Original Article | Thoracic Imaging

eISSN 2005-8330
https://doi.org/10.3348/kjr.2022.0548
Korean J Radiol 2023;24(2):155-165

Korean Journal of Radiology

Effects of Expert-Determined Reference Standards in Evaluating the Diagnostic Performance of a Deep Learning Model: A Malignant Lung Nodule Detection Task on Chest Radiographs

Jung Eun Huh[1, 2*], Jong Hyuk Lee[3*], Eui Jin Hwang[3], Chang Min Park[1, 3, 4, 5]

[1]Institute of Medical and Biological Engineering, Medical Research Center, Seoul National University, Seoul, Korea; [2]Mathematical Institute, University of Oxford, United Kingdom; [3]Department of Radiology, Seoul National University Hospital, Seoul, Korea; [4]Department of Radiology, Seoul National University College of Medicine, Seoul, Korea; [5]Institute of Radiation Medicine, Medical Research Center, Seoul National University, Seoul, Korea

Objective: Little is known about the effects of using different expert-determined reference standards when evaluating the performance of deep learning-based automatic detection (DLAD) models and their added value to radiologists. We assessed the concordance of expert-determined standards with a clinical gold standard (herein, pathological confirmation) and the effects of different expert-determined reference standards on the estimates of radiologists' diagnostic performance to detect malignant pulmonary nodules on chest radiographs with and without the assistance of a DLAD model.
Materials and Methods: This study included chest radiographs from 50 patients with pathologically proven lung cancer and 50 controls. Five expert-determined standards were constructed using the interpretations of 10 experts: individual judgment by the most experienced expert, majority vote, consensus judgments of two and three experts, and a latent class analysis (LCA) model. In separate reader tests, additional 10 radiologists independently interpreted the radiographs and then assisted with the DLAD model. Their diagnostic performance was estimated using the clinical gold standard and various expert-determined standards as the reference standard, and the results were compared using the *t* test with Bonferroni correction.
Results: The LCA model (sensitivity, 72.6%; specificity, 100%) was most similar to the clinical gold standard. When expert-determined standards were used, the sensitivities of radiologists and DLAD model alone were overestimated, and their specificities were underestimated (all p-values < 0.05). DLAD assistance diminished the overestimation of sensitivity but exaggerated the underestimation of specificity (all p-values < 0.001). The DLAD model improved sensitivity and specificity to a greater extent when using the clinical gold standard than when using the expert-determined standards (all p-values < 0.001), except for sensitivity with the LCA model (p = 0.094).
Conclusion: The LCA model was most similar to the clinical gold standard for malignant pulmonary nodule detection on chest radiographs. Expert-determined standards caused bias in measuring the diagnostic performance of the artificial intelligence model.
Keywords: *Deep-learning; Reference standard; Expert-determined standard; Decision-support tool; Chest radiographs*

INTRODUCTION

In recent years, the field of artificial intelligence (AI) for medical applications has rapidly advanced and radically reshaped medicine [1,2]. Automated diagnosis using AI algorithms from medical images, especially deep learning

Received: May 29, 2022 **Revised:** November 30, 2022 **Accepted:** December 19, 2022
*These authors contributed equally to this work.
Corresponding author: Chang Min Park, MD, PhD, Department of Radiology and Institute of Radiation Medicine, Seoul National University College of Medicine, 101 Daehak-ro, Jongno-gu, Seoul 03080, Korea.
• E-mail: cmpark.morphius@gmail.com

(DL)-driven models, is a field that has experienced remarkable progress and has shown excellent diagnostic performance, comparable to or even surpassing human experts or existing diagnostic tests [3,4]. Prime examples include DL models for classifying skin cancers, detecting diabetic retinopathy on retinal fundus photographs and breast cancer on whole-slide pathology images, and detecting and classifying major thoracic diseases on chest radiographs [5-8].

An essential prerequisite to evaluate the performance and clinical applicability of DL models is the rigor of reference standards [2,9]. The clinical gold standard is widely accepted as the best available method of establishing the presence or absence of the target condition [2,9-11]. Therefore, the clinical gold standard is considered the ground truth in most AI research (e.g., pathological confirmation in a malignant lung nodule detection task) [9]. However, expert-determined standards can often be used as proxies in clinical settings due to the unavailability of clinical gold standards (e.g., pathology is not always affordable in clinical practice). The problem is that the performance measurements of the AI model may be substantially affected by expert-determined standards and biased by variability between interpretations of experts [2,12]. Furthermore, the medical imaging field is highly dependent on human interpretation, with significant intra- or inter-observer variability, which biases the evaluation of AI performance [2,12]. In this regard, the consequences of using expert-determined standards, not clinical gold standards, to evaluate the diagnostic performance of AI models and whether they can be used in clinical settings remain an important question [10,11].

Although the impact of imperfect standards on the evaluation of a diagnostic test is a well-known source of bias [13], its impact on radiologists has not been investigated in the AI field. Therefore, we aimed to assess the concordance of various expert-determined standards with the clinical gold standard and to investigate how the diagnostic performance of radiologists alone, radiologists assisted by DL-based automatic detection (DLAD) and DLAD alone for the detection of malignant pulmonary nodules on chest radiographs might change when different standards are used as a reference standard for performance evaluation [14-16].

MATERIALS AND METHODS

This retrospective study was approved by the Institutional Review Board of Seoul National University Hospital (IRB No. H-2112-014-1279), which waived the requirement of written informed consent. The study sample has not been previously reported.

Study Samples and Clinical Gold Standard

For the study dataset, one author (J.H.L., with 10 years of experience in thoracic radiology) randomly selected 50 lung cancers from 50 patients (26 males and 24 females; mean age, 67.4 ± 11.7 years) with pathologically confirmed lung cancers in a tertiary referral hospital (Seoul National University Hospital, Seoul, Korea) between December 2015 and February 2021. The last computed tomography (CT) and chest radiographs of the patients before pathological diagnosis were collected. For the normal control group, 50 normal chest radiographs of 50 individuals (30 males and 20 females; mean age, 56 ± 9.1 years) without any abnormal findings on chest CT examination during the same period were independently selected. The median interval between chest radiographs and CT examination was 16.5 days (interquartile range [IQR]: 4, 31.75 days) and 0 days (IQR: 0, 0 days) for the 50 lung cancer radiographs and normal radiographs, respectively. Therefore, 100 chest radiographs from 100 individuals were included in this study. Detailed information on the study dataset is provided in Table 1.

Using a customized web-based tool (AVIEW, Coreline Soft), the areas of lung cancer on the 50 chest radiographs were marked by two authors (J.H.L. and C.M.P. with 24 years of experience in thoracic radiology) in consensus, according to the lung cancer areas on the corresponding chest CT examinations. These annotated marks served as clinical gold standards. The tool, which was used to construct the clinical gold standard and expert-determined standards and to perform the reader tests, provides the function of an image viewer, annotations of potential lesions, and localization information for the annotation boxes.

Construction of Expert-Determined Standards

Ten experts (board-certified thoracic radiologists) from seven institutions across the country were recruited to construct expert-determined standards (five males and five females; median years of experience in thoracic radiology was 12 years [range: 9–18 years]). The 10 experts did not participate in selecting the study sample and determining the clinical gold standard. They independently evaluated 100 chest radiographs, blinded to clinical and radiological information. If they detected a potentially malignant lung

Table 1. Baseline Clinical, Radiological, and Pathological Characteristics of the Study Sample

Patients with malignant pulmonary nodules (n = 50)	
Age, year	67.4 ± 11.7
Sex, male	26 (52)
Nodule location	
Right upper lobe	13 (26)
Right middle lobe	2 (4)
Right lower lobe	16 (32)
Left upper lobe	9 (18)
Left lower lobe	10 (20)
Nodule size	
≤ 1 cm	8 (16)
> 1 to ≤ 2 cm	23 (46)
> 2 to ≤ 3 cm	9 (18)
> 3 cm	10 (20)
Nodule type	
Solid nodule	38 (76)
Part-solid nodule	12 (24)
Nodule visibility on chest radiographs	
Visible	42 (84)
Non-visible	8 (16)
Obscuration by the adjacent structure*	23 (46)
Pathologic diagnosis	
Adenocarcinoma	41 (82)
Squamous cell carcinoma	6 (12)
Small cell carcinoma	1 (2)
Mixed squamous cell and small cell carcinoma	1 (2)
Sarcomatoid carcinoma	1 (2)
Individuals with normal chest radiographs (n = 50)	
Age, year	56 ± 9.1
Sex, male	30 (60)

Data are mean ± standard deviation or number of patients with % in parentheses. *Pulmonary malignant nodules were obscured by the following structures: bone (n = 12), heart (n = 5), diaphragm (n = 3), bone and mediastinum (n = 1), bone and diaphragm (n = 1), and heart and hilum (n = 1)

lesion, they were instructed to use the same web-based tool to mark the lesion with a box annotation to fit the size of the lesion as closely as possible while including the lesion sufficiently. Multiple box annotations were allowed for a single chest radiograph.

Expert-Determined Standard Construction

Using the reading results of 10 experts, we constructed five image-based expert-determined standards for each chest radiograph: 1) as an individual judgment, the judgment of the most experienced expert (with 18 years of experience in thoracic radiology) was selected and for joint judgment, we combined the initial judgments of the 10 experts in four different ways; 2) majority vote (four combinations of panels composed of three, five, seven, and nine experts); 3) consensus of two experts; 4) consensus of three experts; and 5) latent class analysis (LCA) judgments (Fig. 1A).

Expert-determined standards were defined in binary form (i.e., normal or abnormal radiographs), where an abnormality was defined as an area that contains at least one potential lesion. A radiograph was considered to contain lung cancer based on the majority vote criterion when it received majority agreement from the judgments of the selected experts. To determine the optimal number of experts for majority voting, we conducted 502 majority votes, differing in the composition of the odd number of experts. The following panel combinations were compared: panels composed of 9 (n = 10), 7 (n = 120), 5 (n = 252), and 3 (n = 120) of the 10 experts (Supplementary Material 1). For consensus reading, consensus meetings were organized at two institutions with which two and three experts were affiliated. The experts from each institution were asked to harmonize their initial individual judgments into consensus judgments. Through these consensus meetings at each institution, single expert-determined standards were generated from each of the two institutions. For an advanced summary of the judgments of the 10 experts. The LCA model observes inherent interactive patterns within experts' judgments to statistically estimate the tendencies of individual judgments and conclude which latent class (in our case, considered to imply negativity and positivity in lesion detection) each chest radiograph belongs to [17,18]. Among the various LCA models, we fitted a two-discrete-latent-class model to the collected experts' judgments of the study sample. Detailed information on the LCA model is provided in Supplementary Material 2 and Supplementary Table 1. In addition to image-based analysis, we also constructed lesion-based expert-determined standards (Supplementary Material 3).

Reader Tests

Reader tests were performed to investigate diagnostic performance according to various standards (Fig. 1B). For each reader test, 10 radiologists (five males and five females; median years of experience in reading thoracic imaging, 9 years [range: 5–17 years]) who did not participate in the construction of the clinical gold standard or expert-determined standards were recruited. The reader test consisted of two sessions with a 1-month washout

https://doi.org/10.3348/kjr.2022.0548

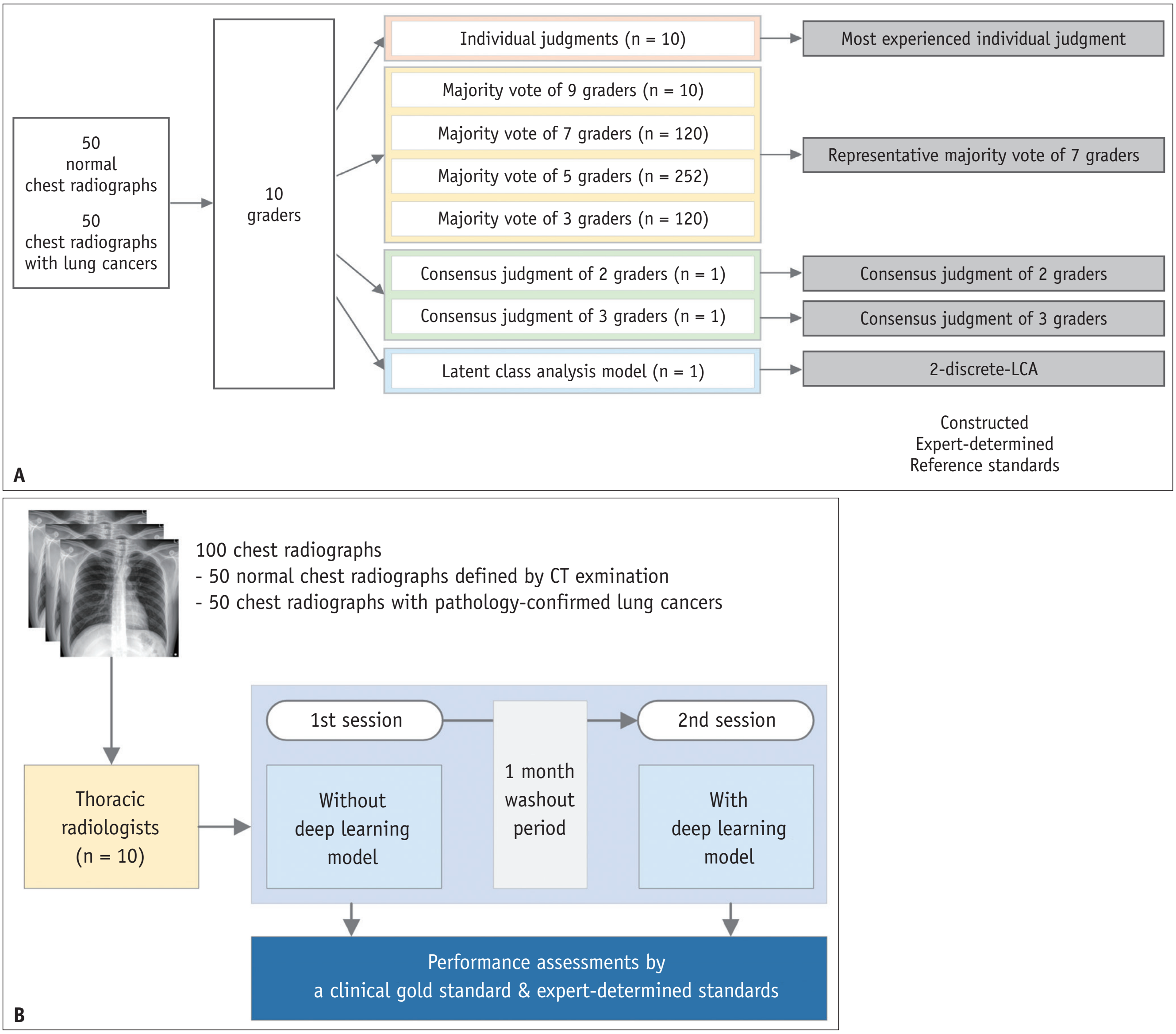
50
normal
chest radiographs
50
chest radiographs
with lung cancers
10
graders
Individual judgments (n = 10)
Majority vote of 9 graders (n = 10)
Majority vote of 7 graders (n = 120)
Majority vote of 5 graders (n = 252)
Majority vote of 3 graders (n = 120)
Consensus judgment of 2 graders (n = 1)
Consensus judgment of 3 graders (n = 1)
Latent class analysis model (n = 1)
Most experienced individual judgment
Representative majority vote of 7 graders
Consensus judgment of 2 graders
Consensus judgment of 3 graders
2-discrete-LCA
Constructed
Expert-determined
Reference standards
A
100 chest radiographs
- 50 normal chest radiographs defined by CT exmination
- 50 chest radiographs with pathology-confirmed lung cancers
Thoracic
radiologists
(n = 10)
1st session
Without
deep learning
model
1 month
washout
period
2nd session
With
deep learning
model
Performance assessments by
a clinical gold standard & expert-determined standards
B

Fig. 1. Construction of expert-determined reference standards and reader tests in this study.
A. Using a study sample consisting of 100 chest radiographs (50 chest radiographs with lung cancer and 50 normal chest radiographs), 10 experts constructed five final expert-determined reference standards: individual judgment of the most experienced expert, the majority vote of seven experts, the consensus of two and three experts, and latent class analysis judgments. The numbers in parentheses indicate the number of combinations that can be made by 10 experts. **B.** Ten radiologists participated in the reader tests, which consisted of two reading sessions. In the first session, they independently reviewed the study sample without a commercially available deep learning-automated detection (DLAD) model. In the second session, they reviewed the chest radiographs and supplied with results of the DLAD model.

period between sessions. In the first session, 10 radiologists consecutively and independently reviewed 100 chest radiographs. In the second session, they reviewed the chest radiographs supplied with the nodule detection results of the commercially available DLAD model (Lunit INSIGHT CXR, version 4.7.2, Lunit) using the toggle function between the chest radiograph alone and the radiograph overlaid with the heatmap. If they detected potential lung cancer, they annotated the lesion with box annotations in the same way used for constructing expert-determined standards.

Statistical Analysis

We assessed the concordance of the above-described five expert-determined standards with the clinical gold standard and compared their sensitivity, specificity, and accuracy. In subsequent reader tests, the diagnostic performance of the 10 radiologists alone or assisted by the DLAD model and the DLAD model alone was calculated. The following comparisons were performed: 1) the diagnostic performance of radiologists between the clinical gold standard and expert-determined standards, 2) when radiologists were assisted by the DLAD model, the change in diagnostic performance between the clinical gold standard and expert-determined standards, and 3) the added value of the DLAD model to the radiologists between the clinical gold standard and expert-determined standards. To increase general validity, bootstrapping was performed with sample sizes of 50 and 100 iterations. All comparisons were performed using the *t* test with Bonferroni correction.

All statistical analyses were performed using R version 4.1.0 (R Project for Statistical Computing) and Mplus version 8.7 (Muthén & Muthén). A *p*-value of < 0.05 was considered statistically significant, and significant *p*-values were determined by Bonferroni correction in each multiple comparison.

RESULTS

Concordance of the Expert-Determined Standards to the Clinical Gold Standard

For the majority vote, seven experts had a sensitivity of 67.3% (95% confidence interval [CI]: 66.9–67.8) and specificity of 100% (95% CI: 100–100) that were equivalent to those of nine experts (sensitivity, 68.0% [95% CI: 66.4–69.6], *p* = 0.403; specificity, 100% [95% CI: 100–100], *p* > 0.999) and significantly higher than those of five experts (sensitivity, 66.3% [95% CI: 65.8–66.8], *p* = 0.002; specificity, 99.5% [95% CI: 99.4–99.6], *p* < 0.001) and three experts (sensitivity, 65.3% [95% CI: 64.3–66.4], *p* < 0.001 [not shown]; specificity 98.4% [95% CI: 97.9–98.8], *p* < 0.001 [not shown]) (Supplementary Table 2). Therefore, a majority vote by seven experts was considered the representative expert-determined standard for the majority vote (unless otherwise noted, hereafter, “majority vote” refers to a majority vote by seven experts). The diagnostic performance of each expert, the consensus judgments of two and three experts, and the LCA model are described in Supplementary Tables 3, 4, and 5, respectively.

The LCA model had a sensitivity of 72.6% (95% CI: 70.9–74.3), which was equivalent to that of individual judgment (70.2% [95% CI: 68.6–71.9], *p* = 0.049) and higher than that of majority vote (68.8% [95% CI: 67.0–70.6], *p* = 0.002) and consensus judgments (two experts: 62.2% [95% CI: 60.3–64.1], *p* < 0.001; three experts: 60.4% [95% CI: 58.5–62.2], *p* < 0.001). The LCA model had a specificity of 100% (95% CI: 100–100), which was significantly higher than that of individual judgment (94.2% [95% CI: 93.4–95.0], *p* < 0.001) and the consensus of three experts (98.2% [95% CI: 97.7–98.7], *p* < 0.001) but equivalent to that of the majority vote and the consensus of two experts (all specificities, 100%, 95% CI: 100–100, *p* > 0.999) (Table 2). The results of the lesion-based analyses are presented in Supplementary Table 6.

Reader Tests

The results of the reader tests are listed in Table 3. When assessed by expert-determined standards, all sensitivities of radiologists alone, radiologists assisted by the DLAD model, and the DLAD model alone were overestimated compared to when the clinical gold standard was applied (range of overestimation: 10.9% to 18.4% for radiologists alone; 7.9% to 15.7% for radiologists assisted by the DLAD model; 4.0% to 14.9% for the DLAD model alone; all *p*-values < 0.05) (Fig. 2A). However, specificities were underestimated when assessed by expert-determined standards (range of underestimation: 4.6% to 9.5% for radiologists alone; 6.9% to 12.7% for radiologists assisted by the DLAD model; 10.0% to 17.9% for the DLAD model alone; all *p*-values < 0.05) (Fig. 2B). The representative cases are shown in Figure 3.

When radiologists were assisted by the DLAD model, the overestimation of sensitivity decreased (individual judgment: 10.9% to 7.9%; majority vote: 17.4% to 13.8%; consensus of two experts: 18.2% to 15.7%; consensus of three experts: 18.4% to 15.7%; LCA model: 16.8% to 15.5%; all *p*-values < 0.001); however, the underestimation of specificity was exaggerated (individual judgment: -6.0% to -8.9%; majority vote: -6.3% to -10.4%; consensus of two experts: -9.5% to -12.7%; consensus of three experts: -9.3% to -12.5%; LCA model: -4.6% to -6.9%; all *p*-values < 0.001).

Using each standard, the added value of the DLAD model was verified for sensitivity (reference standard: 66.4% to 72.8%; individual judgment: 77.3% to 80.7%; majority vote: 83.8% to 86.6%; consensus of two experts: 84.6% to 88.5%; consensus of three experts: 84.8% to 88.5%; LCA model: 83.2% to 88.3%; all *p*-values < 0.001), but the increment was significantly greater with the clinical gold standard than with the expert-determined standards (clinical gold standard: 6.4%; individual judgment: 3.4%; majority vote: 2.8%; consensus of two experts: 3.9%; consensus of three experts: 3.7%; all *p*-values < 0.001), except for the LCA model (5.1%, *p* = 0.094). Although the specificity of radiologists with the clinical gold standard (94.2% to 97.7%,

Table 2. Concordance of Expert-Determined Standards to the Clinical Gold Standard (Pathologic Confirmation) in Terms of Sensitivity, Specificity, and Accuracy and Their Comparison between Different Expert-Determined Standards

Sensitivity					
	Individual judgment*	Majority vote†	Consensus of two experts	Consensus of three experts	LCA model‡
	70.2% (68.6–71.9)	68.8% (67.0–70.6)	62.2% (60.3–64.1)	60.4% (58.5–62.2)	72.6% (70.9–74.3)
Individual judgment*	-	0.233	< 0.001	< 0.001	0.049
Majority vote†	0.233	-	< 0.001	< 0.001	0.002
Consensus of two experts	< 0.001	< 0.001	-	0.163	< 0.001
Consensus of three experts	< 0.001	< 0.001	0.163	-	< 0.001
LCA model‡	0.049	0.002	< 0.001	< 0.001	-
Specificity					
	Individual judgment*	Majority vote†	Consensus of two experts	Consensus of three experts	LCA model‡
	94.2% (93.4–95.0)	100% (100–100)	100% (100–100)	98.2% (97.7–98.7)	100% (100–100)
Individual judgment*	-	< 0.001	< 0.001	< 0.001	< 0.001
Majority vote†	< 0.001	-	> 0.999	< 0.001	> 0.999
Consensus of two experts	< 0.001	> 0.999	-	< 0.001	> 0.999
Consensus of three experts	< 0.001	< 0.001	< 0.001	-	< 0.001
LCA model‡	< 0.001	> 0.999	> 0.999	< 0.001	-
Accuracy					
	Individual judgment*	Majority vote†	Consensus of two experts	Consensus of three experts	LCA model‡
	82.1% (81.1–83.1)	84.3% (83.3–85.4)	81.1% (79.9–82.2)	79.2% (78.1–80.3)	86.3% (85.3–87.2)
Individual judgment*	-	0.003	0.168	< 0.001	< 0.001
Majority vote†	0.003	-	< 0.001	< 0.001	0.007
Consensus of two experts	0.168	< 0.001	-	0.022	< 0.001
Consensus of three experts	< 0.001	< 0.001	0.022	-	< 0.001
LCA model‡	< 0.001	0.007	< 0.001	< 0.001	-

A p-value < 0.005 was considered statistically significant for concordance, according to the Bonferroni correction. Parentheses are 95% confidence intervals. *Individual judgment of expert 5, who had the most experience in thoracic radiology, †The judgment of the majority vote was determined by seven experts, ‡2-D-class LCA. LCA = latent class analysis

p < 0.001) and LCA model (89.6% to 90.8%, p = 0.001) increased significantly with the DLAD model, the other four expert-determined standards did not show significant changes in specificity with the DLAD model (individual judgment: 88.2% to 88.8%, p = 0.094; majority vote: 87.9% to 87.3%, p = 0.134; consensus of two experts: 84.7% to 85%, p = 0.459; and consensus of three experts: 84.9% to 85.2%, p = 0.465). The specificity increment was significantly greater with the clinical gold standard than with the expert-determined standards (clinical gold standard: 3.5%; individual judgment: 0.6%; majority vote: -0.6%; consensus of two experts: 0.3%; consensus of three experts: 0.3%; LCA model: 1.2%; all p-values < 0.001). The results of the lesion-based reader tests using expert-determined standards are presented in Supplementary Material 4, Supplementary Table 7, and Supplementary Figure 1. The correlations between the concordance of expert-determined standards with the clinical gold standard and reader test results are described in Supplementary Material 5.

DISCUSSION

In this study, we assessed the concordance of various expert-determined standards with the clinical gold standard for detecting malignant pulmonary nodules on chest radiographs. Considering both sensitivity and specificity, the LCA model was the most similar to the clinical gold standard (sensitivity, 72.6%; specificity, 100%). In reader tests, the radiologists alone or assisted by the DLAD model tended to show overestimated sensitivity but underestimated specificity when expert-determined standards were applied as ground truth (all p-values < 0.05). These tendencies were diminished for sensitivity but exaggerated for specificity with DLAD assistance (all p-values < 0.001). The

Table 3. Reader Test to Detect Malignant Pulmonary Nodules on Chest Radiographs according to Reference Standards

Session	Diagnostic Measure	Reference Standards					
		Clinical Gold Standard*	Individual Judgment[†]	Majority Vote[‡]	Consensus of Two Experts	Consensus of Three Experts	LCA Model[§]
Reader test without DLAD	Sensitivity	66.4% (65.7–67.2)	77.3% (76.6–78.1)	83.8% (83.0–84.5)	84.6% (83.9–85.3)	84.8% (84.2–85.5)	83.2% (81.5–83.0)
	Specificity	94.2% (93.7–94.7)	88.2% (87.6–88.8)	87.9% (87.3–88.5)	84.7% (84.1–85.3)	84.9% (84.2–85.5)	89.6% (89.4–90.6)
	Accuracy	80.3% (79.9–80.7)	84.1% (83.7–84.6)	86.5% (86.0–86.9)	84.7% (84.3–85.1)	84.8% (84.4–85.3)	87.3% (86.7–87.6)
Reader test with DLAD	Sensitivity	72.8% (71.9–73.7)	80.7% (79.9–81.5)	86.6% (85.7–87.4)	88.5% (87.7–89.3)	88.5% (87.7–89.2)	88.3% (87.5–89.2)
	p-value[‖]	< 0.001	< 0.001	< 0.001	< 0.001	< 0.001	< 0.001
	Specificity	97.7% (97.5–98.0)	88.8% (88.4–89.3)	87.3% (86.8–87.8)	85.0% (84.5–85.6)	85.2% (84.7–85.7)	90.8% (90.4–91.3)
	p-value[‖]	< 0.001	0.094	0.134	0.459	0.465	0.001
	Accuracy	84.9% (84.4–85.4)	85.6% (85.3–85.9)	87.0% (86.7–87.3)	86.1% (85.7–86.5)	86.2% (85.9–86.5)	89.9% (89.6–90.2)
	p-value[‖]	< 0.001	< 0.001	0.069	< 0.001	< 0.001	< 0.001
DLAD standalone	Sensitivity	82.6% (81.9–83.3)	87.0% (86.4–87.7)	94.6% (93.9–95.3)	97.5% (96.8–98.2)	94.1% (93.4–94.8)	97.4% (96.7–98.1)
	Specificity	100% (99.3–100)	86.6% (85.9–87.3)	85.5% (84.9–86.2)	83.5% (82.8–84.2)	82.1% (81.4–82.8)	90.0% (89.3–90.7)
	Accuracy	91.0% (90.3–91.7)	86.7% (86.0–87.4)	88.6% (88.0–89.3)	87.9% (87.3–88.6)	85.9% (85.3–86.6)	92.7% (92.0–93.4)

Parentheses are 95% confidence intervals. The *p*-values of all comparisons between the clinical gold standard and each expert-determined standard were < 0.05. *Clinical gold standard: pathology-proven and CT-proven malignant pulmonary nodules, [†]Individual judgment of expert 5, who had the most experience in thoracic radiology, [‡]The judgment of the majority vote was determined by seven experts, [§]2-D-class LCA, [‖]Comparison of diagnostic performance between radiologists alone and radiologists assisted by the DLAD model. LCA = latent class analysis, DLAD = deep learning-based automatic detection

improvement in radiologists' performance by DLAD assistance was confirmed in terms of sensitivity and specificity, even when expert-determined standards were used as the ground truth. However, the corresponding effects were estimated to be significantly lower than when the clinical gold standard was used as the ground truth (all *p*-values < 0.001), except for the sensitivity with the LCA model ($p = 0.094$).

As little is known about the implications of using expert-determined standards as the ground truth when evaluating the performance and usefulness of DL models, the research results should be cautiously interpreted in terms of which type of reference standard is used as the ground truth [2]. Indeed, while clinical gold standards for the disease have been set as the ground truth in many studies, several other studies have set experts' opinions (e.g., radiological reports from radiologists) as the ground truth even for the same task. For example: 1) DL models detecting malignant pulmonary nodules, which were set based on pathologically proven lesions [16,19] or partially clinically diagnosed malignancies [14,15]; 2) DL models detecting clinically relevant diseases on chest radiographs based on the disease-specific gold standard [8] or expert interpretations [20-24]; and 3) DL models predicting lung nodule malignancy on chest CT based on pathology [25-27] or expert interpretations [28-30]. Of course, clinical gold standards are not always affordable or attempted in real-world clinical settings [31,32]. For example, cancers can be clinically diagnosed and treated without pathological results because patients may have conditions that make them ineligible for biopsy or surgery or due to the potential risk of procedure-related complications [33]. This occasional unavailability of a clinical gold standard requires expert-determined standards by clinical domain experts. In a previous study, lung metastasis was adjudicated by thoracic radiologists who used both clinical and pathological information [15]. Using experts' opinions reflects the real-world clinical setting and is an easy way to set a clinical gold standard; however, the reliability remains questionable regarding whether this method yields a genuine assessment of a DL model's performance. In contrast, another study used only pathologically proven lung cancer as their clinical gold standard; this strategy can avoid measurement bias but is prone to selection bias [16]. Therefore, knowledge of the implications of expert-determined standards for evaluating the diagnostic performance of DL models will strengthen the applicability of DL algorithms to real-world clinical settings.

It should be noted that the LCA model had the highest diagnostic performance among the five constructed expert-determined standards. We suggest a high concordance with the LCA model derived from the model's characteristics of learning the inherent propensities of each expert and automatically making an interactive decision on that basis [17,18]. This is because the LCA model rigorously identified the propensities of all 10 experts and then presented the results, so the model

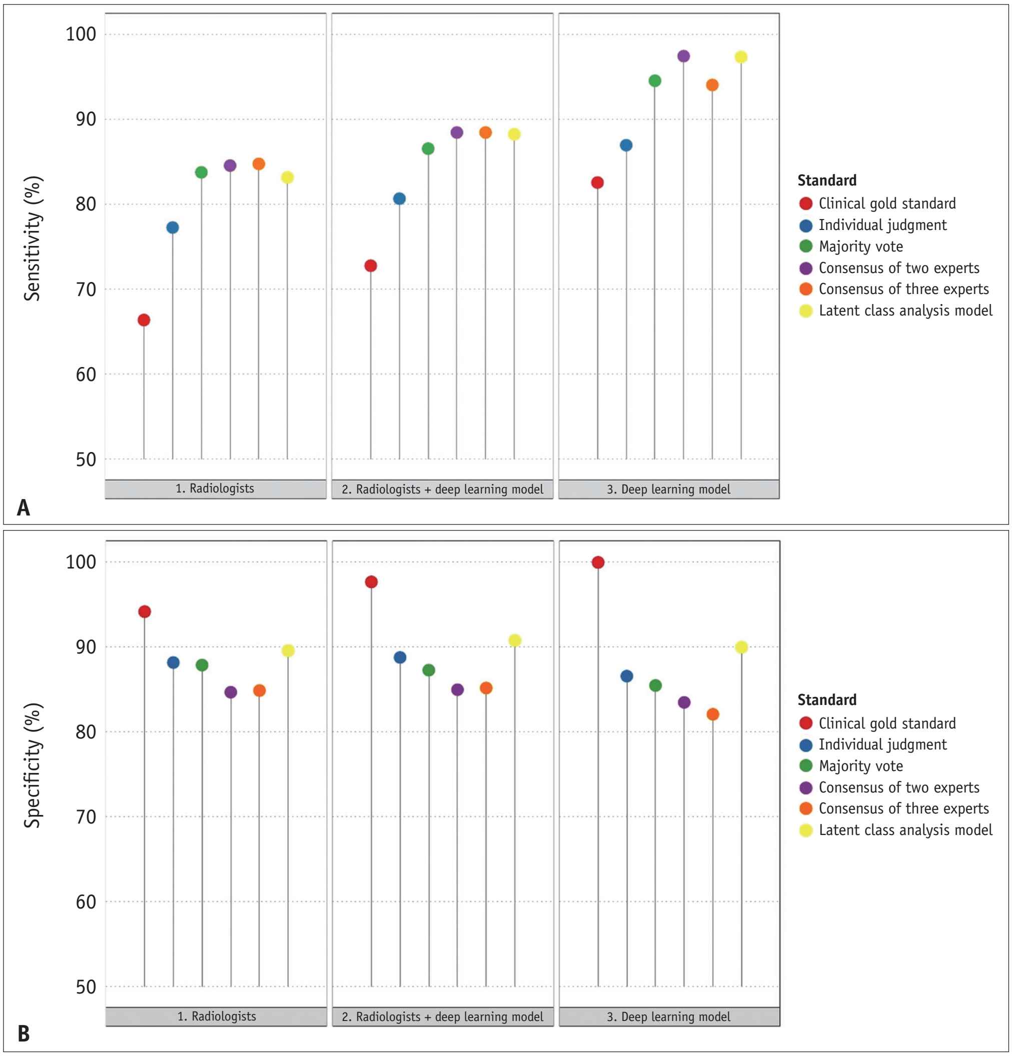

Fig. 2. Plots for the diagnostic performance of radiologists alone, radiologists assisted by a deep learning-based automatic detection (DLAD) model, and the DLAD model alone.
A. Plot for sensitivity. Each color represents each standard. Compared to the clinical gold standard by pathological confirmation (red circle), the sensitivities are overestimated when expert-determined standards are applied. These biases in the sensitivities assessed by the expert-determined standards are diminished when the DLAD model assists radiologists. The added value of the DLAD model is significantly lower when assessed by expert-determined standards than when using the clinical gold standard. **B.** Plot for specificity. Each color represents each standard. Compared to the clinical gold standard by pathological confirmation (red circle), the specificities are underestimated when expert-determined standards are applied. These biases in the specificities assessed by the expert-determined standards are exaggerated when the DLAD model assists radiologists. The added value of the DLAD model is significantly lower when assessed by expert-determined standards than when using the clinical gold standard.

could be free from the biased tendencies of a limited number of experts. We speculate that this might explain the high agreement between the LCA model and the 10 radiologists, leading to the high sensitivity of the reader tests.

In the reader study, we verified the tendency to overestimate sensitivity and underestimate specificity when expert-determined standards were applied as ground truth compared to the clinical gold standard. We speculate that since the expert-determined standards had the inherent characteristics of low sensitivity (60.4%–72.6%) and high specificity (94.2%–100%) in the concordance analyses, they more easily yielded negative results. Readers often misgrade true-positive cases as false positives and false negatives as true-negative cases, leading to these tendencies. Another explanation is that the 10 experts and 10 radiologists participating in the reader tests had common characteristics in reading chest radiographs as human beings (e.g., only detecting clearly visible lung nodules). Interestingly, this tendency was also observed in the DLAD model because it was trained using datasets of chest radiographs with visible lung cancer, as determined by experts [8,14].

This study had several limitations. First, the small

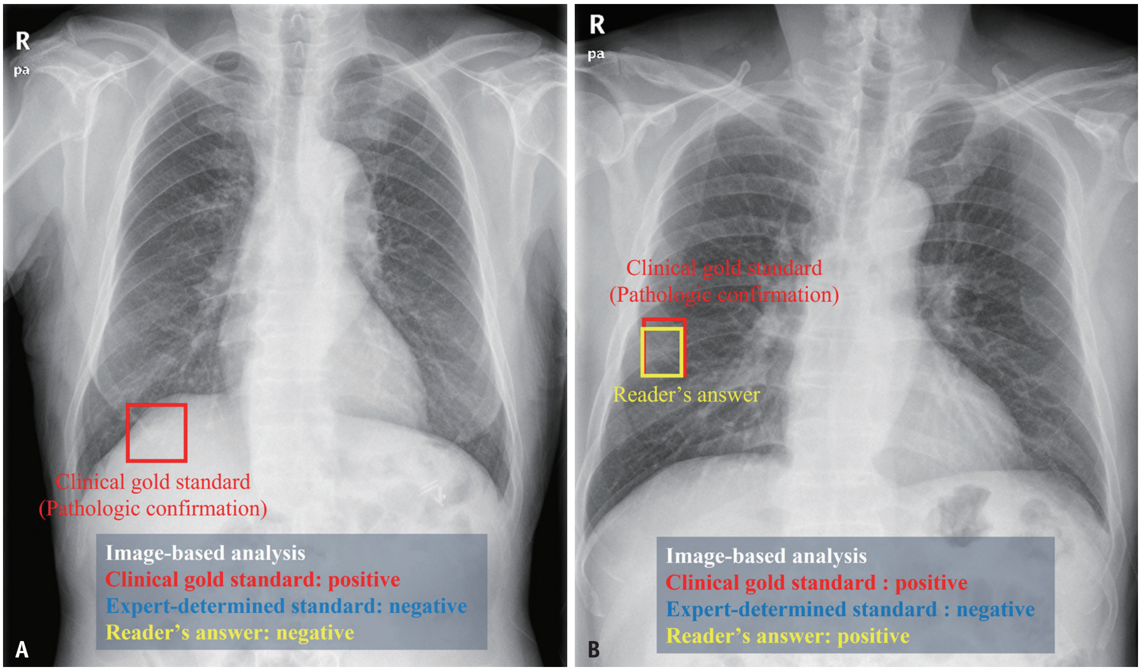
R
pa
Clinical gold standard
(Pathologic confirmation)
Image-based analysis
Clinical gold standard: positive
Expert-determined standard: negative
Reader's answer: negative
A
R
pa
Clinical gold standard
(Pathologic confirmation)
Reader's answer
Image-based analysis
Clinical gold standard : positive
Expert-determined standard : negative
Reader's answer: positive
B

Fig. 3. Representative figures of biased diagnostic performance when expert-determined standards are applied as ground truth in the reader study.
A. Representative cases of overestimating sensitivity. A chest radiograph with lung cancer in the right lower lobe (pathologically proven adenocarcinoma). In both image-based and lesion-based analyses, the radiologist's answer that the radiograph is normal is considered a false negative with the clinical gold standard by pathological confirmation (red box), but a true negative when any expert-determined standard is included. This can lead to an overestimation of sensitivity. **B.** Representative case of underestimating specificity. Another radiograph shows lung cancer in the right lower lobe (pathologically proven adenocarcinoma). In the image-based and lesion-based analyses, the radiologist's answer (yellow box) is a true positive with the clinical gold standard by pathological confirmation (red box) but a false positive with any expert-determined standard. This can lead to an underestimation of specificity.

sample size could have limited the results of this study. Second, although 10 radiologists with sufficient experience in chest imaging participated in the reader tests, their performance might not represent radiologists in general, including radiologists with other specialties or less experienced radiologists (residents or fellows). As less experienced radiologists tend to be more strongly influenced by the DLAD model [8,34], more studies that include these groups are warranted. Third, we only used the malignant pulmonary nodule detection task on chest radiographs as a representative task to simulate a setting in which expert-determined standards were used as the ground truth. However, task characteristics, sample size, disease prevalence, independency and superiority between the reference standard and index test, correction methods (e.g., LCA model), and their interaction terms can affect the results of imperfect reference standards [13]. Therefore, the results of this study should not be interpreted directly in the context of other tasks or other AI algorithms. Further studies investigating the impact of expert-determined standards on various tasks, considering the above influencing factors, are warranted. Finally, we did not use various correction methods other than the LCA model to compensate for expert-determined standards, such as the method of Staquet et al. [13].

In conclusion, the LCA model was most similar to the clinical gold standard for detecting malignant pulmonary nodules on chest radiographs. Expert-determined standards led to a bias in measuring the diagnostic performance of the AI model, highlighting the importance of high-quality reference standards to evaluate the performance of AI.

Supplement

The Supplement is available with this article at https://doi.org/10.3348/kjr.2022.0548.

Availability of Data and Material

The datasets generated or analyzed during the study are available from the corresponding author on reasonable request.

Conflicts of Interest

This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors. Activities not related to the present article: E.J.H. received a research grants from Lunit, Coreline soft, and Monitor corportation. C.M.P. received research grants from Lunit and Coreline Soft, and holds stock in Promedius and stock options in Lunit and Coreline Soft.

Author Contributions

Conceptualization: Jung Eun Huh, Jong Hyuk Lee, Chang Min Park. Data curation: Jung Eun Huh, Jong Hyuk Lee. Formal analysis: Jung Eun Huh, Jong Hyuk Lee. Investigation: Jung Eun Huh, Jong Hyuk Lee. Methodology: Jung Eun Huh, Jong Hyuk Lee. Writing—original draft: Jung Eun Huh, Jong Hyuk Lee. Writing—review & editing: all authors.

ORCID iDs

Jung Eun Huh
https://orcid.org/0000-0001-6565-7036
Jong Hyuk Lee
https://orcid.org/0000-0002-9594-683X
Eui Jin Hwang
https://orcid.org/0000-0002-3697-5542
Chang Min Park
https://orcid.org/0000-0003-1884-3738

Funding Statement

None

Acknowledgments

The authors would like to acknowledge Andrew Dombrowski, Ph.D. (Compecs, Inc.) for his assistance in English editing.

REFERENCES

1. Rajpurkar P, Chen E, Banerjee O, Topol EJ. AI in health and medicine. *Nat Med* 2022;28:31-38
2. Chen PC, Mermel CH, Liu Y. Evaluation of artificial intelligence on a reference standard based on subjective interpretation. *Lancet Digit Health* 2021;3:e693-e695
3. Schlemmer HP, Bittencourt LK, D'Anastasi M, Domingues R, Khong PL, Lockhat Z, et al. Global challenges for cancer imaging. *J Glob Oncol* 2018;4:1-10
4. King BF Jr. Artificial intelligence and radiology: what will the future hold? *J Am Coll Radiol* 2018;15(3 Pt B):501-503
5. Esteva A, Kuprel B, Novoa RA, Ko J, Swetter SM, Blau HM, et al. Dermatologist-level classification of skin cancer with deep neural networks. *Nature* 2017;542:115-118
6. Gulshan V, Peng L, Coram M, Stumpe MC, Wu D, Narayanaswamy A, et al. Development and validation of a deep learning algorithm for detection of diabetic retinopathy in retinal fundus photographs. *JAMA* 2016;316:2402-2410
7. Ehteshami Bejnordi B, Veta M, Johannes van Diest P, van Ginneken B, Karssemeijer N, Litjens G, et al. Diagnostic assessment of deep learning algorithms for detection of lymph node metastases in women with breast cancer. *JAMA* 2017;318:2199-2210
8. Hwang EJ, Park S, Jin KN, Kim JI, Choi SY, Lee JH, et al. Development and validation of a deep learning–based automated detection algorithm for major thoracic diseases on chest radiographs. *JAMA Netw Open* 2019;2:e191095
9. Liu Y, Chen PC, Krause J, Peng L. How to read articles that use machine learning: users' guides to the medical literature. *JAMA* 2019;322:1806-1816
10. Cohen JF, Korevaar DA, Altman DG, Bruns DE, Gatsonis CA, Hooft L, et al. STARD 2015 guidelines for reporting diagnostic accuracy studies: explanation and elaboration. *BMJ Open* 2016;6:e012799
11. U.S. Food & Drug Administration. Statistical guidance on reporting results from studies evaluating diagnostic tests—guidance for industry and FDA staff. FDA.gov Web site. https://www.fda.gov/regulatory-information/search-fda-guidance-documents/statistical-guidance-reporting-results-studies-evaluating-diagnostic-tests-guidance-industry-and-fda. Accessed March 24, 2022
12. Benchoufi M, Matzner-Lober E, Molinari N, Jannot AS, Soyer P. Interobserver agreement issues in radiology. *Diagn Interv Imaging* 2020;101:639-641
13. Umemneku Chikere CM, Wilson KJ, Allen AJ, Vale L. Comparative diagnostic accuracy studies with an imperfect reference standard-a comparison of correction methods. *BMC Med Res Methodol* 2021;21:67
14. Lee JH, Sun HY, Park S, Kim H, Hwang EJ, Goo JM, et al. Performance of a deep learning algorithm compared with radiologic interpretation for lung cancer detection on chest radiographs in a health screening population. *Radiology* 2020;297:687-696
15. Hwang EJ, Lee JS, Lee JH, Lim WH, Kim JH, Choi KS, et al. Deep learning for detection of pulmonary metastasis on chest radiographs. *Radiology* 2021;301:455-463
16. Nam JG, Park S, Hwang EJ, Lee JH, Jin KN, Lim KY, et al. Development and validation of deep learning–based automatic detection algorithm for malignant pulmonary nodules on chest radiographs. *Radiology* 2019;290:218-228
17. Uebersax JS, Grove WM. Latent class analysis of diagnostic agreement. *Stat Med* 1990;9:559-572
18. Dillon WR, Mulani N. A probabilistic latent class model for assessing inter-judge reliability. *Multivariate Behav Res* 1984;19:438-458
19. Sim Y, Chung MJ, Kotter E, Yune S, Kim M, Do S, et al. Deep convolutional neural network–based software improves radiologist detection of malignant lung nodules on chest

radiographs. *Radiology* 2020;294:199-209

20. Jin KN, Kim EY, Kim YJ, Lee GP, Kim H, Oh S, et al. Diagnostic effect of artificial intelligence solution for referable thoracic abnormalities on chest radiography: a multicenter respiratory outpatient diagnostic cohort study. *Eur Radiol* 2022;32:3469-3479
21. Kim JH, Kim JY, Kim GH, Kang D, Kim IJ, Seo J, et al. Clinical validation of a deep learning algorithm for detection of pneumonia on chest radiographs in emergency department patients with acute febrile respiratory illness. *J Clin Med* 2020;9:1981
22. Majkowska A, Mittal S, Steiner DF, Reicher JJ, McKinney SM, Duggan GE, et al. Chest radiograph interpretation with deep learning models: assessment with radiologist-adjudicated reference standards and population-adjusted evaluation. *Radiology* 2020;294:421-431
23. Sung J, Park S, Lee SM, Bae W, Park B, Jung E, et al. Added value of deep learning–based detection system for multiple major findings on chest radiographs: a randomized crossover study. *Radiology* 2021;299:450-459
24. Cui S, Ming S, Lin Y, Chen F, Shen Q, Li H, et al. Development and clinical application of deep learning model for lung nodules screening on CT images. *Sci Rep* 2020;10:13657
25. Venkadesh KV, Setio AAA, Schreuder A, Scholten ET, Chung K, W Wille MM, et al. Deep learning for malignancy risk estimation of pulmonary nodules detected at low-dose screening CT. *Radiology* 2021;300:438-447
26. Massion PP, Antic S, Ather S, Arteta C, Brabec J, Chen H, et al. Assessing the accuracy of a deep learning method to risk stratify indeterminate pulmonary nodules. *Am J Respir Crit Care Med* 2020;202:241-249
27. Ohno Y, Aoyagi K, Yaguchi A, Seki S, Ueno Y, Kishida Y, et al. Differentiation of benign from malignant pulmonary nodules by using a convolutional neural network to determine volume change at chest CT. *Radiology* 2020;296:432-443
28. Shen W, Zhou M, Yang F, Yu D, Dong D, Yang C, et al. Multi-crop convolutional neural networks for lung nodule malignancy suspiciousness classification. *Pattern Recognit* 2017;61:663-673
29. Shen S, Han SX, Aberle DR, Bui AA, Hsu W. An interpretable deep hierarchical semantic convolutional neural network for lung nodule malignancy classification. *Expert Syst Appl* 2019;128:84-95
30. Causey JL, Zhang J, Ma S, Jiang B, Qualls JA, Politte DG, et al. Highly accurate model for prediction of lung nodule malignancy with CT scans. *Sci Rep* 2018;8:9286
31. Umemneku Chikere CM, Wilson K, Graziadio S, Vale L, Allen AJ. Diagnostic test evaluation methodology: a systematic review of methods employed to evaluate diagnostic tests in the absence of gold standard-An update. *PLoS One* 2019;14:e0223832
32. Bertens LC, Broekhuizen BD, Naaktgeboren CA, Rutten FH, Hoes AW, van Mourik Y, et al. Use of expert panels to define the reference standard in diagnostic research: a systematic review of published methods and reporting. *PLoS Med* 2013;10:e1001531
33. Shaikh T, Churilla TM, Murphy CT, Zaorsky NG, Haber A, Hallman MA, et al. Absence of pathological proof of cancer associated with improved outcomes in early-stage lung cancer. *J Thorac Oncol* 2016;11:1112-1120
34. Hwang EJ, Nam JG, Lim WH, Park SJ, Jeong YS, Kang JH, et al. Deep learning for chest radiograph diagnosis in the emergency department. *Radiology* 2019;293:573-580