(6):503-513. doi: 10.1056/NEJMoa1911793
2. Aberle DR, Adams AM, Berg CD; National Lung Screening Trial Research Team. Reduced lung-cancer mortality with low-dose computed tomographic screening. N Engl J Med. 2011;365(5):395-409. doi:10.1056/NEJMoa1102873
3. Aberle DR, DeMello S, Berg CD, et al; National Lung Screening Trial Research Team. Results of the two incidence screenings in the National Lung Screening Trial. N Engl J Med. 2013;369(10):920-931. doi:10.1056/NEJMoa1208962
4. De Koning H, Van Der Aalst C, Ten Haaf K, Oudkerk M. PL02.05 Effects of volume CT lung cancer screening: mortality results of the NELSON randomised-controlled population based trial. J Thorac Oncol. 2018;13(10):S185. doi:10.1016/j.jtho.2018.08.012
5. Church TR, Black WC, Aberle DR, et al; National Lung Screening Trial Research Team. Results of initial low-dose computed tomographic screening for lung cancer. N Engl J Med. 2013;368(21):1980-1991. doi:10.1056/NEJMoa1209120
6. de Hoop B, Schaef Prolup C, Gietema HA, et al. Screening for lung cancer with digital chest radiography: sensitivity and number of secondary work-up CT examinations. Radiology. 2010;255(2):629-637. doi:10.1148/ radiol.09091308
7. Bach PB, Mirkin JN, Oliver TK, et al. Benefits and harms of CT screening for lung cancer: a systematic review. JAMA. 2012;307(22):2418-2429. doi:10.1001/jama.2012.5521
8. Ahmad U, Dettterbeck FC. Current status of lung cancer screening. Semin Thorac Cardiovasc Surg. 2012;24(1):27-36. doi:10.1053/j.semtcvs.2012.01.014
9. Nam JG, Park S, Hwang EJ, et al. Development and validation of deep learning-based automatic detection algorithm for malignant pulmonary nodules on chest radiographs. Radiology. 2019;290(1):218-228. doi: 10.1148/ radiol.2018180237
10. Gavelli G, Giampalma E. Sensitivity and specificity of chest X-ray screening for lung cancer: review article. Cancer. 2000;89(11 Suppl):2453-2456. doi:10.1002/1097-0142(20001201)89:11+<2453::AID-CNCR21>3.0.CO;2-M

11. 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. doi: 10.1148/radiology.148.3.6308709

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

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

14. Quadrelli S, Lyons G, Colt H, Chimondeguy D, Buero A. Clinical characteristics and prognosis of incidentally detected lung cancers. Int J Surg Oncol. 2015;2015:287604. doi:10.1155/2015/287604

15. Ausawalithong W, Thirach A, Marukatat S, Wilaiprasitporn T. Automatic lung cancer prediction from chest x-ray images using the deep learning approach. In: 2018 11th Biomedical Engineering International Conference (BMEICON). Institute of Electrical and Electronics Engineers; 2018:1-5. doi:10.1109/BMEICON.2018.8609997

16. Tataru C, Yi D, Shenoy A, Ma A. Deep learning for abnormality detection in chest x-ray images. Stanford University; 2017. Accessed September 4, 2020. http://cs231n.stanford.edu/reports/2017/pdfs/527.pdf

17. Gang P, Zhen W, Zeng W, et al. Dimensionality reduction in deep learning for chest X-ray analysis of lung cancer. In: 2018 10th International Conference on Advanced Computational Intelligence (ICACI). Institute of Electrical and Electronics Engineers; 2018:878-883. doi:10.1109/ICACI.2018.8377579

18. Hwang EJ, Nam JG, Lim WH, et al. Deep learning for chest radiograph diagnosis in the emergency department. Radiology. 2019;293(3):573-580. doi:10.1148/radiol.2019191125

19. Rajpurkar P, Irvin J, Ball RL, et al. Deep learning for chest radiograph diagnosis: a retrospective comparison of the CheXNet algorithm to practicing radiologists. PLoS Med. 2018;15(11):e1002586. doi:10.1371/journal.pmed.1002586

20. Cicero M, Bilbily A, Colak E, et al. Training and validating a deep convolutional neural network for computer-aided detection and classification of abnormalities on frontal chest radiographs. Invest Radiol. 2017;52(5):281-287. doi:10.1097/RLI.0000000000000341

21. He K, Zhang M, Cotter 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. doi:10.1148/radiol.2019182465

22. Aberle DR, Berg CD, Black WC, et al; National Lung Screening Trial Research Team. The National Lung Screening Trial: overview and study design. Radiology. 2011;258(1):243-253. doi:10.1148/radiol.10091808

23. Patz EF, Jr, Greco E, Gatsonis C, Pinsky P, Kramer BS, Aberle DR. Lung cancer incidence and mortality in National Lung Screening Trial participants who underwent low-dose CT prevalence screening: a retrospective cohort analysis of a randomised, multicentre, diagnostic screening trial. Lancet Oncol. 2016;17(5):590-599. doi:10.1016/S1470-2045(15)00621-X

24. NLST nodule detection. GitHub. Accessed March 29, 2020. https://github.com/lunit-io/nlst-nodule-detection

25. Leisenring W, Alonzo T, Pepe MS. Comparisons of predictive values of binary medical diagnostic tests for paired designs. Biometrics. 2000;56(2):345-351. doi:10.1111/j.0006-341X.2000.00345.x
30. Majkowska A, Mittal S, Steiner DF, et al. Chest radiograph interpretation with deep learning models: assessment with radiologist-adjudicated reference standards and population-adjusted evaluation. *Radiology*. 2020;294(2):421-431. doi:10.1148/radiol.2019191293

31. Purandare NC, Rangarajan V. Imaging of lung cancer: implications on staging and management. *Indian J Radiol Imaging*. 2015;25(2):109-120. doi:10.4103/0971-3026.155831

**SUPPLEMENT.**

eTable 1. Comparison of Positive Predictive Values and Negative Predictive Values of AI Algorithm vs NLST Radiologists in Full TO Data Set

eTable 2. Performance of AI Algorithm vs NLST Radiologists in Balanced Test Set

eFigure 1. Nodule Detection Performance at Various Operating Points in the Nodule Data Set

eFigure 2. Frontal Chest Radiographs of Patients With Malignant Pulmonary Nodules That Were Found by NLST Radiologists but Missed by the AI Algorithm