Fukushima J. Med. Sci.,
Vol. 69, No. 3, 2023

[Original article]

Development of a novel artificial intelligence algorithm to detect pulmonary nodules on chest radiography

Mitsunori Higuchi[1)], Takeshi Nagata[2,3)], Kohei Iwabuchi[3)], Akira Sano[3)], Hidemasa Maekawa[3)], Takayuki Idaka[3)], Manabu Yamasaki[3)], Chihiro Seko[3)], Atsushi Sato[4)], Junzo Suzuki[4)], Yoshiyuki Anzai[5)], Takashi Yabuki[5)], Takuro Saito[6)] and Hiroyuki Suzuki[7)]

[1)]Department of Thoracic Surgery, Aizu Medical Center, Fukushima Medical University, Aizuwakamatsu, Japan, [2)]University of Tsukuba School of Integrative and Global Majors, Tsukuba, Japan, [3)]Mizuho Research and Technologies, Ltd., Tokyo, Japan, [4)]Fukushima Preservative Service Association of Health, Fukushima, Japan, [5)]Aizuwakamatsu Medical Association, Aizuwakamatsu, Japan, [6)]Department of Surgery, Aizu Medical Center, Fukushima Medical University, Aizuwakamatsu, Japan, [7)]Department of Chest Surgery, Fukushima Medical University School of Medicine, Fukushima, Japan

(Received April 11, 2023, accepted September 15, 2023)

Abstract

Background : In this study, we aimed to develop a novel artificial intelligence (AI) algorithm to support pulmonary nodule detection, which will enable physicians to efficiently interpret chest radiographs for lung cancer diagnosis.

Methods : We analyzed chest X-ray images obtained from a health examination center in Fukushima and the National Institutes of Health (NIH) Chest X-ray 14 dataset. We categorized these data into two types : type A included both Fukushima and NIH datasets, and type B included only the Fukushima dataset. We also demonstrated pulmonary nodules in the form of a heatmap display on each chest radiograph and calculated the positive probability score as an index value.

Results : Our novel AI algorithms had a receiver operating characteristic (ROC) area under the curve (AUC) of 0.74, a sensitivity of 0.75, and a specificity of 0.60 for the type A dataset. For the type B dataset, the respective values were 0.79, 0.72, and 0.74. The algorithms in both the type A and B datasets were superior to the accuracy of radiologists and similar to previous studies.

Conclusions : The proprietary AI algorithms had a similar accuracy for interpreting chest radiographs when compared with previous studies and radiologists. Especially, we could train a high quality AI algorithm, even with our small type B data set. However, further studies are needed to improve and further validate the accuracy of our AI algorithm.

Keywords : artificial intelligence (AI), deep learning, chest radiography, lung cancer

Introduction

Recent advances in deep learning and large datasets have enabled algorithms to surpass the performance of medical professionals in a wide variety of medical imaging tasks, including imaging for diabetic retinopathy[1)] and hemorrhage identification[2)]. Lung cancer is the leading cause of cancer-related death worldwide[3)]. Therefore, the control of lung cancer is an urgent problem that needs to be resolved. Early detection of lung cancer is extremely important, and some clinical trials, including the National Lung Screening Trial[4)] and the NELSON trial[5)], have been performed with low-dose computed tomography (CT). Despite the superior ability of CT to detect pulmonary nodules, chest radiography is still widely accepted as the first-line imaging tool to screen for and detect lung le-

Corresponding author : Mitsunori Higuchi, MD, PhD E-mail : higuchi@fmu.ac.jp

sions[6-8]. Pulmonary nodules are common initial radiologic manifestations of lung cancer ; however, they can be easily missed when they are subtle, small, or localized to difficult areas. Pulmonary nodule detection by chest radiography has been the focus of several computer-aided detection (CAD) studies in recent decades[9,10]. However, early solutions were limited due to their low sensitivity and high false-positive rates. In this study, we developed a novel AI algorithm and assessed its ability to detect pulmonary nodules on chest radiography at different levels of detection difficulty with both normal and abnormal control images. We found that the AI algorithm exceeded the average radiologist performance for pulmonary nodule detection. We suggest that automated detection of diseases based on chest radiographs at the level of expert radiologists would confer tremendous benefit in the clinical setting.

Materials and methods

Program formulation

We used the CheXNet model[11], which is a 121-layer convolutional neural network that inputs a chest X-ray image and outputs the probability of pulmonary nodules, and produces a heatmap localizing the areas of the image that are most indicative of pulmonary nodules. We trained the CheXNet model using the National Institutes of Health (NIH) Chest X-ray 14 dataset (Bethesda, MD), which contains 112,120 frontal-view chest X-ray images individually labeled with up to 14 different thoracic diseases, including atelectasis, cardiomegaly, consolidation, edema, effusion, emphysema, fibrosis, hernia, infiltration, mass, nodule, pleural thickening, pneumonia, and pneumothorax. From the NIH Chest X-ray 14 dataset, the data of 2,500 nodules (positive data) and 2,500 normal records (negative data) were used in this study.

Each output was normalized with a sigmoid function to [0,1]. The network was initialized with the pre-trained ImageNet model[12]. First, we focused on the NIH Chest X-ray 14 dataset. The labels consisted of a C dimensional vector [l_1, l_2…l_c], where C = 14 with binary values, representing either the absence (0) or presence (1) of a pathology. As a multi-label problem, we independently treated all labels during the classification by defining the C binary cross-entropy loss function. As the dataset was highly imbalanced, we incorporated additional weights within the loss function based on the label frequency within each batch :

$$L\,(\mathrm{X}, l_n) = -(w_p \cdot l_n \log(\mathrm{p}) + w_n \cdot (1 - l_n) \log(1 - \mathrm{p})),$$

where $w_p = (\mathrm{Pn} + \mathrm{Nn}) \div \mathrm{Pn}$ and $w_n = (\mathrm{Pn} + \mathrm{Nn}) \div \mathrm{Nn}$, with Pn and Nn indicating the number of samples with presence and absence of nodules, respectively.

The weights of the network were initialized with weights from the model that was pre-trained on ImageNet. The network was trained end-to-end using Adam optimization with standard parameters ($\beta_1 = 0.9$ and $\beta_2 = 0.999$). We trained the model using minibatches of size 16. We used an initial learning rate of 0.00001, which was decreased by a factor of 10 each time the validation loss plateaued after an epoch, and we selected the model with the lowest validation loss.

This study was approved by the Institutional Review Board of Fukushima Medical University (IRB-ID : 30290), which is guided by local policy, national law, and the World Medical Association Declaration of Helsinki. The chest radiographs used in this study were acquired by the Fukushima Preservative Service Association of Health in the course of its daily practice, where local lung cancer screening is mainly conducted. The need for written informed consent to use anonymized data was waived by the ethics review board. This study was supported by Grants-in-Aid for Scientific Research in Japan (ID : 21K08890).

Data training

We analyzed the image features as teacher data using 800 chest X-ray images (400 normal images [negative data] and 400 pulmonary nodules [positive data]) from Fukushima Preservative Service Association of Health, as well as 5,000 chest radiographs from the NIH Chest X-ray 14 dataset. The labelling for pulmonary nodules of the Fukushima dataset was assured by a process indicator which guaranteed the accuracy of pulmonary nodule detection in lung cancer screening. We categorized these data into two types : type A included both the Fukushima and NIH datasets and type B included only the Fukushima dataset. Then, we integrated the datasets for deep learning and convolutional neural network analyses using ImageNet to develop the proprietary AI algorithm. We then statistically analyzed the accuracy of radiograph interpretation. For cross-validation, we randomly divided the dataset into five groups, which included positive data and negative data at equal rates. Then, we validated one group

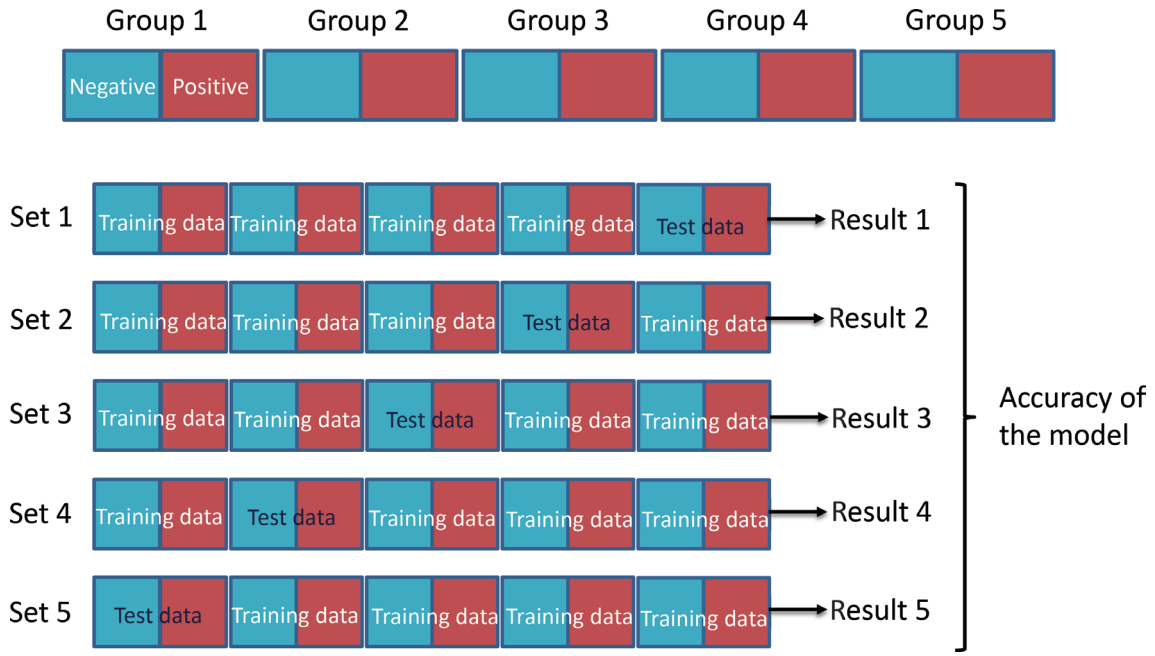

Fig. 1. Schematic view of cross-validation. We randomly divided the dataset into five groups that included positive and negative data at equal rates. Then, we validated one group as test data and used the other groups as training data. Next, we assigned each group as test data and obtained five sets of results (Results 1–5). Finally, we calculated the average accuracy for each set.

as test data and used the other groups as training data. We assigned each group as test data and obtained five sets of results (Figure 1). Finally, we calculated the average accuracy for each set. We compared the receiver operating characteristic (ROC) area under the curve (AUC) of the AI model with values reported previously[11,13,14]. We also showed the accuracy of radiologists' evaluations which were described in the website of L PIXEL Inc., Tokyo[15]. The website showed the method of evaluation of pulmonary nodules by radiologists. Nine radiologists participated in an evaluation test that included 67 radiographs with pulmonary nodule and 253 normal radiographs.

Model interpretation

We demonstrated pulmonary nodules in the form of a heatmap display on each chest radiograph for easy visualization, and we presented the positive probability score as an index value (0.0–1.0), which indicated the possibility of pulmonary nodules using class activation maps (CAMs)[16]. To generate the CAMs, we fed an image into the fully trained network and extracted the feature maps that were output by the final convolutional layer. With f_k as the k^{th} feature map and $w_{c,k}$ as the weight in the final classification layer for feature map k leading to pulmonary nodules, we obtained a map M_c of the most salient features, which were used to classify the images as having pulmonary nodules by taking the weighted sum of the feature maps using their associated weights. The equation is as follows :

$$M_c = \Sigma_k w_{c,k} \cdot f_k$$

We identified the most important features used by the model to predict the presence of pulmonary nodules by upscaling the map M_c to the dimensions of the image and overlaying the image. Our novel AI system underwent mechanical learning of training data, which were obtained using the same radiographic apparatus to eliminate the effects of differences between equipment.

Statistical analysis

The data are described as median and range for continuous variables and as percentages with 95% confidence intervals for quantitative variables. Statistical analyses were performed using SPSS 28.0.1.0 software (IBM Corp., Armonk, NY).

Results

AUC, sensitivity, and specificity

The AUC, sensitivity, and specificity of each of the five groups in both the type A and type B datasets were calculated by cross-validation (Table 1 and Table 2). The ROC curves of both the type A and

Table 1. Accuracy of each result with the type A dataset after cross-validation

Result No.	AUC	Sensitivity	Specificity
1	0.83	0.81	0.63
2	0.61	0.73	0.46
3	0.69	0.79	0.46
4	0.75	0.65	0.73
5	0.82	0.75	0.73
Average	0.74	0.75	0.60

Table 2. Accuracy of each result with the type B dataset after cross-validation

Result No.	AUC	Sensitivity	Specificity
1	0.85	0.81	0.75
2	0.64	0.52	0.70
3	0.83	0.79	0.77
4	0.79	0.72	0.73
5	0.85	0.79	0.77
Average	0.79	0.72	0.74

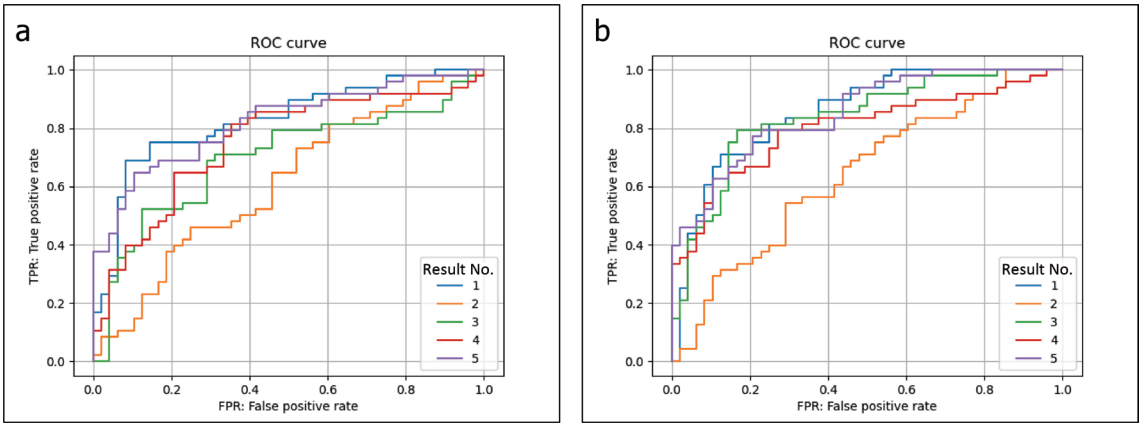

Fig. 2. Receiver operating characteristic (ROC) curves for the type A (a) and type B (b) datasets.

Table 3. Comparison of the accuracy of radiologist detection with AI algorithm detection

	AUC	Sensitivity	Specificity
Radiologists	0.717[16]	0.4710[16]	0.9635[16]
Type A	0.74	0.75	0.60
Type B	0.79	0.72	0.74

Table 4. Comparison of the AUC value of the AI algorithms with previous reports

	Rajpurkar *et al.*[11]	Wang *et al.*[12]	Yao *et al.*[14]	Type A	Type B
AUC	0.78	0.671	0.717	0.74	0.79

B datasets are shown in Figure 2. Our novel AI approach demonstrated an accuracy (AUC) of 0.74, a sensitivity of 0.75, and a specificity of 0.60 for the type A dataset. The respective values for the type B dataset were 0.79, 0.72, and 0.74. The AI algorithm used a positive probability cutoff value of 0.5. The AI algorithm applied to both the type A and B datasets was superior to the accuracy of radiologists (AUC 0.71) and either superior or comparable to previous reports[12,14]. Radiologist detection of pulmonary nodules demonstrated an AUC of 0.7173 ± 0.0344, a sensitivity of 0.4710 ± 0.0611, and a specificity of 0.9635 ± 0.0198[15]. These data are shown in Table 3. Table 4 compares our results with those of previous reports. Overall, our novel AI algorithm (using both the type A and type B datasets) resulted in comparable or superior AUC values to previous reports.

Visual and numerical demonstrations

We demonstrated that heatmaps displayed on the monitor screen clearly corresponding to the location of pulmonary nodules, if each roentgenogram had pulmonary nodules (Figure 3). Each heatmap display expressed the location of pulmonary nodules, with the exception of Figure 3-c-2, which shows a false-positive result. We also evaluated the possibility of chest nodules as a positive probability score (Figure 3). Here, we determined the cutoff value as 0.5, with a value of ≥0.5 suggesting a positive finding. However, the positive probability score of Figure 3-c-2 showed a false-positive result.

Discussion

Chest radiography remains the primary diagnostic imaging modality for thoracic conditions be-

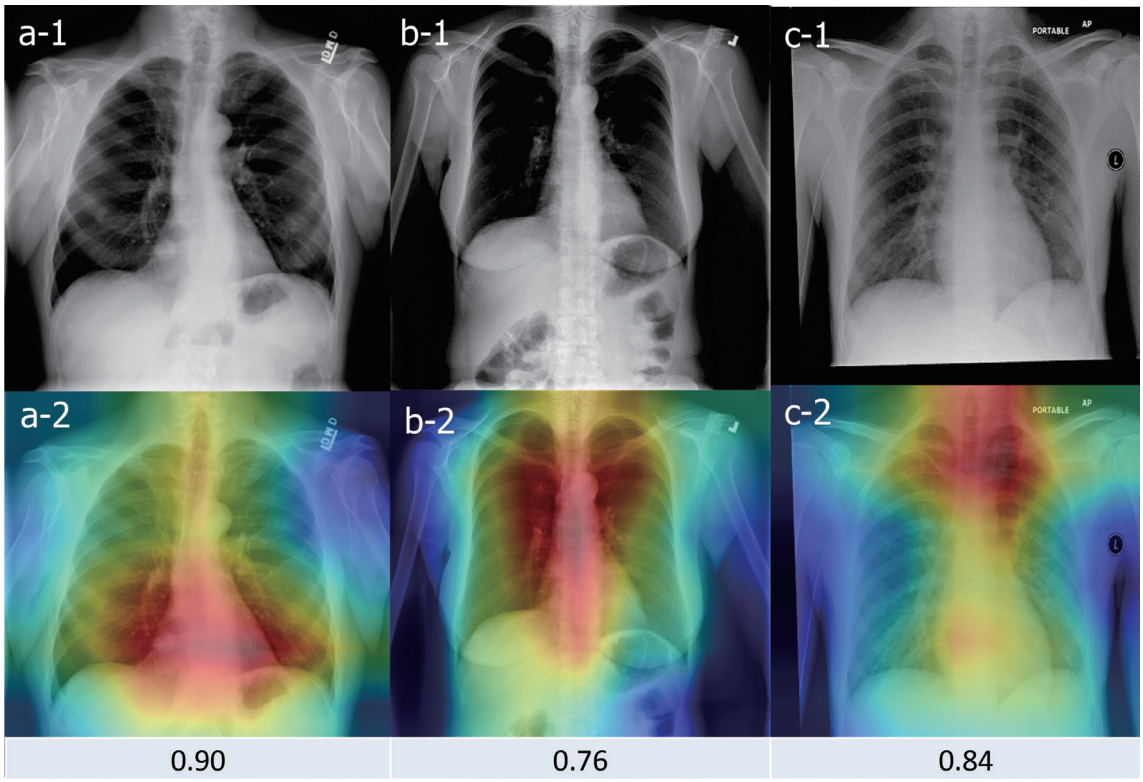

Fig. 3. Three examples from datasets of a health examination center in this study. The proposed AI algorithm correctly detected pulmonary nodules and localized the areas in the image that were most indicative of pulmonary nodules (a-1, a-2). The AI algorithm also detected pulmonary nodules that were missed by the physicians (b-1, b-2). A false-positive display is shown in c-1 and c-2, which requires improvement. The positive probability scores of these cases are shown in a-2, b-2, and c-2, respectively.

cause of its advantages over chest CT, including easier access, lower cost, and lower radiation exposure. However, previous studies have shown that 19%-26% of lung cancers that are visible on chest radiography are actually missed at the time of initial reading[17,18], and low-dose CT as opposed to chest radiography is thus recommended for lung cancer detection[19,20]. Resolving missed abnormal nodules or masses on chest radiography is an urgent problem for both physicians and patients. To date, many studies have reported the development of AI algorithms to read CT and radiography images, and some of these AI algorithms have already been put to clinical use.

In this study, we established two novel AI algorithms (type A and type B), which were constructed based on whether they included the NIH Chest X-ray 14 dataset or not, in addition to the inclusion of 800 chest radiographs from Fukushima Preservative Service Association of Health. The purpose of developing two types of algorithms was to confirm the accuracy of AI derived from a small number of radiographs. The novel AI algorithms, especially the one derived from our small type B data set, were associated with improvements in the AUC and sensitivity of pulmonary nodule detection on chest radiographs compared with the respective values for nodule detection by radiologists, as reported in previous studies[16]. However, specificities of the AI algorithms were inferior to that of radiologist detection[16]. Chest radiography is first used to screen for thoracic diseases ; this is followed by conventional chest CT, positron emission tomography CT, magnetic resonance imaging, and/or other imaging modalities. Therefore, over-detection (false-positive results) of pulmonary nodules on chest radiography is not a crucial problem. In this study, type B yielded better accuracy than type A, which received pre-training with 5,000 radiographs from the NIH Chest X-ray 14 dataset. In general, to acquire a high accuracy by deep learning, massive teaching data are required, such as the NIH Chest X-ray 14 dataset. However, the type A dataset was inferior to the type B dataset. This may have been because the Chest X-ray 14 dataset included images from various types of radiography devices, with imaging environments and examinee postures that may have differed. In contrast, our Type B dataset, while much

smaller than the NIH dataset, is largely derived from regular, standardized health screening that is conducted in Fukushima as part of Japan's system of universal health care. AI developer must be mindful of such aforementioned variations. However, even with such variables, we were able to establish novel AI algorithms that demonstrated comparable or superior accuracy to those reported previously[13,14].

The present study has several limitations that should be noted. First, this study used data with "nodule-positive radiographs" and "normal radiographs" rather than "nodule-negative radiographs" obtained from two datasets. Case-control methods for diagnostic accuracy studies could lead to overestimation with respect to sensitivity and specificity. We compared the AUC-ROC with those of previous studies which examined not only pulmonary nodules but also other pulmonary abnormalities using AI algorithms. Therefore, the AUCs of the current study could be on par with, or surpass, those of previous studies. In real-world situations, chest radiographs can possibly have more than one abnormality. Since clinicians have to detect all abnormalities, including pulmonary nodules, it is still not very clear whether or not an AI algorithm that focuses only on pulmonary nodules is useful. Second, the sample size was small, and images obtained from patients with lung cancer were lacking. We are now planning data collection to resolve these data insufficiencies, especially in terms of obtaining chest radiographs from patients with lung cancer as positive data. Third, we did not use technology that accounted for differences in radiograph apparatus and imaging environments. Therefore, we had to resolve these problems by standardizing the differences. We need to collect more high-quality data from various institutions to overcome this limitation. Third, the accuracy of the novel AI algorithm was evaluated using the NIH Chest X-ray 14 dataset and 800 chest radiographs obtained from Fukushima Preservative Service Association of Health. We did not assess whether this AI algorithm might improve the accuracy of pulmonary nodule detection by chest radiography if the physicians used it for CAD. Therefore, the use of the AI algorithm in CAD should be validated in the future. To evaluate the capability of CAD to assist physicians, a reader performance test comparing the physician performance before and after the use of CAD will be conducted. CAD is certified as a medical software for use by physicians as a second opinion. However, our ultimate goal is to achieve an autonomous AI algorithm to detect pulmonary nodules on chest radiographs, even though we will need to overcome several obstacles, including technical and legal issues, amongst others. The first fully autonomous AI algorithm that is able to perform diagnostic assessment without the supervision of an expert clinician is the IDx-DR AI system, which is used to analyze fundus photographs in the primary care setting to detect diabetic retinopathy. The IDx-DR AI system was approved in 2018 by the Food and Drug Administration[21]. The autonomous AI algorithm should include additional components that help to guarantee the robustness of the AI output.

In conclusion, we developed an AI algorithm to detect pulmonary nodules from frontal-view chest radiographs with an accuracy that exceeds that of radiologists. We hope this technology can improve healthcare delivery and increase access to medical imaging expertise in parts of the world where access to skilled radiologists is limited. Our system also has the potential to support the current high-throughput reading workflow of radiologists by enabling them to gain more confidence in using AI systems to obtain a second opinion. We are now in the process of performing various types of validation to improve the accuracy and to achieve autonomous diagnosis.

Acknowledgements

We thank Emily Woodhouse, PhD, from Edanz (https://jp.edanz.com/ac) for editing a draft of this manuscript.

Conflict of interest disclosure

The authors have no conflicts of interest to declare.

References

1. Gulshan V, Peng L, Coram M, *et al.* Development and validation of a deep learning algorithm for detection of diabetic retinopathy in retinal fundus photographs. JAMA, **316**: 2402-2410, 2016.
2. Grewal M, Srivastava MM, Kumar P, Varadarajan S. RADNET: Radiologist level accuracy using deep learning for hemorrhage detection in CT scans. arXiv preprint, arXiv: 1710.04934, 2017.
3. Howlader N, Noone AM, Krapcho M, *et al.* SEER Cancer Statistics Review, 1975-2010, National Cancer Institute. Bethesda, MD, https://seer.cancer.gov/archive/csr/1975_2010/. Accessed 14 June

2013.
4. Aberle DR, Adams AM, Berg CD, *et al.* The National Lung Screening Trial Research Team. Reduced lung-cancer mortality with low-dose computed tomographic screening. N Engl J Med, **365** : 395-409, 2011.
5. Horeweg N, Scholten ET, de Jong PA, *et al.* Detection of lung cancer through low-dose CT screening (NELSON) : A prespecified analysis of screening test performance and interval cancers. Lancet Oncol, **15** : 1342-1350, 2014.
6. Greene R. Francis H. Williams, MD : Father of chest radiology in North America. Radiographics, **11** : 325-332, 1991.
7. Yoo H, Lee SH, Arru CD, *et al.* AI-based improvement in lung cancer detection on chest radiographs : Results of a multi-reader study in NLST dataset. Eur Radiol, **31** : 9664-9674, 2021.
8. van Beek EJ, Mirsadraee S, Murchison JT. Lung cancer screening : Computed tomography or chest radiographs? World J Radiol, **7** : 189-193, 2015.
9. Chen S, Han Y, Lin J, Zhao X, Kong P. Pulmonary nodule detection on chest radiographs using balanced convolutional neural network and classic candidate detection. Artif Intell Med, **107** : 101881, 2020.
10. Liang CH, Liu YC, Wu MT, Garcia-Castro F, Alberich-Bayarri A, Wu FZ. Identifying pulmonary nodules or masses on chest radiography using deep learning : External validation and strategies to improve clinical practice. Clin Radiol, **75** : 38-45, 2020.
11. Rajpurkar P, Irvin J, Zhu K, *et al.* CheXNet : Radiologist-level pneumonia detection on chest X-rays with deep learning. arXiv preprint, arXiv : 1711.05225, 2017.
12. Deng J, Dong W, Socher R, Li LJ, Fei-Fei L. ImageNet : A large-scale hierarchical image database. In : Computer Vision and Pattern Recognition, 248-255, 2009.
13. Wang X, Peng Y, Lu L, Lu Z, Bagheri M, Summers RM. ChestX-ray8 : Hospital-scale chest x-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases. arXiv preprint, arXiv : 1705.02315, 2017.
14. Yao L, Poblenz E, Dagunts D, Covington B, Bernard D, Lyman K. Learning to diagnose from scratch by exploiting dependencies among labels. arXiv preprint, arXiv : 1710.10501, 2017.
15. https://eirl.ai/eirl-chest_nodule/
16. Zhou B, Khosla A, Lapedriza A, Oliva A, Torralba A. Learning deep features for discriminative localization. In : Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2021-2929, 2016.
17. Gavelli G, Giampalma E. Sensitivity and specificity of chest X-ray screening for lung cancer : Review article. Cancer, **89** : 2453-2456, 2000.
18. Quekel LG, Kessels AG, Goei R, van Engelshoven JM. Miss rate of lung cancer on the chest radiograph in clinical practice. Chest, **115** : 720-724, 1999.
19. Manser R, Lethaby A, Irving LB, *et al.* Screening for lung cancer. Cochrane Database Syst Rev, 2013 : CD001991, 2013.
20. Aberle DR, Adams AM, Berg CD, *et al.* ; The National Lung Screening Trial Research Team. Reduced lung-cancer mortality with low-dose computed tomographic screening. N Eng J Med, **365** : 395-409, 2011.
21. Abramoff MD, Lavin PT, Birch M, Shah N, Folk JC. Pivotal trial of an autonomous AI-based diagnostic system for detection of diabetic retinopathy in primary care officers. NPJ Digit Med, **1** : 39, 2018.