

# Development and Validation of Deep Learning–based Automatic Detection Algorithm for Malignant Pulmonary Nodules on Chest Radiographs



Ju Gang Nam, MD\* • Sunggyun Park, PhD\* • Eui Jin Hwang, MD • Jong Hyuk Lee, MD • Kwang-Nam Jin, MD, PhD • Kun Young Lim, MD, PhD • Thienkai Huy Vu, MD, PhD • Jae Ho Sohn, MD • Sangbeum Hwang, PhD • Jin Mo Goo, MD, PhD • Chang Min Park, MD, PhD

From the Department of Radiology and Institute of Radiation Medicine, Seoul National University Hospital and College of Medicine, 101 Daehak-ro, Jongno-gu, Seoul 03080, Republic of Korea (J.G.N., E.J.H., J.M.G., C.M.P.); Lunit Incorporated, Seoul, Republic of Korea (S.P.); Department of Radiology, Armed Forces Seoul Hospital, Seoul, Republic of Korea (J.H.L.); Department of Radiology, Seoul National University Boramae Medical Center, Seoul, Republic of Korea (K.N.J.); Department of Radiology, National Cancer Center, Goyang, Republic of Korea (K.Y.L.); Department of Radiology and Biomedical Imaging, University of California, San Francisco, San Francisco, Calif (T.H.V., J.H.S.); and Department of Industrial & Information Systems Engineering, Seoul National University of Science and Technology, Seoul, Republic of Korea (S.H.). Received January 30, 2018; revision requested March 20; revision received July 29; accepted August 6. **Address correspondence to C.M.P.** (e-mail: [cmpark.morphius@gmail.com](mailto:cmpark.morphius@gmail.com)).

Study supported by SNUH Research Fund and Lunit (06–2016–3000) and by Seoul Research and Business Development Program (FI170002).

\*J.G.N. and S.P. contributed equally to this work.

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

Radiology 2018; 00:1–11 • <https://doi.org/10.1148/radiol.2018180237> • Content codes:  

**Purpose:** To develop and validate a deep learning–based automatic detection algorithm (DLAD) for malignant pulmonary nodules on chest radiographs and to compare its performance with physicians including thoracic radiologists.

**Materials and Methods:** For this retrospective study, DLAD was developed by using 43 292 chest radiographs (normal radiograph–to–nodule radiograph ratio, 34 067:9225) in 34 676 patients (healthy-to-nodule ratio, 30 784:3892; 19 230 men [mean age, 52.8 years; age range, 18–99 years]; 15 446 women [mean age, 52.3 years; age range, 18–98 years]) obtained between 2010 and 2015, which were labeled and partially annotated by 13 board-certified radiologists, in a convolutional neural network. Radiograph classification and nodule detection performances of DLAD were validated by using one internal and four external data sets from three South Korean hospitals and one U.S. hospital. For internal and external validation, radiograph classification and nodule detection performances of DLAD were evaluated by using the area under the receiver operating characteristic curve (AUROC) and jackknife alternative free-response receiver-operating characteristic (JAFROC) figure of merit (FOM), respectively. An observer performance test involving 18 physicians, including nine board-certified radiologists, was conducted by using one of the four external validation data sets. Performances of DLAD, physicians, and physicians assisted with DLAD were evaluated and compared.

**Results:** According to one internal and four external validation data sets, radiograph classification and nodule detection performances of DLAD were a range of 0.92–0.99 (AUROC) and 0.831–0.924 (JAFROC FOM), respectively. DLAD showed a higher AUROC and JAFROC FOM at the observer performance test than 17 of 18 and 15 of 18 physicians, respectively ( $P < .05$ ), and all physicians showed improved nodule detection performances with DLAD (mean JAFROC FOM improvement, 0.043; range, 0.006–0.190;  $P < .05$ ).

**Conclusion:** This deep learning–based automatic detection algorithm outperformed physicians in radiograph classification and nodule detection performance for malignant pulmonary nodules on chest radiographs, and it enhanced physicians' performances when used as a second reader.

© RSNA, 2018

Online supplemental material is available for this article.

Chest radiography, one of the most common diagnostic imaging tests in medicine, is used for screening, diagnostic work-ups, and monitoring of various thoracic diseases (1,2). One of its major objectives is detection of pulmonary nodules because pulmonary nodules are often the initial radiologic manifestation of lung cancers (1,2). However, to date, pulmonary nodule detection on chest radiographs has not been completely satisfactory, with a reported sensitivity ranging between 36%–84%, varying widely according to the tumor size and study population (2–6). Indeed, chest radiography has been shown to be prone to many reading errors with low interobserver and

intraobserver agreements because of its limited spatial resolution, noise from overlapping anatomic structures, and the variable perceptual ability of radiologists. Recent work shows that 19%–26% of lung cancers visible on chest radiographs were in fact missed at their first readings (6,7). Of course, hindsight is always perfect when one knows where to look.

For this reason, there has been increasing dependency on chest CT images over chest radiographs in pulmonary nodule detection. However, even low-dose CT scans require approximately 50–100 times higher radiation dose than single-view chest radiographic examinations (8,9)

## Abbreviations

AUROC = area under the ROC curve, DLAD = deep learning–based automatic detection, FOM = figure of merit, JAFROC = jackknife alternative free-response ROC, ROC = receiver operating characteristic

## Summary

Our deep learning–based automatic detection algorithm outperformed physicians in radiograph classification and nodule detection performance for malignant pulmonary nodules on chest radiographs, and when used as a second reader, it enhanced physicians' performances.

## Implications for Patient Care

- Our deep learning–based automatic detection algorithm showed excellent detection performances on both a per-radiograph and per-nodule basis in one internal and four external validation data sets.
- Our deep learning–based automatic detection algorithm demonstrated higher performance than the thoracic radiologist group.
- When accompanied by our deep learning–based automatic detection algorithm, all physicians improved their nodule detection performances.

in addition to unavoidable limitations in accessibility and cost (10,11). Therefore, various computer-aided diagnosis techniques have been proposed, resulting in variable levels of performance. Whereas computer-aided diagnosis has shown remarkable improvements in its detection performance, it still hardly approaches that of a typical radiologist (12), and it is not yet accepted in routine clinical practice. One of the major issues with the acceptance of computer-aided diagnosis at chest radiography has been the large number of false-positive marks that need to be assessed.

Recently, deep learning technology has been adopted to address many unsolved difficult scientific and technical problems. In particular, a convolutional neural network has shown promise as a high-capacity parametric model for image analysis by using a large number of parameters derived from training data (13–16). Indeed, convolutional neural network has enabled the performance of deep learning–based visual recognition tasks in daily life, reaching near human-level object recognition performance (16). Herein, we assessed whether interpretation of medical imaging studies, particularly radiologic examinations, can be one such promising application of this deep learning technology (17,18).

The purpose of our study was to develop a deep learning–based automatic detection (DLAD) algorithm for the detection of malignant pulmonary nodules on chest radiographs, and to validate its detection performance with that of physicians including thoracic radiologists.

## Materials and Methods

For our retrospective study, ethics review and institutional review board approval were obtained from all participating institutions and the requirement for informed consent was waived. The study was supported by grants from Seoul National University Hospital (grant number 04–2016–3000), Lunit (Seoul, Korea), and the Seoul Research and Business Development Program (grant number FI170002). The authors who are not

corporate employees had control of the data and the information submitted for publication.

## Data Sets

For algorithm development, we retrospectively collected 43 292 chest radiographs obtained between January 2010 and December 2015 at our hospital (Seoul National University Hospital). These chest radiographs consisted of 34 067 normal chest radiographs without any abnormal findings from 30 784 patients (male-to-female patient ratio, 16 986:13 798; mean age for men vs women, respectively: 51.3 years  $\pm$  16.5 [standard deviation] vs 51.3 years  $\pm$  15.8) and 9225 chest radiographs with malignant pulmonary nodules (hereafter referred to as nodule chest radiographs) from 3892 patients (male-to-female patient ratio, 2244:1648; mean age for men vs women, respectively: 64.6 years  $\pm$  12.0 vs 61.1 years  $\pm$  12.1). Nodule chest radiographs were obtained from patients with malignant pulmonary nodules proven at pathologic analysis and normal chest radiographs on the basis of their radiology reports, and all chest radiographs were carefully reviewed by thoracic radiologists. All chest radiography was deidentified according to the Health Insurance Portability and Accountability Act Safe Harbor standard. Thereafter, the chest radiography data were randomly assigned into one of the following three data sets: a training data set that consisted of 42 092 chest radiographs (33 467 normal and 8625 nodule chest radiographs) to optimize network weights, a tuning data set (600 chest radiographs consisting of 300 normal and 300 nodule chest radiographs) to optimize hyperparameters, and an internal validation data set (600 chest radiographs consisting of 300 normal and 300 nodule chest radiographs) to validate the detection performance of the trained network. The patients in the three data sets were different and exclusive to each of the other data sets.

Four additional temporally or spatially independent data sets were prepared for external validation from three different Korean hospitals (Seoul National University Hospital, 181 patients with 62 normal and 119 nodule chest radiographs; Boramae Hospital, 182 patients with 59 normal and 123 nodule chest radiographs; and National Cancer Center, 181 patients with 70 normal and 111 nodule chest radiographs) and one US hospital (University of California San Francisco Medical Center, 149 patients with 60 normal and 89 nodule chest radiographs). Patients with available chest radiography and referential chest CTs performed within 1 month were included; for nodule chest radiographs, either pathologic analysis–proven or clinically confirmed malignant pulmonary nodules were included, and all normal chest radiography showed no abnormal findings at CT. All chest radiography in the external validation data sets were mutually exclusive, performed in independent patients, and did not overlap with the development data set. In establishing a reference standard for nodule presence, nodules smaller than 5 mm at CT were excluded from analysis. Radiographs with more than 3 nodules, lung consolidation, or pleural effusion obscuring lung nodules were also excluded. Every chest radiograph was obtained via posterior-anterior projection by using a digital radiography technique with various machines. The external validation data set from Seoul National University Hospital was

**Table 1: Patient Information and Nodule Characteristics from the Four External Validation Data Sets**

Characteristic	Seoul National University Hospital	Boramae Hospital	National Cancer Center	University of California San Francisco Medical Center
<b>Patient information</b>				
No. of chest radiographs	181	182	181	149
No. of normal radiographs	62	59	70	60
No. of nodule chest radiographs	119	123	111	89
Patients with nodules	119	123	111	89
No. of men	72	89	52	54
No. of women	47	34	59	35
Mean age*	53.1 ± 11.6	71.2 ± 11.2	63.0 ± 9.8	61.1 ± 7.5
No. of healthy patients	62	59	70	60
No. of men	27	26	48	30
No. of women	35	33	22	30
Mean age*	63.0 ± 12.0	53.6 ± 10.5	63.0 ± 9.8	49.2 ± 16.5
<b>Nodule information</b>				
Total no. of nodules	143	139	115	104
No. of nodules per chest radiograph				
One nodule	99	109	107	77
Two nodules	16	12	4	9
Three nodules	4	2	0	3
<b>Disease entity</b>				
Primary lung cancer	102 (71.3)	117 (84.2)	95 (82.6)	86 (82.7)
Pulmonary metastasis	41 (28.7)	22 (15.8)	20 (17.4)	18 (17.4)
<b>Diagnostic mode</b>				
Pathologically confirmed	110 (76.9)	122 (80.6)	113 (98.3)	89 (85.6)
Clinically diagnosed	33 (23.1)	17 (12.2)	2 (1.7)	15 (14.4)
Mean nodule size (cm) <sup>†</sup>	2.56 ± 1.49 (0.6–8.2)	3.96 ± 2.28 (1.0–12.0)	2.04 ± 0.69 (0.7–4.7)	3.56 ± 2.68 (0.7–17.0)
<b>No. of nodules according to size</b>				
≤1.0 cm	18 (12.6)	4 (2.9)	8 (7.0)	6 (5.8)
1.1–1.5 cm	20 (14.0)	11 (7.9)	26 (22.6)	13 (12.5)
1.6–2.0 cm	25 (17.5)	17 (12.2)	20 (17.4)	17 (16.3)
2.1–3.0 cm	31 (21.7)	33 (23.7)	59 (51.3)	21 (20.2)
>3.0 cm	49 (34.3)	74 (53.2)	2 (1.7)	47 (45.2)
<b>Lobar distribution</b>				
Right upper	35	38	40	34
Right middle	9	9	10	13
Right lower	35	33	19	8
Left upper	43	39	34	35
Left lower	21	20	12	14
<b>Transaxial location</b>				
Medial half of lung	74 (51.7)	65 (46.8)	32 (27.8)	52 (50.0)
Lateral half of lung	69 (48.3)	74 (53.2)	83 (72.2)	52 (50.0)
<b>Craniocaudal location</b>				
Superior to the carina	73 (51.0)	57 (41.0)	47 (40.9)	51 (49.0)
Between the carina and inferior pulmonary vein	60 (42.0)	50 (36.0)	45 (39.1)	24 (23.1)
Inferior to inferior pulmonary vein	10 (7.0)	32 (23.0)	23 (20.0)	29 (27.9)
Overlapped or masked by heart	11 (7.7)	13 (9.4)	4 (3.5)	1 (1.0)
Overlapped or masked by diaphragm	9 (6.3)	15 (10.8)	3 (2.6)	3 (2.9)
Overlapped or masked by hilar vessels	21 (14.7)	25 (18.0)	1 (0.9)	14 (13.5)
Overlapped or masked by clavicle/first rib	26 (18.2)	21 (15.1)	9 (7.8)	8 (7.7)

**Table 1 (continues)**

**Table 1 (continued): Patient Information and Nodule Characteristics from the Four External Validation Data Sets**

Characteristic	Seoul National University Hospital	Boramae Hospital	National Cancer Center	University of California San Francisco Medical Center
Internal characteristics at CT				
Solid	128 (89.5)	121 (87.1)	87 (75.7)	102 (98.1)
Subsolid	7 (4.9)	7 (5.0)	22 (19.1)	0 (0)
Cavitation	8 (5.6)	9 (6.5)	6 (5.2)	1 (1.0)
Calcification	0 (0)	2 (1.4)	0 (0)	1 (1.0)

Note.—Unless otherwise indicated, data are numbers of nodules or patients and data in parentheses are percentages.

\* Data are  $\pm$  standard deviation.

† Data are  $\pm$  standard deviation; data in parentheses are range.

obtained between October 2016 and January 2017, and an observer performance test was additionally performed by using this data set (Fig E1 [online]). Detailed demographic information is provided in Table 1 and acquisition technique of chest radiographs in Table E1 (online).

### Labeling and Annotation

In the developmental data sets, chest radiographs were labeled as either normal or nodule chest radiographs (image-level labeling), and the location of nodules on the nodule chest radiographs were annotated (pixel-level annotation) in 37.2% (3213 of 8625) of nodule chest radiographs in the training data set (for semisupervised learning) and in all chest radiographs from the tuning and internal validation data sets by five board-certified radiologists with 7–14 years of experience, blinded to other radiologists' labeling and annotation. Up to five nodules were annotated per chest radiograph, and a simple majority decision served as the reference standard for the presence and location of pulmonary nodules. CTs performed within 1 week were used as the reference for indeterminate radiographs.

For the four external validation data sets, four thoracic radiologists (C.M.P., G.N.J., K.Y.L., and T.H.V., each with 14–18 years of experience), one each from the four participating institutions, labeled the chest radiographs and annotated the location of the malignant nodules on chest radiographs on the basis of chest CTs performed within 1 month, and also evaluated the characteristics of the nodules including size, location, and internal characteristics. For the data set from Seoul National University Hospital, where the observer performance test was performed, the conspicuity of each lesion on the chest radiographs was decided by two thoracic radiologists (S.Y.A. and C.M.P., with 6 and 16 years of experience, respectively) in consensus by using a five-point scale: score of 5, clearly visible at initial look; score of 4, moderately visible at initial look; score of 3, possibly neglected but true nodule at retrospective focused evaluation; score of 2, 50%–70% confidence at retrospective focused evaluation; score of 1, less than 50% confidence even with review at CT (Fig E2 [online]).

### Development of DLAD

DLAD uses the pixel intensity of chest radiography as an input and then outputs its location and the presence of malignant

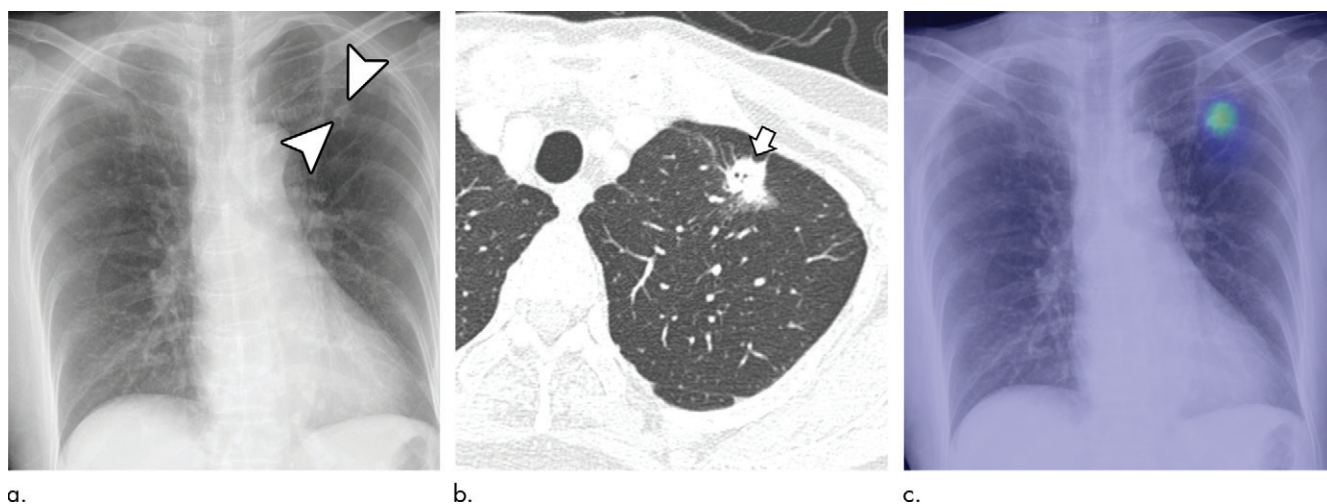
nodules. For training data set chest radiography, as previously mentioned, both the image-based information and the positional information of malignant nodules were used. Instead of a fully supervised learning algorithm, DLAD was trained in a semisupervised learning manner by using all of the image-level labels, but only part of the pixel-level annotations in the training data set: 37.2% (3213 of 8625) of nodule chest radiographs underwent pixel-level annotation. This semisupervised learning method is not only cost effective but can also enable the trained model to learn features of nodules that may be missed by radiologists.

A deep convolutional neural network with 25 layers and eight residual connections was designed, and the residual connection proposed by He et al (16) was used. Residual mapping is known to be effective in training a deeper neural network, resulting in better generalization performance. Brightness, contrast, and image size on input chest radiographs were randomly adjusted to make DLAD irrelevant to the variations. After finding a lung area by using the lung segmentation algorithm based on network-wise training, DLAD was trained to focus on the lung area. To speed up the training, the batch normalization technique was used. The outputs of three networks trained on the same data but with different hyperparameters were averaged for the final prediction. Detailed information is provided in Appendix E1 (online).

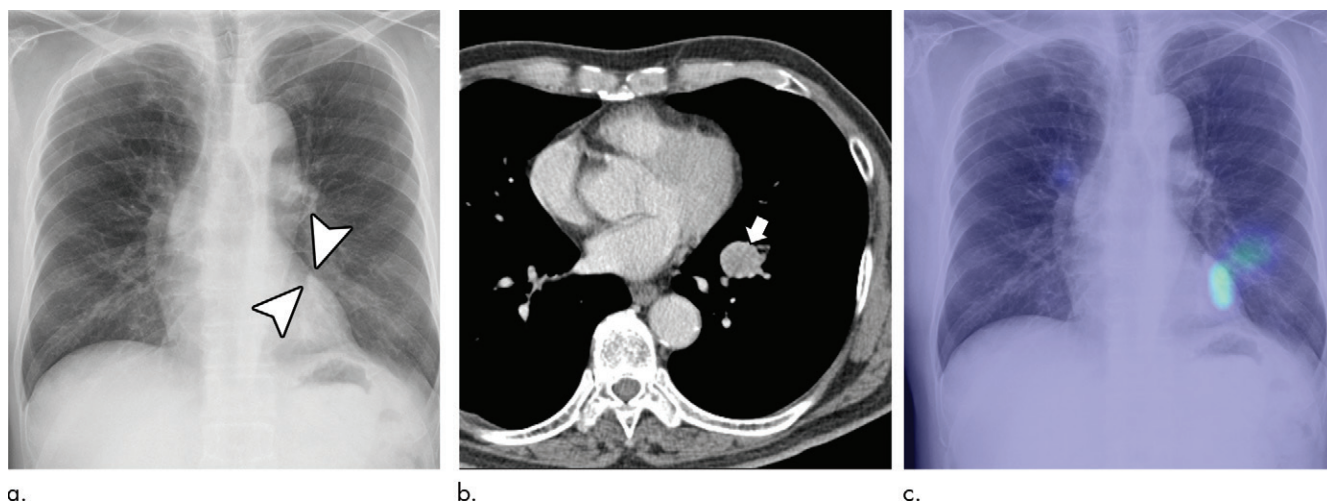
### Evaluation of DLAD

The trained DLAD generated continuous probability values between 0 and 1 for each chest radiograph corresponding to the probability that there would be nodules on the chest radiograph. In addition, DLAD produced continuous activation values ranging from 0 to 1 for each pixel of the images, derived from an activation function of the neural network. For internal and external validation, radiograph classification and nodule detection performances were evaluated on per-patient and per-nodule basis, respectively. Regarding consideration for lack of CT confirmation for normal chest radiographs in our development data set, we retrained another algorithm by using a new data set with all CT-confirmed chest radiography and evaluated its performance by using four external validation data sets (Appendix E1 [online]).





**Figure 1:** Images in a 78-year-old female patient with a 1.9-cm part-solid nodule at the left upper lobe. **(a)** The nodule was faintly visible on the chest radiograph (arrowheads) and was detected by 11 of 18 observers. **(b)** At contrast-enhanced CT examination, biopsy confirmed lung adenocarcinoma (arrow). **(c)** DLAD reported the nodule with a confidence level of 2, resulting in its detection by an additional five radiologists and an elevation in its confidence by eight radiologists.



**Figure 2:** Images in a 64-year-old male patient with a 2.2-cm lung adenocarcinoma at the left upper lobe. **(a)** The nodule was faintly visible on the chest radiograph (arrowheads) and was detected by seven of 18 observers. **(b)** Biopsy confirmed lung adenocarcinoma in the left upper lobe on contrast-enhanced CT image (arrow). **(c)** DLAD reported the nodule with a confidence level of 2, resulting in its detection by an additional two radiologists and an elevated confidence level of the nodule by two radiologists.

An observer performance test was conducted by using the data set from Seoul National University Hospital to compare the radiograph classification and nodule detection performances of DLAD with those of physicians. Eighteen physicians from four different subgroups (three nonradiology physicians, six radiology residents, five board-certified radiologists, and four subspecialty trained thoracic radiologists) participated as observers. In test 1, each observer independently reviewed each chest radiograph to discriminate normal chest radiographs from nodule chest radiographs (radiograph classification) and localized lung nodules (nodule detection) on a five-point confidence scale without DLAD as follows: 1, potential lesion with low degree of suspicion; 2, dubious lesion; 3, possible lesion with more than 50% confidence; 4, probable lesion with high confidence; and 5, definite lesion (1,2). In test 2, each observer rescored the nodules from test 1 after DLAD (Fig E3 [online]). The results of DLAD

were shown as color-coded saliency maps (Figs 1c, 2c, and Fig E3b [online]), and each observer was asked to decide whether or not to change their previous decision and rerate the score of each nodule (Appendix E1 [online]).

### Statistical Analysis

For the validation data sets, per-patient–based analysis was performed from the probability values by using the area under the receiver operating characteristic (ROC) curve (AUROC) analysis for the radiograph classification task. Per-nodule–based analysis was performed from activation values by using figures of merit (FOMs) from jackknife alternative free-response receiver-operating characteristic analysis (JAFROC version 4.2.1; <http://www.devchakraborty.com>) for the nodule detection task, defined as the probability that true lesions are rated higher than nonlesion marks on normal chest radiographs

**Table 2: Detection Performance of DLAD from the Four External Validation Data Sets**

Parameter	Radiograph Classification Performance				Nodule Detection Performance		
	AUROC	Sensitivity (%)	Specificity (%)	F1 Score (Precision, Recall)	JAFROC FOM	Sensitivity (%)*	Rate of FP Findings (%)†
Seoul National University Hospital	0.92	79.0 (94/119) [70.8, 85.4]	95 (59/62) [86.2, 98.9]	87.0 [94.9, 79.0]	0.885	69.9 (100/143) [62.0, 76.9]	0.34 [61/181]
Boramae Hospital	0.99	91.1 (112/123) [84.6, 95.1]	98 (58/59) [80.2, 100]	94.9 [99.1, 91.1]	0.924	82.0 (114/139) [74.7, 87.6]	0.30 [54/182]
National Cancer Center	0.94	71.2 (79/111) [62.1, 78.9]	100 (70/70) [93.8, 100]	83.2 [100, 71.2]	0.831	69.6 (80/115) [60.6, 77.3]	0.02 [3/181]
University of California San Francisco Medical Center	0.96	88 (78/89) [79.0, 93.1]	93 (56/60) [83.6, 97.8]	91.2 [95.1, 87.6]	0.880	75.0 (78/104) [65.8, 82.4]	0.25 [37/149]

Note.—Data in parentheses are numerator and denominator; data in brackets are 95% confidence intervals. AUROC = area under the receiver operating characteristic curve, DLAD = deep learning–based automatic detection algorithm, FOM = figure-of-merit, FP = false positive, JAFROC = jackknife alternative free-response receiver operating characteristic.

\* Per-nodule-based nodule detection sensitivities were calculated by using the number of detected nodules divided by the number of total malignant nodules, for which the threshold of the activation-value was set at 0.3.

† Rates of false-positive findings were calculated as the total number of nodules with false-positive findings divided by the total number of chest radiographs, for which the threshold of activation-value was set at 0.3.

(19). Regarding the observer performance test, the radiograph classification and nodule detection performances of DLAD, physicians (test 1), and physicians with DLAD (test 2) were evaluated and compared by using pairwise comparison ROC curve analysis (20) and JAFROC FOMs, respectively: the random-case, fixed-reader method was used for individual-observer comparison and for group-averaged comparison (21). The random-case, random-reader method was used for averaged comparison among all 18 observers. Nodule detection performance was additionally calculated by assuming all nodules from each physician and DLAD were accounted for; the larger confidence level was selected when both the physician and DLAD detected a nodule. Detailed subgroup analyses were performed on the basis of the characteristics of the nodules, including nodule size and conspicuity. The correlation between activation values of DLAD and averaged confidence levels of nine board-certified radiologists were evaluated by using Spearman rank correlation.

ROC analysis was performed by using software (MedCalc version 15.8; MedCalc, Mariakerke, Belgium) and the Dorfman-Berbaum-Metz model was applied by using JAFROC version 4.2.1. For all tests, a *P* value less than .05 was considered to indicate statistical significance, and *P* value correction was performed following the Bonferroni method; *P* values were multiplied by 8 for grouped-observers comparison and by 2 for comparisons among all 18 observers (22).

## Results

### Internal and External Validation Tests of DLAD Performance

For the internal validation data set of 600 chest radiographs (300 normal and 300 nodule chest radiographs), the AUROC

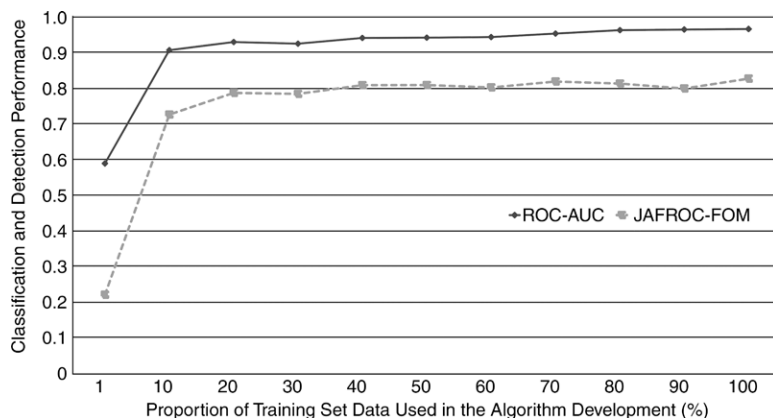
analysis and JAFROC FOM analysis were 0.96 and 0.852, respectively, similar to the AUROC and JAFROC FOM calculated from the tuning data set (0.98 and 0.874, respectively). At external validation, AUROCs were 0.92, 0.99, 0.94, and 0.96, and JAFROC FOMs were 0.870, 0.924, 0.831, and 0.880 for Seoul National University Hospital, Boramae Hospital, National Cancer Center, and University of California San Francisco Medical Center, respectively (Table 2). The nodule detection false-positive rate of DLAD ranged between 0.02 and 0.34 in the four external data sets.

### Subsampling Experiment

To determine the sufficiency of the size of the training set, we randomly selected a certain proportion of chest radiographs in the training data set (ie, 1% and  $N \times 10\%$ ,  $N$  ranging from 1–10) and the detection performances of the algorithms developed were evaluated. There was only a 0.003 AUROC increase and 0.014 JAFROC FOM increase between the algorithms developed from 80% to 100% of the data set (Appendix E1 [online], Fig 3).

### Comparison of Radiograph Classification and Nodule Detection Performances of DLAD

Regarding radiograph classification performance comparison (DLAD vs test 1), DLAD showed an excellent AUROC of 0.91, which was higher than 16 of 18 physicians (*P* value range, <.001 to .52), with statistical significance demonstrated in 11 physicians (Tables 3, E2 [online]). Regarding nodule detection performance comparison (DLAD vs test 1), DLAD exhibited a JAFROC FOM of 0.885, higher than all physicians and significantly higher than those of 15 of 18 physicians (*P* < .05; Table 3). The random case, random reader analysis from JAFROC revealed that the performance of DLAD was higher than the



**Figure 3:** Per-radiograph classification and per-nodule detection performances of the algorithm trained with a limited proportion of chest radiographs in the training set. ROC = receiver operating characteristic, AUC = area under the curve, JAFROC = jackknife alternative free-response receiver-operating characteristic, FOM = figure of merit.

grouped JAFROC FOM of the 18 physicians (JAFROC FOM, 0.885 vs 0.794, respectively;  $P = .002$ ). DLAD demonstrated exceptionally high specificity of 95.2%, even when the threshold was set at score of 1, while preserving high sensitivity of 80.7% (96 of 119) whereas the averaged sensitivity of the physicians was 70.4% (1507 of 2142; Table E2 [online]). The rate of false-positive findings per image of DLAD was 0.30, whereas that of the pooled data of physicians was 0.25.

Regarding the added value of DLAD as a second reader (test 1 vs test 2), the radiograph classification performance of physicians improved with DLAD (17 of 18 physicians; mean AUROC improvement of 0.04 [range,  $-0.0001$  to  $0.14$ ]) and the changes were statistically significant in 15 of 18 physicians. Regarding nodule detection performance, all physicians showed improved detection performances by using DLAD (mean JAFROC FOM improvement, 0.043 [range,  $0.006$ – $0.190$ ]) with significant differences in 14 physicians (Table 3). Two physicians (observers 15 and 16) achieved a higher JAFROC FOM than that of DLAD alone (0.878 and 0.872, respectively) on test 2 (Table 3). JAFROC FOM of nonradiology physicians, radiology residents, board-certified radiologists, and thoracic radiologists subgroups was 0.691, 0.796, 0.821, and 0.833, respectively, and in all four subgroups, their nodule detection performances improved with DLAD (JAFROC FOM for nonradiology physicians, radiology residents, board-certified radiologists, and thoracic radiologists, respectively: 0.828, 0.829, 0.840, and 0.854; corrected  $P < .05$ ; Table 3).

When the maximum confidence level of each physician and DLAD was selected for each nodule, the corresponding nodule detection performance was superior to test 2 for most observers (16 of 18 physicians; mean JAFROC FOM difference, 0.027 [range,  $-0.035$  to  $0.094$ ]) except for two nonradiologist physicians (JAFROC FOM range, 0.801–0.896), and was higher than DLAD alone in 10 of 18 physicians (Table E4 [online]).

DLAD detected all nodules with conspicuity score of 4 or higher (74 of 74) and most nodules greater than 3 cm (96%; 47 of 49) whereas the pooled data of all physicians detected

only 86.0% (449 of 522) and 99.5% (806 of 810) of nodules with conspicuity scores of 4 and 5, respectively, and 85.7% (756 of 882) of nodules greater than 3 cm. Table 4 and Table E3 (online) show how the DLAD and physicians performed over the ranges of nodule conspicuity. For nodules with conspicuity score of 3 or less, DLAD demonstrated a detection rate of 38% (26 of 69), higher than that of the pooled data of thoracic radiologists (27.2%; 75 of 276). Regarding nodules smaller than 1 cm, however, DLAD detected only 11% (two of 18) whereas the pooled data of thoracic radiologists detected 41% (33 of 80) of these lesions. Compared with thoracic radiologists, the confidence score of DLAD tended to exceed that of thoracic radiologists for nodules larger than 1.5 cm (Table E3 [online]). At test 2, the nonradiology physician group showed a higher acceptance rate for initially missed but DLAD-detected

nodules with true-positive findings than did the other groups: the nonradiology physician group dismissed 37% (30 of 82) of DLAD-corrected nodules whereas the other three radiologist subgroups dismissed 67% (41 of 61) to 70% (59 of 84) of them (Appendix E1 [online]; Tables 4, E5 [online]). For nodules overlapped by other structures, DLAD successfully detected more nodules than did the pooled data of thoracic radiologists (60% [40 of 67] vs 56.0% [150 of 268], respectively), but did not for subphrenic nodules (none of nine; Table E6 [online]). For all 143 nodules, the averaged confidence level of the nine board-certified radiologists and the activation values provided from DLAD showed a good correlation (0.824; 95% confidence interval: 0.763, 0.870;  $P < .001$ ). Representative cases are shown in Figures 1 and 2.

### Retrained Algorithm Analysis with CT-confirmed Normal Chest Radiographs

Another algorithm was retrained by using 27 100 chest radiographs, confirmed at CT to be normal, from 27 100 patients and the same nodule chest radiographs from DLAD. The results, including AUROCs, JAFROC FOMs, and F1 scores (ie, the harmonic mean of precision and recall) remained consistent with our results from DLAD among four external validation data sets (Appendix E1 [online]; Table E7 [online]).

### Discussion

Our study results demonstrated that DLAD could accurately detect malignant pulmonary nodules on chest radiographs with better performance than that of physicians, and that it enhanced performance of physicians when used as a second reader. More specifically, DLAD showed high specificity and was able to detect 100% of high conspicuity nodules (score of  $\geq 4$ ), most large ( $> 3$  cm) nodules, and more nodules in overlapped areas than the four groups of physicians in our study.

Compared with previously reported conventional image processing-based computer-aided diagnoses, DLAD showed a markedly decreased rate of false-positive findings and provided high specificity while preserving sensitivity (1,12,23–26),



**Table 3: Patient Classification and Nodule Detection at the Observer Performance Test**

Observer	Test 1		DLAD versus Test 1 ( <i>P</i> Value)		Test 2		Test 1 versus Test 2 ( <i>P</i> Value)	
	Radiograph Classification (AUROC)	Nodule Detection (JAFROC FOM)	Radiograph Classification	Nodule Detection	Radiograph Classification (AUROC)	Nodule Detection (JAFROC FOM)	Radiograph Classification	Nodule Detection
<b>Nonradiology physicians</b>								
Observer 1	0.77	0.716	<.001	<.001	0.91	0.853	<.001	<.001
Observer 2	0.78	0.657	<.001	<.001	0.90	0.846	<.001	<.001
Observer 3	0.80	0.700	<.001	<.001	0.88	0.783	<.001	<.001
Group		0.691		<.001*		0.828		<.001*
<b>Radiology residents</b>								
Observer 4	0.78	0.767	<.001	<.001	0.80	0.785	.02	.03
Observer 5	0.86	0.772	.001	<.001	0.91	0.837	.02	<.001
Observer 6	0.86	0.789	.05	.002	0.86	0.799	.08	.54
Observer 7	0.84	0.807	.01	.003	0.91	0.843	.003	.02
Observer 8	0.87	0.797	.10	.003	0.90	0.845	.03	.001
Observer 9	0.90	0.847	.52	.12	0.92	0.867	.04	.03
Group		0.790		<.001*		0.867		<.001*
<b>Board-certified radiologists</b>								
Observer 10	0.87	0.836	.05	.01	0.90	0.865	.004	.002
Observer 11	0.83	0.804	<.001	<.001	0.84	0.817	.03	.04
Observer 12	0.88	0.817	.18	.005	0.91	0.841	.01	.01
Observer 13	0.91	0.824	>.99	.02	0.92	0.836	.51	.24
Observer 14	0.88	0.834	.14	.03	0.88	0.840	.87	.23
Group		0.821		.02*		0.840		.01*
<b>Thoracic radiologists</b>								
Observer 15	0.94	0.856	.15	.21	0.96	0.878	.08	.03
Observer 16	0.92	0.854	.60	.17	0.93	0.872	.34	.02
Observer 17	0.86	0.820	.02	.01	0.88	0.838	.14	.12
Observer 18	0.84	0.800	<.001	<.001	0.87	0.827	.02	.02
Group		0.833		.08*		0.854		<.001*

Note.—Observer 4 had 1 year of experience; observers 5 and 6 had 2 years of experience; observers 7–9 had 3 years of experience; observers 10–12 had 7 years of experience; observers 13 and 14 had 8 years of experience; observer 15 had 26 years of experience; observer 16 had 13 years of experience; and observers 17 and 18 had 9 years of experience. Observers 1–3 were 4th-year residents from obstetrics and gynecology, orthopedic surgery, and internal medicine, respectively. Radiograph classification performance was described by using an area under the receiver operating characteristic curve analysis, and associated *P* values were calculated by using comparison receiver operating characteristic analysis. Nodule detection performance was described by using a figure of merit from jackknife free-response receiver operating characteristic analysis, and associated *P* values were calculated by using the random-case, fixed-reader method for individual-to-observer comparison by using jackknife free-response receiver operating characteristic analysis. AUROC = area under the receiver operating characteristic curve, DLAD = deep learning-based automatic detection algorithm, FOM = figure of merit, JAFROC = jackknife alternative free-response receiver operating characteristic.

\* For group averaged comparison, corrected *P* values are presented (multiplied by 8).

resulting in better detection performance than thoracic radiologists. Whereas previous computer-aided diagnoses exhibited a range of rates of false-positive findings per image of 0.9–3.3 and a specificity range of 32%–81%, DLAD showed a rate of false-positive findings per image of around 0.02–0.34 with a higher specificity (range, 93.3%–100%; 95% confidence interval, 90.9%–100%). This may result in reducing unnecessary additional work ups, which can induce radiation hazard. The strength of DLAD was in finding conspicuous and/or large nodules and

detecting nodules in overlapped areas (eg, hilar and apex). These results suggest that DLAD can help reduce human error and improve the accuracy of chest radiograph interpretation (27). Considering that conventional computer-aided diagnoses are known to produce frequent false-positive results, our DLAD results may have possibly been overlooked by the observers, resulting in underestimation of the added value of DLAD in our study. Indeed, detection performance was improved simply by taking the maximum confidence level of each physician and DLAD,



**Table 4: Percentage of Detected Nodules by DLAD and Observers According to Nodule Characteristics**

Parameter	Detected by DLAD (%)	Nodules Detected by the Pooled Data of Observers (Test 1)			
		Nonradiology Physicians ( <i>n</i> = 3) (%)	Radiology Residents ( <i>n</i> = 6) (%)	Board-certified Radiologists ( <i>n</i> = 5) (%)	Thoracic Radiologists ( <i>n</i> = 4) (%)
Conspicuity					
5 ( <i>n</i> = 45)	100 (45/45) [90.6, 100]	99.3 (134/135) [96, 100]	99.6 (269/270) [97.7, 100]	99.1 (223/225) [96.6, 100]	100.0 (180/180) [97.5, 100]
4 ( <i>n</i> = 29)	100 (29/29) [86, 100]	66 (57/87) [55, 75]	85.6 (149/174) [79.6, 90.1]	91.7 (133/145) [86.0, 95.3]	94.8 (110/116) [88.9, 97.8]
3 ( <i>n</i> = 30)	57 (17/30) [39, 73]	23 (21/90) [16, 33]	51.7 (93/180) [44.4, 58.9]	50.0 (75/150) [42.1, 57.9]	49.2 (59/120) [40.4, 58.0]
2 ( <i>n</i> = 23)	30 (7/23) [15, 51]	9 (6/69) [4, 18]	17.4 (24/138) [11.9, 24.6]	14.8 (17/115) [9.3, 22.5]	17 (16/92) [11, 26]
1 ( <i>n</i> = 16)	12 (2/16) [2, 37]	4 (2/48) [0, 15]	3 (3/96) [1, 9]	0 (0/80) [0, 6]	0 (0/64) [0, 7]
Size					
>30 mm ( <i>n</i> = 49)	96 (47/49) [86, 100]	79.6 (117/147) [72.3, 85.4]	85.4 (251/294) [80.9, 89.0]	86.9 (213/245) [82.1, 90.6]	89.3 (175/196) [84.1, 93.0]
20–30 mm ( <i>n</i> = 31)	74 (23/31) [56, 86]	60 (56/93) [50, 70]	67.2 (125/186) [60.2, 73.6]	61.9 (96/155) [54.1, 69.2]	68.5 (85/124) [59.9, 76.1]
15–20 mm ( <i>n</i> = 25)	68 (17/25) [48, 83]	44 (33/75) [33, 55]	62.7 (94/150) [54.7, 70.0]	64.0 (80/125) [55.3, 71.9]	59.0 (59/100) [49.2, 68.1]
10–15 mm ( <i>n</i> = 20)	55 (11/20) [34, 74]	23 (14/60) [14, 36]	37.5 (45/120) [29.3, 46.4]	40.0 (40/100) [30.9, 49.8]	41 (33/80) [31, 52]
<10 mm ( <i>n</i> = 18)	11 (2/18) [2, 34]	0 (0/54) [0, 8]	21.3 (23/108) [14.6, 30.0]	21 (19/90) [14, 31]	18 (13/72) [11, 29]
Score*					
4 ( <i>n</i> = 75)	100 (75/75) [94, 100]	88.9 (200/225) [84.1, 92.4]	95.3 (429/450) [92.9, 97.0]	97.6 (366/375) [95.4, 98.8]	100 (300/300) [98.5, 100]
3 ( <i>n</i> = 11)	73 (8/11) [43, 91]	27 (9/33) [15, 44]	74 (49/66) [64, 83]	58 (32/55) [45, 70]	75 (33/44) [60, 86]
2 ( <i>n</i> = 11)	46 (5/11) [21, 72]	9 (3/33) [2, 24]	47 (31/66) [35, 59]	42 (23/55) [30, 55]	50 (22/44) [36, 64]
1 ( <i>n</i> = 10)	50 (5/10) [24, 76]	13 (4/30) [5, 30]	33 (20/60) [23, 46]	42 (21/50) [29, 56]	25 (10/40) [14, 40]
0 ( <i>n</i> = 36)	19 (7/36) [10, 35]	3.7 (4/108) [1.1, 9.4]	4.2 (9/216) [2.1, 7.8]	3.3 (6/180) [1.4, 7.2]	0 (0/144) [0, 3.1]
Summation	70 (100/143) [62, 77]	51.3 (220/429) [46.6, 56.0]	62.7 (538/858) [59.5, 66.0]	62.7 (448/715) [59.1, 66.1]	63.8 (365/572) [59.8, 67.7]

Note.—Data in parentheses are numerator and denominator; data in brackets are 95% confidence interval. DLAD = deep learning–based automatic detection algorithm.

\* The score was defined as the number of thoracic radiologists who successfully detected the nodule (confidence  $\geq 1$ ).

with the JAFROC FOM reaching as high as 0.896 for one observer. Thus, we surmise that nodule detection performance will improve further when observers are fully informed regarding the low false-positive rate of DLAD (Table E4 [online]).

These results are further bolstered by the fact that the performance of DLAD remained relatively consistent among the tuning, internal validation, and four external validation data sets in our study, except for the exceptionally high performance of the data set from Boramae Hospital. The high performance in the data set can be attributed to their different nodule population because 53.2% (74 of 139) of their nodules were greater than 3 cm (Table 1). By providing a sufficient number of fully

supervised data combined with weakly supervised data and by performing regularization during training, successful generalization was able to be achieved by avoiding the overfitting of our algorithm (13,28). It is also promising that the performance of DLAD was able to be reproduced even in the validation data set from a different country (data set from University of California San Francisco Medical Center).

DLAD had limitations. DLAD did not accurately detect small (<1 cm; detection rate, 11% [two of 18]) and less conspicuous nodules. Because DLAD was trained under the supervision of labeling and annotation information provided by radiologists, the limitations of human perception are reflected in

the DLAD performance. In the future, establishing an algorithm supervised by annotations on chest radiographs containing radiologically undetectable nodules by using CT as the reference standard is warranted. DLAD also did not detect any retrophrenic nodules, whereas thoracic radiologists detected 31% (11 of 36) retrophrenic nodules. Because DLAD was trained to focus on the lung area to prevent false-positive findings detected outside the lungs, DLAD was undertrained for the retrophrenic area. With further optimization from focused training by using retrophrenic nodules of both the posterior-anterior and lateral views, we expect the performance of DLAD to improve. In addition, the cut off of activation values may need to be better optimized. The activation values produced from DLAD are derived from an activation function of the neural network and may not necessarily be proportional to the possibility of abnormal density at chest radiography. We identified the correlation of human confidence of abnormality and activation values by using the averaged confidence scores of nine radiologists and the activation values of nodules.

Our study has limitations. First, we only included malignant nodules, most of which were confirmed at pathologic analysis. Second, DLAD may have been undertrained for small nodules (ie, nodules < 1 cm). Third, DLAD was trained only for malignant nodule detection. Benign lung nodules and the differentiation of benign from malignant nodules was not optimized. Fourth, pneumonias and interstitial lung disease were not considered. Fifth, in routine practice, interval change plays a key role in the interpretation of chest radiographs, however, this was not integrated into the current algorithm. Furthermore, in this study, we only dealt with chest radiographs obtained at posterior-anterior projections. Inclusion of the lateral view images or bone suppressed and bone view images, or dual-energy chest radiographs by using deep learning technology might further enhance the performance. Finally, this was a retrospective study and did not exactly represent the real-world setting. Further research is warranted to determine the applicability of this DLAD in a prospective multi-institutional study, or it should be evaluated with prospectively collected large multi-institutional data sets (eg, National Lung Cancer Screening Trial).

In conclusion, we developed a deep learning-based automatic detection algorithm that outperformed physicians at per-radiograph classification and malignant nodule detection on chest radiographs. When this algorithm was used as a second reader, physicians demonstrated enhanced performance for malignant nodule detection.

**Acknowledgments:** The DLAD development and evaluation group included 29 contributors. Labeling and annotating the development data set was performed by Dong Hyeon Kim (Republic of Korea Air Force), Sungmin Woo, Wonseok Choi, Inpyeong Hwang, Hyungjin Kim, and Yong Sub Song (Seoul National University College of Medicine); Jiyeon Lim (Research Institute of Radiological Science, Yonsei University College of Medicine); Jung Im Kim (Kyung Hee University Hospital at Gangdong, Kyung Hee University); So Young Choi (Eulji University Hospital); Nyoung Keun Lee (Sungmin hospital); and Jae Yeon Wi. Preparation of the development data set was performed by Su Suk Oh (Seoul National University College of Medicine). Development of the algorithm was performed by Geonhwan Ju, Minsung Kim, Min Jae Kang, and Beomseok Suh (Lunit). The observer performance test was conducted by Taek Min Kim, Ji Hee Kang, Yun Soo Jeong, Sanghyup Lee, Junghoan Park, and Jae Won Choi (Seoul National University College of Medicine); Mi-Jin Kang (Inje University Sanggye-

paik Hospital); Jin Young Yoo (Chungbuk National University Hospital); Hyunju Lim (National Cancer Center); and Jung Wee Park, Jiyeon Han, and JinJu Choi (Seoul National University Hospital and College of Medicine). Evaluation of nodule characteristics used in the observer performance test was performed by Su Yeon Ahn (Konkuk University Medical Center).

**Author contributions:** Guarantor of integrity of entire study, 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.G.N., S.P., E.J.H., J.H.L., J.M.G., C.M.P.; clinical studies, S.P., E.J.H., K.N.J., K.Y.L., T.H.V., J.H.S., J.M.G., C.M.P.; experimental studies, J.G.N., S.P., K.N.J., T.H.V., J.H.S., S.H.; statistical analysis, J.G.N., S.P., T.H.V., S.H.; and manuscript editing, J.G.N., S.P., E.J.H., T.H.V., J.M.G., C.M.P.

**Disclosures of Conflicts of Interest:** J.G.N. disclosed no relevant relationships. S.P. Activities related to the present article: disclosed no relevant relationships. Activities not related to the present article: disclosed employment by Lunit. Other relationships: disclosed no relevant relationships. E.J.H. disclosed no relevant relationships. J.H.L. disclosed no relevant relationships. K.N.J. disclosed no relevant relationships. K.Y.L. disclosed no relevant relationships. T.H.V. disclosed no relevant relationships. J.H.S. disclosed no relevant relationships. S.H. Activities related to the present article: disclosed no relevant relationships. Activities not related to the present article: disclosed stock options in Lunit. Other relationships: disclosed no relevant relationships. J.M.G. Activities related to the present article: disclosed money paid to author by Lunit for a research grant. Activities not related to the present article: disclosed no relevant relationships. Other relationships: disclosed no relevant relationships. C.M.P. Activities related to the present article: disclosed money paid to author by Lunit for a research grant. Activities not related to the present article: disclosed no relevant relationships. Other relationships: disclosed no relevant relationships.

## References

- 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.
- de Hoop B, De Boer DW, Gietema HA, et al. Computer-aided detection of lung cancer on chest radiographs: effect on observer performance. *Radiology* 2010;257(2):532–540.
- Li F, Arimura H, Suzuki K, et al. Computer-aided detection of peripheral lung cancers missed at CT: ROC analyses without and with localization. *Radiology* 2005;237(2):684–690.
- Potchen EJ, Cooper TG, Sierra AE, et al. Measuring performance in chest radiography. *Radiology* 2000;217(2):456–459.
- Toyoda Y, Nakayama T, Kusunoki Y, Iso H, Suzuki T. Sensitivity and specificity of lung cancer screening using chest low-dose computed tomography. *Br J Cancer* 2008;98(10):1602–1607.
- Gavelli G, Giampalma E. Sensitivity and specificity of chest X-ray screening for lung cancer: review article. *Cancer* 2000;89(11 Suppl):2453–2456.
- 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.
- Mettler FA Jr, Huda W, Yoshizumi TT, Mahesh M. Effective doses in radiology and diagnostic nuclear medicine: a catalog. *Radiology* 2008;248(1):254–263.
- 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.
- Gerritsen MG, Willemink MJ, Pompe E, et al. Improving early diagnosis of pulmonary infections in patients with febrile neutropenia using low-dose chest computed tomography. *PLoS One* 2017;12(2):e0172256.
- den Harder AM, Willemink MJ, van Hamersvelt RW, et al. Pulmonary nodule volumetry at different low computed tomography radiation dose levels with hybrid and model-based iterative reconstruction: a within patient analysis. *J Comput Assist Tomogr* 2016;40(4):578–583.
- Schalekamp S, van Ginneken B, Karssemeijer N, Schaefer-Prokop CM. Chest radiography: new technological developments and their applications. *Semin Respir Crit Care Med* 2014;35(1):3–16.
- LeCun Y, Bengio Y, Hinton G. Deep learning. *Nature* 2015;521(7553):436–444.
- Krizhevsky A, Sutskever I, Hinton GE. Imagenet classification with deep convolutional neural networks. In: *NIPS'12: Proceedings of the 25th International Conference on Neural Information Processing Systems - Volume 1*. Red Hook, NY: Curran Associates, 2012; 1097–1105.
- LeCun Y, Bottou L, Bengio Y, Haffner P. Gradient-based learning applied to document recognition. *Proc IEEE* 1998;86(11):2278–2324.
- He K, Zhang X, Ren S, Sun J. Deep residual learning for image recognition. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016; 770–778.
- Liao S, Gao Y, Oto A, Shen D. Representation learning: a unified deep learning framework for automatic prostate MR segmentation. In: *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Berlin, Germany: Springer, 2013; 254–261.

18. Kumar D, Wong A, Clausi DA. Lung nodule classification using deep features in CT images. In: 2015 12th Conference on Computer and Robot Vision. LOCATION: IEEE, 2015; 133–138.
19. Chakraborty DP. Recent developments in imaging system assessment methodology, FROC analysis and the search model. Nucl Instrum Methods Phys Res A 2011;648(Supplement 1):S297–S301.
20. DeLong ER, DeLong DM, Clarke-Pearson DL. Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach. Biometrics 1988;44(3):837–845.
21. Fletcher JG, Yu L, Li Z, et al. Observer performance in the detection and classification of malignant hepatic nodules and masses with CT image-space denoising and iterative reconstruction. Radiology 2015;276(2):465–478.
22. Bland JM, Altman DG. Multiple significance tests: the Bonferroni method. BMJ 1995;310(6973):170.
23. Novak RD, Novak NJ, Gilkeson R, Mansoori B, Aandal GE. A comparison of computer-aided detection (CAD) effectiveness in pulmonary nodule identification using different methods of bone suppression in chest radiographs. J Digit Imaging 2013;26(4):651–656.
24. Dellios N, Teichgraber U, Chelaru R, Malich A, Papageorgiou IE. Computer-aided Detection Fidelity of Pulmonary Nodules in Chest Radiograph. J Clin Imaging Sci 2017;7(1):8.
25. Li F, Engelmann R, Armato SG 3rd, MacMahon H. Computer-aided nodule detection system: results in an unselected series of consecutive chest radiographs. Acad Radiol 2015;22(4):475–480.
26. Schalekamp S, van Ginneken B, Heggelman B, et al. New methods for using computer-aided detection information for the detection of lung nodules on chest radiographs. Br J Radiol 2014;87(1036):20140015.
27. de Groot PM, Carter BW, Abbott GF, Wu CC. Pitfalls in chest radiographic interpretation: blind spots. Semin Roentgenol 2015;50(3):197–209.
28. Bengio Y, Courville A, Vincent P. Representation learning: a review and new perspectives. IEEE Trans Pattern Anal Mach Intell 2013;35(8):1798–1828. \* The score was defined as the number of thoracic radiologists who successfully detected the nodule (confidence  $\geq 1$ ).