의학 박사 학위논문

검진 코호트에서 흉부 X 선 영상에 대한 딥러닝 기반 인공지능 모델의 유용성 평가

2022년 2월

서울대학교 대학원

의학과 영상의학 전공

이 종 혁

검진 코호트에서 흉부 X선 영상에 대한 딥러닝 기반 인공지능 모델의 유용성 평가

2022년

이종혁

의학 박사 학위논문

검진 코호트에서 흉부 X 선 영상에 대한 딥러닝 기반 인공지능 모델의 유용성 평가

2022년 2월

서울대학교 대학원

의학과 영상의학 전공

이 종 혁

Doctoral Thesis

Validation of Deep-Learning Algorithms in Chest Radiographs in Screening Cohorts

February 2022

The Graduate School, Seoul National University

Medicine – Radiology

Jong Hyuk Lee

검진 코호트에서 흉부 X 선 영상에 대한 딥러닝 기반 인공지능 모델의 유용성 평가

지도 교수 박 창 민

이 논문을 의학박사 학위논문으로 제출함
2021년 10월

서울대학교 대학원
의학과 영상의학 전공
이 종 혁

이종혁의 의학박사 학위논문을 인준함
2022년 1월

위 원 장	구 진 모	(인)
부위원장	박 창 민	(인)
위 원	임 재 준	(인)
위 원	이 재 성	(인)
위 원	이 상 민	(인)

ABSTRACT

Objectives: To validate deep-learning (DL) algorithms for detecting active pulmonary tuberculosis and lung cancers in screening chest radiographs and optimizing candidate selection for lung cancer CT screening (LCS).

Methods: Validation of DL algorithms were performed using chest radiographs from the following cohorts: 1) a cohort undergoing systematic screening for tuberculosis between January 2013 and July 2018, 2) a cohort in a single check-up center between January 2008 and December 2012 (for detecting lung cancers), and 3) a cohort in the same health check-up center between January 2004 and June 2018 (for optimizing selection for lung cancer CT screening candidates). The area under the receiver operating characteristic curves (AUC) for detecting tuberculosis and lung cancers and prediction of lung cancers were measured. For lesion-detection tasks, accuracy measures including sensitivities, specificities, positive predictive values (PPVs), negative predictive values (NPVs) were calculated at pre-defined operating thresholds (for tuberculosis: high sensitivity threshold=0.16, high specificity threshold= 0.46; for lung cancers: high sensitivity threshold=0.16). For identifying LCS candidates, discrimination and calibration of the model for incident lung cancer and its added value to

the 2021 US Preventive Services Task Force (USPSTF) recommendations were evaluated in terms of the lung cancer detection rate, the proportion of selected CT screening candidates, and PPV.

Results: In a systematic screening cohort for tuberculosis of 20,235 chest radiographs from 19,686 asymptomatic individuals (21±2 years, 19,475 men), all five radiographs from four individuals with active pulmonary tuberculosis were correctly classified as having abnormal findings by the DL algorithm with specificities of 95.9% and 99.7%, PPVs of 0.6% and 6.8%, and NPVs of both 100% at high sensitivity and specificity thresholds, respectively. With high specificity thresholds, DL algorithm showed comparable diagnostic measures to the pooled radiologists (P-values>0.05). As for lung cancers, in a subset comprising 10,285 chest radiographs from 10,202 individuals (54±11 years, 5,857 men) with 10 radiographs of visible lung cancers, the algorithm's AUC was 0.989 (95% confidence interval [CI]: 0.968 – 0.999), and it showed comparable sensitivity (90% [9 of 10]) to the radiologists (60% [6 of 10], P=0.248) with a lower specificity (96.9% [9,956 of 10,275] vs. 99.8% [10,249 of 10,275], P<0.001). In the screening cohort of 100,525 radiographs from 50,070 individuals (53±11 years, 28,090 men) with 47 radiographs of visible lung cancers, the algorithm's AUC was 0.969 (95% CI: 0.946 – 0.992),

and its sensitivity and specificity were 83% (39 of 47) and 97% (97,479 of 100,478), respectively. For optimization of candidate selection for LCS in the entire population and the subset of USPSTF-eligible individuals, the AUCs were 0.677 (95% CI: 0.623 – 0.731) and 0.745 (95% CI: 0.677 – 0.813), respectively. In individuals with pack-year information (n=17,390), when the model-driven optimization strategy was applied to the USPSTF-eligible population by excluding low-to-indeterminate risk, the proportion of selected CT screening candidates decreased to 35.8% (6,233 of 17,390) from 45.1% (7,835 of 17,390; P<0.001) with 3 missed lung cancers (0.19% [3 of 1,602]). The lung cancer detection rate (0.3% [53 of 17,390]; P=0.848) and PPV (0.9% [53 of 6,233]; P=0.416) remained unaffected.

Conclusion: Deep-learning algorithms can be a promising tool in real-world screening chest radiographs in terms of detecting active pulmonary TB and lung cancers, and optimizing candidate selection for lung cancer CT screening.

Keywords: deep learning; diagnosis; screening; tuberculosis, lung cancer, chest radiographs, computer-assisted

Student Number: 2020-36356

목 차

List of Tables

List of Figures

List of Abbreviations

DL = Deep learning

TB = Tuberculosis

AUC = area under the receiver operating characteristic curve

PPV = positive predictive value

NPV = negative predictive value

USPSTF =US Preventive Services Task Force

PLCO = Prostate, Lung, Colorectal, and Ovarian Cancer Screening Trial

NLST = National Lung Screening Trial

LCS = Lung cancer CT screening

INTRODUCTION

Deep learning (DL) algorithms have demonstrated excellent performance in various fields of medicine, including detecting active pulmonary tuberculosis (TB) and lung cancers on chest radiographs and predicting lung cancers (1-5). TB is the single most common life-threatening infectious disease and one of the leading causes of death worldwide (6-8). To reduce transmission of and mortality from TB, the World Health Organization (WHO) recommends systematic screening for active TB in high-risk groups, and a symptom screen has been a major initial test (7-11). However, up to half of bacteriologically-confirmed TB cases do not have any symptom (7-9). In a recent review of TB prevalence surveys conducted since 2000, more than 50% of all confirmed cases reported no symptoms at the time of diagnosis (8). It is increasingly clear that symptom screening is insufficiently sensitive as an initial triage tool in mass screening (12). Chest radiographs have long been used as an alternative screening test, because of their high sensitivity for pulmonary TB (7, 8). However, there have been limitations associated with their use, including low specificity for pulmonary TB (46% to 86%) and intra-reader and inter-reader variabilities (6-8). Additionally, in resource-

constrained environments, there is often a lack of trained experts who can interpret the radiographs (6, 7, 13). In the context of mass screening, interpretation of chest radiographs is time-consuming and labor-intensive; involvement of experts may be costly for programs that are intended to be cost-efficient (7, 8). To overcome these issues, computer-aided detection (CAD) has been developed and applied to the detection of TB (14-18). While these have shown promise in some studies, CAD systems have had limited accuracy in screening of asymptomatic individuals (7, 15).

As for TB, multiple previous studies dealing with DL algorithms have already reported the excellent classification performance of DL algorithm on chest radiographs in disease-enriched datasets (1, 2, 4). Hwang et al. (2) reported in their external validation study consisting of six independent test datasets (TB proportion, 39% to 60%) that their DL algorithm showed very excellent detection performance of active pulmonary TB on radiographs, in which the sensitivities and specificities ranged between 94.3 - 100% and 91.1 - 100%, respectively at high sensitivity threshold and 84.1 - 99.0% and 99.1 - 100%, respectively at high specificity threshold. In addition, Qin et al. also showed that their DL algorithm had excellent classification performance (area under the curve [AUC] of 0.94) of active pulmonary TB in a high-

prevalence setting (TB prevalence of 9.1%) (4).

Lung cancer is a leading cause of cancer death worldwide, accounting for up to one-quarter of all cancer deaths (19). Since lung cancers are diagnosed in an advanced stage in most cases, screening of early-stage lung cancer has emerged as a strategy for reducing lung cancer mortality (20-22). Indeed, lung cancer CT screening (LCS) for high-risk smokers reduced lung cancer mortality by 20% in the National Lung Screening Trial (NLST) and by 24% in male participants in the Nederlands-Leuvens Longkanker Screenings Onderzoek (NELSON) trial (23-25). In contrast, the value of using chest radiographs as a screening modality could not be proven for either early lung cancer detection or lung cancer mortality reduction (26). Thus, it is controversial whether lung cancer screening should be performed using chest radiographs. Nonetheless, chest radiographs are widely used as an initial screening tool for several important thoracic diseases, including lung cancer, in the general population thanks to their low cost, easy accessibility, negligible radiation dose, and reasonable diagnostic capability (27-29). Nam et al. reported that a DL algorithm achieved a sensitivity of 71 – 91%, a specificity of 93 – 100%, and an AUC of 0.92 – 0.99 in their validation datasets with a lung cancer prevalence of approximately 60 – 68% (3). Sim

et al. reported that another DL algorithm had comparable diagnostic performance to that of radiologists in the detection of lung cancer in their study population composed of 75% lung cancer-containing chest radiographs (30). Also, they showed that assistance from their algorithm improved sensitivity (from 65% to 73%) and reduced the false-positive rate (FPR) (from 20% to 18%) (30). But those previous studies (1-4, 30) validated their DL algorithms with arbitrarily-selected test datasets, instead of datasets reflecting real-world clinical practice.

Meanwhile, concerns regarding LCS include the considerable number of negative screening examinations (i.e., normal baseline or follow-up screens) with unnecessary radiation exposure and medical expenditures, and false-positive results that potentially lead to invasive diagnostic procedures (23, 24, 31-38). Lu et al. (5) recently developed and validated a deep-learning model to identify high-risk candidates for LCS. The model uses easily-obtainable inputs including age, sex, smoking status, and a chest radiograph image, and it showed a higher AUC and sensitivity than the Centers for Medicare & Medicaid Services eligibility criteria (AUC, 0.755 vs. 0.634, $P<0.001$; sensitivity, 74.9% vs. 63.8%, $P=0.012$), while missing 30.7% fewer incident lung cancers (5). However, model validation in the prior study was

based on publicly available chest radiographs from 2 US clinical trials which first enrolled participants in 1993 (24, 39). Therefore, for generalizability, external validation with recent real-world data including non-US data is necessary (5). In addition, the potential interaction of the model output with the recently updated 2021 US Preventive Services Task Force (USPSTF) recommendations was not analyzed (40).

The purpose of this study was to evaluate the usefulness of DL algorithms on chest radiographs in real-world screening settings in terms of detecting active pulmonary TB and lung cancers and optimizing candidate selection for LCS. This dissertation contains the contents previously published in two journals at the time of the examination (41, 42).

MATERIALS AND METHODS

This retrospective studies were approved by the institutional review board of the Armed Forces Medical Command of Korea (IRB number: AFMC-18028-IRB-18-025) and Seoul National University Hospital (IRB number: 1808-038-964, 2010-174-1169), and the requirement for informed consent was waived.

Study population

Systematic screening for TB

This study was performed at an armed forces hospital which covers a majority of military personnel in the capital city, Seoul. We collected all chest radiographs from servicepersons who visited the hospital for the routine medical check-up and, underwent chest radiographs for TB screening between January 2013 and July 2018. Screening was done systematically as part of routine evaluation for TB, rather than prompted by evaluation for clinical symptoms (7, 8). If servicepersons had clinical symptoms such as fever, cough, or sputum, they were referred to clinical interview with medical doctors and excluded from this screening program.

Detection of lung cancers on chest radiographs

We collected all chest radiographs from individuals who participated in a health check-up program at Healthcare System Gangnam Center, Seoul, Korea between January 2008 and December 2012. The center provides a comprehensive medical check-up and screening program for non-communicable diseases such as malignancies (43), and chest radiographs are a core test in this screening program to detect any lung disease requiring further diagnostic tests or treatments (43). The participants in this study paid the screening costs at their own expense and they were not assessed based on the predefined lung cancer risk factors. In this regard, the study population was an average-risk general population, rather than an LCS population that was carefully selected in terms of age and history of cigarette smoking. All individuals underwent chest radiographs as part of the health check-up, not for an evaluation of specific symptoms or signs.

The following individuals were excluded from the study population: (a) individuals with a past history of lung cancer; (b) those with lung lesions pathologically confirmed as pre-invasive lesions of lung cancer, not definitive lung cancer.

Optimization of candidate selection for LCS

All chest radiographs from consecutive participants in the same center (Healthcare System Gangnam Center) between January 2004 and June 2018 were collected. If an individual had several radiographs during the study period, we included a chest radiograph taken at the time the individual's smoking status was first recorded. Inclusion criteria were then applied: (a) individuals who had posteroanterior chest radiographs, (b) individuals aged 50 to 80 years as recently recommended by the USPSTF (40), and (c) current or former smokers (5). The exclusion criteria were: (a) individuals with pre-invasive lesions (e.g., adenocarcinoma in situ, atypical adenomatous hyperplasia) or metastasis from extrathoracic malignancy (n=4), and (b) those with presumptive lung cancers that were not pathologically or clinically confirmed, as the ground truth for those nodules was indeterminate (n=4).

In subgroup analyses, we aimed to validate the CXR-LC model for predicting incident lung cancer within 3 and 5 years, respectively, from the date of chest radiographs. Among those diagnosed with lung cancer, individuals with less than 3 and 5 year intervals between chest radiographs and the date of diagnosis were included. For those without incident lung cancer, we included individuals who had low-dose chest CT examinations

at least 3 and 5 years after their radiographs, respectively, to strictly guarantee the absence of incident lung cancer. Keyword searching of the CT reports was then performed to identify and exclude individuals with indeterminate lung nodules or who had undergone lung resection.

Reference standard

Systematic screening for TB

The primary task was detection of active pulmonary TB, as defined by a positive microbiological test (smear microscopy, culture, or TB polymerase chain reaction) (7). The second task was classification of radiologically-identifiable relevant abnormalities. The “radiologically-identifiable relevant abnormalities” in the present study refer to abnormalities detected on chest radiographs which require further diagnostic or therapeutic actions. We defined the reference standards of the second task with a three-step process applied to each radiograph. First, we considered the original radiological report of each radiograph as the first read, which was originally read by one of seven board-certified radiologists. Second, one experienced radiologist (J.H.L. with seven years of experience in thoracic radiology) blind to original radiological reports and available patients’ information, read all chest

radiographs independently, which was the second read. Third, for the chest radiographs read as having abnormal radiological findings either on the first or second read, two experienced radiologists (C.M.P. and J.M.G with 21 and 29 years of experience in thoracic radiology, respectively) performed a consensus reading and jointly determined whether or not each radiograph contained “radiologically-identifiable relevant abnormalities.” In this step, lesion conspicuity and relevant level of each radiologic abnormality were labeled using 3 levels: no visible, visible but uncertain, certainly visible for lesion conspicuity; non relevant, equivocally relevant, and certainly relevant for relevant level. Chest radiographs that were judged as having ‘no visible’ or ‘non relevant’ abnormalities in the third step were classified as not having radiologically-identifiable relevant abnormalities and the remainder were classified as having abnormalities.

Detection of lung cancers on chest radiographs and optimization of candidate selection for LCS

Individuals diagnosed with lung cancer by November 2020 were identified through a search of electronic medical records. The health check-up center has a patient referral system through which individuals requiring further

diagnostic or therapeutic management are referred to a tertiary referral hospital (Seoul National University Hospital, Seoul, Korea). Therefore, individuals who had chest radiographs at the health check-up center and were subsequently diagnosed with lung cancer at further follow-up visits could be recognized.

For the task of detection of lung cancers, we determined cancer-positive chest radiographs from individuals diagnosed with lung cancers using the following criteria: (a) lung cancer present on a chest CT scan taken within 3 months of the chest radiograph, (b) if chest CT was not available, a chest radiograph taken within 15 months before being diagnosed with lung cancer.

In contrast, if lung cancer or any significant but not confirmed as a benign nodule (i.e., non-calcified nodules of 6 mm or larger) was not present on a chest CT within 3 months of the chest radiographs, the radiographs were classified as cancer-negative. In addition, if chest CT was not available and follow-up radiographs after 12 months or longer revealed cancer-negative results, the prior chest radiographs were regarded as cancer-negative.

For cancer-positive chest radiographs, two board-certified radiologists (J.H.L. and E.J.H., with 7 and 9 years of experience in thoracic radiology,

respectively) independently assessed the visibility of lung cancer on each chest radiograph, referring to the available chest CT examinations. In this assessment, the lung cancers on chest radiographs were dichotomized as visible or invisible. Finally, lung cancers on chest radiographs designated as visible by either radiologist were classified as visible lung cancers on chest radiographs, and chest radiographs concordantly judged as visible by both radiologists were categorized as clearly visible lung cancers on chest radiographs. Lung cancers determined as invisible by both radiologists were classified as invisible lung cancers on chest radiographs.

Deep learning algorithms

Systematic screening for TB and detection of lung cancers on chest radiographs

A commercially-available DL algorithm (Lunit INSIGHT for Chest Radiography, version 4.7.2; Lunit) was used. The algorithm was developed for the detection of major thoracic diseases (1). The DL algorithm provides both an image-wise probability value of a chest radiograph being abnormal, and a per-pixel localization map overlaid on the input chest radiograph identifying the location of abnormalities. All localization maps of chest

radiographs with positive results from the DL algorithm were checked to ensure that the algorithm adequately localized each lung cancer lesion on the chest radiographs (SUPPLEMENT).

Optimization of candidate selection for LCS

The CXR-LC model was developed based on a convolutional neural network to predict long-term (up to 12-year) incident lung cancer using data from the Prostate, Lung, Colorectal, and Ovarian Cancer Screening Trial (PLCO) and externally validated in the NLST (5). The inputs are clinical information (age, sex, and smoking status) and chest radiograph image, and the model calculates a probability for incident lung cancer ranging between 0% and 100%. More detailed information can be found in the original study (5), and the model is available online (https://github.com/vineet1992/CXR-LC). According to Lu et al. (5), the CXR-LC risk probabilities could be converted to the following ordinal risk categories : (a) low risk (< 2%), (b) indeterminate risk (2% to < 3.297%), (c) high risk (3.297% to < 8%), and (d) very high risk (≥ 8%) (SUPPLEMENT).

Reader study

Systematic screening for TB

Based on the radiographs' original radiologic reports, we calculated the diagnostic performances of the seven board-certified radiologists in terms of detection of active pulmonary TB and radiologically-identifiable relevant abnormalities on chest radiographs. Decision on active pulmonary TB on each radiograph was determined at the discretion of each radiologist, and the following image features were comprehensively checked to determine whether each radiograph indicated active pulmonary TB or not: abnormalities in the upper lobes of one or both lungs, presence of centrilobular nodules, cavitary lesions, consolidation, or miliary nodules. Diagnostic performance of pooled radiologists and per-radiologist were calculated, and compared with those of DL algorithm.

Detection of lung cancers on chest radiographs

A subset of this cohort was sampled for a reader study who participated in this program between July 2008 and December 2008 (hereafter, validation set). The chest radiographs from this validation set were reviewed once by one of three board-certified radiologists (J.H.K. with 7 years of experience in reading chest radiographs, H.I.C. and J.P. with 6 years of experience in

reading chest radiographs; their subspecialties were not thoracic radiology). The three radiologists were blinded to all clinical information, and asked to determine whether each radiograph had suspicious abnormalities for lung cancer or not.

Statistical Analysis

Systematic screening for TB and detection of lung cancers on chest radiographs

To measure the diagnostic performance of the algorithm and radiologists, receiver operating characteristic (ROC) curve analyses were performed and AUC was used as its measure for detection of active pulmonary TB. As for lung cancers, the ROC curve analyses of the DL algorithm with AUC calculation were appraised through the following three tasks: (a) detection of clearly visible lung cancers on chest radiographs, (b) detection of visible lung cancers on chest radiographs, and (c) discrimination between cancer-positive chest radiographs and cancer-negative chest radiographs. These three evaluations were performed independently.

For each task, diagnostic measures such as sensitivity, specificity, positive predictive value (PPV), negative predictive value (NPV), and accuracy were

also calculated with pre-defined thresholds (for TB, high sensitivity threshold, 0.16, high specificity threshold, 0.46; for lung cancer, high sensitivity threshold, 0.16), at which the DL algorithm demonstrated 95% sensitivity and 95% specificity, respectively in previous study (1). With these thresholds, the chest radiographs of which probability scores are of these cutoff values or higher are allocated as positive test result by the algorithm, and the chest radiographs with probability scores less than the thresholds are designated as negative results. McNemar's tests were performed to compare classification between the algorithm and pooled radiologists, and comparison of PPVs and NPVs were performed using the method suggested by Moskowitz et al (44).

For analysis of detection performance of pulmonary TB, Mann-Whitney U tests were performed for the comparison of DL algorithm-assigned probability scores according to lesion conspicuity and relevant levels. For lung cancer, we calculated the threshold value where the specificity of the algorithm matched that of the pooled radiologists. The corresponding sensitivity, NPV, and PPV at this threshold were also calculated and compared with those of the radiologists.

Model calibration was investigated by plotting the observed versus

predicted probabilities and by using the P-value for the Spiegelhalter statistic (45, 46)

Optimization of candidate selection for LCS

Baseline characteristics were compared between this study population and the development dataset of the CXR-LC model with the Student t-test for continuous variables and the Pearson chi-square test for categorical variables.

We evaluated the CXR-LC model performance for incident lung cancer prediction in 3 tasks: (a) incident lung cancer within the entire follow-up period by November 2020; (b) lung cancer within 3 years after chest radiographs; (c) lung cancer within 5 years after chest radiographs. In each task, lung cancer incidence was stratified according to CXR-LC risk categories (low, indeterminate, high, and very high risk) and compared with the Pearson chi-square tests. In addition, ROC curve analysis was conducted for these 3 tasks, and the AUC values were calculated. As additional diagnostic measures, sensitivity, specificity, PPV, and NPV were evaluated with a cutoff value of 3.297%, corresponding to the threshold between the low-to-indeterminate risk and the high to very-high-risk

categories (5). The aforementioned analyses at 3 time horizons were conducted for the total study population and USPSTF-eligible individuals, respectively. Model calibration was analyzed by plotting the observed versus predicted probabilities and by using the P-value for the Spiegelhalter statistic (45, 46). A statistically significant value of the Spiegelhalter Z-test indicated poor calibration. The calibration slope and intercept were also calculated. Considering that the follow-up interval of this study population is shorter (see Results) than the intended time horizon of the CXR-LC (i.e., 12 years), the predicted probability was linearly transformed to 6-year incident lung cancer risk (i.e., the probability was divided by two) for the calibration analyses.

We also investigated the added value of applying the CXR-LC model to the 2021 USPSTF recommendations for LCS among individuals with available pack-year (PY) information. Specifically, we hypothesized that participants in low-to-indeterminate (cutoff, 3.297%) CXR-LC risk category could be excluded from the LCS, even if they met the current USPSTF criteria (40). This approach aimed to minimize negative CT screening examinations and enrich the screening population with higher risk individuals to potentially reduce false-positive results. The lung cancer detection rate, proportion of

selected CT screening candidates, and PPV were assessed. The lung cancer detection rate was defined as the proportion of CT screening candidates diagnosed with incident lung cancer among all individuals, and PPV was defined as the proportion of individuals with lung cancer among the USPSTF-eligible or CXR-LC model-positive individuals. These measures were compared between the USPSTF-eligible candidates with and without the CXR-LC risk category-based optimization.

Data were collected and saved in a spreadsheet (Excel 2016; Microsoft Corporation, Redmond, WA, USA). All statistical analyses were performed using R version 4.1.0 (R Project for Statistical Computing, Vienna, Austria), and a P-value of<0.05 was considered to indicate statistical significance.

RESULTS

Part 1. Systematic screening for TB

Study population

A total of 20,136 chest radiographs from 19,687 servicepersons were performed and collected. Among these, one radiograph was excluded for poor image quality. The final study population was comprised of 20,135 chest radiographs from 19,686 asymptomatic military servicepersons (19.475 men; 211 women; mean age, 21±2 years; median number of chest radiographs per individual, 1; range, 1-6) (Figure 1).

Five chest radiographs from four individuals had radiologic abnormalities confirmed as active pulmonary TB, and 28 chest radiographs from 26 individuals were judged as having radiologically-identifiable relevant abnormalities. On chest radiographs confirmed with active pulmonary TB, the following image features were present: patchy opacities with nodules in left upper lung field (n=3), cavitary lesion in right upper lung field (n=1), and centrilobular nodules in left upper lung field (n=1). In 18 out of these 28 chest radiographs with radiologically-identifiable relevant abnormalities, their abnormalities on chest radiographs were supported by clinical

diagnosis, bacteriological confirmation, or CT examination. Details about the radiological abnormalities identified are provided in Table 1.

Figure 1. Flowchart for study population and determining reference standards

Reference standards of tuberculosis were defined as bacteriologically confirmed cases (n=5). Any radiologically-identifiable relevant abnormalities were judged by consensus reading of two experienced radiologists (n=28). This figure contains the contents previously published at the time of the examination (41).

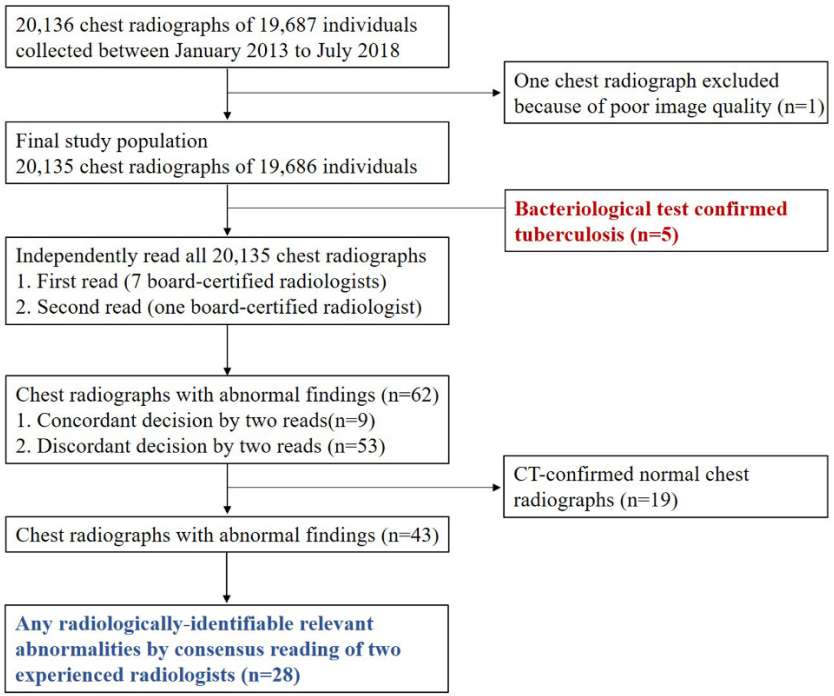


Table 1. Description and classification of positive cases determined by the reference standards as determined by further clinical and diagnostic testing. This table contains the contents previously published at the time of the examination (41).

Task	Diagnoses of positive cases	Gold standards	Radiological findings
Detection of active pulmonary tuberculosis (n=5)	Bacteriologically-confirmed tuberculosis (n=5)	Positive microbiological test	Patchy opacities with nodules in left upper lung field (n=3) Cavitary lesion in right upper lung field (n=1) Centrilobular nodules in left upper lung field (n=1)
Detection of radiologically-identifiable relevant abnormalities (n=28)	Bacteriologically-confirmed tuberculosis (n=5)	Positive microbiological test	
	Bacteriologically-confirmed pneumonia (n=2)	Positive microbiological test	Patchy opacity in left lower lung field (n=1) Centrilobular nodules in left middle lung field (n=1)
	Clinically-diagnosed pneumonia (n=2)	A clinical course including a response to antibiotics and initial and follow-up CT findings	Patchy opacities in left lower lung filed (n=1) and right lower lung field (n=1)
	Pulmonary sequestration (n=1)	CT findings	Mass in left lower lung field (n=1)
	Simple pulmonary eosinophilia (n=1)	A clinical course including blood eosinophilia and initial and follow-up CT findings	Well-defined nodule in right upper lung field (n=1)
	CT-based fibrosis and/or calcified nodules suggestive of healing and previous tuberculosis (n=7)	CT findings	Linear opacities and multifocal calcifications in both upper lung field (n=7)

	Non-specified relevant radiologic abnormalities (n=10)	Not specified*	Nodules in right upper lung field (n=4), right lower lung field (n=2), and left middle lung field (n=1) Linear opacities in left upper lung field (n=2) and right upper lung field (n=1)

* Not specified: Follow-up clinical course or images (chest radiographs or CT) did not confirm the diagnosis of the entities (n=6). The patients did not have further medical examination including radiological examinations (n=4)

Diagnostic performance of the DL algorithm

Table 2 reports the diagnostic performances of the DL algorithm at the high sensitivity and high specificity thresholds. The algorithm classified 832 and 74 chest radiographs as potential TB cases at the high sensitivity and high specificity thresholds, respectively.

For detection of active pulmonary TB, the AUC of the DL algorithm was 0.999 (95% confidence interval [CI]: 0.999 – 1.000) (Figure 2). The algorithm correctly classified all five chest radiographs with active pulmonary TB as abnormal chest radiographs (sensitivity of 100%; 95% CI: 56.6% – 100%) with both the high sensitivity and high specificity thresholds (Figure 3). With the high sensitivity threshold, specificity, PPV, and NPV of the algorithm were 95.9% (95% CI: 95.6% - 96.2%), 0.6% (95% CI: 0.3% - 1.4%), and 100% (95% CI: 100% - 100%), respectively. The algorithm's specificity, PPV, and NPV with the high specificity threshold were 99.7% (95% CI: 99.6% - 99.7%), 6.8% (95% CI: 2.9% - 14.9%), and 100% (95% CI: 100% - 100%), respectively.

For classifying radiologically-identifiable relevant abnormalities, the performance of the algorithm had an AUC of 0.967 (95% CI: 0.938 – 0.996) (Figure 2). Its sensitivity, specificity, PPV, and NPV were 82.1% (95% CI:

64.4% - 92.1%), 96.0% (95% CI: 95.7% - 96.2%), 2.8% (95% CI: 1.8% - 4.1%), 99.9% (95% CI: 99.9% - 100%) at the high sensitivity threshold, and 67.9% (95% CI: 49.3% - 82.1%), 99.7% (95% CI: 99.6% - 99.8%), 25.7% (95%: 17.1% - 36.7%), and 99.9% (95% CI: 99.9% - 100%) at the high specificity threshold.

The model calibration was poor for detecting active pulmonary TB and radiologically-identifiable relevant abnormalities (P<0.001 for both), and the model overestimated the predicted value of both types of diseases.

Table 2. Diagnostic performance of the DL algorithm on 20,135 screening chest radiographs for detection of active pulmonary TB. This table contains the contents previously published at the time of the examination (41).

Task	Probability threshold	Sensitivity	P-value	Specificity	P-value	Positive predictive value	P-value	Negative predictive value	P-value	Accuracy
Detection of active pulmonary tuberculosis	High sensitivity threshold (0.16)	100% (5 of 5)	0.999	95.9% (19,303 of 20,130)	<0.001	0.6% (5 of 832)	<0.001	100% (19,303 of 19,303)	0.327	96.0%
	High specificity threshold (0.46)	100% (5 of 5)	0.999	99.7% (20,061 of 20,130)	0.151	6.8% (5 of 74)	0.936	100% (20,061 of 20,061)	0.317	99.7%
	Pooled radiologists	80.0% (4 of 5)	NA	99.7% (20,076 of 20,130)	NA	6.9% (4 of 58)	NA	99.9% (20,076 of 20,135)	NA	99.7%
Detection of radiologically-identifiable relevant abnormalit	High sensitivity threshold (0.16)	82.1% (23 of 28)	0.999	96.0% (19,298 of 20,107)	0.001	2.8% (23 of 832)	<0.001	99.9% (19,298 of 19,303)	0.936	96.0%
	High specificity threshold (0.46)	67.9% (19 of 28)	0.289	99.7% (20,052 of 20,107)	0.043	25.7% (19 of 74)	0.010	99.9% (20,052 of 20,061)	0.157	99.7%

ies	Pooled radiologists	82.1% (23 of 28)	NA	99.8% (20,072 of 20,107)	NA	39.7% (23 of 58)	NA	99.9% (20,072 of 20,077)	NA	99.8%

DL=deep learning; TB=tuberculosis

Parenthesis of sensitivity, the number of true positive of actual positive cases

Parenthesis of specificity, the number of true negative of actual negative cases

Parenthesis of negative predictive value, true negative of predicted negative cases

Parenthesis of positive predictive value, true positive of predicted positive cases.

P-value: comparison of the DL algorithm’s performance with pooled radiologists.

Figure 2. Receiver operating characteristic (ROC) curves of the deep-learning (DL) detection algorithm

(A) ROC curve of DL algorithm for TB. The area under the ROC curve (AUC) was 0.999 (95% confidence interval [CI]: 0.999 – 1.000). (B) ROC curve of the DL algorithm for any radiologically-identifiable relevant abnormalities. The AUC of the algorithm was 0.967 (95% CI: 0.938 – 0.996). This figure contains the contents previously published at the time of the examination (41).

.

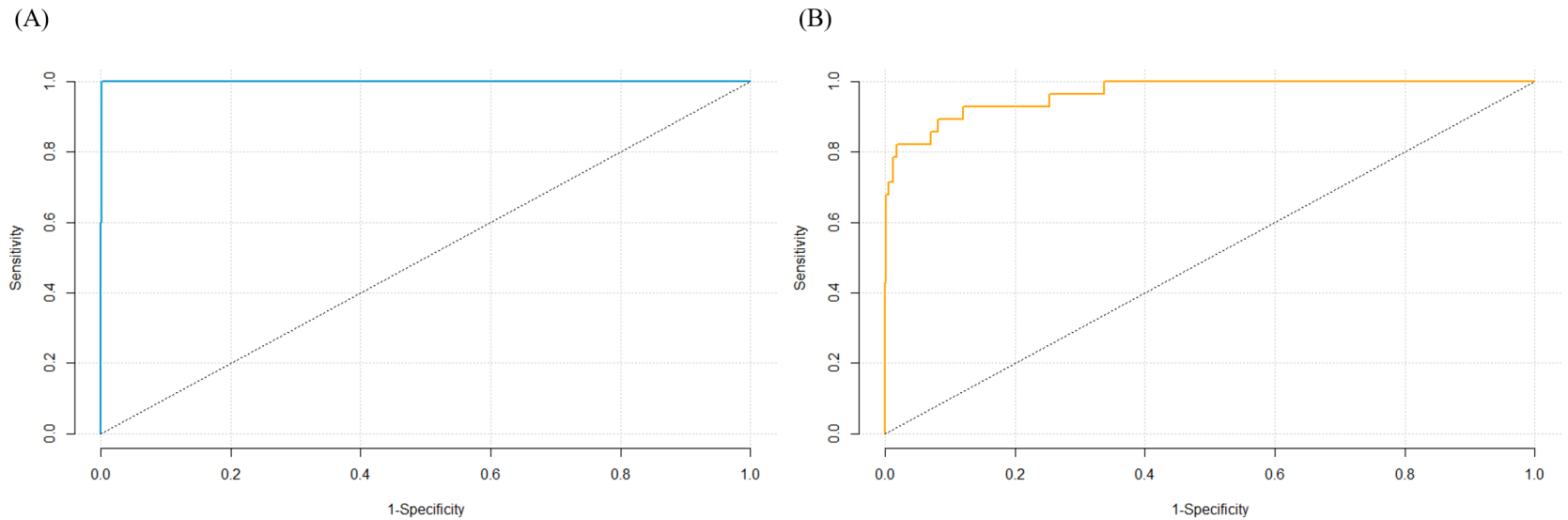
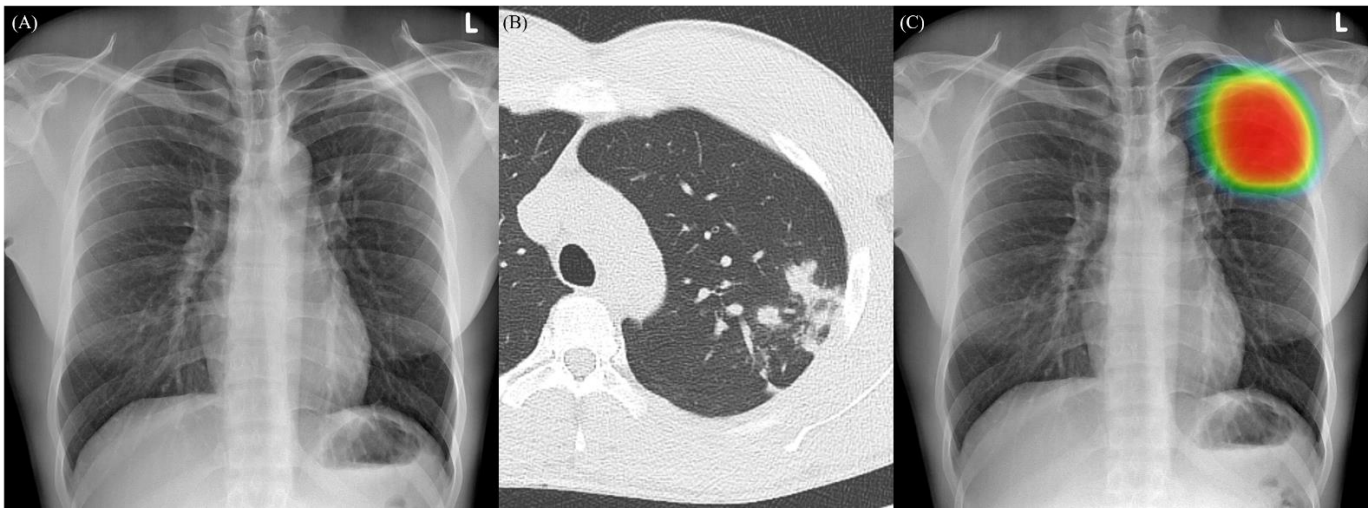


Figure 3. Representative cases of the DL algorithm to detect active pulmonary TB on chest radiographs

(A) A chest radiograph of a 21-years-old male taken for routine medical check-up. Ill-defined consolidations and linear infiltrations were identified in the left upper lung field, suggesting a typical finding of active pulmonary TB. (B) Chest computed tomography taken for further diagnostic evaluation demonstrated nodular infiltrations and consolidations in the left upper lobe. The patient was confirmed as having tuberculosis by a polymerase chain reaction and then underwent tuberculosis treatment. (C) The DL algorithm provided a probability value of 0.989 as positive case and correctly localized the lesions in the left upper lung field. This figure contains the contents previously published at the time of the examination (41).



Comparison of the probability scores of abnormal chest radiographs according to lesion conspicuity/relevance levels

For the 28 radiographs with radiologically-identifiable relevant abnormalities, lesion conspicuity and relevant levels and the algorithm's reported probabilities were compared. For lesion conspicuity, chest radiographs with 'certainly visible' labels (n=16; median probability, 0.840; interquartile range [IQR]: 0.639 – 0.951) had significantly higher probabilities than those with 'visible but uncertain' labels (n=12; median, 0.334; IQR: 0.104 – 0.684) (P=0.003). In terms of relevant levels, chest radiographs with 'certainly relevant' labels (n=18; median, 0.782; IQR: 0.548 – 0.948) showed significantly higher probabilities than those of 'equivocally relevant' labels (n=10; median, 0.261; IQR: 0.114 – 0.689) (P=0.016).

Diagnostic performance compared with 7 board-certified radiologists

Sensitivities, specificities, PPVs, and NPVs of pooled radiologists are described in Table 2, and the performance of individual radiologist is summarized in Table 3. For the detection of active pulmonary TB, the

algorithm showed comparable diagnostic measures to the pooled radiologists with both high sensitivity (sensitivity, 100% vs. 80%, P>0.999; NPV, 100% vs. 99.9%; P=0.327) and high specificity thresholds (sensitivity, 100% vs. 80%, P>0.999; specificity, 99.7% vs. 99.7%, P=0.151; PPV, 6.8% vs. 6.9%, P=0.936; NPV, 100% vs. 99.9%; P=0.317); however, the algorithm had lower specificity (95.9% vs. 99.7%, P<0.001) and PPV (0.6% vs. 6.9%, P<0.001) at the high sensitivity threshold.

For radiologically-identifiable relevant abnormalities, the algorithm had comparable sensitivities and NPVs to those of the pooled radiologists with both high sensitivity (sensitivity, 82.1% vs. 82.1%, P >0.999; NPV, 99.9% vs. 99.9%, P=0.936) and high specificity (sensitivity, 67.9% vs. 82.1%, P=0.289; NPV, 99.9% vs. 99.9%, P=0.157) thresholds. However, the algorithm showed lower specificities and PPVs than those of the radiologists with both high sensitivity (specificity, 96.0% vs. 99.8%, P<0.001; PPV, 2.8% vs. 39.7%, P<0.001) and high specificity (specificity, 99.7% vs. 99.8%, P=0.043; PPV, 25.7% vs. 39.7%, P=0.010) thresholds.

Table 3. Diagnostic performance of the original radiological reports consisting of 7 board-certified radiologists. This table contains the contents previously published at the time of the examination (41).

Task	Reader	Sensitivity	Specificity	Positive predictive value	Negative predictive value	Accuracy
Detection of active pulmonary tuberculosis	Pooled readers	80% (4 of 5)	99.7% (20,076 of 20,130)	6.9% (4 of 58)	99.9% (20,076 of 20,135)	99.7%
	Reader 1	NA	NA	NA	NA	NA
	Reader 2	100% (1 of 1)	99.8% (3,268 of 3,275)	12.5% (1 of 8)	100% (3,268 of 3,268)	99.8%
	Reader 3	NA	NA	NA	NA	NA
	Reader 4	NA	NA	NA	NA	NA
	Reader 5	NA	NA	NA	NA	NA
	Reader 6	50% (1 of 2)	99.7% (2,255 of 2,261)	14.3% (1 of 7)	99.9% (2,255 of 2,256)	99.7%
	Reader 7	100% (2 of 2)	99.4% (3,838 of 3,871)	8% (2 of 25)	100% (3,848 of 3,848)	99.4%
Detection of radiologically-identifiable relevant abnormalities	Pooled readers	0.821 (23 of 28)	99.8% (20,072 of 20,107)	39.7% (23 of 58)	99.9% (20,072 of 20,077)	99.8%
	Reader 1	NA	NA	NA	NA	NA
	Reader 2	1.000 (6 of 6)	99.9% (3,268 of 3,270)	75% (6 of 8)	100% (3,268 of 3,268)	99.9%
	Reader 3	0.250 (1 of 4)	99.8% (4,009 of 4,017)	11.1% (1 of 9)	99.9% (4,009 of 4,012)	99.7%
	Reader 4	1.000 (3 of 3)	99.9% (2,492 of 2,493)	75% (3 of 4)	100% (2,492 of 2,492)	100%
	Reader 5	0.500 (1 of 2)	99.9% (3,599 of 3,603)	20% (1 of 5)	99.9% (3,599 of 3,600)	99.9%
	Reader 6	0.667	99.8%	28.6%	99.9%	99.7%

		(2 of 3)	(2,255 of 2,260)	(2 of 7)	(2,255 of 2,256)	
	Reader 7	1.000 (10 of 10)	99.6% (3,848 of 3,863)	40.0% (10 of 25)	100% (3,848 of 3,848)	99.6%

Number of chest radiographs per each radiologist: reader 1 (n=601), reader 2 (n=3,276), reader 3 (n=4,021), reader 4 (n=2,496), reader 5 (n=3,605), reader 6 (n=2,263), reader 7 (n=3,873)

Part 2. Detection of lung cancers on chest radiographs

Study population

For the validation set, 13,640 individuals with 13,760 chest radiographs were initially included and 3,434 individuals with 3,471 chest radiographs were excluded due to the exclusion criteria. Finally, 10,206 individuals (5,859 men and 4,347 women; mean age, 54±11 years; age range, 18–95 years) with 10,289 chest radiographs were included in the task of detecting cancer-positive radiographs. For detecting visible lung cancer, 10,285 chest radiographs from 10,202 individuals (5,857 men and 4,345 women; mean age, 54±11 years) were included. Four individuals were excluded because their lung cancers were invisible on chest radiographs (Figure 4 and Table 4).

For the screening cohort, 71,951 individuals (37,938 men and 34,013 women; mean age, 50±12 years) underwent 132,207 chest radiographs, and 21,853 individuals with 31,631 chest radiographs were excluded based on the exclusion criteria. Thus, 50,098 individuals (28,105 men and 21,993 women; mean age, 53±11 years; age range, 18–99 years) with 100,576 chest radiographs were included in the task of detecting cancer-positive radiographs. For the analysis of visible lung cancer detection, 51

radiographs of invisible lung cancers from 37 individuals were excluded, and 50,070 individuals (28,090 men and 21,980 women; mean age, 53±11 years; age range, 18–99 years) with 100,525 chest radiographs were used. In the analysis of clearly visible lung cancer, 19 additional radiographs from 15 individuals were excluded because these radiographs were judged as having invisible lung cancer by one of the two radiologists (J.H.L. and E.J.H.).

Figure 4. Flowchart for study of detection of lung cancers on chest radiographs. (A) Validation set and (B) screening cohort. This figure contains the contents previously published at the time of the examination (42).

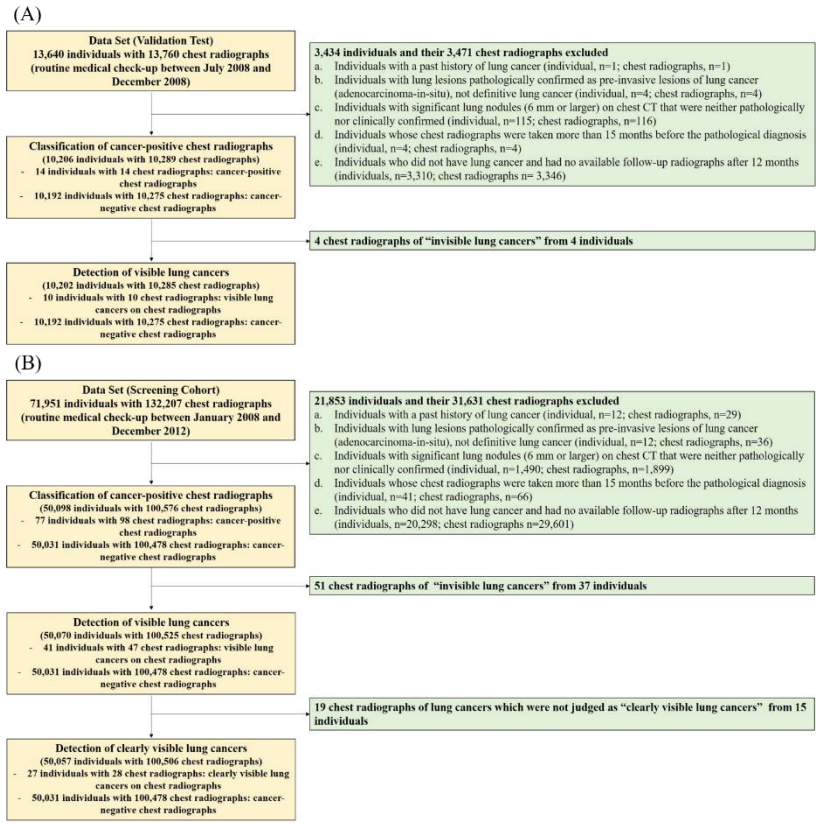


Table 4. Baseline clinical characteristics of Individuals and chest radiographs in the validation set and screening cohort. This table contains the contents previously published at the time of the examination (42).

	Validation test	Screening cohort
Number of individuals	10,206	50,098
Number of chest radiographs	10,289	100,576
Mean age ± standard deviation in years (range)	54±11 (18 – 95)	53±11 (18 – 99)
Sex		
Men	5,859	28,105
Women	4,347	21,993
Number of chest radiographs per individual (median and range)	1 (1 – 4)	1 (1 – 20)
Number of cancer-positive chest radiographs	14 (0.1% from 10,289 chest radiographs)	98 (0.1% from 100,576 chest radiographs)
Number of chest radiographs with visible lung cancers on chest radiographs*	10 (0.1% from 10,285 chest radiographs of 10,202 individuals)	47 (0.05% from 100,525 chest radiographs of 50,070 individuals)
Number of chest radiographs with clearly visible lung cancers**	Not evaluated	28 (0.03% from 100,506 chest radiographs of 50,057 individuals)

* In the analysis of visible lung cancer, 4 and 51 chest radiographs of invisible lung cancers were excluded in validation test and screening cohort, respectively.

** In the analysis of clearly visible lung cancer, 70 chest radiographs were excluded because these radiographs were judged as having invisible lung cancer by at least one of the two radiologists

Prevalence of lung cancer in study populations

In the validation set, 14 individuals (0.1% of 10,206 individuals) with 14 radiographs (0.1% of 10,289 chest radiographs) were confirmed to have lung cancers, and 10,192 individuals (99.9%) with 10,275 radiographs (99.9%) were judged to have no lung cancers. The 10 radiographs (0.1% of 10,285 radiographs) of 10 individuals (0.1% of 10,202 individuals) judged to have visible lung cancers.

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

radiograph taken 12 months after the initial radiograph had visible lung cancer.

Lung cancer detection performance in the validation set

The detection performances of the DL algorithm and pooled radiologists for visible lung cancers on chest radiographs are shown in Table 5. The algorithm's AUC was 0.989 (95% CI: 0.968 – 0.999), and it detected three more lung cancers (9 of 10 radiographs; sensitivity, 90%) than the radiologists (6 of 10 radiographs; sensitivity, 60%) (Figure 5). However, this difference was not statistically significant (P=0.248). The algorithm had an equivalent NPV to the radiologists (99.9% vs. 99.9%, P=0.091), but lower specificity and PPV (specificity, 96.9% vs.99.8%, P<0.001; PPV, 2.7% vs. 18.8%, P<0.001). At the threshold where the algorithm's specificity matched that of the radiologists (0.847), the algorithm's sensitivity, NPV, and PPV were 70%, 99.9%, and 21.2%, respectively, and all diagnostic measures of the algorithm were comparable to those of the radiologists (sensitivity, P>0.999; NPV, P=0.563; PPV, P=0.264).

The classification of cancer-positive chest radiographs is presented in Table 6 and Figure 5. The algorithm's AUC was 0.892 (95% CI: 0.794 –

0.989). It detected three more lung cancers (9 of 14 radiographs; sensitivity, 64.3%) than the radiologists (6 of 14 radiographs; sensitivity 42.9%) (P=0.248). However, it had a lower specificity (96.9% vs. 99.8%; P<0.001) (Figure 6).

Table 5. Comparison between the diagnostic performance of the DL algorithm and that of three board-certified radiologists for the detection of visible lung cancers on chest radiographs in the validation set. This table contains the contents previously published at the time of the examination (42).

	Threshold	Sensitivity	P-value*	Specificity	P-value*	Negative predictive value	P-value*	Positive predictive value	P-value*	Accuracy
Visible lung cancers on chest radiographs	Pooled performance of three radiologists	60% [26%, 88%] (6 of 10)	NA	99.8% [99.7%, 100%] (10,249 of 10,275)	NA	99.9% [99.9%, 100%] (10,249 of 10,253)	NA	18.8% [7%, 36%] (6 of 32)	NA	99.7% [99.7%, 100%] (10,255 of 10,285)
	Deep-learning algorithm **	90% [55%, 100%] (9 of 10)	0.248	96.9% [96.8%, 97%] (9,956 of 10,275)	<0.001	99.9% [99.9%, 100%] (9,956 of 9,957)	0.091	2.7% [1.3%, 5.1%] (9 of 328)	<0.001	96.9% [96.9%, 97%] (9,965 of 10,285)
	Matched threshold†, 0.847	70% [35%, 93%] (7 of 10)	>0.999	99.8% [99.8%, 100%] (10,249 of 10,275)	NA	99.9% [99.9%, 100%] (10,249 of 10,252)	0.563	21.2% [9%, 39%] (7 of 33)	0.264	99.7% [99.7%, 100%] (10,256 of 10,285)

* P-values are for comparisons with the pooled diagnostic performance of three board-certified radiologists.

** A pre-defined threshold of 0.16 was used.

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

Prevalence of visible lung cancers: 10 individuals (0.1% of 10,202 individuals) with 10 radiographs (0.1% of 10,285 radiographs)

95% confidence intervals are presented in square brackets

Parentheses for sensitivity: the number of true positives of actual positive cases; parentheses for specificity: the number of true negatives of actual negative cases; parentheses of negative predictive value: true negatives of predicted negative cases; parentheses of positive predictive value: true positives of predicted positive cases.

Figure 5. ROC curves of the DL algorithm for (A) the detection of visible lung cancer on chest radiographs and (B) cancer-positive chest radiographs compared with board-certified radiologists in the validation set.

(A) In the validation set composed of 10,285 chest radiographs including 10 chest radiographs with visible lung cancer, the algorithm had an AUC of 0.989 (95% CI: 0.968 – 0.999) and the radiologists showed a sensitivity of 60% and a specificity of 99.8%. In the magnified illustration, a red dot that represents the performance of the radiologists is below the ROC curve of the algorithm.

(B) In the validation set composed of 10,289 chest radiographs including 14 cancer-positive chest radiographs, the DL algorithm had an AUC of 0.892 (95% CI: 0.794 – 0.989). In comparison, the three board-certified radiologists showed a sensitivity of 42.9% and a specificity of 99.8% for this task. In the magnified figure, a red dot that represents the performance of the radiologists is below the ROC curve of the algorithm.

This figure contains the contents previously published at the time of the examination (42).

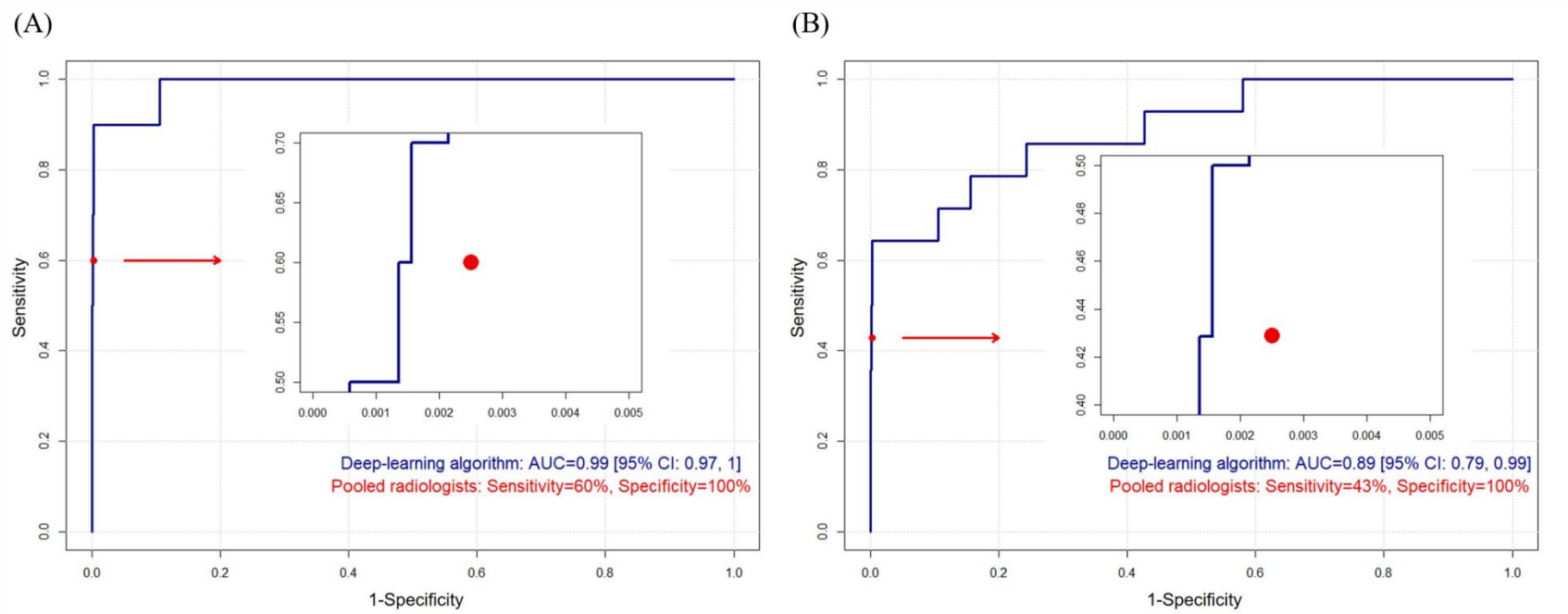

(A)
Sensitivity
1-Specificity
Deep-learning algorithm: AUC=0.99 [95% CI: 0.97, 1]
Pooled radiologists: Sensitivity=60%, Specificity=100%
(B)
Sensitivity
1-Specificity
Deep-learning algorithm: AUC=0.89 [95% CI: 0.79, 0.99]
Pooled radiologists: Sensitivity=43%, Specificity=100%

Table 6. Comparison between the diagnostic performance of the DL algorithm and that of three board-certified radiologists for the detection of cancer-positive chest radiographs in the validation set. This table contains the contents previously published at the time of the examination (42).

	Threshold	Sensitivity	P-value*	Specificity	P-value*	Negative predictive value	P-value*	Positive predictive value	P-value*	Accuracy
Cancer-positive chest radiographs	Pooled performance of three radiologists	42.9% [18%, 71%] (6 of 14)		99.8% [99.8%, 100%] (10,249 of 10,275)		99.9% [99.9%, 100%] (10,249 of 10,257)		18.8% [7%, 36%] (6 of 32)		99.7% [99.7%, 100%] (10,255 of 10,285)
	Deep-learning algorithm **	64.3% [35%, 87%] (9 of 14)	0.248	96.9% [96.9%, 97%] (9,956 of 10,275)	<0.001	99.9% [99.9%, 100%] (9,956 of 9,961)	0.105	2.7% [1.3%, 5.1%] (9 of 328)	<0.001	96.9% [96.9%, 97%] (9,965 of 10,285)
	Matched threshold†, 0.808	64.3% [35%, 87%] (9 of 14)	0.248	99.8% [99.8%, 100%] (10,249 of 10,275)	NA	99.9% [99.9%, 100%] (10,249 of 10,254)	0.089	25.7% [12%, 43%] (9 of 35)	0.192	99.7% [99.7%, 100%] (10,256 of 10,285)

* P-values are for comparisons with the pooled diagnostic performance of three board-certified radiologists.

** A pre-defined threshold of 0.16 was used.

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

Prevalence of cancer-positive chest radiographs: 14 individuals (0.1% of 10,206 individuals) with 14 radiographs (0.1% of 10,289 chest radiographs)

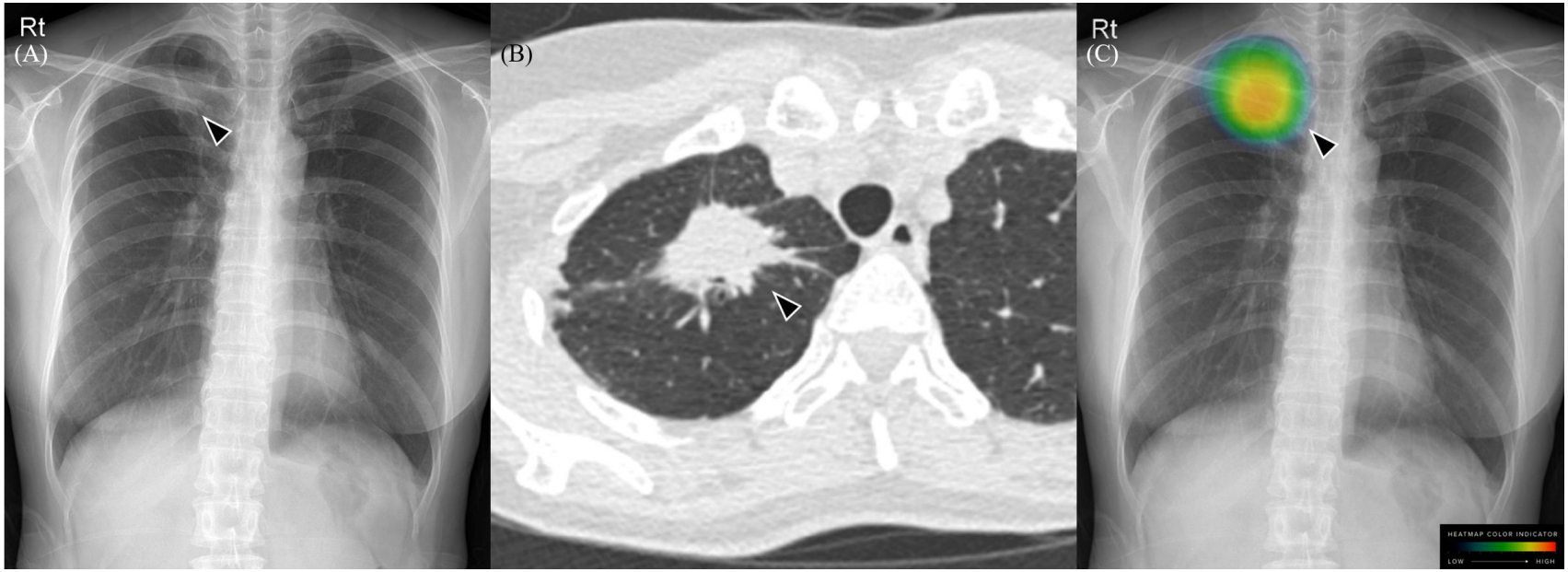
95% confidence intervals are presented in square brackets.

Parentheses for sensitivity: the number of true positives of actual positive cases; parentheses for specificity: the number of true negatives of actual negative cases; parentheses of negative predictive value: true negatives of predicted negative cases; parentheses of positive predictive value: true positives of predicted positive case

Figure 6. Representative case of the DL algorithm correctly detecting visible lung cancer on a chest radiograph in a health check-up.

A 56-year-old woman with a radiograph taken as part of a comprehensive health check-up and screening. (a) The radiograph showed an ill-defined lesion with a diameter of 3.5 cm (arrowhead) that was faintly identified in the right upper lung apex, which was obscured by the bony thorax. (b) A non-contrast chest CT scan taken on the same day as the radiograph demonstrated a 4.1-cm lung mass (arrowhead) with a spiculated margin in the right upper lobe apex on axial-plane. Right upper lobe lobectomy was then performed and the mass was pathologically proven to be invasive adenocarcinoma with an acinar and bronchioloalveolar pattern. (c) The DL algorithm provided a probability value of 0.85 for this being a positive case and correctly localized the lesions in the right upper lung apex (arrowhead). This lung mass was missed by a board-certified radiologist in the reader study.

This figure contains the contents previously published at the time of the examination (42).


Rt
(A)
(B)
Rt
(C)
HEATMAP COLOR INDICATOR
LOW
HIGH

Lung cancer detection performance of the DL algorithm in the entire screening cohort

The performance metrics of the DL algorithm for lung cancer detection in the entire screening cohort are tabulated in Table 7. The algorithm classified 3% (3,038 of 100,576 radiographs for cancer-positive radiographs, 3,038 of 100,525 radiographs for visible lung cancers on chest radiographs, and 3,027 of 100,506 radiographs for clearly visible lung cancers on chest radiographs) of chest radiographs as abnormal.

For the classification of cancer-positive radiographs, the AUC of the algorithm was 0.78 (95% CI: 0.728 – 0.833) (Figure 7). The algorithm correctly classified 39 of the 98 cancer-positive radiographs (sensitivity, 39.8%). The specificity, NPV, and PPV of the algorithm were 97%, 99.9%, and 1.3%.

In the detection of visible lung cancers on chest radiographs, the algorithm had an AUC of 0.969 (95% CI: 0.946 – 0.992) (Figure 7). Visible lung cancers were correctly detected on 39 out of 47 radiographs (sensitivity, 83%). The specificity, NPV, and PPV of the algorithm for detecting visible lung cancers were 97%, 99.9%, and 1.3%.

For the detection of clearly visible lung cancers on chest radiographs, the

algorithm showed an AUC of 0.998 (95% CI: 0.997 – 0.999) (Figure 7). Clearly visible lung cancers were correctly detected on all 28 out of 28 chest radiographs (sensitivity, 100%). The specificity, NPV, and PPV of the algorithm were 97%, 100%, and 0.9% (Figure 8). When the three ROC curves were compared with each other, the performance of the algorithm improved with increased visibility (all P-values between the AUCs of the three ROC <0.05).

The model calibration was poor for classifying cancer-positive, visible, and clearly visible lung cancers on chest radiographs (all P<0.001), and the model overestimated the risk of lung cancers.

Table 7. Diagnostic performance of the DL algorithm for detection of lung cancers on health screening cohort chest radiographs. This table contains the contents previously published at the time of the examination (42).

	Sensitivity	Specificity	Negative predictive value	Positive predictive value	Accuracy
Cancer-positive chest radiographs	39.8% [30%, 50%] (39 of 98)	97% [97%, 97%] (97,479 of 100,478)	99.9% [99.9%, 100%] (97,479 of 97,538)	1.3% [0.9%, 1.8%] (39 of 3,038)	97% [97%, 97%] (97,518 of 100,576)
Visible cancers on chest radiographs	83% [69%, 92%] (39 of 47)	97% [97%, 97%] (97,479 of 100,478)	99.9% [99.9%, 100%] (97,479 of 97,487)	1.3% [0.9%, 1.8%] (39 of 3,038)	97% [97%, 97%] (97,518 of 100,525)
Clearly visible cancers on chest radiographs	100% [88%, 100%] (28 of 28)	97% [97%, 97%] (97,479 of 100,478)	100% [100%, 100%] (97,479 of 97,479)	0.9% [0.7%, 1.3%] (28 of 3,027)	97% [97%, 97%] (97,507 of 100,506)

A pre-defined threshold of 0.16 was used

Prevalence of cancer-positive chest radiographs: 77 individuals (0.2% of 50,098 individuals) with 98 radiographs (0.1% of 100,576 chest radiographs)

Prevalence of visible lung cancers: 41 individuals (0.1% of 50,070 individuals) with 47 radiographs (0.05% of 100,525 chest radiographs)

Prevalence of clearly visible lung cancers: 27 individuals (0.05% of 50,057 individuals) with 28 radiographs (0.03% of 100,506 chest radiographs)

95% confidence intervals are presented in square brackets.

Parentheses for sensitivity: the number of true positives of actual positive cases; parentheses for specificity: the number of true negatives of actual

negative cases; parentheses of negative predictive value: true negatives of predicted negative cases; parentheses of positive predictive value: true positives of predicted positive cases.

Figure 7. ROC curves of the DL algorithm for the detection of lung cancer on chest radiographs in a health check-up screening cohort.

(A) ROC curve of the DL algorithm for the classification of cancer-positive chest radiographs in a health check-up screening. The AUC was 0.78 (95% CI: 0.728 – 0.833). (B) ROC curve of the DL algorithm for visible lung cancers on chest radiographs, with an AUC of 0.969 (95% CI: 0.946 – 0.992). (C) ROC curve of the DL algorithm for the detection of clearly visible lung cancers on chest radiographs. The AUC of the algorithm was 0.999 (95% CI: 0.997 – 0.999).

This figure contains the contents previously published at the time of the examination (42).

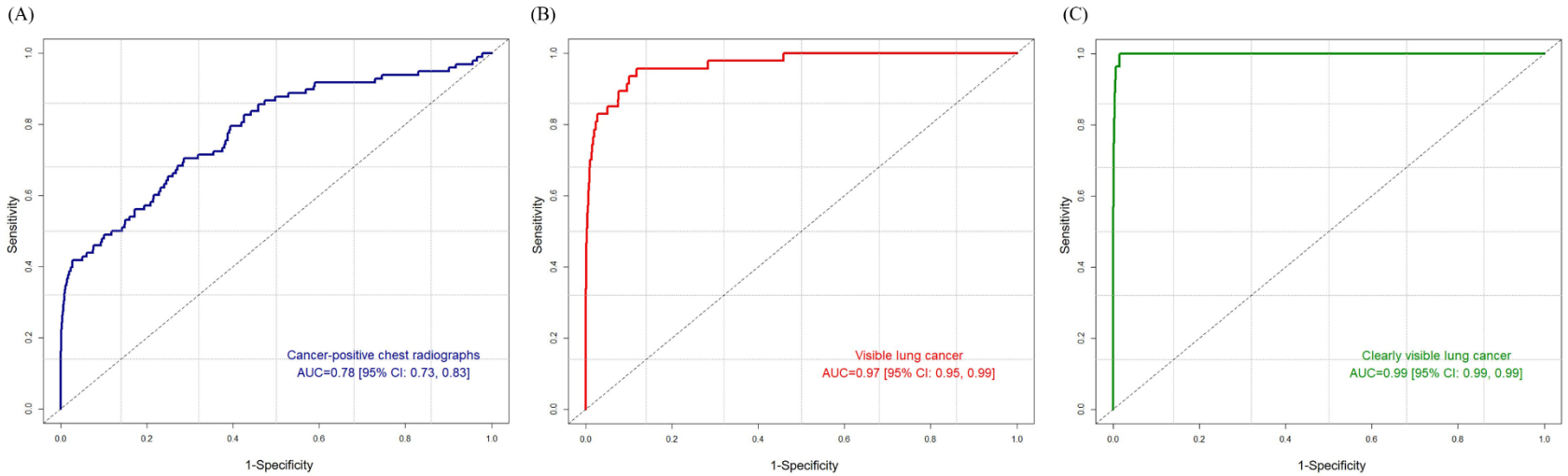
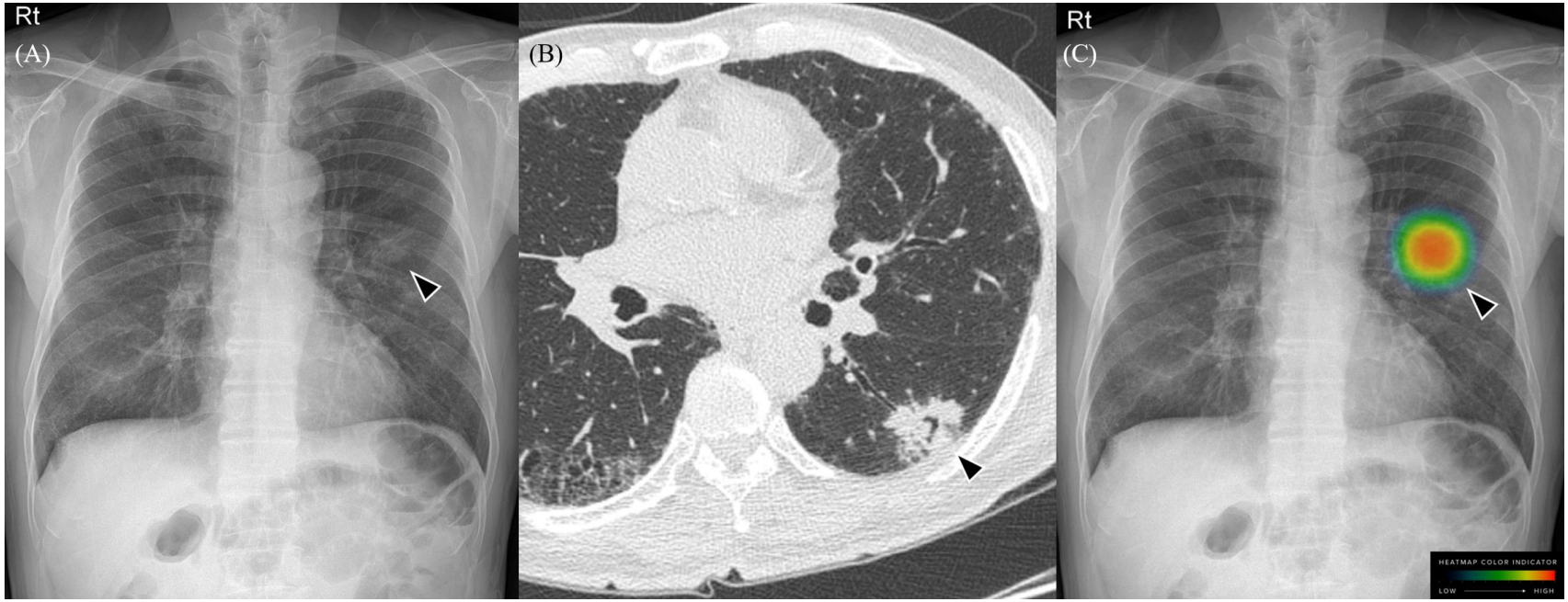

(A)
Sensitivity
1-Specificity
Cancer-positive chest radiographs
AUC=0.78 [95% CI: 0.73, 0.83]
(B)
Sensitivity
1-Specificity
Visible lung cancer
AUC=0.97 [95% CI: 0.95, 0.99]
(C)
Sensitivity
1-Specificity
Clearly visible lung cancer
AUC=0.99 [95% CI: 0.99, 0.99]

Figure 8. Representative case of the DL algorithm detecting clearly visible lung cancer on a chest radiograph in a health check-up screening.

A 67-year-old man with chest radiograph taken as part of a comprehensive health check-up and screening. (A) The radiograph showed a faintly visible lung mass (arrowhead) with a diameter of 3.5 cm was present in the left middle lung field. (B) A non-contrast chest CT scan taken on the same day as the chest radiograph demonstrated a 3.3-cm lung mass (arrowhead) with a spiculated margin and an air-bronchogram in the left lower lobe on axial plane. The patient underwent left lower lobe lobectomy and this mass was pathologically proven to be a squamous cell carcinoma. (C) The deep-learning algorithm provided a probability value of 0.91 for the patient having lung cancer and correctly localized the lesion in the left middle lung field (arrowhead).

This figure contains the contents previously published at the time of the examination (42).


Rt
(A)
(B)
Rt
(C)
HEATMAP COLOR INDICATOR
LOW
HIGH

Part 3. Optimization of candidate selection for LCS

Study Population

A total of 19,488 individuals (18,467 men and 1,021 women; mean age, 57.7±6.4 years) were included. The flow diagram is given in Figure 9. For the study population, clinical information (age, sex, smoking status [current or former smoker] with PY) collected as part of the medical check-up, was obtained. Among the 19,488 individuals, PY information was available for 17,390 (16,635 men, 755 women; mean age, 57.6±6.4 years), and 7,835 (7,699 men, 136 women; mean age, 57.4±6.1 years) met the updated USPSTF guidelines for LCS (i.e., 20-PY smoking history and currently smoking or having quit within the past 15 years).

In the subgroup analysis, 6,768 (6,513 men and 255 women; mean age 57.1±6.2 years) and 4,874 (4,689 men and 185 women, mean age 57.2±6.1 years) individuals were included in the subgroup analyses to predict incident lung cancer within 3 and 5 years, respectively (Table 8). Of these individuals, 5,932 (5,770 men and 162 women; mean age 57.0±6.1 years) and 4,192 (4,082 men and 110 women; mean age 57.0±6.0 years) had PY information and 2,751 (2,715 men and 36 women; mean age 56.9±5.9 years) and 1,972 (1,944 men and 28 women; mean age 56.9±5.8

years) were USPSTF-eligible, respectively.

The baseline characteristics of our study and the development cohort of the CXR-LC are compared in Table 8. Age and sex were significantly different between these 2 populations (age, mean 57.7 years vs. 62.4 years, P<0.001; men, 94.8% vs. 51.7%, P<0.001). The study population contained a significantly higher proportion of current smokers than the development dataset (31.1% vs. 19.3%, P<0.001). Lung cancer incidence was lower in the study population (0.6%; 107 of 19,488) than in the model development dataset (2.3%; 962 of 41,856) (P<0.001). The lung cancer incidence rates within 3 and 5 years were 0.7% (49 of 6,768; P<0.001) and 1.3% (64 of 4,874; P<0.001), respectively, in our study. Other clinical variables including age, sex, and smoking status also differed significantly between the 2 datasets (P<0.001 for all variables).

The median follow-up interval between the chest radiographs and the individuals' last radiographs was 11.6 months (IQR: 0 – 58.8 months). The median interval between the chest radiographs and the date of lung cancer diagnosis was 40 months (IQR: 0 – 85 months).

Figure 9. Flowchart of the study of optimization of candidate selection for LCS population.

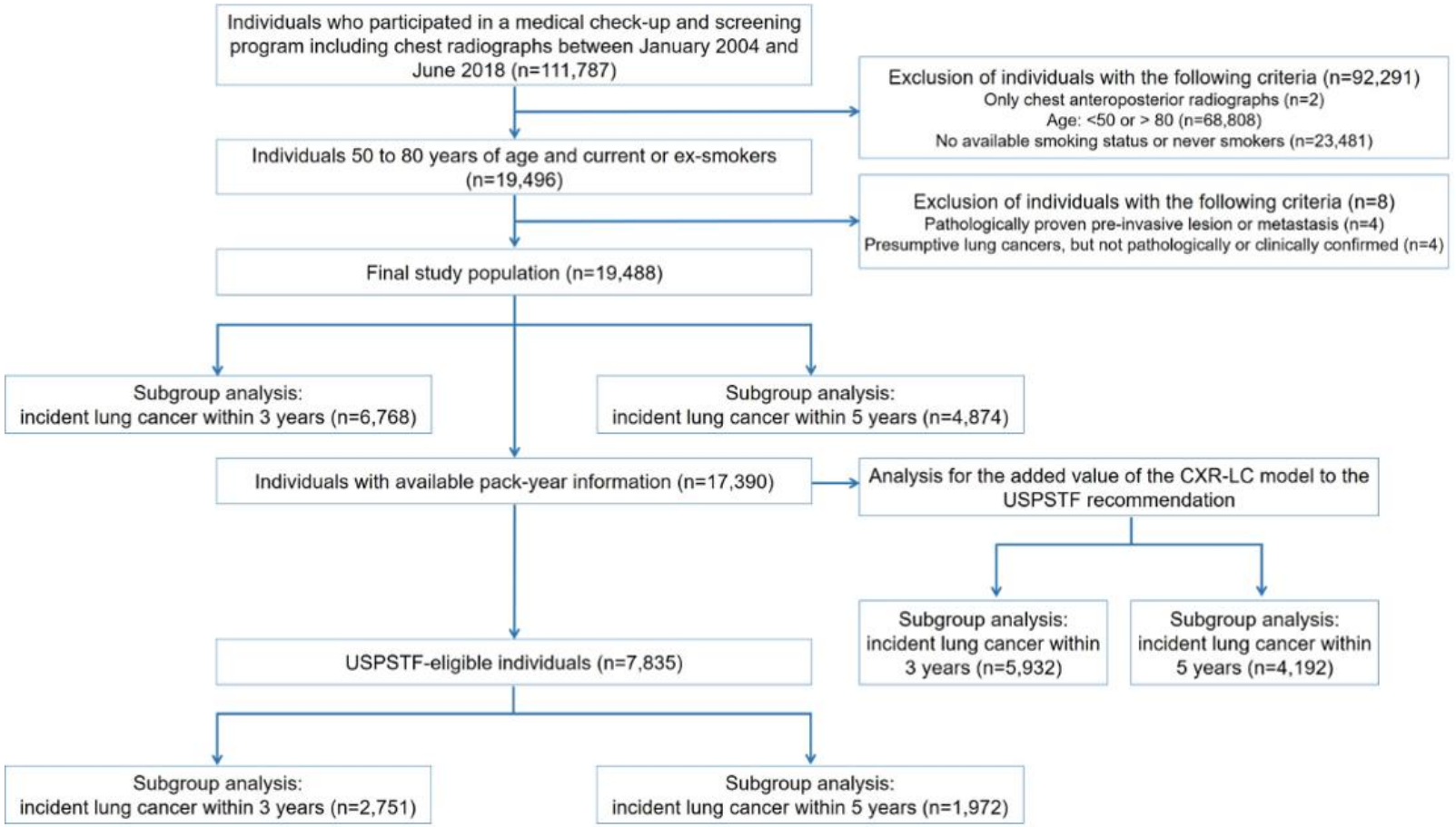


Table 8. Baseline characteristics of the study population and development dataset of the CXR-LC model.

	CXR-LC development dataset	Study population	Subgroup analysis	
			Incident lung cancer within 3 years	Incident lung cancer within 5 years
Number of individuals	41,856	19,488	6,768	4,874
Number of chest radiographs	85,478*	19,488	6,768	4,874
Age (years)	62.4 ± 5.4	57.7 ± 6.4§	57.1 ± 6.2§	57.2 ± 6.1§
Sex				
Male	21,648 (51.7%)	18,467 (94.8%)§	6,513 (96.2%)§	4,689 (96.2%)§
Female	20,208 (48.3%)	1,021 (5.2%)	255 (3.8%)	185 (3.8%)
Smoking status†				
Current smoker	4,392 (19.3%)	6,067 (31.1%)§	2,062 (30.5%)§	1,413 (29.0%)§
Former smoker	18,319 (80.7%)	13,421 (68.9%)	4,706 (69.5%)	3,461 (71.0%)
Lung cancer incidence	2.3% (962 of 41,856)‡	0.6% (107 of 19,488)§ ‖	0.7% (49 of 6,768)§	1.3% (64 of 4,874)§

* The development dataset included both enrollment (T0) and the first annual (T1) chest radiographs.

† CXR-LC model: a deep-learning model to predict incident lung cancer. This convolutional neural network model was first developed in smokers (current smokers, n=4,392; former smokers, n=18,319) and nonsmokers (n=19,145) from the Prostate, Lung, Colorectal, and Ovarian Cancer Screening Trial (PLCO), then fine-tuned in the subset of smokers. The final CXR-LC model was validated in smokers from the PLCO

independent data sets and NLST external validation data sets.

‡ Lung cancer incidence reported at 12 years.

§ P-values for the comparison of variables between the development dataset and this study population were below 0.001

|| While 105 cancers were pathologically confirmed, 2 lung cancers were clinically diagnosed.

Diagnostic Performance of the CXR-LC Model

The incidence of lung cancer was proportional to the CXR-LC risk categories: low risk, 0.2% (5 of 2,707); indeterminate risk, 0.2% (6 of 2,500); high risk, 0.4% (31 of 7,786); and very high risk, 1.0% (65 of 6,495) (P<0.001). For the USPSTF-eligible individuals, the model risk category also stratified the incidence of lung cancer: low risk, 0.1% (1 of 834); indeterminate risk, 0.3% (2 of 768); high risk, 0.3% (10 of 2,979); and very high risk, 1.3% (43 of 3,254) (P<0.001). Consistent associations were found for incident lung cancers within 3 and 5 years (P<0.001 for both) (Table 9).

The AUCs of the CXR-LC model were 0.677 (95% CI: 0.623 – 0.731) and 0.745 (95% CI: 0.677 – 0.813) for the entire study population and the USPSTF-eligible individuals, respectively. For lung cancer within 3 and 5 years, the AUCs of the model were 0.761 (95% CI: 0.693 – 0.829) and 0.739 (95% CI: 0.677 – 0.801) for the total study population, and 0.783 (95% CI: 0.695 – 0.871) and 0.782 (95% CI: 0.700 – 0.782) for the USPSTF-eligible individuals, respectively (Table 10 and Figure 10).

With the cutoff value of 3.297%, the model showed sensitivity of 89.7% (95% CI: 82.4% - 94.8%), specificity of 26.8% (95% CI: 26.2% - 27.4%), PPV of 0.7% (95% CI: 0.6% - 0.7%), and NPV of 99.8% (95% CI: 99.6% -

99.9%). For the USPSTF-eligible individuals, it had sensitivity of 94.6% (95% CI: 85.1% - 98.9%), specificity of 20.6% (95% CI: 19.7% - 21.5%), PPV of 0.9% (95% CI: 0.8% - 0.9%), and NPV of 99.8% (95% CI: 99.4% - 99.9%) The diagnostic results of the CXR-LC model for lung cancer within 3 and 5 years are shown in Table 10.

The model calibration was poor for the entire study population and the USPSTF-eligible individuals, respectively (P<0.001 for both), and the model overestimated the risk of lung cancer as would be expected as the CXR-LC model was originally calibrated for 12-year lung cancer (Table 11). Similar results were observed for incident lung cancer within 3 and 5 years.

Table 9. Lung cancer occurrence stratified by the CXR-LC risk categories.

			CXR-LC risk categories*				
Task	Group	Incidence of lung cancer	Low risk	Indeterminate risk	High risk	Very high risk	P-value
Incident lung cancer	50-80 y, smokers	107 of 19,488 (0.5%)	5 of 2,707 (0.2%)	6 of 2,500 (0.2%)	31 of 7,786 (0.4%)	65 of 6,495 (1.0%)	<0.001
	USPSTF eligible†	56 of 7,835 (0.7%)	1 of 834 (0.1%)	2 of 768 (0.3%)	10 of 2,979 (0.3%)	43 of 3,254 (1.3%)	<0.001
Incident lung cancer within 3 years	50-80 y, smokers	49 of 6,768 (0.7%)	1 of 1,012 (0.1%)	1 of 924 (0.1%)	13 of 2,750 (0.5%)	34 of 2,082 (1.6%)	<0.001
	USPSTF eligible†	29 of 2,751 (1.1%)	0 of 335 (0%)	1 of 287 (0.3%)	6 of 1,083 (0.6%)	22 of 1,046 (2.1%)	<0.001
Incident lung cancer within 5 years	50-80 y, smokers	64 of 4,874 (1.3%)	1 of 703 (0.1%)	2 of 679 (0.3%)	18 of 2,015 (0.9%)	43 of 1,477 (2.9%)	<0.001
	USPSTF eligible†	34 of 1,972 (1.7%)	0 of 237 (0%)	1 of 215 (0.5%)	7 of 777 (0.9%)	26 of 743 (3.5%)	<0.001

CXR-LC model: a deep-learning model to predict incident lung cancer

* CXR-LC lung cancer risk over 12 years: Low risk: < 2%, indeterminate risk: 2 to <3.297%, high risk: 3.297 to <8%, very high risk: ≥8%

† 2021 US Preventive Services Task Force (USPSTF) eligibility criteria for lung cancer screening: adults aged 50 to 80 years who have a 20 pack-year smoking history and currently smoke or have quit within the past 15 years.

‡ The median interval between the chest radiographs and the date of lung cancer diagnosis was 40 months (interquartile range: 0 to 85 months)

Table 10. Discrimination performance of the CXR-LC model for incident lung cancer.

Task	Group	AUC	Sensitivity*	Specificity*	PPV*	NPV*
Incident lung cancer	50-80 y, smokers	0.677 (0.623 – 0.731)	89.7% [96 of 107] (82.4% to 94.8%)	26.8% [5,196 of 19,381] (26.2% to 27.4%)	0.7% [96 of 14,281] (0.6% to 0.7%)	99.8% [5,196 of 5,207] (99.6% to 99.9%)
	USPSTF eligible†	0.745 (0.677 – 0.813)	94.6% [53 of 56] (85.1% to 98.9%)	20.6% [1,599 of 7,779] (19.7% to 21.5%)	0.9% [53 of 6,233] (0.8% to 0.9%)	99.8% [1,599 of 1,602] (99.4% to 99.9%)
Incident lung cancer within 3 years	50-80 y, smokers	0.761 (0.693 – 0.829)	95.9% [47 of 49] (86.0% to 99.5%)	28.8% [1,934 of 6,719] (27.7% to 29.9%)	1.0% [47 of 4,832] (0.9% to 1.0%)	99.9% [1,934 of 1.936] (99.6% to 100%)
	USPSTF eligible†	0.783 (0.695 – 0.871)	96.6% [28 of 29] (82.2% to 99.9%)	22.8% [621 of 2,722] (21.1% to 24.4%)	1.3% [28 of 2,129] (1.2% to 1.4%)	99.8% [621 of 622] (98.9% to 100%)
Incident lung cancer within 5 years	50-80 y, smokers	0.739 (0.677 – 0.801)	95.3% [61 of 64] (86.9% to 99.0%)	28.7% [1,379 of 4,810] (27.4% to 30.0%)	1.7% [61 of 3,492] (1.7% to 1.8%)	99.8% [1,379 of 1,382] (99.3% to 99.9%)
	USPSTF eligible†	0.782 (0.700 – 0.863)	97.1% [33 of 34] (84.7% to 99.9%)	23.3% [451 of 1,487] (21.4% to 25.2%)	2.2% [33 of 1,520] (2.0% to 2.3%)	99.8% [451 of 452] (98.5% to 100%)

CXR-LC model: a deep-learning model to predict incident lung cancer; AUC: area under the curve; PPV: positive predictive value; NPV: negative predictive value; Numbers in brackets are raw data. Numbers in parentheses are 95% confidence intervals.

* A cutoff value of 3.297% 12-year lung cancer risk was used for the CXR-LC model.

† 2021 US Preventive Services Task Force (USPSTF) eligibility criteria for lung cancer screening: adults aged 50 to 80 years who have a 20 pack-year smoking history and currently smoke or have quit within the past 15 years.

Figure 10. ROC curves of the CXR-LC model for the following three tasks: (A) incident lung cancer within the entire follow-up period, (B) incident lung cancers within 3 years from chest radiographs; (C) incident lung cancer within 5 years from chest radiographs. CXR-LC model: a deep-learning model to predict incident lung cancer.

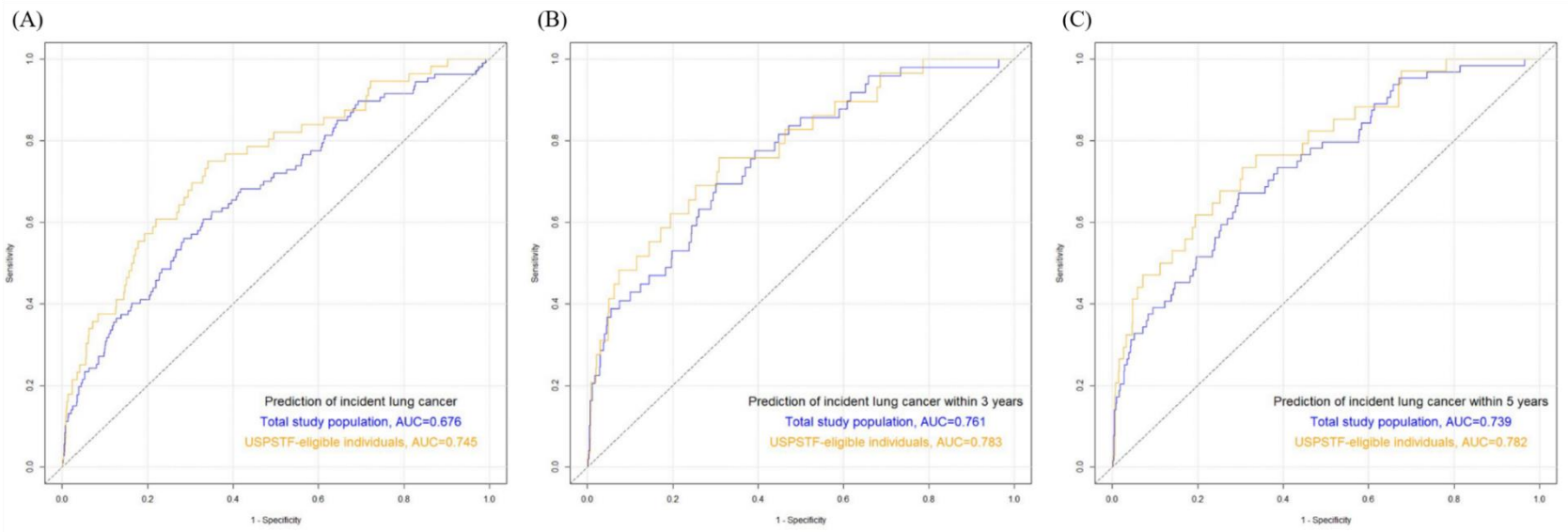


Table 11. Calibration performance of the CXR-LC model for incident lung cancer.

Task	Group	Slope	Intercept	P-value†
Incident lung cancer	50-80 y, smokers	0.94	-2.06	< 0.001
	USPSTF eligible*	1.46	-0.42	< 0.001
Incident lung cancer within 3 years	50-80 y, smokers	1.57	0.19	< 0.001
	USPSTF eligible*	1.83	1.07	< 0.001
Incident lung cancer within 5 years	50-80 y, smokers	1.41	0.32	< 0.001
	USPSTF eligible*	1.83	1.60	< 0.001

CXR-LC model: a deep-learning model to predict incident lung cancer

* 2021 US Preventive Services Task Force (USPSTF) eligibility criteria for lung cancer screening: adults aged 50 to 80 years who have a 20 pack-year smoking history and currently smoke or have quit within the past 15 years.

† Results of the Spiegelhalter Z-test

Added Value of CXR-LC to the 2021 USPSTF Recommendations

When the USPSTF recommendations were applied to 17,390 individuals in whom PY information was available, 7,835 LCS candidates were selected; however, 37 individuals with lung cancer (0.47%, 37 of 7,835) were missed. The lung cancer detection rate, proportion of selected CT screening candidates, and PPV were 0.3% (56 of 17,390), 45.1% (7,835 of 17,390), and 0.7% (56 of 7,835), respectively.

When the CXR-LC model was used to exclude the low-to-indeterminate risk categories, the LCS candidates decreased by 20.4% (1,602 of 7,835), while 3 individuals with lung cancer (incidence of 0.19% [3 of 1,602] in the excluded population) were missed (Table 12). The proportion of selected CT screening candidates (35.8% [6,233 of 17,390]) was lower than that of the entire USPSTF-eligible group ($P<0.001$). However, the lung cancer detection rate and PPV were still maintained at 0.3% (53 of 17,390, $P=0.848$) and 0.9% (53 of 6,233, $P=0.416$), respectively. In the analyses predicting incident lung cancer within 3 and 5 years, significant decreases in the proportion of selected CT screening candidates were consistently demonstrated ($P<0.001$ in all cases) with the lung cancer detection rate and PPV maintained ($P>0.05$ in all cases) (Table 12).

Table 12. Added value of the CXR-LC model to the 2021 US Preventive Services Task Force (USPSTF) Recommendations in smokers aged 50 to 80 years with available pack-year information.

Task	Incidence of lung cancer	Group	Missed lung cancer	CT screening candidates	Lung cancer detection rate	Proportion of selected CT screening candidates	Positive predictive value
Incident lung cancer	0.5% (93 of 17,390)	USPSTF eligible*	0.39% (37 of 9,555)	7,835	0.3% (56 of 17,390)	45.1% (7,835 of 17,390)	0.7% (56 of 7,835)
		USPSTF eligibility with CXR-LC model ≥ indeterminate risk †	0.30% (40 of 13,157)	6,233	0.3% (53 of 17,390)	35.8% (6,233 of 17,390)	0.9% (53 of 6,233)
		Differences or their P-values	0.19% (3 of 1,602)	1,602	0.848	<0.001	0.416
Incident lung cancer within 3 years	0.7% (42 of 5,932)	USPSTF eligible*	0.41% (13 of 3,181)	2,751	0.5% (29 of 5,932)	46.4% (2,751 of 5,932)	1.1% (29 of 2,751)
		USPSTF eligibility with CXR-LC model ≥ indeterminate risk †	0.37% (14 of 3,803)	2,129	0.5% (28 of 5,932)	35.9% (2,129 of 5,932)	1.3% (28 of 2,129)
		Differences or their P-values	0.16% (1 of 622)	622	>0.999	<0.001	0.479
Incident lung cancer within 5	1.3% (56 of 4,192)	USPSTF eligible*	0.99% (22 of 2,220)	1,972	0.8% (34 of 4,192)	47.0% (1,972 of 4,192)	1.7% (34 of 1,972)
		USPSTF eligibility with	0.86% (23 of 2,672)	1,520	0.8% (33 of 4,192)	36.3% (1,520 of 4,192)	2.2%(33 of 1,520)

years		CXR-LC model ≥ indeterminate risk †					
		Differences or their P-values	0.22% (1 of 452)	452	>0.999	<0.001	0.407

CXR-LC model: a deep-learning model to predict incident lung cancer; Lung cancer detection rate (the number of lung cancer cases / the number of test cases); Proportion of selected CT screening candidates (the number of test-positive cases / the number of test cases); Positive predictive value (the number of lung cancer cases / the number of test-positive cases)

* 2021 US Preventive Services Task Force (USPSTF) eligibility criteria for lung cancer screening: adults aged 50 to 80 years who have a 20 pack-year smoking history and currently smoke or have quit within the past 15 years.

† Low-to-indeterminate risk: <3.297% 12-year lung cancer risk based on CXR-LC

DISCUSSION

In this study, we validated the DL algorithm for detection of active pulmonary TB in the systematic screening chest radiographs and lung cancers in the health check-up radiographs. The algorithm correctly identified all five chest radiographs from individuals with microbiologically-confirmed active pulmonary TB, while maintaining high specificities (95.9% - 99.7%) and NPVs (100%). The algorithm showed comparable diagnostic performances to the pooled radiologists for screening chest radiographs with active pulmonary TB, especially at the high specificity threshold. Regarding lung cancer detection, the algorithm had an AUC of 0.989 and a comparable sensitivity (90% vs. 60%, P=0.245) to radiologists, with a lower specificity (96.9% vs. 99.8%. P<0.001). Finally, we externally validated the CXR-LC model to predict incident lung cancer in a medical check-up population. Lung cancer incidence was positively associated with the CXR-LC risk categories in the total study population and the USPSTF-eligible individuals. We corroborated the added value of the model to the updated USPSTF recommendations for LCS candidate selection by excluding the low-to-indeterminate risk subset with only few lung cancers.

Prior studies of DL algorithms for the detection of pulmonary TB and lung cancers on chest radiographs had limitations in their applicability to real-world settings for the following reasons (47-49): (a) they were tested using disease-enriched datasets (TB prevalence, 14.4% – 50%; lung cancer prevalence, 16% - 75%), which were clearly unrealistic; (b) their test datasets were arbitrarily selected in terms of the size, number, and location of the lung lesions; and (c) their test datasets comprised clearly dichotomized cases (chest radiographs with TB or lung cancer vs. normal chest radiographs), which intentionally excluded any indeterminate chest radiographs or radiographs with other pathologies. By contrast, we performed our study in a real-world screening setting. Therefore, we believe that the algorithm analyzed in our study could be reasonably applied for the detection of lung lesions on chest radiographs in a systematic screening for TB or health check-up population with an average risk of lung cancer.

Generally, systematic screening for active pulmonary TB involves a greater workload and costs compared with passive detection of symptomatic individuals presenting for medical care, in whom the number needed to test to identify a case is much lower (7-9). Therefore, the use of a sensitive screening tool which can identify high-risk individuals for further screening

can reduce costs while maintaining high overall sensitivity of the program (7, 8). In the WHO systematic review of screening approaches for TB, chest radiographs had the highest sensitivity of any screening tool (7, 8). However, using chest radiographs in systematic screening for active pulmonary TB generally requires skilled radiologists, which can be an obstacle for its wide utilization, particularly in resource-constrained settings (6-9). The DL algorithm in this study had comparable diagnostic performance for detection of active pulmonary TB as board-certified radiologists. We thought DL algorithm can be a potential option or a powerful triaging tool for mass screening program of active pulmonary TB in such resource-constrained settings. Indeed, the WHO recommends a sensitive tool to identify individuals who should undergo bacteriological examinations in systematic screening strategy (7, 8), and the DL algorithm could be used in the process of this strategy, thanks to its high sensitivity for TB. In settings where expert radiologists are available, the DL algorithm could be used as a screening aid to focus the efforts of radiologists, decreasing their workload and the program costs (7-9).

The sensitivity for the detection of visible lung cancers on chest radiographs has been reported to be highly variable with 20% to 92%, and radiologists'

perceptual errors reported to be the most common and preventable cause for failure to diagnose lung cancers, or missed lung cancers on chest radiographs (50-53). Given that the DL algorithm classified only 3% of chest radiographs as having a high probability of being abnormal in this study, it can help reduce diagnostic errors caused by simple mistakes or perceptual errors due to radiologists' insufficient expertise, as this algorithm showed a consistent and high detection performance for lung cancers on chest radiographs and was not vulnerable to the perceptual errors of human readers.

In LCS, negative screening examinations impose unnecessary radiation exposure and false-positive CT examinations increase diagnostic work-ups, including invasive procedures, leading to increased healthcare expenditures (23, 24, 31-33, 54). When the CXR-LC low-risk exclusion scheme was added to the USPSTF recommendations, proportions of selected CT screening candidates decreased significantly without reduction in lung cancer detection rates and PPVs, implying a decrease of negative and, potentially, false-positive CT screening examinations. As the goal of LCS is to detect lung cancers cost-effectively, the CXR-LC model may be suitable for this goal. Indeed, 20.4% (1,602 of 7,835) of the LCS candidates

could be reduced by excluding the low-to-indeterminate risk categories at the expense of only few missed lung cancers.

Several limitations of this study should be mentioned. First, this study was performed retrospectively at a single center per each investigation. Second, despite the DL algorithm showing outstanding performance in detecting active pulmonary TB or lung cancers on chest radiographs, we did not evaluate the added value of the DL algorithm to the diagnostic performances of the radiologists in this study. Further studies evaluating the added value of the algorithm in a screening setting would be warranted. Third, for lung cancers detection or prediction tasks, not all participants had contemporaneous chest CT exams at the time of their chest radiographs. This fact is problematic, as there could be nodules that are missed by our method of relying on longitudinal follow-up as the gold standard. Very slowly-growing lung cancers might not be found using our method for follow-up. Fourth, although the CXR-LC model was developed to predict 12-year lung cancer incidence, the median follow-up interval in this study was only 11.6 months, and we predicted incident lung cancers within a period shorter than 12 years. This was why the calibrations of the models were poor.

In conclusion, DL algorithms detected active pulmonary TB and lung cancers on chest radiographs with performance comparable to that of radiologists in the screening settings, and it will be helpful for optimization of candidate selection for LCS.

References

1. Hwang EJ, Park S, Jin K-N, Im Kim J, Choi SY, Lee JH, Goo JM, Aum J, Yim J-J, Cohen JG. Development and validation of a deep learning–based automated detection algorithm for major thoracic diseases on chest radiographs. JAMA network open 2019;2(3):e191095-e191095.

2. Hwang EJ, Park S, Jin K-N, Kim JI, Choi SY, Lee JH, Goo JM, Aum J, Yim J-J, Park CM. Development and validation of a deep learning–based automatic detection algorithm for active pulmonary tuberculosis on chest radiographs. Clinical Infectious Diseases 2019;69(5):739-747.

3. Nam JG, Park S, Hwang EJ, Lee JH, Jin K-N, Lim KY, Vu TH, Sohn JH, Hwang S, Goo JM. Development and validation of deep learning–based automatic detection algorithm for malignant pulmonary nodules on chest radiographs. Radiology 2019;290(1):218-228.

4. Qin ZZ, Sander MS, Rai B, Titahong CN, Sudrungrot S, Laah SN, Adhikari LM, Carter EJ, Puri L, Codlin AJ. Using artificial intelligence to read chest radiographs for tuberculosis detection: A multi-site evaluation of the diagnostic accuracy of three deep learning systems. Scientific reports 2019;9(1):1-10.

5. Lu MT, Raghu VK, Mayrhofer T, Aerts HJ, Hoffmann U. Deep learning using chest radiographs to identify high-risk smokers for lung cancer screening computed tomography: development and validation of a prediction model. Annals of Internal Medicine 2020;173(9):704-713.

6. (2019) WHO. Global tuberculosis repot 2019.

https://www.who.int/tb/publications/global_report/en/.

7. (2016) WHO. Chest radiograpehy in tuberculosis detection: summary of current WHO recommendations and guidance on programmatic approaches. https://www.who.int/tb/publications/chest-radiography/en/.

8. (2013) WHO. Systematic screening for active tuberculosis: principles and recommendations. . https://www.who.int/tb/tbscreening/en/.

9. (2011) WHO. Early detection of tuberculosis: an overview of approaches, guidelines and tools. https://apps.who.int/iris/handle/10665/70824.

10. Organization WH. The global plan to stop TB, 2016-2020. http://www.stoptb.org/global/plan/plan2/.

11. (2010) WHO. Public-private mix for TB care and control. https://www.who.int/tb/publications/tb-publicprivate-toolkit/en/.

12. Yoon C, Dowdy DW, Esmail H, MacPherson P, Schumacher SG. Screening for tuberculosis: time to move beyond symptoms. The Lancet Respiratory Medicine 2019;7(3):202-204.

13. Dara M, Solovic I, Sotgiu G, D'Ambrosio L, Centis R, Tran R, Goletti D, Duarte R, Aliberti S, De Benedictis FM. Tuberculosis care among refugees arriving in Europe: a ERS/WHO Europe Region survey of current practices. European Respiratory Journal 2016;48(3):808-817.

14. Melendez J, Sánchez CI, Philipsen RH, Maduskar P, Dawson R, Theron G, Dheda K, Van Ginneken B. An automated tuberculosis screening strategy combining X-ray-based computer-aided detection and clinical information. Scientific reports 2016;6(1):1-8.

15. Pande T, Cohen C, Pai M, Ahmad Khan F. Computer-aided detection of

pulmonary tuberculosis on digital chest radiographs: a systematic review. The International Journal of Tuberculosis and Lung Disease 2016;20(9):1226-1230.

16. Hogeweg L, Mol C, de Jong PA, Dawson R, Ayles H, van Ginneken B. Fusion of local and global detection systems to detect tuberculosis in chest radiographs. International conference on medical image computing and computer-assisted intervention: Springer, 2010; p. 650-657.

17. Rahman MT, Codlin AJ, Rahman MM, Nahar A, Reja M, Islam T, Qin ZZ, Khan MAS, Banu S, Creswell J. An evaluation of automated chest radiography reading software for tuberculosis screening among public-and private-sector patients. European Respiratory Journal 2017;49(5).

18. Jaeger S, Karargyris A, Candemir S, Folio L, Siegelman J, Callaghan F, Xue Z, Palaniappan K, Singh RK, Antani S. Automatic tuberculosis screening using chest radiographs. IEEE transactions on medical imaging 2013;33(2):233-245.

19. Siegel RL MK, Jemal A. Cancer statistics, 2020. CA Cancer J Clin 2020;70(1):7-30.

20. Van Iersel CA, De Koning HJ, Draisma G, Mali WP, Scholten ET, Nackaerts K, Prokop M, Habbema JDF, Oudkerk M, Van Klaveren RJ. Risk-based selection from the general population in a screening trial: selection criteria, recruitment and power for the Dutch-Belgian randomised lung cancer multi-slice CT screening trial (NELSON). International Journal of Cancer 2007;120(4):868-874.

21. Wille MM, Dirksen A, Ashraf H, Saghir Z, Bach KS, Brodersen J,

Clementsen PF, Hansen H, Larsen KR, Mortensen J. Results of the randomized Danish lung cancer screening trial with focus on high-risk profiling. American journal of respiratory and critical care medicine 2016;193(5):542-551.

22. Hocking WG, Hu P, Oken MM, Winslow SD, Kvale PA, Prorok PC, Ragard LR, Commins J, Lynch DA, Andriole GL. Lung cancer screening in the randomized Prostate, Lung, Colorectal, and Ovarian (PLCO) cancer screening trial. JNCI: Journal of the National Cancer Institute 2010;102(10):722-731.

23. de Koning HJ, van der Aalst CM, de Jong PA, Scholten ET, Nackaerts K, Heuvelmans MA, Lammers J-WJ, Weenink C, Yousaf-Khan U, Horeweg N. Reduced lung-cancer mortality with volume CT screening in a randomized trial. New England Journal of Medicine 2020;382(6):503-513.

24. Team NLSTR. Reduced lung-cancer mortality with low-dose computed tomographic screening. New England Journal of Medicine 2011;365(5):395-409.

25. Pinsky PF. Lung cancer screening with low-dose CT: a world-wide view. Translational lung cancer research 2018;7(3):234.

26. Dominioni L, Poli A, Mantovani W, Pisani S, Rotolo N, Paolucci M, Sessa F, Conti V, D'Ambrosio V, Paddeu A. Assessment of lung cancer mortality reduction after chest X-ray screening in smokers: a population-based cohort study in Varese, Italy. Lung Cancer 2013;80(1):50-54.

27. Shankar A, Saini D, Dubey A, Roy S, Bharati SJ, Singh N, Khanna M, Prasad CP, Singh M, Kumar S. Feasibility of lung cancer screening in

developing countries: challenges, opportunities and way forward. Translational lung cancer research 2019;8(Suppl 1):S106.

28. Gossner J. Lung cancer screening-don't forget the chest radiograph. World Journal of Radiology 2014;6(4):116.

29. Dominioni L, Rotolo N, Mantovani W, Poli A, Pisani S, Conti V, Paolucci M, Sessa F, Paddeu A, D'Ambrosio V. A population-based cohort study of chest x-ray screening in smokers: lung cancer detection findings and follow-up. BMC cancer 2012;12(1):1-12.

30. Sim Y, Chung MJ, Kotter E, Yune S, Kim M, Do S, Han K, Kim H, Yang S, Lee D-J. Deep convolutional neural network–based software improves radiologist detection of malignant lung nodules on chest radiographs. Radiology 2020;294(1):199-209.

31. Jonas DE, Reuland DS, Reddy SM, Nagle M, Clark SD, Weber RP, Enyioha C, Malo TL, Brenner AT, Armstrong C. Screening for lung cancer with low-dose computed tomography: updated evidence report and systematic review for the US Preventive Services Task Force. Jama 2021;325(10):971-987.

32. Henschke CI, Yip R, Yankelevitz DF, Smith JP. Definition of a positive test result in computed tomography screening for lung cancer: a cohort study. Annals of internal medicine 2013;158(4):246-252.

33. Yip R, Henschke CI, Yankelevitz DF, Smith JP. CT screening for lung cancer: alternative definitions of positive test result based on the national lung screening trial and international early lung cancer action program databases. Radiology 2014;273(2):591-596.

34. Rampinelli C, De Marco P, Origgi D, Maisonneuve P, Casiraghi M, Veronesi G, Spaggiari L, Bellomi M. Exposure to low dose computed tomography for lung cancer screening and risk of cancer: secondary analysis of trial data and risk-benefit analysis. bmj 2017;356.

35. Gareen IF, Duan F, Greco EM, Snyder BS, Boiselle PM, Park ER, Fryback D, Gatsonis C. Impact of lung cancer screening results on participant health-related quality of life and state anxiety in the National Lung Screening Trial. Cancer 2014;120(21):3401-3409.

36. van den Bergh KA, Essink-Bot M-L, Borsboom GJ, Scholten ET, Prokop M, de Koning HJ, van Klaveren RJ. Short-term health-related quality of life consequences in a lung cancer CT screening trial (NELSON). British journal of cancer 2010;102(1):27-34.

37. Byrne MM, Weissfeld J, Roberts MS. Anxiety, fear of cancer, and perceived risk of cancer following lung cancer screening. Medical Decision Making 2008;28(6):917-925.

38. Pinsky PF. Assessing the benefits and harms of low-dose computed tomography screening for lung cancer. Lung cancer management 2014;3(6):491-498.

39. Oken MM, Hocking WG, Kvale PA, Andriole GL, Buys SS, Church TR, Crawford ED, Fouad MN, Isaacs C, Reding DJ. Screening by chest radiograph and lung cancer mortality: the Prostate, Lung, Colorectal, and Ovarian (PLCO) randomized trial. Jama 2011;306(17):1865-1873.

40. Krist AH, Davidson KW, Mangione CM, Barry MJ, Cabana M, Caughey AB, Davis EM, Donahue KE, Doubeni CA, Kubik M. Screening for lung

cancer: US preventive services task force recommendation statement. JAMA 2021;325(10):962-970.

41. Lee JH, Park S, Hwang EJ, Goo JM, Lee WY, Lee S, Kim H, Andrews JR, Park CM. Deep learning–based automated detection algorithm for active pulmonary tuberculosis on chest radiographs: diagnostic performance in systematic screening of asymptomatic individuals. European Radiology 2021;31(2):1069-1080.

42. Lee JH, Sun HY, Park S, Kim H, Hwang EJ, Goo JM, Park CM. 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(3):687-696.

43. Lee C, Choe EK, Choi JM, Hwang Y, Lee Y, Park B, Chung SJ, Kwak M-S, Lee J-E, Kim JS. Health and Prevention Enhancement (H-PEACE): a retrospective, population-based cohort study conducted at the Seoul National University Hospital Gangnam Center, Korea. BMJ open 2018;8(4):e019327.

44. Moskowitz CS, Pepe MS. Comparing the predictive values of diagnostic tests: sample size and analysis for paired study designs. Clinical trials 2006;3(3):272-279.

45. Tammemägi MC, Ten Haaf K, Toumazis I, Kong CY, Han SS, Jeon J, Commins J, Riley T, Meza R. Development and validation of a multivariable lung cancer risk prediction model that includes low-dose computed tomography screening results: a secondary analysis of data from the National lung screening trial. JAMA network open 2019;2(3):e190204-

e190204.

46. Walsh CG, Sharman K, Hripcsak G. Beyond discrimination: a comparison of calibration methods and clinical usefulness of predictive models of readmission risk. Journal of biomedical informatics 2017;76:9-18.

47. Ting DSW, Tan T-E, Lim CT. Development and validation of a deep learning system for detection of active pulmonary tuberculosis on chest radiographs: Clinical and technical considerations. Oxford University Press US, 2019.

48. Ting DS, Yi PH, Hui F. Clinical applicability of deep learning system in detecting tuberculosis with chest radiography. Radiology 2018;286(2):729-731.

49. Park SH. Diagnostic case-control versus diagnostic cohort studies for clinical validation of artificial intelligence algorithm performance. Radiology 2019;290(1):272-273.

50. Quekel L, Goei R, Kessels A, van Engelshoven J. Detection of lung cancer on the chest radiograph: impact of previous films, clinical information, double reading, and dual reading. Journal of clinical epidemiology 2001;54(11):1146-1150.

51. Berlin NI, Buncher CR, Fontana RS, Frost JK, Melamed MR. The National Cancer Institute Cooperative Early Lung Cancer Detection Program: Results of the Initial Screen (Prevalence) Early Lung Cancer Detection: Introduction. American Review of Respiratory Disease 1984;130(4):545-549.

52. Gavelli G, Giampalma E. Sensitivity and specificity of chest x-ray screening for lung cancer. Cancer 2000;89(S11):2453-2456.

53. Muhm JR, Miller W, Fontana R, Sanderson D, Uhlenhopp M. Lung cancer detected during a screening program using four-month chest radiographs. Radiology 1983;148(3):609-615.

54. van Klaveren RJ, Oudkerk M, Prokop M, Scholten ET, Nackaerts K, Vernhout R, van Iersel CA, van den Bergh KA, van't Westeinde S, van der Aalst C. Management of lung nodules detected by volume CT scanning. New England Journal of Medicine 2009;361(23):2221-2229.

SUPPLEMENT

Supplement Text. Information about the deep-learning algorithms used in systematic screening for active pulmonary TB and detection of lung cancers on chest radiographs (Lunit INSIGHT) and Optimization of candidate selection for LCS (CXR-LC model).

The commercially available deep-learning algorithm (Lunit INSIGHT for Chest Radiography, version 4.7.2; Lunit) used in this study was designed for the detection of major thoracic diseases (pulmonary malignancy, active pulmonary tuberculosis, pneumonia, and pneumothorax) on the chest radiographs. It was developed with 54,221 normal chest radiographs and 35,613 chest radiographs in patients with major thoracic diseases (13,926 chest radiographs with pulmonary malignancy; 6,768 chest radiographs with active pulmonary tuberculosis; 6,903 chest radiographs with pneumonia; 8,016 chest radiographs with pneumothorax). For these training data sets, chest radiographs were collected in the patients with pathological (pulmonary malignancy), bacteriological (active pulmonary tuberculosis and pneumonia) confirmation, or relevant radiological report (pneumothorax), and then these chest radiographs were reviewed by the board-certified radiologists. For

each input chest radiograph, the algorithm provided probability score as continuous value between 0 and 1 as the image-level probability of abnormal CR. Heat maps overlaid on the input chest radiograph with per-pixel localization probability maps were provided to the users.

The CXR-LC model was developed with chest radiographs from both smokers and nonsmokers from the Prostate, Lung, Colorectal, and Ovarian Cancer Screening Trial (PLCO) dataset to predict incident lung cancers (up to 12 years). It was validated using chest radiographs from only smokers in the PCLO and National Lung Screening Trial datasets. The inputs are clinical information (age, sex, and smoking status) and chest radiographs, and the model calculates a probability for incident lung cancer ranging between 0% and 100%. In the original study and this study, the CXR-LC risk probabilities were converted to the following ordinal 12-year lung cancer risk categories: (a) low risk (< 2%), (b) indeterminate risk (2% to < 3.297%), (c) high risk (3.297% to < 8%), and (d) very high risk (≥ 8%). The threshold of 3.297% was determined to correspond to the Centers for Medicare & Medicaid Services-defined screening population size, and the < 2% and ≥ 8% thresholds were chosen to represent a low and very high 12-year risk, respectively.

초 록

서론: 인공지능 기반 병변 검출 딥러닝 (DL) 알고리듬의 검진 흉부 X 선 검사에서 활동성 결핵과 폐암 검출능을 확인하고, 폐암 발생을 예측하는 인공지능 알고리듬 (CXR-LC 모델)을 이용하여 폐암 검진 CT 수검자 선택 최적화에 관한 유용성을 검증하고자 한다.

방법: 병변 검출 DL 모델과 CXR-LC 모델의 유용성을 다음의 코호트들에서 평가하고자 한다. 1) 2013 년 1 월부터 2018 년 7 월까지 단일 군병원에서 시행 된 결핵검진프로그램에 참여한 코호트 (폐 결핵 검출능 확인), 2) 2008 년 1 월부터 2012 년 12 월까지 단일 건강검진 기관에 참여한 코호트 (폐암 검출능 확인), 3) 2004 년 1 월부터 2018 년 6 월까지 동일한 건강검진 기관에 참여한 코호트 (CXR-LC 의 폐암 예측능 확인). 병변검출 DL 모델의 결핵 및 폐암 검출에 대한 진단능을 area under the receiver operating characteristic curves (AUC), sensitivity, specificity, positive predictive value, negative predictive value, accuracy 를 기결정 된 thresholds (폐 결핵: high sensitivity threshold=0.16, high specificity threshold= 0.46; 폐암:

high sensitivity threshold=0.16)을 기준으로 계산하고, 영상의학 전문의와 그 결과를 비교하였다. CXR-LC 모델의 폐암 검진 CT 수검자 선택능은 폐암 발생 예측에 대한 구별 (discrimination)과 보정(calibration)을 확인하고, 추가로 2021 년 US Preventive Services Task Force (USPSTF) recommendations 에 추가적인 이득이 있는지를 lung cancer detection rate, proportion of selected CT screening candidates, positive predictive value 를 이용하여 확인하였다.

결과: 총 19,686 명의 결핵 검진군의 20,235 흉부 X 선 중, 4 명의 5 장의 흉부 X 선이 활동성 결핵으로 확인되었다. 병변 검출 DL 모델은 이에 대해서 high sensitivity, high specificity thresholds 모두에서 폐결핵을 찾아내었다. 이때 각각의 specificities 는 95.9% 와 99.7%, PPVs 는 0.6% 와 6.8%, NPVs 는 모두 100%였다. 특히 high specificity threshold 에서 이러한 진단능은 영상의학전문의와 차이가 없었다 (P>0.05). DLAD 의 폐암검출의 위한 코호트는 50,070 명의 검진군의 100,525 흉부 X 선이 포함되었고 그 중 47 장에서 폐암이 보였다. 그 중, reader study 를 위한 validation set 으로 10,202 명의

검진군의 10,285 장의 흉부 X 선이 선별되었고, 그 중 10 장에서 폐암이 보였다. 이 validation set 에서 보이는 폐암에 대한 DLAD 의 AUC 는 0.989 로 나타났고, 영상의학전문의와 비슷한 수준의 sensitivity 를 보였지만 (P=0.248), 유의하게 낮은 specificity 를 보였다 (96.9% vs. 99.8%, P<0.001). 전체 검진군에서는 보이는 폐암에 대해 AUC 0.969 을 가졌고, sensitivity 는 83%, specificity 는 97%를 나타내었다. 폐암 검진 CT 수검자를 위한 코호트는 총 19,488 명의 검진군의 19,488 장의 흉부 X 선이 포함되었고, 그 중 폐암은 107 명에게서 발생하였다. CXR-LC 모델은 폐암 발생에 대한 AUC 0.676 을 가졌고, 특히 USPSTF-eligible 검진군에게서는 AUC 0.745 를 가졌다. 흡연량이 조사 된 17,390 명에게서 USPSTF-eligible 검진자에게 추가로 CXR-LC 를 적용하여 low-to-indeterminate risk 에 해당하는 검진자를 제외하였을 때, proportion of selected CT screening candidates 가 45.1%에서 35.8%로 유의하게 감소하였고 (P<0.001), 동시에 lung cancer detection rate(P=0.848)와 positive predictive value(0.416)는 변화가 없었다.

결론: 병변 검출 DL 모델은 검진 흉부 X 선 검사에서 활동성 결핵과 폐암 검출을 영상의학 전문의 수준으로 할 수 있고, 폐암 발생을 예측하는 CXR-LC 모델은 유의한 정도의 폐암 검진 CT 수검자를 줄이는 동시에 폐암 검출 능은 감소하지 않았다.

주요어: 딥러닝; 진단; 검진; 결핵; 폐암; 흉부 X 선; computer-assisted

학번: 2020-36356