

Supplement

Data curation

The chest radiographs (CRs) with normal or abnormal findings were initially collected and classified in terms of whether they had each of the 10 abnormalities based on radiology reports from routine clinical practice made at Seoul National University Hospital between March 2004 and December 2017. For data curation, all CRs were reviewed once again by at least one of 20 board-certified radiologists (labeling group; 7-14 years of experience in reading CRs). The data curation process was performed in two steps: image-level classification and pixel-level annotation. During the image-level classification, radiologists from the labeling group confirmed whether each CR was categorized correctly for each abnormality. Afterwards, during the pixel-level annotation, they annotated the exact location of each abnormal finding on the CR.

During the data curation process performed by the labeling group, CRs originally designated as normal that actually showed an abnormality and CRs initially classified as abnormal that did not have detectable abnormal findings were excluded from the dataset. Finally, 146,717 CRs (143,768 postero-anterior projection CRs and 2,949 antero-posterior projection CRs; 90,317 normal and 56,400 abnormal CRs) from 108,053 patients (55,394 men and 52,659 women; mean age, 56.1 ± 14.5 years) were used as our development dataset. For each abnormality, 64.4-82.5% of the total dataset had an image-level label, and among the positively labeled images, 32.0-100% had a pixel-level annotation. Overall, 51.5% (29,018/56,400) abnormal CRs had a pixel-level annotation for at least one of the 10 abnormalities. Detailed numbers of the CRs with labeling or annotations for each abnormality are provided in Table E2.

Algorithm development

We used a deep convolutional neural network with a ResNet-34 backbone [1] with pre-trained weights from ImageNet for DLAD-10. The final layer outputs 10 different abnormality-specific channels, each representing the probability for each abnormality. As the indeterminate features of convolutional neural network may result in limitation of receptive fields [2], we inserted an Attend-and-Compare module into the indeterminate layers to increase the detection performance [3]. For data augmentation, we used the combination method of standard ImageNet augmentation techniques [4] and AutoAugment [5], excluding the color-related operations. Additionally, conventional image processing techniques such as brightness and contrast adjustment, blurring, and random cropping were applied. DICOM raw pixel values were standardized into (0, 1) range and resized to have maximum 2,000 pixels for either height or width, without changing aspect ratio. DLAD-10 generates coarse probability maps with multiple channels, where each channel corresponds to each target abnormality. A weakly-supervised localization technique was used, in which the image-level probability scores were max-pooled from the output probability maps. When a ground truth image-level label was available, the loss was computed with image-level probability scores and the label. When a ground truth pixel-level annotation was available, the loss was computed with probability maps and annotation maps. Binary cross entropy was used for both procedures. Probability maps were used to localize possible regions representing the target abnormality. During the inference process, each CR was split into patches, and the network prediction of the image patches was aggregated to create a prediction result for the whole image. We used stochastic gradient descent as the optimizer, with a mini-batch size of 128. We applied a learning rate of 0.01 first, and then decreased it to 0.001 after 30 epochs. The models were trained up to 40 epochs.

1. He K, Zhang X, Ren S, Sun J. Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition; 2016; 2016. p. 770-778.
2. Luo W, Li Y, Urtasun R, Zemel R. Understanding the effective receptive field in deep convolutional neural networks. In: Advances in neural information processing systems; 2016; 2016. p. 4898-4906.
3. Kim M, Park J, Na S, Park CM, Yoo D. Learning Visual Context by Comparison. *arXiv preprint arXiv:200707506* 2020.
4. Szegedy C, Vanhoucke V, Ioffe S, Shlens J, Wojna Z. Rethinking the inception architecture for computer vision. In: Proceedings of the IEEE conference on computer vision and pattern recognition; 2016; 2016. p. 2818-2826.
5. Cubuk ED, Zoph B, Mane D, Vasudevan V, Le QV. Autoaugment: Learning augmentation strategies from data. In: Proceedings of the IEEE conference on computer vision and pattern recognition; 2019; 2019. p. 113-123.

Supplementary Figures and Figure Legends

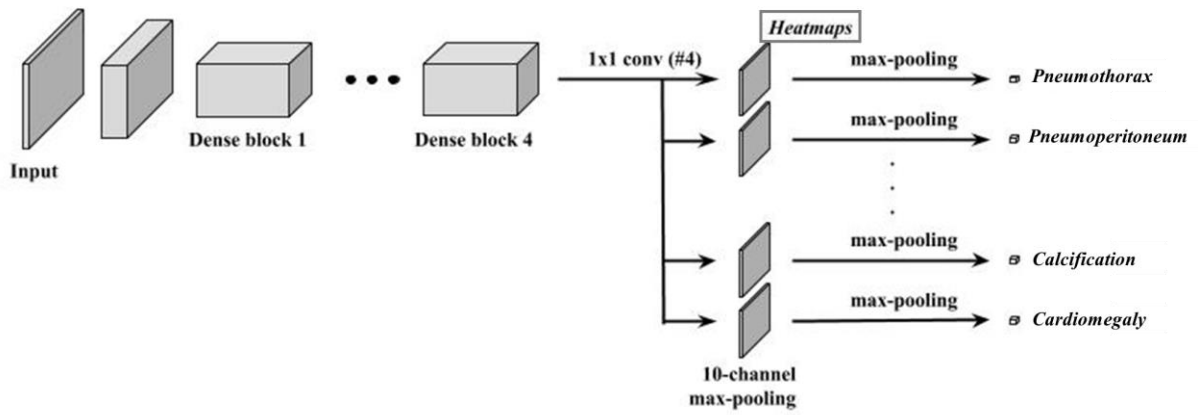
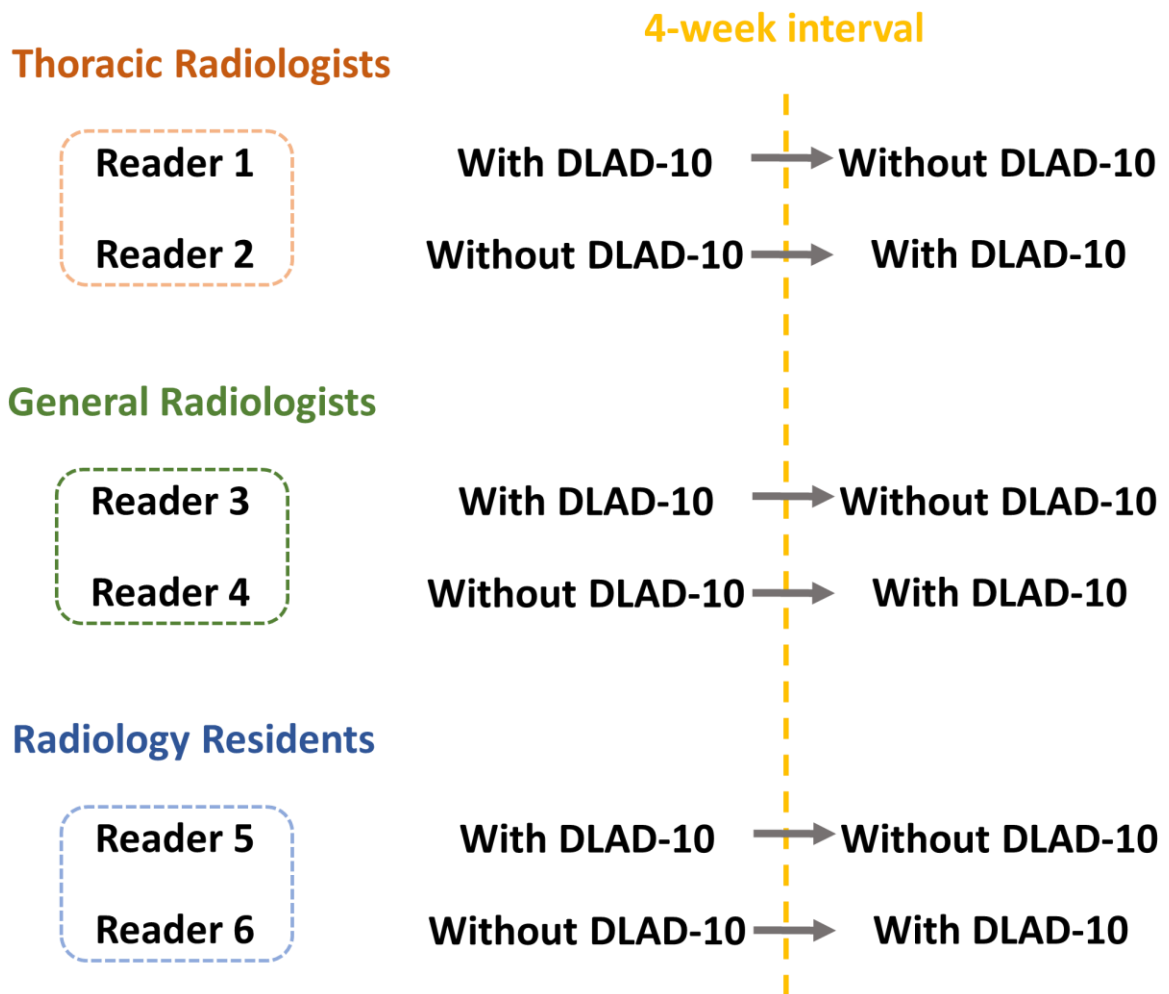


Figure E1: Architecture of the DLAD-10 algorithm.

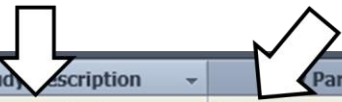
Study Scheme of the Simulation Test



**Displayed DLAD-10 results on PACS
worklist during simulated reading test**

Abnormality type

Abnormality score (%)



	Pri	Requ	Study description	Part	Conclusion	Repo
▶ 28			urgent: consolidation	94.46		Wait
▶ 29			urgent: consolidation	95.79		Wait
▶ 30			urgent: consolidation	19.5		Wait
▶ 31			urgent: consolidation	95.92		Wait
▶ 32			urgent: consolidation	17.29		Wait
▶ 33			urgent: consolidation	98.07		Wait
▶ 34			urgent: consolidation	88.83		Wait
▶ 35			urgent: consolidation	49.82		Wait
▶ 36			urgent: consolidation	19.91		Wait
▶ 37			urgent: consolidation	94.52		Wait
▶ 38			urgent: consolidation	98.21		Wait
▶ 39			urgent: consolidation	76.93		Wait
▶ 40			urgent: consolidation	88.32		Wait
▶ 41			urgent: consolidation	95.54		Wait
▶ 42			urgent: consolidation	49.54		Wait
▶ 43			urgent: consolidation	98.72		Wait
▶ 44			urgent: consolidation	97.53		Wait
▶ 45			urgent: consolidation	98.82		Wait
▶ 46			critical: pneumothorax	96.88		Wait
▶ 47			critical: pneumothorax	39.49		Wait
▶ 48			critical: pneumothorax	93.98		Wait

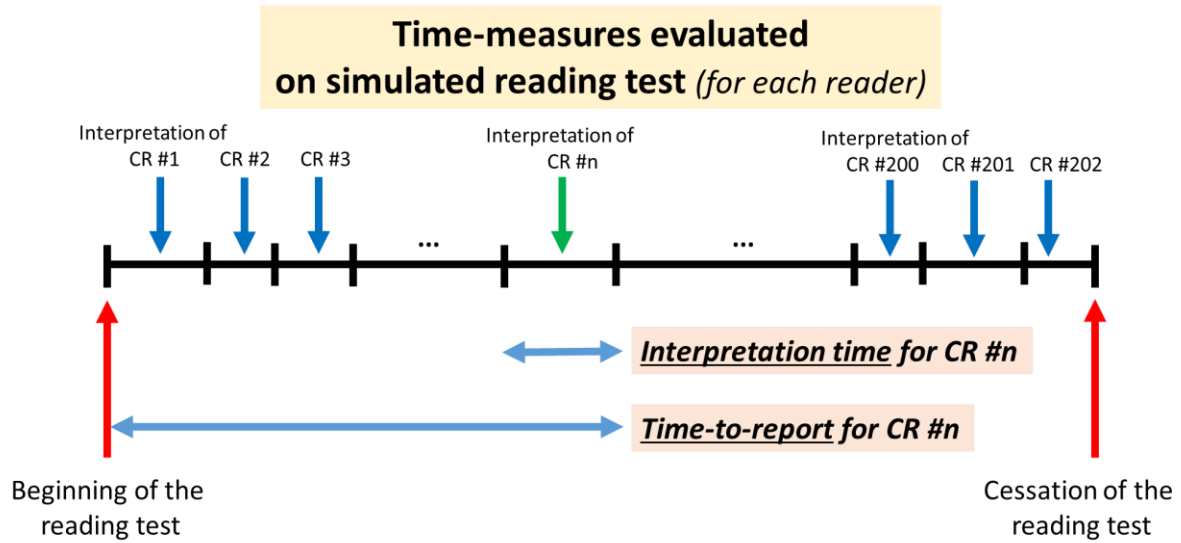


Figure E2: Details of the simulated reading test for emergency department visit patients. (A) Six readers, including two thoracic radiologists, two general radiologists, and two radiology residents performed the simulation test with DLAD-10 (the DLAD-10-aided reading session) and without DLAD-10 (the conventional reading session) at a 4-week interval. Three radiologists conducted the DLAD-10-aided reading session first, while the others performed the conventional reading session first. (B) In the DLAD-10-aided reading session, a list of the radiographs were displayed with abnormality types, urgency, and scores on the worklist. The readers were able to rearrange the order of the radiographs according to those findings. In the DLAD-10-aided reading session, each case was shown as a set of two images: an original image followed by a localized map of abnormalities from DLAD-10. (C) During each session, the time taken for the interpretation of each radiograph by each reader was recorded on the PACS. From these recordings, interpretation time taken for each radiograph and the time taken from the start of the reading session to the interpretation of each radiograph (time-to-report) were calculated.

DLAD-10 = deep learning-based abnormality detection algorithm

Table E1. Manufacturer and technique-related information for the chest radiographs

Manufacturer		Manufacture model name	Technique	Development dataset	Internal validation dataset	External validation dataset*	Simulated reading test dataset
Philips Healthcare	The Netherlands	VS, VT etc.	DR	68,963	1,301	117	
FUJIFILM	Japan	5000, VELOCITY-T	DR	5,924	265		
GE Healthcare	USA	Global 1 Platform	DR	13,447	291	3	
Canon	Japan	CXDI, CN	DR	4,329	165		
Siemens	USA	Fluorospot Compact FD	DR	12,341	221	58	176
Samsung	Korea	DGR-C22M2A/KR	DR	860	18	9	
DongKang	Korea	INNOVISION	DR	4,260	14		

Konica	Japan	CS-7	DR	552	49		
Carestream	USA	DRX- Revolution	DR	4,761	198	3	26
Listem	Korea	DRS	DR	109	1		
Unknown			DR	31,171			
Total				146,717	2,523	190	202

*This dataset refers to the SNUH dataset used for external validation.

DR = digital radiography

Table E2. Development dataset information

	Pneumo thorax	Pneumop eritoneum	Mediastinal widening	Nodule	Consoli dation	Pleural effusion	Atelectasis	Fibrosis	Calcificati on	Cardio megaly	Abnormal
Negative label	112,008	92,194	93,583	102,731	87,940	109,947	90,662	91,375	91,691	91,015	90,317
Positive label	7,254	4,936	471	12,408	13,004	11,089	10,559	3,856	2,840	3,399	56,400
Annotation +	4,324	1,860	471	6,029	11,282	5,972	3,378	3,506	2,455	2,421	29,018
Annotation -	2,930	3,076	0	6,379	1,722	5,117	7,181	350	385	978	27,382
Not labeled	27,455	49,587	52,663	31,578	45,773	25,681	45,496	51,486	52,186	52303	0
Total	146,717	146,717	146,717	146,717	146,717 7	146,717	146,717	146,717	146,717	146,717	146,717

Table E3. Details of the internal and external validation datasets

	Internal validation dataset	SNUH dataset	PadChest open dataset
Total CRs		190	673
Reference standard	Radiologists (labeling group)	Cardiothoracic ratio on CR (for cardiomegaly), Same-day CT (others)	Radiologists (labeling group)
Projection	PA: 2,311 AP: 212	PA: 169 AP: 21	all PA
Pneumothorax	384	23	11
Pneumoperitoneum	152	19	24
Mediastinal widening	86	18	4
Nodule	507	23	32
Consolidation	414	34	119
Pleural effusion	164	37	54
Atelectasis	208	28	58
Fibrosis	218	19	29
Calcification	208	21	36
Cardiomegaly	215	18	90
Normal	747	55	334

Cardiomegaly was only evaluated on postero-anterior images.

CR = chest radiograph, PA = postero-anterior, AP = antero-posterior

Table E4. Details of the dataset used for the simulated reading test for emergency department patients

	Number	Other information
Total	202	Manufacturer
Postero-anterior	176	Fluorospot Compact FD (Siemens, USA)
Antero-posterior	26	DXR-Revolution (Carestream, USA)
Disease entity		Reference standards
Critical	4	
Pneumothorax	2	Same-day CT scan
Pneumomediastinum	1	Same-day CT scan
Acute aortic syndrome	1	Same-day CT scan: aortic dissection
Urgent	52	
Pneumonia	20	Same-day CT scan
Pulmonary edema	6	Same-day CT + clinically supported
Active tuberculosis	4	Same-day CT scan + PCR-confirmed
Interstitial lung disease	3	Same-day CT scan: UIP pattern
Lung nodule	10	Same-day CT scan, confirmed to be lung cancer (n=5), metastasis (n=4), and granuloma (n=1)
Isolated pleural effusion	7	Same-day CT scan
Mediastinal mass	1	Same-day CT scan, confirmed thymic epithelial tumor
Rib fracture	1	Same-day CT scan
Non-urgent/normal	146	Same-day CT scan

CT = computed tomography, *PCR* = polymerase chain reaction, *UIP* = usual interstitial pneumonia

Table E5. Detailed results from the simulated reading test: urgency categorization

Performance of DLAD-10									
	Non-urgent	Urgent	Critical	Accuracy					
Critical	0	0	4	100%					
Pneumothorax	0	0	2	100%					
Pneumoperitoneum	0	0	1	100%					
Aortic dissection	0	0	1	100%					
Urgent	5	34	13	65.4%					
Pneumonia	2	10	8	50.0%					
Pulmonary edema	0	4	2	66.7%					
Active tuberculosis	0	4	0	100%					
ILD	0	2	1	66.7%					
Nodule	1	8	1	80.0%					
Pleural effusion	1	6	0	85.7%					
Mediastinal mass	0	0	1	0.0%					
Rib fracture	1	0	0	0.0%					
Non-urgent/normal	130	7	9	89.0%					
	Conventional reading session					DLAD-10-aided reading session			
Pooled readers (readers 1-6)									
	Non-urgent	Urgent	Critical	Accuracy		Non-urgent	Urgent	Critical	Accuracy
Critical	13	4	7	29.2%		6	1	17	70.8%
Pneumothorax	3	2	7	58.3%		1	1	10	83.3%
Pneumoperitoneum	6	0	0	0.0%		0	0	6	100.0%
Aortic dissection	4	2	0	0.0%		5	0	1	16.7%
Urgent	68	244	0	78.2%		49	260	3	83.3%
Pneumonia	24	96	0	80.0%		23	94	3	78.3%
Pulmonary edema	1	35	0	97.2%		1	35	0	97.2%
Active tuberculosis	8	16	0	66.7%		3	21	0	87.5%
ILD	0	18	0	100.0%		0	18	0	100.0%
Nodule	18	42	0	70.0%		9	51	0	85.0%

Pleural effusion	7	35	0	83.3%		6	36	0	85.7%
Mediastinal mass	4	2	0	33.3%		1	5	0	83.3%
Rib fracture	6	0	0	0.0%		6	0	0	0.0%
Non-urgent/normal	801	72	3	91.4%		817	59	0	93.3%
Thoracic radiologists (readers 1 and 2)									
	Non-urgent	Urgent	Critical	Accuracy		Non-urgent	Urgent	Critical	Accuracy
Critical	3	1	4	50.0%		2	0	6	75.0%
Pneumothorax	0	0	4	100.0%		0	0	4	100.0%
Pneumoperitoneum	2	0	0	0.0%		0	0	2	100.0%
Aortic dissection	1	1	0	0.0%		2	0	0	0.0%
Urgent	16	88	0	84.6%		14	90	0	86.5%
Pneumonia	5	35	0	87.5%		6	34	0	85.0%
Pulmonary edema	0	12	0	100.0%		0	12	0	100.0%
Active tuberculosis	2	6	0	75.0%		1	7	0	87.5%
ILD	0	6	0	100.0%		0	6	0	100.0%
Nodule	3	17	0	85.0%		2	18	0	90.0%
Pleural effusion	2	12	0	85.7%		2	12	0	85.7%
Mediastinal mass	2	0	0	0.0%		1	1	0	50.0%
Rib fracture	2	0	0	0.0%		2	0	0	0.0%
Non-urgent/normal	258	31	3	88.4%		268	24	0	91.8%
General radiologists (readers 3 and 4)									
	Non-urgent	Urgent	Critical	Accuracy		Non-urgent	Urgent	Critical	Accuracy
Critical	5	1	2	25.0%		2	1	5	62.5%
Pneumothorax	1	1	2	50.0%		1	1	2	50.0%
Pneumoperitoneum	2	0	0	0.0%		0	0	2	100.0%
Aortic dissection	2	0	0	0.0%		1	0	1	50.0%
Urgent	36	68	0	65.4%		23	79	2	76.0%
Pneumonia	12	28	0	70.0%		10	28	2	70.0%
Pulmonary edema	1	11	0	91.7%		1	11	0	91.7%
Active tuberculosis	4	4	0	50.0%		2	6	0	75.0%

ILD	0	6	0	100.0%		0	6	0	100.0%
Nodule	12	8	0	40.0%		6	14	0	70.0%
Pleural effusion	3	11	0	78.6%		2	12	0	85.7%
Mediastinal mass	2	0	0	0.0%		0	2	0	100.0%
Rib fracture	2	0	0	0.0%		2	0	0	0.0%
Non-urgent/normal	282	10	0	96.6%		281	11	0	96.2%
Radiology residents (readers 5 and 6)									
	Non-urgent	Urgent	Critical	Accuracy		Non-urgent	Urgent	Critical	Accuracy
Critical	5	2	1	12.5%		2	0	6	75.0%
Pneumothorax	2	1	1	25.0%		0	0	4	100.0%
Pneumoperitoneum	2	0	0	0.0%		0	0	2	100.0%
Aortic dissection	1	1	0	0.0%		2	0	0	0.0%
Urgent	16	88	0	84.6%		13	89	2	85.6%
Pneumonia	7	33	0	82.5%		6	32	2	80.0%
Pulmonary edema	0	12	0	100.0%		0	12	0	100.0%
Active tuberculosis	2	6	0	75.0%		0	8	0	100.0%
ILD	0	6	0	100.0%		0	6	0	100.0%
Nodule	3	17	0	85.0%		3	17	0	95.0%
Pleural effusion	2	12	0	85.7%		2	12	0	85.7%
Mediastinal mass	0	2	0	100.0%		0	2	0	100.0%
Rib fracture	2	0	0	0.0%		2	0	0	0.0%
Non-urgent/normal	261	31	0	89.4%		273	19	0	93.5%

ED = emergency department, *DLAD-10* = deep learning-based abnormality detection algorithm, *ILD* = interstitial lung disease

Table E6. Analysis of interpretation time in the simulated reading test for emergency department patients

		Non-urgent (n=146)	Urgent (n=52)	Critical (n=4)	Total (n=202)
Thoracic radiologists					
Reader 1	Conventional	21.8 ± 24.7 (6-203)	48.8 ± 30.3 (8-129)	24.0 ± 16.1 (5-44)	28.8 ± 28.6 (5-203)
	DLAD-10-aided	19.6 ± 20.4 (4-127)	52.7 ± 31.1 (5-138)	42.5 ± 30.7 (6-81)	28.6 ± 27.8 (4-138)
	*P-values	.40	.40	.43	.94
Reader 2	Conventional	25.8 ± 16.7 (9-90)	52.7 ± 46.3 (12-344)	30.3 ± 24.6 (15-67)	32.8 ± 29.9 (9-344)
	DLAD-10-aided	<u>21.2 ± 21.3</u> (6-147)	45.9 ± 29.1 (9-161)	33.0 ± 25.1 (13-69)	<u>27.8 ± 25.9</u> (6-161)
	*P-values	.03	.29	.61	.02
General radiologists					
Reader 3	Conventional	14.4 ± 10.9 (7-78)	<u>29.1 ± 15.4</u> (8-80)	27.3 ± 5.9 (21-35)	18.5 ± 13.7 (7-80)
	DLAD-10-aided	11.8 ± 15.8 (2-102)	46.5 ± 29.3 (3-134)	53.0 ± 41.1 (23-113)	21.5 ± 25.9 (2-134)
	*P-values	.07	<.001	.31	.06
Reader 4	Conventional	10.8 ± 7.5 (4-58)	34.8 ± 23.9 (7-107)	16.3 ± 16.1 (6-40)	17.1 ± 13.3 (4-107)

	DLAD-10-aided	<u>6.8 ± 7.2</u> (1-47)	<u>23.4 ± 13.6</u> (3-68)	30.0 ± 9.1 (19-40)	<u>11.5 ± 12.0</u> (1-68)
	*P-values	<.001	.001	.18	<.001
Radiology residents					
Reader 5	Conventional	15.5 ± 10.0 (6-52)	27.4 ± 29.8 (8-220)	20.5 ± 14.9 (8-42)	18.6 ± 18.1 (6-220)
	DLAD-10-aided	14.0 ± 14.0 (4-65)	26.9 ± 16.3 (4-94)	25.0 ± 7.8 (16-35)	16.9 ± 14.0 (4-94)
	*P-values	.06	.89	.71	.21
Reader 6	Conventional	19.1 ± 17.5 (7-129)	43.0 ± 35.45 (8-198)	19.5 ± 13.9 (9-40)	25.3 ± 25.5 (1-198)
	DLAD-10-aided	<u>8.32 ± 13.1</u> (2-74)	39.3 ± 23.6 (2-97)	36.5 ± 22.3 (16-68)	<u>16.9 ± 21.5</u> (2-97)
	*P-values	<.001	.38	.07	<.001
Pooled radiologists (readers 1-6)					
	Conventional	17.9 ± 16.4	39.3 ± 32.8	<u>23.0 ± 15.2</u>	23.5 ± 23.7
	DLAD-10-aided	<u>13.5 ± 16.5</u>	39.1 ± 26.8	36.7 ± 24.4	<u>20.5 ± 22.8</u>
	*P-values	<.001	.92	.01	<.001

Data are presented as average ± SD time [sec] (range) taken to interpret each CR. Significant differences in the interpretation time between the conventional and DLAD-10-aided reading sessions according to the paired t-test are underlined.

*P values were calculated from the paired t-test

DLAD-10 = deep learning-based abnormality detection algorithm