ORIGINAL ARTICLE
LUNG IMAGING

Development and validation of a deep learning algorithm detecting 10 common abnormalities on chest radiographs

Ju Gang Nam[1,2], Minchul Kim[3], Jongchan Park[3], Eui Jin Hwang[1,2], Jong Hyuk Lee[1,2], Jung Hee Hong[1,2], Jin Mo Goo[1,2,4] and Chang Min Park[1,2,4]

Affiliations: [1]Dept of Radiology, Seoul National University Hospital, Seoul, Republic of Korea. [2]College of Medicine, Seoul National University, Seoul, Republic of Korea. [3]Lunit Inc., Seoul, Republic of Korea. [4]Institute of Radiation Medicine, Seoul National University Medical Research Center, Seoul, Republic of Korea.

Correspondence: Chang Min Park, Dept of Radiology and Institute of Radiation Medicine, Seoul National University College of Medicine, 101 Daehak-ro, Jongno-gu, Seoul 03080, Republic of Korea.
E-mail: cmpark.morphius@gmail.com

@ERSpublications
A deep learning algorithm detecting 10 common abnormalities was trained with 146 717 images and showed excellent performance on chest radiographs, helping radiologists improve their performance and advance the reporting time for critical or urgent cases https://bit.ly/3k8tZ5P

Cite this article as: Nam JG, Kim M, Park J, *et al*. Development and validation of a deep learning algorithm detecting 10 common abnormalities on chest radiographs. *Eur Respir J* 2021; 57: 2003061 [https://doi.org/10.1183/13993003.03061-2020].

ABSTRACT We aimed to develop a deep learning algorithm detecting 10 common abnormalities (DLAD-10) on chest radiographs, and to evaluate its impact in diagnostic accuracy, timeliness of reporting and workflow efficacy.

DLAD-10 was trained with 146 717 radiographs from 108 053 patients using a ResNet34-based neural network with lesion-specific channels for 10 common radiological abnormalities (pneumothorax, mediastinal widening, pneumoperitoneum, nodule/mass, consolidation, pleural effusion, linear atelectasis, fibrosis, calcification and cardiomegaly). For external validation, the performance of DLAD-10 on a same-day computed tomography (CT)-confirmed dataset (normal:abnormal 53:147) and an open-source dataset (PadChest; normal:abnormal 339:334) was compared with that of three radiologists. Separate simulated reading tests were conducted on another dataset adjusted to real-world disease prevalence in the emergency department, consisting of four critical, 52 urgent and 146 nonurgent cases. Six radiologists participated in the simulated reading sessions with and without DLAD-10.

DLAD-10 exhibited area under the receiver operating characteristic curve values of 0.895–1.00 in the CT-confirmed dataset and 0.913–0.997 in the PadChest dataset. DLAD-10 correctly classified significantly more critical abnormalities (95.0% (57/60)) than pooled radiologists (84.4% (152/180); p=0.01). In simulated reading tests for emergency department patients, pooled readers detected significantly more critical (70.8% (17/24) *versus* 29.2% (7/24); p=0.006) and urgent (82.7% (258/312) *versus* 78.2% (244/312); p=0.04) abnormalities when aided by DLAD-10. DLAD-10 assistance shortened the mean±SD time-to-report critical and urgent radiographs (640.5±466.3 *versus* 3371.0±1352.5 s and 1840.3±1141.1 *versus* 2127.1±1468.2 s, respectively; all p<0.01) and reduced the mean±SD interpretation time (20.5±22.8 *versus* 23.5±23.7 s; p<0.001).

DLAD-10 showed excellent performance, improving radiologists' performance and shortening the reporting time for critical and urgent cases.

This article has supplementary material available from erj.ersjournals.com

Received: 7 Aug 2020 | Accepted: 3 Nov 2020

Introduction

Chest radiographs are the most frequently performed radiological examination [1, 2], and the large work burden hampers timely diagnoses, interferes with the clinical workflow and increases the misinterpretation rate [3]. In recent years, deep learning technology has been widely applied for chest radiograph interpretation [4–9]. Various algorithms have shown good performance in specific task-based analyses, including detection of lung nodules/masses, pneumothorax and pulmonary tuberculosis [5–9]. These algorithms may increase radiologists' detection performance and improve their confidence, but it remains unclear whether these algorithms could decrease radiologists' work burden and facilitate timely diagnoses.

Radiologists' interpretation of a radiograph can be typically divided into four processes: 1) detection and localisation of clinically relevant abnormalities, 2) comparison with previous radiographs (if any), 3) final interpretation with differential diagnoses, and 4) generation of a radiology report. In this study, we focused on the first step, and developed an automated algorithm that can detect and localise common abnormal findings on chest radiographs. Several algorithms covering multiple abnormalities have been reported, but the coverage of findings was limited [10, 11] or the performance was unsatisfactory compared with radiologists [12, 13].

Thus, the purpose of our study was to develop a deep learning-based algorithm for 10 common radiological abnormalities (DLAD-10), and to evaluate and compare its performance with that of radiologists. In addition, we investigated whether DLAD-10 could boost the detection performance and workflow efficacy of radiologists on simulated reading tests for patients visiting the emergency department.

Materials and methods

This retrospective study was approved by our institutional review boards and the requirement for patients' informed consent was waived.

Development of the DLAD-10

DLAD-10 was developed for 10 abnormalities, selected to cover a majority of thoracic diseases [14]: pneumothorax, mediastinal widening, pneumoperitoneum, nodule/mass, consolidation, pleural effusion, linear atelectasis, fibrosis, calcification and cardiomegaly. These abnormalities were defined in accordance with the Fleischer Society glossary [15]. Specifically, "mediastinal widening" indicated enlargement of the aortic shadow, suggesting aortic disease [16], and "fibrosis" indicated focal fibrotic change rather than diffuse reticular opacities, suggesting interstitial lung disease (ILD) [15]. The study design is summarised in figure 1.

Development dataset

For the development of DLAD-10, 146 717 chest radiographs (143 768 posteroanterior and 2949 anteroposterior projection; 90 317 normal and 56 400 abnormal) from 108 053 patients (55 394 males and 52 659 females; mean±SD age 56.1±14.5 years) taken between March 2004 and December 2017 were retrospectively collected from Seoul National University Hospital (Seoul, Republic of Korea) (see supplementary table E1 for chest radiograph scanner information). Some of the dataset was used in our previous studies [6, 11], but the algorithm was re-designed and trained to perform a different task. Every chest radiograph was reviewed by at least one of 20 board-certified radiologists (labelling group; 7–14 years of experience) and image-level labels were obtained for each of the 10 abnormalities. Each abnormality was then localised (pixel-level annotation) by the labelling group (details described in the supplementary material). Training was conducted in a semisupervised manner, in which all radiographs were assigned at least one label for the 10 abnormalities, but some were not annotated for the exact location. Details on the numbers of chest radiographs are provided in supplementary table E2.

Deep learning algorithm

The model used a ResNet34-based deep convolutional neural network [17]. The final layer output 10 different abnormality-specific channels, each representing the probability map for the corresponding abnormality (supplementary figure E1). We inserted an Attend-and-Compare Module in the intermediate layers to improve detection performance [18]. During the training, the AutoAugment algorithm [19] combined with conventional image processing techniques such as brightness/contrast adjustment, blurring and random cropping was applied to augment the training dataset. During the inference process, each chest radiograph image was split into patches and the network prediction of the image patches was aggregated to create a prediction result for the whole image. Binary cross-entropy was used as the loss function, stochastic gradient descent was used as the optimiser, the learning rate was 0.01–0.001 and up to 40 epochs were used (details described in the supplementary material).

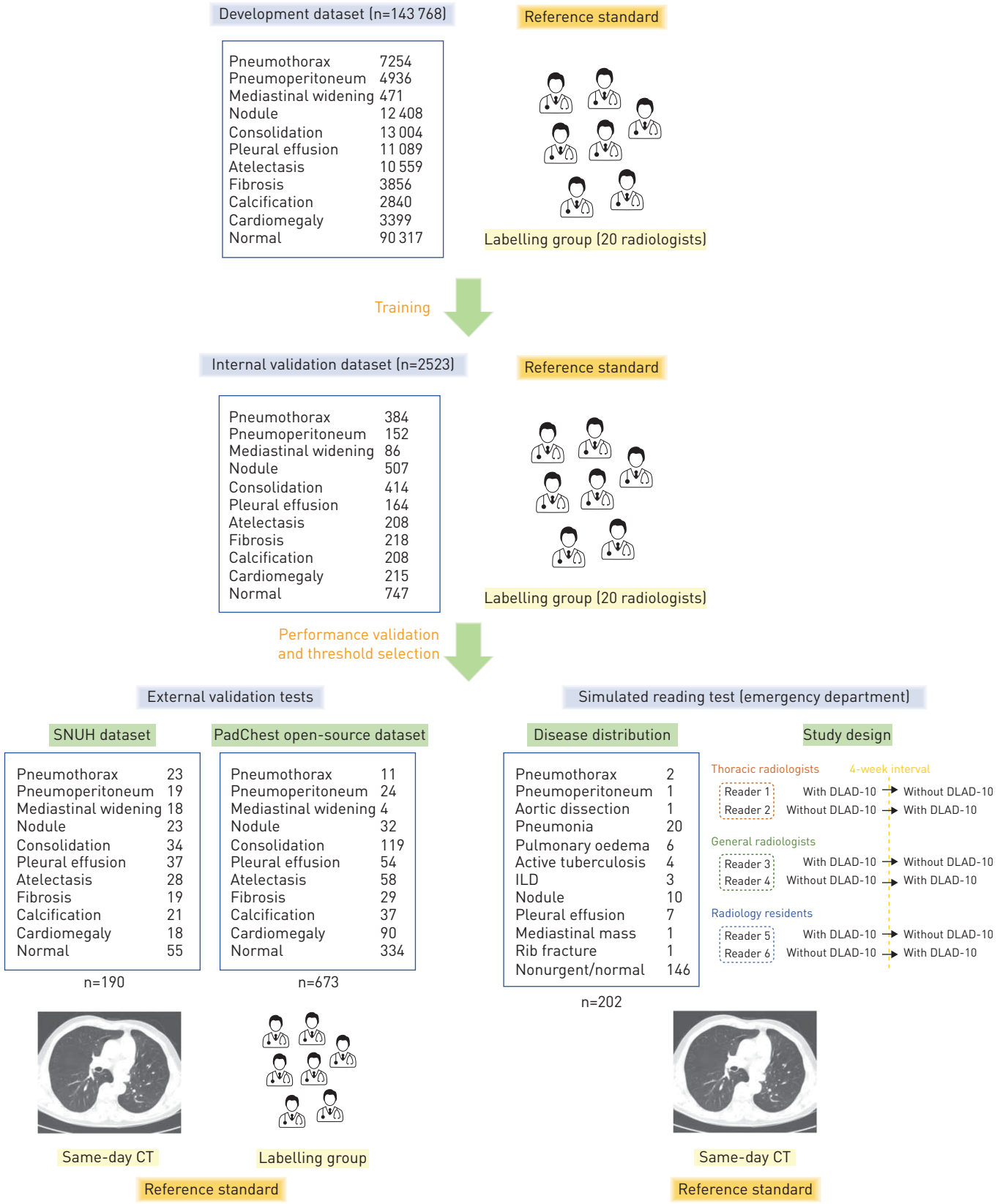

FIGURE 1 Development and validation of DLAD-10. SNUH: Seoul National University Hospital; ILD: interstitial lung disease; CT: computed tomography. See main text and supplementary figure E1 for details of the training stage.

Internal validation

For internal validation, 2523 chest radiographs (2311 posteroanterior and 212 anteroposterior projection; 747 normal and 1776 abnormal) from 2523 patients not included in the development dataset were collected, all with image-level labels for the 10 abnormalities (table 1). The area under the receiver operating characteristic curve (AUROC) was calculated for each abnormality, and binary classification cut-offs yielding 90% and 95% specificity and sensitivity were obtained.

Urgency categorisation of chest radiographs

We categorised the 10 abnormalities covered by DLAD-10 according to their clinical urgency following a previous study [20]. "Critical" abnormalities were defined as findings requiring immediate management within 12–24 h (pneumothorax, pneumoperitoneum and mediastinal widening), "urgent" abnormalities were defined as nonemergent findings that nonetheless require a prompt evaluation of the disease aetiology (nodule/mass, consolidation and pleural effusion) and "nonurgent" abnormalities were defined as those that do not change patients' management (linear atelectasis, fibrosis and calcification). Cardiomegaly was not categorised, as its clinical significance is generally not solely decided by chest radiography. Chest radiographs containing multiple abnormalities with different urgency categories were classified as belonging to the most urgent category (figure 2a).

External validation

External validation was conducted using two independent datasets. First, a temporally independent dataset consisting of 190 chest radiographs (169 posteroanterior and 21 anteroposterior projection) taken from January to December 2018 from 190 patients (101 males and 89 females; mean±SD age 59.4±14.5 years), accompanied with same-day computed tomography (CT) scans as the reference standard, was collected from Seoul National University Hospital (SNUH dataset). The dataset was curated and labelled by one thoracic radiologist (J.G.N.; 6 years of experience) to contain 20–40 chest radiographs for each abnormality; the cases were consecutively selected for each abnormality (supplementary table E3). The reference standard of cardiomegaly was only applied to posteroanterior images (n=169) according to the cardiothoracic ratio (cut-off 0.5) [21, 22], while CT was referenced for the other nine abnormalities. Additionally, an open dataset (PadChest) consisting of 673 chest radiographs labelled by the labelling group was used for the other external validation test [23]. The numbers of chest radiographs for individual abnormalities are presented in supplementary table E3.

Using the 190 same-day CT-confirmed dataset (SNUH dataset), we performed a reader test, in which three thoracic radiologists (J.H.H., J.H.L. and E.J.H.; 7–10 years of experience) participated in a comparative analysis with DLAD-10. The three radiologists were not involved in the labelling process

TABLE 1 Results of DLAD-10 and selected thresholds for each abnormality from the internal validation test

	Critical			Urgent			Nonurgent			Cardiomegaly
	Pneumothorax	Pneumoperitoneum	Mediastinal widening	Nodule	Consolidation	Pleural effusion	Atelectasis	Fibrosis	Calcification	
Positive cases n	384	152	86	507	414	164	208	218	208	215
Negative cases n	2139	2371	2437	2016	2109	2359	2315	2305	2315	2308
AUROC	0.996	0.996	0.966	0.936	0.925	0.933	0.935	0.893	0.952	0.963
Threshold selection										
Optimal	0.38#	0.15#	0.21#	0.45	0.54	0.13	0.37	0.18	0.43	0.15#
Sensitivity %	96.4	97.4	93.0	84.8	88.7	97.6	86.5	89.5	88.0	92.1
Specificity %	98.1	98.9	91.0	88.2	81.2	78.8	85.1	77.5	91.5	89.2
Sensitivity 90%				0.31	0.48	0.30	0.20	0.16	0.14	
Sensitivity 95%			0.08	0.21#	0.32#	0.13#	0.10	0.04	0.11	0.10
Specificity 90%				0.50	0.78	0.63	0.53	0.70#	0.37	0.23
Specificity 95%			0.74	0.70	0.91	0.84	0.73#	0.86	0.67#	0.66

AUROC: area under the receiver operating characteristic curve. #: selected threshold value for each abnormality. The optimal thresholds corresponding to the Youden index were selected for critical abnormalities and cardiomegaly, as they yielded satisfactory sensitivity and specificity. For urgent abnormalities, threshold values yielding 95% sensitivity were selected. The specificities at the corresponding thresholds were 71.4%, 72.5% and 79.4% for nodules, consolidation and pleural effusion, respectively. For nonurgent abnormalities, 95% specificity was selected for atelectasis and calcification (corresponding sensitivities 61.5% and 81.3%, respectively). For fibrosis, the threshold yielding 90% specificity was selected (corresponding sensitivity 60.6%), as the threshold of 95% specificity yielded suboptimal sensitivity (39.9%).

during the algorithm development. Each radiologist reviewed 190 chest radiographs independently and decided whether each abnormality was present on each chest radiograph.

Simulated reading test for emergency department patients

Dataset for the simulated reading test

To investigate the boosting effect of diagnostic accuracy, the timely diagnosis of clinically relevant diseases and the workflow efficacy of DLAD-10 in real clinical situations, chest radiographs taken from patients who visited the emergency department of Seoul National University Hospital in 2018 and had same-day CT scans as a reference standard were collected (supplementary table E4). Among the 1455 chest radiographs from 1178 patients, 202 chest radiographs from 202 patients (95 males and 107 females; mean±SD age 57.6±17.9 years) were selected to match the previously reported disease prevalence of patients visiting the emergency department [24]. Of these chest radiographs, 72.3% (146/202) were clinically insignificant cases and 27.7% (56/202) were clinically relevant cases, including pneumonia (35.7% (20/56)), pulmonary oedema (10.7% (6/56)), active tuberculosis (7.1% (4/56)), ILD (5.4% (3/56)), nodule/mass (17.9% (10/56)), pleural effusion without any other abnormality (12.5% (7/56)), mediastinal mass (1.8% (1/56)), rib fracture (1.8% (1/56)), pneumothorax (3.6% (2/56)), acute aortic syndrome (1.8% (1/56)) and pneumoperitoneum (1.8% (1/56)). The corresponding CT images served as a reference standard for all diseases, while PCR results were additionally used for active pulmonary tuberculosis. The clinically relevant cases were categorised into critical (pneumothorax, aortic dissection and pneumoperitoneum) and urgent (pneumonia, pulmonary oedema, active tuberculosis, ILD, isolated pleural effusion, mediastinal mass and rib fracture) diseases following the same criteria used for abnormality categorisation [25].

Integration of DLAD-10 into the picture archiving and communication system and reader test

The results of DLAD-10 were integrated into our institution's picture archiving and communication system (PACS) (Gx; Infinitt Healthcare, Seoul, Republic of Korea), so that readers could adjust their worklist on the PACS and rearrange the order of chest radiographs according to abnormal findings or probability scores yielded by DLAD-10 at their discretion. For the abnormal findings, the most emergent finding of a chest radiograph image was displayed on the worklist along with its probability (supplementary figure E2). When a reader opened the image, two chest radiographs were displayed: one without the DLAD-10 results (original chest radiograph) and the other with all abnormal findings localised by DLAD-10 with their probability scores (figure 2).

Six readers, including two thoracic radiologists (7 years of experience), two board-certified general radiologists (6 years of experience) and two radiology residents who had experience in reading emergency department chest radiographs, participated in the reader test. None of the readers was involved in the labelling process on DLAD-10 development. Each reader interpreted the 202 chest radiographs twice at a 4-week interval, once with DLAD-10 results (DLAD-10-aided reading session) and once without DLAD-10 results (conventional reading session). In the conventional reading session, 202 chest radiographs were listed in the PACS worklist in random order and the readers interpreted them sequentially. In the DLAD-10-aided reading session, the readers were able to rearrange the list of chest radiographs according to the urgency and probability score provided by DLAD-10. They were instructed to interpret the more urgent cases first (supplementary figure E2). Reporting was conducted in the same manner as the routine reading process performed in the emergency department. After reviewing the images, each reader made formal reports of three to four lines of text, including abnormal findings and possible differential diagnoses. Three of the six readers performed the conventional reading session before the DLAD-10-aided reading session, while the other three performed the DLAD-10-aided reading session first (supplementary figure E2). The time taken for interpretation of each chest radiograph by each reader was recorded on the PACS. From these recordings, the interpretation time taken for each chest radiograph and the time taken from the start of the reading session to the interpretation of each chest radiograph (time-to-report) were calculated (supplementary figure E2c).

Statistical analyses

The AUROCs of DLAD-10 in classifying each abnormality in the internal validation dataset and two external validation datasets were calculated. The optimal thresholds corresponding to the Youden index [26] and thresholds yielding 90% and 95% sensitivity and specificity for each abnormality were obtained from the internal validation test, and were applied in the external validation and simulated reading test. The sensitivity and specificity of DLAD-10 were compared with those of the pooled three radiologists in the external validation test using generalised estimated equations. For the simulated reading test, the urgency categorisation accuracy for each disease was calculated for DLAD-10 and the readers. The accuracy of the readers in two reading sessions was compared using the McNemar test. Interpretation time and

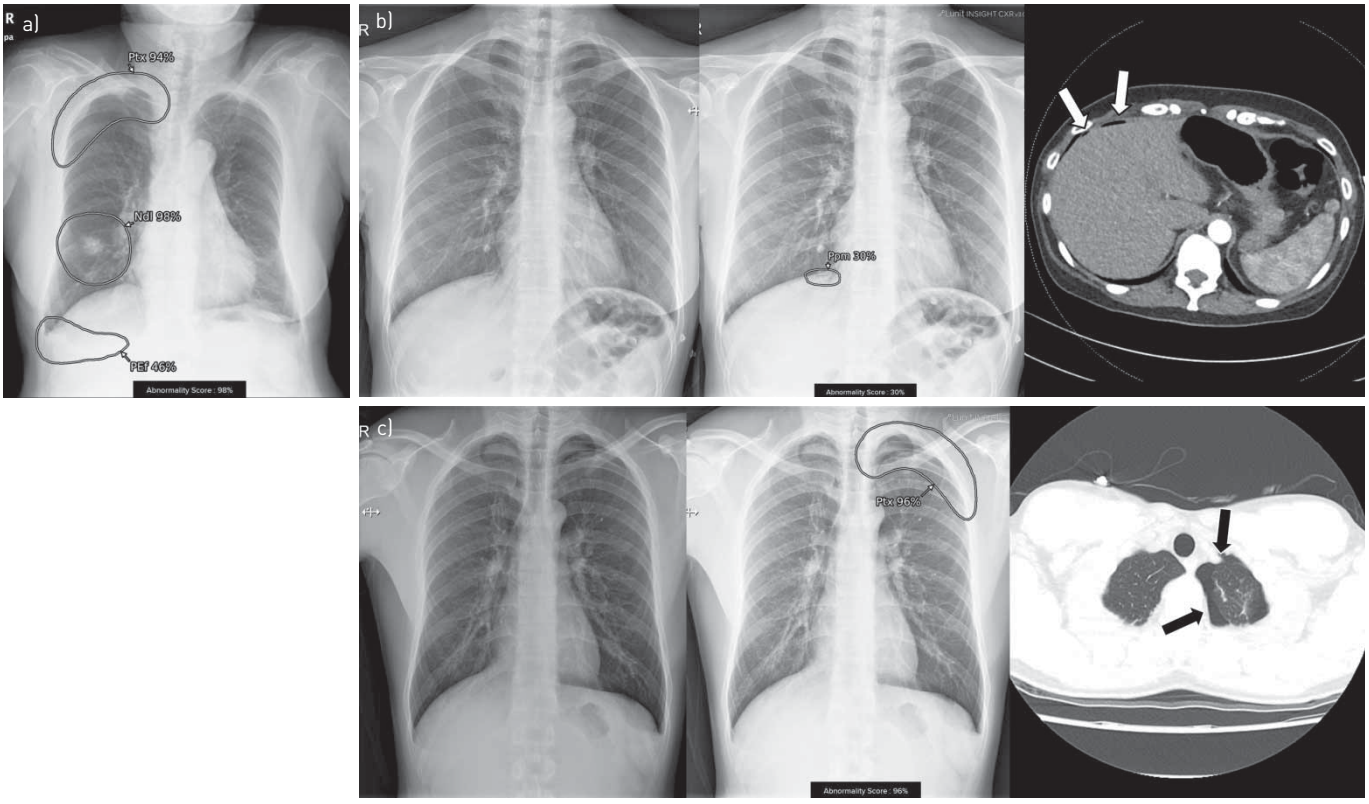

FIGURE 2 Examples of DLAD-10 output. a) Each of 10 possible abnormalities was localised and displayed with its probability score. Urgency categorisation was performed based on the most urgent abnormality. This image was categorised as critical as it contained pneumothorax (Ptx) (in addition to nodule (Ndl) and pleural effusion (PEf)). b) A 47-year-old female patient visited the emergency department complaining of vague chest pain. A small pneumoperitoneum (Ppm) was detected by DLAD-10, while no readers detected the lesion in the conventional reading session. In the DLAD-10-aided reading session, all readers detected pneumoperitoneum. c) A 24-year-old male patient visited the emergency department due to left chest pain. A small left pneumothorax (Ptx) was detected by DLAD-10. Three readers reported pneumothorax in the conventional reading session and all six readers reported it in the DLAD-10-aided reading session. The arrows on the computed tomography scans in b) and c) indicate the corresponding abnormalities visualised on the chest radiographs.

time-to-report taken for each chest radiograph were measured and compared between the two reading sessions using the paired t-test. Statistical analyses were performed with scikit-learn version 0.19.0 [27], MedCalc version 15.8 (MedCalc, Ostend, Belgium) and SPSS version 25 (IBM, Armonk, NY, USA).

Results

Internal validation test

DLAD-10 showed AUROCs of 0.893–0.996 in the internal validation dataset (table 1). The threshold for each abnormality was selected based on its clinical significance and sensitivity/specificity: pneumothorax 0.38, pneumoperitoneum 0.15, mediastinal widening 0.21, nodule/mass 0.32, consolidation 0.32, pleural effusion 0.13, linear atelectasis 0.73, fibrosis 0.70, calcification 0.67 and cardiomegaly 0.15. High sensitivity thresholds (sensitivity >93%) were selected for critical (pneumothorax, pneumoperitoneum and mediastinal widening) or urgent (nodule/mass, consolidation and pleural effusion) abnormalities, and high specificity thresholds (specificity >90%) were selected for nonurgent abnormalities (linear atelectasis, fibrosis and calcification) (table 1 and figure 1).

External validation tests

DLAD-10 showed AUROCs of 0.895 (cardiomegaly) to 1.00 (pneumoperitoneum) for each abnormality in the CT-confirmed SNUH dataset and 0.913 (linear atelectasis) to 0.997 (pneumothorax) in the PadChest dataset (table 2). Compared with thoracic radiologists, DLAD-10 generally showed higher sensitivities, while radiologists were more specific (table 3). DLAD-10 showed comparable performance to the radiologists in terms of AUROCs for most abnormalities, while the performance of most radiologists was located below DLAD-10's performance curve for critical abnormalities (figure 3). DLAD-10 correctly categorised chest radiographs containing critical abnormalities better than the pooled radiologists (95.0%

TABLE 2 External validation results of DLAD-10

	SNUH dataset	PadChest open dataset
Total chest radiographs	190	673
Reference standard	Same-day CT or cardiothoracic ratio (cardiomegaly)	Radiologists (labelling group)
Pneumothorax	0.999 (100, 98.2)	0.997 (100, 95.2)
Pneumoperitoneum	1.00 (100, 98.8)	0.994 (87.5, 98.8)
Mediastinal widening	0.978 (83.3, 93.6)	0.953 (100, 80.1)
Nodule	0.943 (95.7, 71.9)	0.932 (90.6, 74.6)
Consolidation	0.916 (82.4, 78.2)	0.967 (98.3, 74.7)
Pleural effusion	0.944 (86.5, 87.6)	0.981 (98.1, 85.1)
Atelectasis	0.909 (67.9, 94.4)	0.913 (87.9, 78.7)
Fibrosis	0.972 (78.9, 95.3)	0.971 (96.6, 86.8)
Calcification	0.923 (76.2, 97.0)	0.966 (91.7, 89.3)
Cardiomegaly	0.895 (61.1, 93.4)	0.913 (87.8, 81.8)

Data are presented as n or area under the receiver operating characteristic curve (sensitivity %, specificity %). CT: computed tomography.

TABLE 3 Comparison of the performance of DLAD-10 and three thoracic radiologists in the external validation test

	DLAD-10	Pooled thoracic radiologists	p-value
Sensitivity and specificity for detecting each abnormality			
Pneumothorax (n=23)			
Sensitivity	100 (23/23)	91.3 (63/69)	<0.001
Specificity	98.2 (164/167)	99.6 (499/501)	0.10
Pneumoperitoneum (n=19)			
Sensitivity	100 (19/19)	94.7 (54/57)	0.25
Specificity	98.2 (168/171)	99.8 (512/513)	<0.01
Mediastinal widening (n=18)			
Sensitivity	83.3 (15/18)	61.1 (33/54)	0.03
Specificity	93.6 (161/172)	98.1 (506/516)	<0.001
Nodule (n=23)			
Sensitivity	95.7 (22/23)	71.0 (49/69)	0.04
Specificity	71.9 (120/167)	90.6 (454/501)	<0.001
Consolidation (n=34)			
Sensitivity	82.4 (28/34)	60.8 (62/102)	0.01
Specificity	78.2 (122/156)	91.2 (427/468)	<0.001
Pleural effusion (n=37)			
Sensitivity	86.5 (32/37)	74.8 (83/111)	0.03
Specificity	87.6 (134/153)	95.4 (438/459)	<0.001
Atelectasis or fibrosis (n=45[#])			
Sensitivity	75.6 (34/45)	68.9 (93/135)	0.29
Specificity	90.3 (131/145)	83.9 (365/435)	0.02
Calcification (n=21)			
Sensitivity	76.2 (16/21)	58.7 (37/63)	0.02
Specificity	97.0 (164/169)	96.8 (491/507)	0.89
Cardiomegaly (n=18)			
Sensitivity	61.1 (11/18)	35.2 (19/54)	0.02
Specificity	93.4 (141/151)	98.5 (446/453)	0.002
Urgency categorisation accuracy[¶]			
Critical (n=60)	95.0 (57/60)	84.4 (152/180)	0.01
Critical or urgent (n=110)	95.5 (105/110)	91.2 (301/330)	0.09
Normal/nonurgent (n=80)	80.0 (64/80)	88.3 (212/240)	0.03

Data are presented as % (n/N), unless otherwise stated. [#]: some patients had both atelectasis and fibrosis; [¶]: accuracy of correctly classifying chest radiographs according to their urgency category. p-values were calculated using generalised estimating equations.

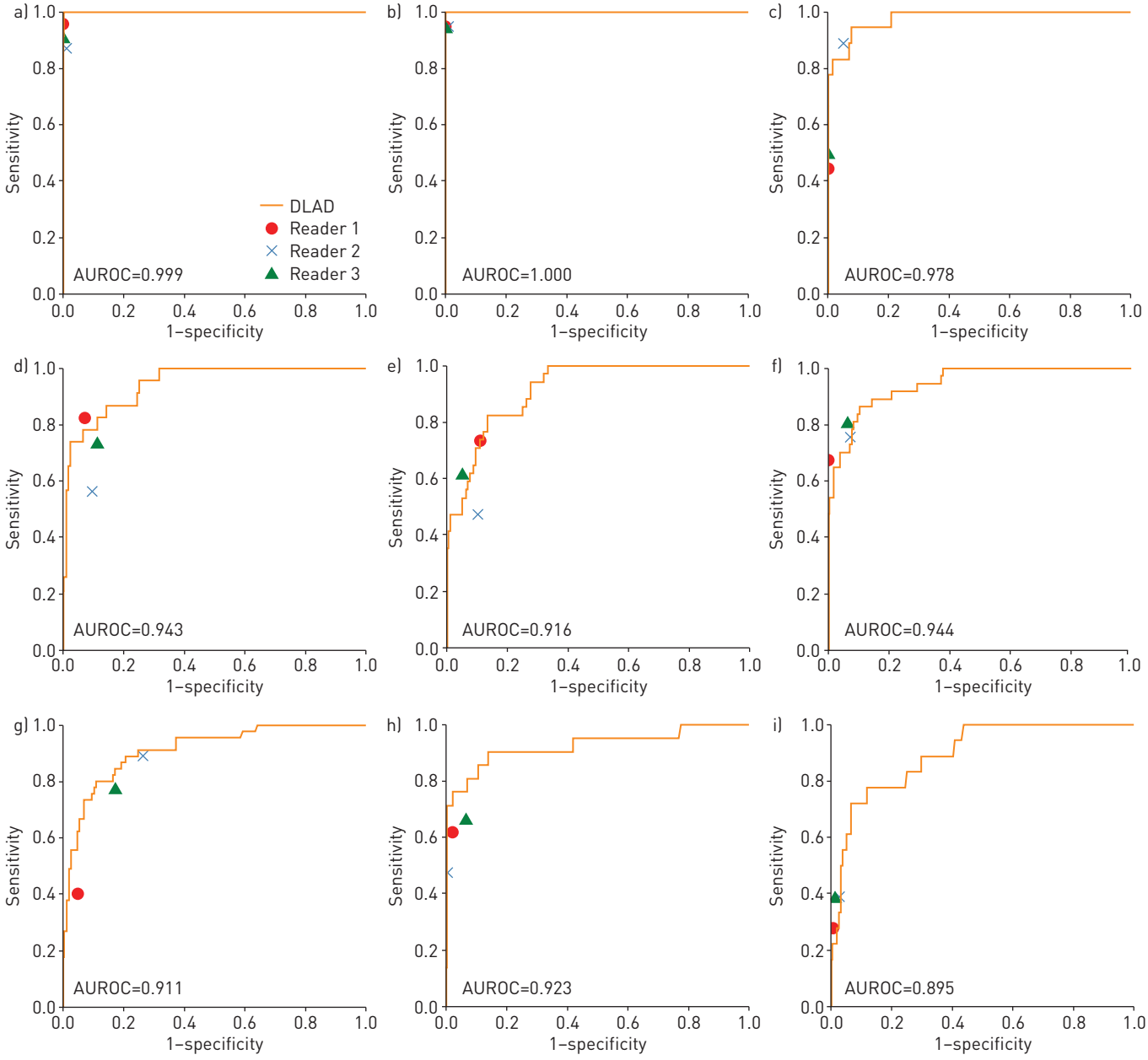

FIGURE 3 Results of DLAD-10 and three thoracic radiologists for the Seoul National University Hospital external validation dataset: the area under the receiver operating characteristic curve (AUROC) of DLAD-10 and the performance of each radiologist are presented for each abnormality. a) Pneumothorax, b) pneumoperitoneum, c) mediastinal widening, d) nodule, e) consolidation, f) pleural effusion, g) atelectasis or fibrosis, h) calcification and i) cardiomegaly.

(57/60) *versus* 84.4% (152/180); p=0.01) (table 3). However, DLAD-10 was inferior to the pooled radiologists for classifying normal or nonurgent cases (80.0% (64/80) *versus* 88.3% (212/240); p=0.03).

Simulated reading test for emergency department patients

Urgency categorisation accuracy

The performance of urgency categorisation was evaluated in terms of whether the readers detected critical or urgent abnormalities corresponding to specific disease entities (urgency categorisation accuracy) (table 4 and supplementary table E5). Without DLAD-10 (conventional reading), the pooled readers correctly detected only 29.2% (7/24) of critical cases, whereas they detected 70.8% (17/24) of critical cases in the DLAD-10-aided reading session (p=0.03). DLAD-10 detected all critical cases, but some were ignored by readers, particularly for mediastinal widening (supplementary table E5). For urgent cases, DLAD-10-aided reading increased the detection rate (82.7% (258/312)) compared with conventional reading (78.2% (244/

TABLE 4 Comparison of the urgency categorisation accuracy between two reading sessions in the simulated reading test

Disease of patient	Conventional reading session				DLAD-10-aided reading session				p-value*
	Nonurgent n	Urgent n	Critical n	Accuracy %	Nonurgent n	Urgent n	Critical n	Accuracy %	
Critical (n=4)	13	4	7	29.2	6	1	17	70.8	0.006*
Pneumothorax (n=2)	3	2	7	58.3	1	1	10	83.3	0.38
Pneumoperitoneum (n=1)	6	0	0	0.0	0	0	6	100.0	
Aortic dissection (n=1)	4	2	0	0.0	5	0	1	16.7	1.00
Urgent (n=52)	68	244	0	78.2	50	258	4	82.7	0.04*
Pneumonia (n=20)	24	96	0	80.0	22	94	4	78.3	0.81
Pulmonary oedema (n=6)	1	35	0	97.2	1	35	0	97.2	1.00
Active tuberculosis (n=4)	8	16	0	66.7	3	21	0	87.5	0.13
ILD (n=3)	0	18	0	100.0	0	18	0	100.0	
Nodule (n=10)	18	42	0	70.0	11	49	0	81.7	0.04*
Pleural effusion (n=7)	7	35	0	83.3	6	36	0	85.7	1.00
Mediastinal mass (n=1)	4	2	0	33.3	1	5	0	83.3	0.25
Rib fracture (n=1)	6	0	0	0.0	6	0	0	0.0	
Nonurgent/normal (n=146)	801	72	3	91.4	822	54	0	93.8	0.03*

ILD: interstitial lung disease. p-values were calculated using the McNemar test. *: p<0.05.

312); p=0.04). The performance increment was the steepest for lung nodules/masses (81.7% (49/60) *versus* 70.0% (42/60); p=0.04). Interestingly, the categorisation accuracy for nonurgent/normal cases also improved with DLAD-10 assistance (93.8% (822/876) *versus* 91.4% (801/876); p=0.03). Examples are shown in figure 2b and c.

Time-to-report

In the conventional reading session, the mean±SD time-to-report for critical, urgent and nonurgent/normal categories was 3371.0±1352.5, 2127.1±1468.2 and 2815.4±1475.9 s, respectively. In the DLAD-10-aided reading session, in which chest radiograph prioritisation was done by embedding DLAD-10 results into the PACS worklist, the time-to-report substantially decreased for critical (640.5±466.3 s; p<0.001) and urgent (1840.3±1141.1 s; p=0.002) cases (table 5), while it significantly increased for nonurgent/normal cases (3267.1±1265.7 s; p=0.007).

TABLE 5 Comparison of time-to-report between two reading sessions in the simulated reading test

Disease of patient	Time-to-report s		p-value
	Conventional reading session	DLAD-10-aided reading session	
Critical (n=4)	3371.0±1352.5 (1473–6186)	640.5±466.3 (25–1562)	<0.001
Pneumothorax (n=2)	4305.3±1131.9 (3105–6186)	644.3±577.7 (25–1562)	<0.001
Pneumoperitoneum (n=1)	2898.7±799.3 (2163–3912)	641.5±383.4 (261–1190)	0.001
Aortic dissection (n=1)	1975.0±511.1 (1473–2546)	632.0±345.0 (261–1085)	<0.001
Urgent (n=52)	2127.1±1468.2 (123–6227)	1840.3±1141.1 (44–5722)	0.002
Pneumonia (n=20)	1968.7±1262.2 (123–6090)	1512.8±1142.0 (66–5389)	0.002
Pulmonary oedema (n=6)	768.3±340.9 (154–1471)	1310.2±813.6 (133–2856)	<0.001
Active tuberculosis (n=4)	2950.6±1669.7 (750–6009)	2248.7±844.9 (845–3603)	0.04
ILD (n=3)	2822.9±1162.0 (1419–5090)	1412.5±706.7 (166–2673)	<0.001
Nodule (n=10)	2834.2±1514.7 (645–6227)	2496.8±1049.1 (206–5031)	0.12
Pleural effusion (n=7)	2349.0±1596.0 (268–5667)	2157.0±991.2 (44–4258)	0.47
Mediastinal mass (n=1)	1115.2±255.4 (845–1387)	611.8±364.7 (248–1225)	0.01
Rib fracture (n=1)	454.0±115.8 (318–635)	3663.5±1136.4 (2251–5772)	0.001
Nonurgent/normal (n=146)	2815.4±1475.9 (7–6624)	3267.1±1265.7 (15–5776)	<0.001

Data are presented as mean±SD (range) time taken to report the corresponding chest radiographs since the initialisation of each reading session. ILD: interstitial lung disease. p-values were calculated using the paired t-test.

Interpretation time of each radiograph

The mean±SD interpretation time of the pooled readers decreased in the DLAD-10-aided reading session compared with the conventional reading session (time per chest radiograph 20.5±22.8 *versus* 23.5±23.7 s; p<0.001) and five of the six readers had a shorter mean interpretation time. With DLAD-10 assistance, the pooled readers spent a significantly shorter time for nonurgent/normal cases (13.5±16.5 *versus* 17.9±16.4 s; p<0.001) and a significantly longer interpretation time for critical cases (36.7±24.4 *versus* 23.0±15.2 s; p=0.01) (supplementary table E2).

Discussion

In our study, DLAD-10 successfully detected 10 common abnormalities in two external validation datasets with high AUROCs, ranging from 0.895 to 1.00. On a CT-referenced external validation dataset, DLAD-10 showed better sensitivity than the thoracic radiologists for most abnormalities (eight out of 10). On the simulated reading test for emergency department patients, the pooled readers increased their accuracy for identifying critical and urgent cases when aided with DLAD-10, and had a lower false-positive rate for nonurgent/normal cases. With DLAD-10 assistance, the readers spent a significantly shorter time-to-report for critical and urgent cases. Pooled readers took a shorter interpretation time for nonurgent/normal cases, resulting in an overall decrease in the mean reading time.

DLAD-10 was developed to assist radiologists or physicians in routine clinical practice. The training data of DLAD-10 were curated by radiologists mostly without CT reference, intended to resemble radiologists' performance, resulting in reasonable output for the readers [28, 29]. Another strength of DLAD-10 is that it can localise most thoracic abnormalities with high accuracy. This characteristic of DLAD-10 can be further modified to make an end-to-end algorithm generating a preliminary radiology report from a radiograph, which may drastically reduce radiologists' workload. No previous deep learning algorithms have been capable of covering most clinically relevant abnormalities on chest radiographs with radiologist-level performance. Most previously reported deep learning algorithms for chest radiographs focused on specific tasks [5, 6, 9, 10], had insufficient coverage of abnormalities [7, 8] or showed limited detection performance compared with radiologists [12, 13].

In this study, we integrated DLAD-10 results into a PACS worklist and tested the potential of a deep learning algorithm as a prioritisation tool. We found that rearrangement of chest radiographs by DLAD-10 pre-analysis enabled earlier reporting of critical or urgent chest radiographs. Further prospective studies investigating the turnaround time are needed, but our study is meaningful as a pioneering report showing the potential role of a deep learning algorithm as a prioritisation tool.

There was a substantial difference in the radiologists' performance between the datasets. The detection rate of critical chest radiographs by the thoracic radiologists was 84.4% (152/180) in the external validation test SNUH dataset, while that in the simulated reading test for emergency department patients was 50.0% (4/8) (29.2% (7/24) for pooled six readers). This difference may reflect a discrepancy between an experimentally designed reader test and real clinical situations. The SNUH dataset included 31.6% (60/190) critical chest radiographs, while the simulated reading test dataset contained few critical chest radiographs (2.0% (4/202)). The higher performance gap between DLAD-10 and the radiologists in the simulated reading test suggests that DLAD-10 may have a clinical impact in real-world situations.

DLAD-10 showed lower specificity than the radiologists in both the external validation and simulated reading test. As the threshold values of DLAD-10 for critical and urgent abnormalities were selected to be sensitive, its specificity was inevitably lower than the radiologists. However, DLAD-10 assistance reduced the false-positive rate of the readers for nonurgent/normal cases (6.2% (54/876) *versus* 8.6% (75/876); p=0.03), probably because the false-positive results of DLAD-10 were easy to discard (*e.g.* misclassifying linear atelectasis as consolidation or fibrosis as a nodule).

Further improvements and modifications of DLAD-10 are warranted. Some important abnormalities, including rib/vertebral fractures and central line/tube malposition, were not covered. Furthermore, DLAD-10 did not differentiate diffuse reticular opacities representing ILD from consolidation. Although DLAD-10 successfully detected most ILD cases as diffuse consolidation in the simulated reading test (66.7% (2/3)), differentiation of reticular opacities from consolidation would be beneficial, as the clinical management is different. Additionally, differential diagnosis and interval change evaluations should be included in the next steps.

Our study has some other limitations. First, our validation datasets were retrospectively collected and could have been affected by selection bias. Second, the criteria for urgency classification that we used could be disputed by other researchers. Third, DLAD-10 did not cover lateral images. Last, the worklist rearrangement on PACS based on chest radiograph urgency is a novel feature, which could be unfamiliar

to the readers. Becoming accustomed to this function could contribute to further improvements of efficacy.

In conclusion, our DLAD-10 deep learning algorithm detecting 10 common abnormalities showed excellent performance on chest radiographs, helping radiologists to improve their performance and advance the reporting time for critical and urgent cases.

Acknowledgements: The authors would like to express their appreciation for Hyewon Choi, Seung-Jin Yoo, Sewoo Kim, Seungchul Han, Jihyuk Lee and Yuna Lee from Seoul National University Hospital (Seoul, Republic of Korea) for participating in the simulated reading test. We also appreciate Lunit Inc. (Seoul, Republic of Korea) and Infinitt Healthcare (Seoul, Republic of Korea) for providing technical support for our validation tests.

Author contributions: J.G. Nam: data curation, statistical analysis, manuscript writing; M. Kim: algorithm development, manuscript writing (supporting); J. Park: algorithm development, manuscript writing (supporting); E.J. Hwang: data curation, validation test, manuscript editing; J.H. Lee: data curation, validation test, manuscript editing; J.H. Hong: validation test, manuscript editing; J.M. Goo: supervising, manuscript editing; C.M. Park: study conceptualisation and organisation, supervising, manuscript writing.

Conflict of interest: J.G. Nam reports grants from the National Research Foundation of Korea funded by the Ministry of Science and ICT (NRF-2018R1A5A1060031), and from Seoul National University Hospital Research Fund (03-2019-0190), during the conduct of the study. M. Kim is an employee of Lunit Inc., and was involved in the development of the algorithm and writing the corresponding part of the manuscript, but did not have control over any of the validation data submitted for publication. J. Park is an employee of Lunit Inc., and was involved in the development of the algorithm and writing the corresponding part of the manuscript, but did not have control over any of the validation data submitted for publication. E.J. Hwang has nothing to disclose. J.H. Lee has nothing to disclose. J.H. Hong has nothing to disclose. J.M. Goo has nothing to disclose. C.M. Park reports grants from the National Research Foundation of Korea funded by the Ministry of Science and ICT (NRF-2018R1A5A1060031), and from Seoul National University Hospital Research Fund (03-2019-0190), during the conduct of the study.

Support statement: This work was supported by the National Research Foundation of Korea grant funded by the Ministry of Science and ICT (grant NRF-2018R1A5A1060031) and the Seoul National University Hospital Research Fund (grant 03-2019-0190). Funding information for this article has been deposited with the Crossref Funder Registry.

References

1 Mettler FA Jr, Mahesh M, Bhargavan-Chatfield M, *et al.* Patient exposure from radiologic and nuclear medicine procedures in the United States: procedure volume and effective dose for the period 2006–2016. *Radiology* 2020; 295: 418–427.
2 United Nations Scientific Committee on the Effects of Atomic Radiation. Sources and Effects of Ionizing Radiation. Annex D. New York, United Nations, 2000.
3 White CS, Flukinger T, Jeudy J, *et al.* Use of a computer-aided detection system to detect missed lung cancer at chest radiography. *Radiology* 2009; 252: 273–281.
4 Topol EJ. High-performance medicine: the convergence of human and artificial intelligence. *Nat Med* 2019; 25: 44–56.
5 Lakhani P, Sundaram B. Deep learning at chest radiography: automated classification of pulmonary tuberculosis by using convolutional neural networks. *Radiology* 2017; 284: 574–582.
6 Nam JG, Park S, Hwang EJ, *et al.* Development and validation of deep learning-based automatic detection algorithm for malignant pulmonary nodules on chest radiographs. *Radiology* 2019; 290: 218–228.
7 Hwang EJ, Park S, Jin K-N, *et al.* Development and validation of a deep learning-based automatic detection algorithm for active pulmonary tuberculosis on chest radiographs. *Clin Infect Dis* 2019; 69: 739–747.
8 Park S, Lee SM, Kim N, *et al.* Application of deep learning-based computer-aided detection system: detecting pneumothorax on chest radiograph after biopsy. *Eur Radiol* 2019; 29: 5341–5348.
9 Hwang EJ, Hong JH, Lee KH, *et al.* Deep learning algorithm for surveillance of pneumothorax after lung biopsy: a multicenter diagnostic cohort study. *Eur Radiol* 2020; 30: 3660–3671.
10 Park S, Lee SM, Lee KH, *et al.* Deep learning-based detection system for multiclass lesions on chest radiographs: comparison with observer readings. *Eur Radiol* 2020; 30: 1359–1368.
11 Hwang EJ, Park S, Jin K-N, *et al.* Development and validation of a deep learning-based automated detection algorithm for major thoracic diseases on chest radiographs. *JAMA Netw Open* 2019; 2: e191095.
12 Rajpurkar P, Irvin J, Ball RL, *et al.* Deep learning for chest radiograph diagnosis: a retrospective comparison of the CheXNeXt algorithm to practicing radiologists. *PLoS Med* 2018; 15: e1002686.
13 Rajpurkar P, Irvin J, Zhu K, *et al.* Chexnet: radiologist-level pneumonia detection on chest x-rays with deep learning. *arXiv* 2017; preprint [https://arxiv.org/abs/1711.05225].
14 Ferkol T, Schraufnagel D. The global burden of respiratory disease. *Ann Am Thorac Soc* 2014; 11: 404–406.
15 Hansell DM, Bankier AA, MacMahon H, *et al.* Fleischner Society: glossary of terms for thoracic imaging. *Radiology* 2008; 246: 697–722.
16 Lai V, Tsang WK, Chan WC, *et al.* Diagnostic accuracy of mediastinal width measurement on posteroanterior and anteroposterior chest radiographs in the depiction of acute nontraumatic thoracic aortic dissection. *Emerg Radiol* 2012; 19: 309–315.
17 He K, Zhang X, Ren S, *et al.* Deep residual learning for image recognition. 2016. https://ieeexplore.ieee.org/document/7780459 Date last accessed: 13 November 2020.
18 Kim M, Park J, Na S, *et al.* Learning visual context by comparison. *arXiv* 2020; preprint [https://arxiv.org/abs/2007.07506].

19 Cubuk ED, Zoph B, Mane D, *et al.* Autoaugment: learning augmentation strategies from data. 2019. https://openaccess.thecvf.com/content_CVPR_2019/html/Cubuk_AutoAugment_Learning_Augmentation_Strategies_From_Data_CVPR_2019_paper.html Date last accessed: 13 November 2020.
20 Annarumma M, Withey SJ, Bakewell RJ, *et al.* Automated triaging of adult chest radiographs with deep artificial neural networks. *Radiology* 2019; 291: 196–202.
21 Hemingway H, Shipley M, Christie D, *et al.* Is cardiothoracic ratio in healthy middle aged men an independent predictor of coronary heart disease mortality? Whitehall study 25 year follow up. *BMJ* 1998; 316: 1353–1354.
22 Zaman MJS, Sanders J, Crook AM, *et al.* Cardiothoracic ratio within the "normal" range independently predicts mortality in patients undergoing coronary angiography. *Heart* 2007; 93: 491–494.
23 Bustos A, Pertusa A, Salinas J-M, *et al.* PadChest: a large chest x-ray image dataset with multi-label annotated reports. *Medical Image Analysis* 2020; 66: 101797.
24 Hwang EJ, Nam JG, Lim WH, *et al.* Deep learning for chest radiograph diagnosis in the emergency department. *Radiology* 2019; 293: 573–580.
25 Raven MC, Lowe RA, Maselli J, *et al.* Comparison of presenting complaint vs discharge diagnosis for identifying "nonemergency" emergency department visits. *JAMA* 2013; 309: 1145–1153.
26 Youden WJ. Index for rating diagnostic tests. *Cancer* 1950; 3: 32–35.
27 Pedregosa F, Varoquaux G, Gramfort A, *et al.* Scikit-learn: machine learning in Python. *J Mach Learn Res* 2011; 12: 2825–2830.
28 Majkowska A, Mittal S, Steiner DF, *et al.* Chest radiograph interpretation with deep learning models: assessment with radiologist-adjudicated reference standards and population-adjusted evaluation. *Radiology* 2020; 294: 421–431.
29 McBee MP, Awan OA, Colucci AT, *et al.* Deep learning in radiology. *Acad Radiol* 2018; 25: 1472–1480.