Original Article | Thoracic Imaging

eISSN 2005-8330
https://doi.org/10.3348/kjr.2022.0189
Korean J Radiol 2022;23(10):1009-1018

Korean Journal of Radiology
KJR

Artificial Intelligence-Based Identification of Normal Chest Radiographs: A Simulation Study in a Multicenter Health Screening Cohort

Hyunsuk Yoo[1, 2]*, Eun Young Kim[3]*, Hyungjin Kim[2], Ye Ra Choi[4], Moon Young Kim[4], Sung Ho Hwang[5], Young Joong Kim[6], Young Jun Cho[6], Kwang Nam Jin[4]

[1]Lunit Inc, Seoul, Korea; [2]Department of Radiology, Seoul National University College of Medicine, Seoul National University Hospital, Seoul, Korea; [3]Department of Radiology, Gil Medical Center, Gachon University College of Medicine, Incheon, Korea; [4]Department of Radiology, Seoul National University-Seoul Metropolitan Government Boramae Medical Center, Seoul, Korea; [5]Department of Radiology, Korea University Anam Hospital, Seoul, Korea; [6]Department of Radiology, Konyang University Hospital, Konyang University College of Medicine, Daejeon, Korea

Objective: This study aimed to investigate the feasibility of using artificial intelligence (AI) to identify normal chest radiography (CXR) from the worklist of radiologists in a health-screening environment.
Materials and Methods: This retrospective simulation study was conducted using the CXRs of 5887 adults (mean age ± standard deviation, 55.4 ± 11.8 years; male, 4329) from three health screening centers in South Korea using a commercial AI (Lunit INSIGHT CXR3, version 3.5.8.8). Three board-certified thoracic radiologists reviewed CXR images for referable thoracic abnormalities and grouped the images into those with visible referable abnormalities (identified as abnormal by at least one reader) and those with clearly visible referable abnormalities (identified as abnormal by at least two readers). With AI-based simulated exclusion of normal CXR images, the percentages of normal images sorted and abnormal images erroneously removed were analyzed. Additionally, in a random subsample of 480 patients, the ability to identify visible referable abnormalities was compared among AI-unassisted reading (i.e., all images read by human readers without AI), AI-assisted reading (i.e., all images read by human readers with AI assistance as concurrent readers), and reading with AI triage (i.e., human reading of only those rendered abnormal by AI).
Results: Of 5887 CXR images, 405 (6.9%) and 227 (3.9%) contained visible and clearly visible abnormalities, respectively. With AI-based triage, 42.9% (2354/5482) of normal CXR images were removed at the cost of erroneous removal of 3.5% (14/405) and 1.8% (4/227) of CXR images with visible and clearly visible abnormalities, respectively. In the diagnostic performance study, AI triage removed 41.6% (188/452) of normal images from the worklist without missing visible abnormalities and increased the specificity for some readers without decreasing sensitivity.
Conclusion: This study suggests the feasibility of sorting and removing normal CXRs using AI with a tailored cut-off to increase efficiency and reduce the workload of radiologists.
Keywords: *Artificial intelligence; Chest radiograph; Screening; Lung cancer; Normal triage*

INTRODUCTION

Chest radiography (CXR) is one of the most common radiological examinations. CXR images are obtained during the initial work-up for various respiratory or cardiac symptoms and for screening purposes [1]. According to a recent study, the detection rate for tuberculosis was 1.8–5.3 per 100000 persons and that for lung cancer was 9.1–24.4 per 100000 persons on annual health examination using CXRs in Japan [2]. A substantial proportion of CXR findings,

Received: March 20, 2022 **Revised:** August 11, 2022 **Accepted:** August 12, 2022
*These authors contributed equally to this work.
Corresponding author: Kwang Nam Jin, MD, PhD, Department of Radiology, Seoul National University-Seoul Metropolitan Government Boramae Medical Center, 20 Boramae-ro 5-gil, Dongjak-gu, Seoul 07061, Korea.
• E-mail: wlsrhkdska@gmail.com

such as calcified nodules, are normal or insignificant and do not have clinical implications [3-5]. This is problematic because there is a shortage of radiologists relative to the high volume of CXRs, which contributes to excessive workloads, eventually causing burnout [6].

Multiple artificial intelligence (AI) algorithms have recently been developed and have shown excellent standalone performance comparable to that of radiologists in many clinical settings [7-16]. These results suggest that AI may serve as an independent reader for triaging normal images from radiologists' worklists to reduce their burden of reading CXRs. This would be invaluable for improving workflow efficiency, especially in settings with considerable delays in report generation owing to a shortage of radiologists relative to the high volume of diagnostic images [17,18].

Annarumma et al. [19] developed and simulated an AI system for automated triaging of CXRs and suggested that AI-based workflow can reduce the turnaround time for critical and urgent findings. However, this simulation is likely to overestimate the benefits of AI because it is based on an internal dataset similar to the developed model. To analyze the true advantages and weaknesses of an AI-based workflow, it is important to evaluate the system using multiple external clinical cohorts in real-world practice [20]. Furthermore, to the best of our knowledge, no studies have evaluated the efficiency of triage AI for identifying normal CXRs in a health-screening environment. Therefore, this study aimed to investigate the feasibility of using AI to remove normal CXRs from the worklist for CXR reading in a health-screening environment through a simulation.

MATERIALS AND METHODS

In this study, we used the same cohort as that used in a previously published study [21] that evaluated the standalone performance of the AI algorithm. However, the scope of the present study was different as it focused on using the AI algorithm to improve radiologists' efficiency by sorting and removing normal CXRs.

The present study was approved by the Institutional Review Boards of the three participating institutions (IRB No. GBIRB2020-414 for Gil Medical Center, 30-2020-265 for Boramae Medical Center, 2020-11-006 for Konyang University Hospital). All data were de-identified, and the requirement for written informed consent was waived.

Cohort Description

CXR images were consecutively collected from participants who underwent medical check-ups and screening programs at Boramae Hospital (BRMH, Seoul, South Korea), Gachon University Gil Medical Center (GUGMC, Incheon, South Korea), and Konyang University Hospital (KYUH, Daejeon, South Korea) between January and December 2018. Among them, 5887 patients who underwent chest computed tomography (CT) examinations within 1 month of the CXR examination were included in this study as the whole dataset. A random subsample of 480 patients was used as the observer performance test (OPT) dataset. The selection of the patients for this study is shown in Figure 1.

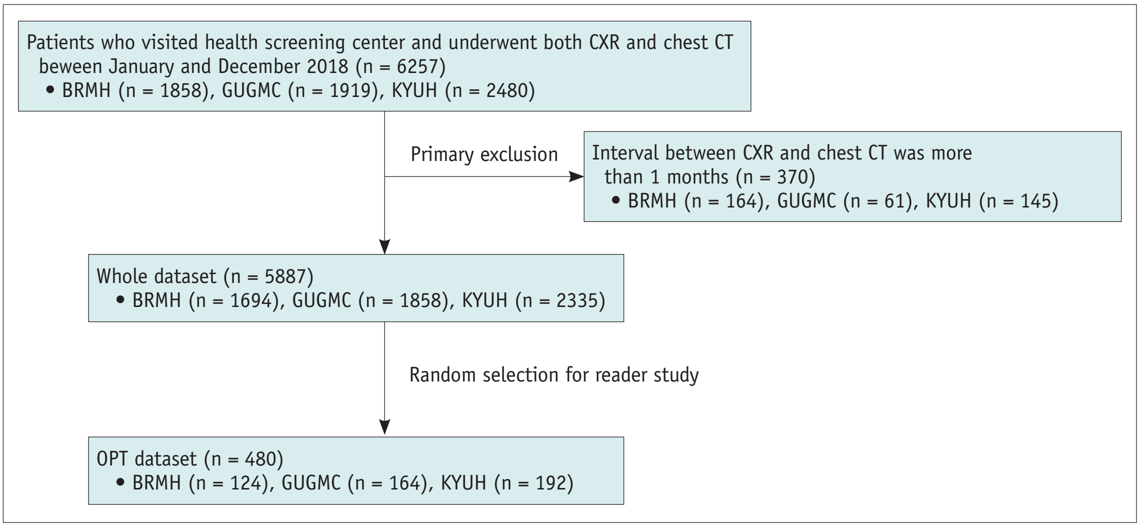
Patients who visited health screening center and underwent both CXR and chest CT beween January and December 2018 (n = 6257)
• BRMH (n = 1858), GUGMC (n = 1919), KYUH (n = 2480)
Primary exclusion
Interval between CXR and chest CT was more than 1 months (n = 370)
• BRMH (n = 164), GUGMC (n = 61), KYUH (n = 145)
Whole dataset (n = 5887)
• BRMH (n = 1694), GUGMC (n = 1858), KYUH (n = 2335)
Random selection for reader study
OPT dataset (n = 480)
• BRMH (n = 124), GUGMC (n = 164), KYUH (n = 192)

Fig. 1. Flow chart for selecting CXRs to be included in the whole dataset and OPT dataset. BRMH = Boramae Hospital, CXR = chest radiography, GUGMC = Gachon University Gil Medical Center, KYUH = Konyang University Hospital, OPT = observer performance test

https://doi.org/10.3348/kjr.2022.0189

Reference Standards for CXR with Referable Thoracic Abnormalities

CXR with referable thoracic abnormalities, defined as abnormalities requiring further diagnostic evaluation or management, was performed in two steps. First, three board-certified radiologists (with 19, 13, and 12 years of experience in thoracic imaging, respectively) identified patients with referable thoracic abnormalities according to the findings on chest CT and clinical information from medical records. They marked the location of the abnormalities and classified the abnormalities into one of 11 categories (nodule or mass, consolidation, atelectasis, pleural effusion, pneumothorax, cardiomegaly, bronchiectasis, interstitial lung disease, pericardial effusion, mediastinal lesion, and others) [22-24]. The final clinical diagnosis was obtained from relevant electronic medical records [25].

Subsequently, CXRs of patients identified in the first step (n = 618, 10.5%) were independently reviewed by three board-certified radiologists (with 7, 10, and 13 years of experience in thoracic radiology, respectively). A CXR image was regarded as positive for visible abnormalities and clearly visible abnormalities if at least one reader and at least two readers, respectively, identified it as abnormal.

AI Algorithm

In our study, we used a commercially available AI solution (Lunit INSIGHT CXR3, version 3.5.8.8) that covers 10 common abnormalities in CXR: atelectasis, calcification, cardiomegaly, consolidation, fibrosis, mediastinal widening, nodule or mass, pleural effusion, pneumoperitoneum, and pneumothorax. The AI algorithm is based on the residual neural network 34 (ResNet-34), which takes the raw pixel map of a DICOM file and generates a probability map and scores (range, 0–100) for each individual lesion [16]. The AI algorithm was trained with 103405 normal CXR and 64651 abnormal CXR images, 39090 of which were annotated by at least one of the 15 board-certified radiologists (with 7–14 years of experience) (Supplementary Fig. 1).

We used all 10 abnormalities for the AI triage (i.e., human reading of only those rendered abnormal by AI), while only three of the 10 abnormalities (nodule or mass, consolidation, and pneumothorax) were used for AI-assisted reading (i.e., AI assistance as a concurrent reader). For AI triage, the maximum score of the 10 abnormalities was used to classify a CXR image as normal or not.

Thresholds for AI Triage and AI-Assisted Reading

To determine the optimal cut-off values for the AI triage, 1280 images with complete contour-level annotations for 10 target abnormalities were randomly sampled from the internal validation dataset, separate from the current simulation study dataset. One radiologist reviewed false-negative cases (15 of 384 positive cases) that had abnormality scores below the commercially recommended cut-off (abnormality score = 15) [16]. Two cut-off values for triaging were selected based on the number of missed lesions: sensitivity-weighted threshold (abnormality score = 5) and triaging efficiency-weighted threshold (abnormality score = 10), at which none of the images with visible and clearly visible abnormalities were missed in the internal validation set.

The threshold for AI-assisted reading was set at 15, which was the commercially recommended cut-off found using Youden criteria in the internal validation set; at this operating point, the AI model had a sensitivity of 97.3% and specificity of 78.2% [16].

OPT

An OPT was conducted using the OPT dataset to assess the performance of readers in detecting visible referable thoracic abnormalities. A total of nine readers (three physicians with 15, 14, and 12 years of clinical experience; three board-certified radiologists with 9, 6, and 2 years of experience in radiology; and three subspecialty-trained thoracic radiologists with 15, 12, and 11 years of experience in thoracic radiology) participated in OPT.

During OPT, readers were asked to find referable thoracic abnormalities and disregard insignificant findings. If the readers answered yes, they were prompted to draw a region of interest for the abnormality and select confidence ratings from 1 (confidence level 0%–20%) to 5 (confidence level 80%–100%). The readers reviewed each CXR twice in two separate sessions. In the first session, all images were read by human readers without assistance from AI. After at least a 4-week wash-out period, all images were read again by the readers using AI results as a concurrent reader. In the second session, the readers were allowed to toggle between the original CXR and AI heatmap.

Statistical Analysis

The percentages of normal CXRs filtered out and CXRs with visible and clearly visible referable abnormalities erroneously removed by the simulated triage by AI were obtained for the entire dataset.

We evaluated the diagnostic performance of the 1st session (AI-unassisted reading) and 2nd session (AI-assisted reading) using the area under the receiver operating characteristic curve (ROC) (see Supplementary Table 1 for corresponding results). We used the paired version of DeLong's test to compare the ROC curves.

We evaluated the effect of simulated AI triage on the sensitivity and specificity of the readers for diagnosing visible referable abnormalities. The sensitivity and specificity results were compared among AI-unassisted reading (i.e., all images read by human readers without AI), reading with AI assistance (i.e., all images read by human readers with AI assistance as concurrent readers), and reading with AI triage (i.e., human reading of only those rendered abnormal by AI) using McNemar's test.

All statistical analyses were performed using MedCalc version 19.5.1 (MedCalc Software) or R software, version

Table 1. Characteristics of Patients Included in the Whole Dataset and OPT Dataset

	Whole Dataset				OPT Dataset (n = 480)
	Individual Institution			Total (n = 5887)	
	BRMH (n = 1694)	GUGMC (n = 1858)	KYUH (n = 2335)		
Age, year	55.6 ± 11.4	53.1 ± 11.1	54.5 ± 12.6	55.4 ± 11.8	53.9 ± 11.3
Sex					
Male	996 (58.8)	1458 (78.5)	1875 (80.3)	4329 (73.5)	364 (75.8)
Female	698 (41.2)	400 (21.5)	460 (19.7)	1558 (26.5)	116 (24.2)
Risk for lung cancer*					
High risk	129 (8.0)	350 (19.4)	314 (13.5)	793 (13.8)	59 (12.5)
Average risk	1485 (92.0)	1450 (80.6)	2009 (86.5)	4944 (86.2)	412 (87.5)
Percentages of abnormal lesions					
Visible thoracic abnormalities	62 (3.7)	182 (9.8)	161 (6.9)	405 (6.9)	28 (5.8)
Clearly visible thoracic abnormalities	37 (2.2)	95 (5.1)	95 (4.1)	227 (3.9)	14 (2.9)
CT-based radiologic diagnosis					
Nodule or mass	44 (2.6)	65 (3.5)	94 (4.0)	203 (3.4)	13 (2.7)
Consolidation	5 (0.3)	3 (0.2)	23 (1.0)	31 (0.5)	4 (0.8)
Atelectasis	20 (1.2)	21 (1.1)	22 (1.0)	63 (1.1)	5 (1.0)
Pleural effusion	1 (0.1)	1 (0.1)	2 (0.1)	4 (0.1)	0 (0.0)
Pneumothorax	0 (0.0)	0 (0.0)	1 (0.0)	1 (0.0)	0 (0.0)
Cardiomegaly	4 (0.2)	10 (0.5)	27 (1.2)	41 (0.7)	1 (0.2)
Bronchiectasis	35 (2.1)	70 (3.8)	21 (0.9)	126 (2.1)	6 (1.3)
Interstitial lung disease	4 (0.2)	6 (0.3)	10 (0.4)	20 (0.3)	1 (0.2)
Pericardial effusion	1 (0.1)	0 (0.0)	1 (0.0)	2 (0.0)	0 (0.0)
Mediastinal lesion	1 (0.1)	0 (0.0)	3 (0.1)	4 (0.1)	1 (0.2)
Others[†]	13 (0.8)	101 (5.5)	48 (2.1)	162 (2.8)	16 (3.3)
Clinical diagnosis					
Lung cancer	5 (0.3)	10 (0.5)	9 (0.4)	24 (0.4)	1 (0.2)
Pulmonary tuberculosis	3 (0.2)	5 (0.3)	8 (0.3)	16 (0.3)	1 (0.2)
Pneumonia	5 (0.3)	4 (0.2)	18 (0.8)	27 (0.5)	4 (0.8)
Interstitial lung disease	2 (0.1)	6 (0.3)	10 (0.4)	18 (0.3)	0 (0.0)
Bronchiectasis	35 (2.1)	90 (4.8)	20 (0.9)	145 (2.5)	7 (1.5)
Pneumothorax	0 (0.0)	0 (0.0)	1 (0.0)	1 (0.0)	0 (0.0)
Rib fracture, benign rib diseases or rib malignancy	1 (0.1)	26 (1.4)	1 (0.0)	28 (0.5)	7 (1.5)
Others[‡]	3 (0.2)	50 (2.7)	34 (1.5)	87 (1.5)	5 (1.0)

Age is in mean ± standard deviation. All other data are number of patients with percentage in parentheses. *Patients aged 55–74 years with a smoking history of 30 pack-years or more are classified as high-risk for lung cancer, [†]Other radiologic findings include emphysema, multiple scattered tiny nodules, rib or vertebral lesions, and lesions in the soft tissue, [‡]Other clinical diseases include pneumoconiosis, emphysema, pulmonary vascular malformations, or other congenital lung lesions, pleural diseases, and vascular diseases. BRMH = Boramae Hospital, GUGMC = Gachon University Gil Medical Center, KYUH = Konyang University Hospital, OPT = observer performance test

3.6.1 (R Foundation for Statistical Computing). For all tests, statistical significance was set at $p < 0.05$.

RESULTS

Cohort Characteristics

Demographic characteristics of the participants are shown in Table 1. Among the entire dataset (n = 5887), 405 (6.9%) and 227 (3.9%) CXRs were positive for visible and clearly visible thoracic abnormalities, respectively. Twenty-four patients (0.4%) had lung cancer, and 16 patients (0.3%) had pulmonary tuberculosis. In the OPT dataset (n = 480), 28 (5.8%) and 14 (2.9%) CXRs contained visible and clearly visible abnormalities, respectively.

Simulation of AI Triage on Whole Dataset

The percentages of normal images that were successfully removed and abnormal images that were incorrectly removed from the radiologist's worklist according to the two different thresholds for AI triage are presented in Table 2. At the sensitivity and triaging efficiency-weighted thresholds, 19.7% and 42.9% of normal images, respectively, were successfully removed from the worklist at the cost of incorrect removal of 1.2% and 3.5% of CXRs with visible abnormalities, respectively, and 0.9% and 1.8% of CXRs with clearly visible abnormalities, respectively (Fig. 2). All visible cases of lung cancer (n = 18) and active tuberculosis (n = 15) were correctly triaged by AI as abnormal at both the sensitivity- and triaging efficiency-weighted thresholds (Table 3).

Table 2. Percentages of Normal and Abnormal CXRs Removed by AI-Based Triage for the Whole Dataset and OPT Dataset

Whole Dataset (n = 5887)			
Threshold for AI	% of Normal CXRs Removed	% of Abnormal CXRs Removed	
		Visible (n = 405)	Clearly Visible (n = 227)
Sensitivity-weighted threshold	19.7 [1082/5482] (18.7–20.8)	1.2 [5/405] (0.2–2.3)	0.9 [2/227] (0.0–2.1)
Triaging efficiency-weighted threshold	42.9 [2354/5482] (41.6–44.3)	3.5 [14/405] (1.7–5.2)	1.8 [4/227] (0.1–3.5)
OPT Dataset (n = 480)			
Threshold for AI	% of Normal CXRs Removed	% of Abnormal CXRs Removed	
		Visible (n = 28)	Clearly Visible (n = 14)
Sensitivity-weighted threshold	21.2 [96/452] (17.5–25.0)	0.0 [0/28] (0.0–0.0)	0.0 [0/14] (0.0–0.0)
Triaging efficiency-weighted threshold	41.6 [188/452] (37.0–46.1)	0.0 [0/28] (0.0–0.0)	0.0 [0/14] (0.0–0.0)

Data are % [raw number] (95% confidence interval). AI = artificial intelligence, CXR = chest radiography, OPT = observer performance test

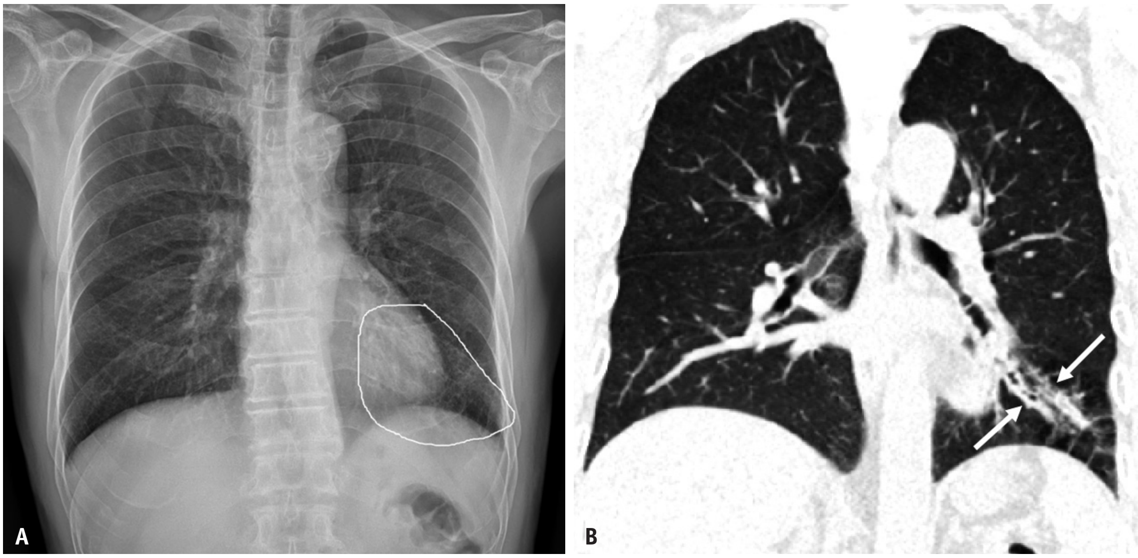

Fig. 2. A 70-year-old male patient who underwent chest radiography and CT in health screening center.
A. Bronchiectasis (circle) was a clearly visible abnormality (two of three radiologists annotated this as an abnormality). The score of triage artificial intelligence was 0.07. Therefore, this abnormality was misclassified as normal by the triaging efficiency-weighted threshold (0.10) but was correctly classified as an abnormality by the sensitivity-weighted threshold (0.05). **B.** Low-dose CT shows bronchiectasis in the left lower lobe (arrows).

Table 3. Percentages of Correct Triage of CXR according to Two Different Thresholds in the Clinical Diagnosis

Final Clinical Diagnosis of Patients with Visible Abnormalities in Their CXR	% of Patients Correctly Triaged at Sensitivity-Weighted Threshold	% of Patients Correctly Triaged at Triaging Efficiency-Weighted Threshold
All visible abnormalities (n = 405)	98.8 [400/405]	96.5 [391/405]
Lung cancer (n = 18)	100.0 [18/18]	100.0 [18/18]
Pulmonary tuberculosis (n = 15)	100.0 [15/15]	100.0 [15/15]
Pneumonia (n = 16)	100.0 [16/16]	93.8 [15/16]
Interstitial lung disease (n = 13)	100.0 [13/13]	100.0 [13/13]
Bronchiectasis (n = 81)	100.0 [81/81]	97.5 [79/81]
Pneumothorax (n = 1)	100.0 [1/1]	100.0 [1/1]
Rib diseases (n = 13)	92.3 [12/13]	92.3 [12/13]
Others* (n = 53)	96.2 [51/53]	94.3 [50/53]
Clinically normal, but radiologic abnormalities† (n = 195)	99.0 [193/195]	96.4 [188/195]

Data are % [raw number]. *Others include pneumoconiosis, emphysema, pulmonary vascular malformations or other congenital lung lesions, pleural diseases, and vascular diseases, †Clinically normal, but radiologic abnormalities included atelectasis or focal fibrosis, benign pulmonary nodules or pulmonary nodule with indeterminate nature, nonspecific parenchymal opacities, and cardiomegaly. CXR = chest radiography

Effect of AI Triage on Reader Performance in OPT Dataset

In the OPT dataset, 41.6% of normal images were successfully removed from the worklist at the triaging efficiency-weighted threshold, and AI triage did not lead to additional missed referable thoracic abnormalities (Table 2). Compared to AI-unassisted reading, the simulated AI triage yielded a significant increase in average specificity for thoracic radiologists (84.6% vs. 88.8%, $p = 0.001$), general radiologists (87.2% vs. 90.6%, $p = 0.005$), and non-radiology physicians (79.1% vs. 84.2%, $p < 0.001$) in the detection of visible thoracic abnormalities using the triaging efficiency-weighted threshold (Table 4). None of the CXRs with visible abnormalities (n = 28) were removed from the worklist, and the sensitivity of all readers did not decrease with the use of triage AI.

Compared with AI-assisted reading (all images read by human readers with AI assistance as a concurrent reader), the simulated triage AI also showed a significant increase in average specificity for thoracic radiologists (83.0% vs. 88.8%, $p < 0.001$) and non-radiology physicians (76.5% vs. 84.2%, $p < 0.001$), but not for general radiologists (89.1% vs. 90.6%, $p = 0.196$). No significant increase in the detection of visible abnormalities was observed (Table 4). The changes in the readers' performance for clearly visible abnormalities are described in Supplementary Table 2.

DISCUSSION

In this study, we simulated how an AI algorithm would improve the efficiency of radiologists in a multicenter health-screening cohort. At the triaging efficiency-weighted operating point, 42.9% of all normal images were successfully removed from the worklist, although 3.5% of CXRs with visible abnormalities and 1.8% of CXRs with clearly visible abnormalities were erroneously removed. However, all visible cases of active tuberculosis and all lung cancers were correctly classified as abnormal by triage AI.

Two of the most active AI research and marketing fields in medical imaging are mammography and CXR. Although several studies have reported the benefit of AI triage of mammography in breast cancer screening in terms of effectiveness of breast cancer detection and reducing radiologists' workload [26-29], no study has evaluated the AI prescreening effect of CXR in the health screening setting. CXR has quite different characteristics from mammography; mammography is a proven method for the screening of breast cancer, but CXR is not considered an efficient mass screening method for lung cancer or tuberculosis in asymptomatic adults in public health. CXR presents with many other abnormal chest findings in addition to cancer or tuberculosis. To our knowledge, this is the first study to analyze the efficiency of triage AI in identifying normal CXR using a multicenter cohort dataset. By simulating an AI-based triaging system, we demonstrated that a significant proportion of normal CXRs can be successfully removed from the worklist, with the potential to improve the reading efficiency of radiologists.

For this system to be successfully implemented, it is important to minimize the number of abnormal images that are incorrectly removed from the radiology worklist because of normal triaging [26,28]. Thus, we used a commercial algorithm that covers a wide variety of radiological

Table 4. Sensitivity and Specificity of the Readers for the Detection of Visible Abnormalities with and without the AI Triage Using the Triaging Efficiency-Weighted Threshold

Readers	Sensitivity					Specificity				
	AI-Unassisted*	AI-Assisted†	AI Triage‡	*P*§	*P*‖	AI-Unassisted*	AI-Assisted†	AI Triage‡	*P*§	*P*‖
Thoracic radiologist 1	64.3 (44.1–81.4)	57.1 (37.2–75.5)	64.3 (44.1–81.4)	1.000	0.585	93.6 (90.9–95.7)	94.5 (91.9–96.4)	94.5 (91.9–96.4)	0.568	1.000
Thoracic radiologist 2	71.4 (51.3–86.8)	75.0 (55.1–89.3)	71.4 (51.3–86.8)	1.000	0.763	92.3 (89.4–94.5)	88.1 (84.7–90.9)	94.7 (92.2–96.6)	0.144	< 0.001
Thoracic radiologist 3	92.9 (76.5–99.1)	92.9 (76.5–99.1)	92.9 (76.5–99.1)	1.000	1.000	67.9 (63.4–72.2)	66.6 (62.0–70.9)	77.2 (73.1–81.0)	0.002	< 0.001
General radiologist 1	78.6 (59.0–91.7)	75.0 (55.1–89.3)	78.6 (59.0–91.7)	1.000	0.752	83.0 (79.2–86.3)	83.4 (79.7–86.7)	88.3 (84.9–91.1)	0.023	0.035
General radiologist 2	75.0 (55.1–89.3)	71.4 (51.3–86.8)	75.0 (55.1–89.3)	1.000	0.763	83.2 (79.4–86.5)	88.5 (85.2–91.3)	86.7 (83.2–89.7)	0.141	0.412
General radiologist 3	53.6 (33.9–72.5)	53.6 (33.9–72.5)	53.6 (33.9–72.5)	1.000	1.000	95.6 (93.2–97.3)	95.4 (93.0–97.1)	96.7 (94.6–98.1)	0.390	0.316
Non-radiology physician 1	75.0 (55.1–89.3)	78.6 (59.0–91.7)	75.0 (55.1–89.3)	1.000	0.752	72.1 (67.7–76.2)	69.2 (64.8–73.5)	81.6 (77.8–85.1)	< 0.001	< 0.001
Non-radiology physician 2	75.0 (55.1–89.3)	85.7 (67.3–96.0)	75.0 (55.1–89.3)	1.000	0.318	81.0 (77.0–84.5)	74.8 (70.5–78.7)	84.7 (81.1–87.9)	0.140	< 0.001
Non-radiology physician 3	67.9 (47.6–84.1)	64.3 (44.1–81.4)	67.9 (47.6–84.1)	1.000	0.078	84.3 (80.6–87.5)	85.6 (82.0–88.7)	86.3 (82.8–89.3)	0.396	0.762
Thoracic radiologist average	76.2 (65.7–84.8)	75.0 (64.4–83.8)	76.2 (65.7–84.8)	1.000	0.857	84.6 (82.6–86.5)	83.0 (80.9–85.0)	88.8 (87.0–90.4)	0.001	< 0.001
General radiologist average	69.0 (58.0–78.7)	66.7 (55.5–76.6)	69.0 (58.0–78.7)	1.000	0.750	87.2 (85.3–89.0)	89.1 (87.2–90.6)	90.6 (88.9–92.1)	0.005	0.196
Non-radiology physician average	72.6 (61.8–81.8)	76.2 (65.7–84.8)	72.6 (61.8–81.8)	1.000	0.594	79.1 (76.9–81.3)	76.5 (74.2–78.8)	84.2 (82.2–86.1)	< 0.001	< 0.001

Data are % with 95% confidence intervals in parentheses. *All images read by human readers without any AI assistance, †All images read by human readers with AI assistance as concurrent reader, ‡Human reading of only those rendered abnormal by triage AI, §AI-triage vs. AI-unassisted reading, ‖AI-triage vs. AI-assisted reading. AI = artificial intelligence abnormalities that may be encountered in clinical practice. In addition, the operating thresholds were carefully tailored to minimize the number of missed abnormalities, with more emphasis on sensitivity rather than specificity. We selected two operating thresholds based on the performance of the AI algorithm in the internal validation set: the sensitivity-weighted threshold and the triaging efficiency-weighted threshold. Of the 405 visible abnormalities, only five were incorrectly classified as normal with the sensitivity-weighted threshold, but 14 were incorrectly classified as normal with the triaging efficiency-weighted threshold. However, 23.2% of the additional normal images were successfully removed from the worklist when compared to triaging at the sensitivity-weighted threshold. This shows that a trade-off exists between increasing classification efficiency and additional abnormality loss and that the operating threshold of the triage AI should be optimized for the clinical scenario in which an AI-based workflow will be deployed.

In the reader study, 41.6% of normal CXRs were successfully removed from the worklist without any additional missed abnormal CXR cases, which increased the specificity without decreasing the sensitivity when triage AI was simulated. In the simulation of triage AI, there was a significant increase in specificity for the detection of visible abnormalities for some readers with various experience levels compared to all images read by them with or without AI assistance as a concurrent reader.

The AI system to identify normal CXR is not perfect and remains controversial. Rare diseases such as pulmonary vascular malformations or other congenital lung diseases, pneumoconiosis, and emphysema could be erroneously classified as normal using the triage AI software. This limitation is mainly due to insufficient AI training in these rare cases. However, radiologists often find it difficult to make a correct diagnosis because of the limited resolution of CXR examinations. Therefore, AI software is recommended as a complementary tool to assist in diagnosis or prioritize reading orders rather than strictly applying the results.

Our study has a few limitations. First, the inclusion criteria included subjects with paired chest CT examinations; therefore, the entire dataset may not represent the whole population undergoing CXR examinations for a check-up and contained an oversampled high-risk population. The disease prevalence in the study dataset is likely higher than that in the actual health-screening setting; thus, the true benefit of the AI-based workflow may have been underestimated. Second, the sample size calculation for the OPT was not designed to evaluate the efficiency of AI triage. The sample of 480 CXRs with 28 positive cases of referable thoracic abnormalities was too small to evaluate the efficiency of triage AI. With the small number of positive cases correctly classified as abnormal using triage AI, the specificity increases without a change in the sensitivity by decreasing false positive calls. Therefore, careful interpretation is required. Third, the workflow was simulated only in the South Korean population, and the results may differ for populations with different ethnic and racial compositions.

In conclusion, this simulation study showed that triage AI effectively identified normal CXR in a health screening cohort, reducing the workload by approximately 40% and increasing specificity in some readers. Further prospective trials are required to validate our findings.

Supplement

The Supplement is available with this article at https://doi.org/10.3348/kjr.2022.0189.

Availability of Data and Material

The datasets generated or analyzed during the study are included in this published article and its supplement.

Conflicts of Interest

Hyunsuk Yoo was an employee of Lunit, South Korea. Hyungjin Kim and Kwang Nam Jin received research fund from Lunit for unrelated works. Other authors had neither financial interests nor conflicts of interest to declare.

Author Contributions

Conceptualization: Hyunsuk Yoo, Kwang Nam Jin. Data curation: Eun Young Kim, Young Jun Cho, Kwang Nam Jin. Formal analysis: Hyunsuk Yoo, Eun Young Kim, Young Jun Cho, Sung Ho Hwang, Kwang Nam Jin. Funding acquisition: Young Jun Cho. Investigation: Hyunsuk Yoo, Eun Young Kim, Kwang Nam Jin. Methodology: Hyunsuk Yoo, Eun Young Kim, Hyungjin Kim, Kwang Nam Jin. Project administration: Young Jun Cho, Kwang Nam Jin. Resources: Eun Young Kim, Young Jun Cho, Kwang Nam Jin. Software: Hyunsuk Yoo, Hyungjin Kim. Supervision: Young Jun Cho, Kwang Nam Jin. Validation: Hyunsuk Yoo, Eun Young Kim. Visualization: Hyunsuk Yoo, Eun Young Kim. Writing—original draft: Hyunsuk Yoo, Eun Young Kim. Writing—review & editing: Hyunsuk Yoo, Eun Young Kim, Hyungjin Kim, Ye Ra Choi, Moon Young Kim, Young Joong Kim, Young Jun Cho, Kwang Nam Jin.

ORCID iDs

Hyunsuk Yoo
https://orcid.org/0000-0003-3255-7439
Eun Young Kim
https://orcid.org/0000-0002-2101-7982
Hyungjin Kim
https://orcid.org/0000-0003-0722-0033
Ye Ra Choi
https://orcid.org/0000-0002-2455-1718
Moon Young Kim
https://orcid.org/0000-0003-3025-0409
Sung Ho Hwang
https://orcid.org/0000-0003-1850-0751
Young Joong Kim
https://orcid.org/0000-0002-7084-0289
Young Jun Cho
https://orcid.org/0000-0002-0632-1011
Kwang Nam Jin
https://orcid.org/0000-0001-5494-9113

Funding Statement

This work was supported by a grant from the Korea Health Industry Development Institute (Grant number: HI19C0847) and by the Gachon University Gil Medical Center (Grant number: FRD2021-11).

Acknowledgments

Lunit provided technical support in building a customized web-based image review tool for this study. We thank Soo Young Kwak for providing advice on statistical interpretation of the data and Sang Hyup Lee for reviewing the images in the internal validation set.

REFERENCES

1. Speets AM, van der Graaf Y, Hoes AW, Kalmijn S, Sachs AP, Rutten MJ, et al. Chest radiography in general practice: indications, diagnostic yield and consequences for patient management. *Br J Gen Pract* 2006;56:574-578
2. Watanabe Y, Nakagawa T, Fukai K, Honda T, Furuya H, Hayashi T, et al. Descriptive study of chest x-ray examination in mandatory annual health examinations at the workplace in Japan. *PLoS One* 2022;17:e0262404
3. Basi SK, Marrie TJ, Huang JQ, Majumdar SR. Patients admitted to hospital with suspected pneumonia and normal chest radiographs: epidemiology, microbiology, and outcomes. *Am J Med* 2004;117:305-311
4. Pinsky PF, Freedman M, Kvale P, Oken M, Caporaso N, Gohagan J. Abnormalities on chest radiograph reported in subjects in a cancer screening trial. *Chest* 2006;130:688-693
5. Joshi R, Patil S, Kalantri S, Schwartzman K, Menzies D, Pai M. Prevalence of abnormal radiological findings in health care workers with latent tuberculosis infection and correlations with T cell immune response. *PLoS One* 2007;2:e805
6. Harolds JA, Parikh JR, Bluth EI, Dutton SC, Recht MP. Burnout of radiologists: frequency, risk factors, and remedies: a report of the ACR Commission on Human Resources. *J Am Coll Radiol* 2016;13:411-416
7. Courtiol P, Maussion C, Moarii M, Pronier E, Pilcer S, Sefta M, et al. Deep learning-based classification of mesothelioma improves prediction of patient outcome. *Nat Med* 2019;25:1519-1525
8. Hwang EJ, Nam JG, Lim WH, Park SJ, Jeong YS, Kang JH, et al. Deep learning for chest radiograph diagnosis in the emergency department. *Radiology* 2019;293:573-580
9. Hwang EJ, Park S, Jin KN, Kim JI, Choi SY, Lee JH, et al. Development and validation of a deep learning–based automated detection algorithm for major thoracic diseases on chest radiographs. *JAMA Netw Open* 2019;2:e191095
10. Jang S, Song H, Shin YJ, Kim J, Kim J, Lee KW, et al. Deep learning-based automatic detection algorithm for reducing overlooked lung cancers on chest radiographs. *Radiology* 2020;296:652-661
11. Khan FA, Majidulla A, Tavaziva G, Nazish A, Abidi SK, Benedetti A, et al. Chest x-ray analysis with deep learning-based software as a triage test for pulmonary tuberculosis: a prospective study of diagnostic accuracy for culture-confirmed disease. *Lancet Digit Health* 2020;2:e573-e581
12. Lee JH, Park S, Hwang EJ, Goo JM, Lee WY, Lee S, et al. Deep learning-based automated detection algorithm for active pulmonary tuberculosis on chest radiographs: diagnostic performance in systematic screening of asymptomatic individuals. *Eur Radiol* 2021;31:1069-1080
13. Lee JH, Sun HY, Park S, Kim H, Hwang EJ, Goo JM, et al. Performance of a deep learning algorithm compared with radiologic interpretation for lung cancer detection on chest radiographs in a health screening population. *Radiology* 2020;297:687-696
14. Nam JG, Park S, Hwang EJ, Lee JH, Jin KN, Lim KY, et al. Development and validation of deep learning–based automatic detection algorithm for malignant pulmonary nodules on chest radiographs. *Radiology* 2019;290:218-228
15. Sim Y, Chung MJ, Kotter E, Yune S, Kim M, Do S, et al. Deep convolutional neural network–based software improves radiologist detection of malignant lung nodules on chest radiographs. *Radiology* 2020;294:199-209
16. Yoo H, Kim KH, Singh R, Digumarthy SR, Kalra MK. Validation of a deep learning algorithm for the detection of malignant pulmonary nodules in chest radiographs. *JAMA Netw Open*

2020;3:e2017135

17. Recht M, Bryan RN. Artificial intelligence: threat or boon to radiologists? *J Am Coll Radiol* 2017;14:1476-1480
18. Thrall JH, Li X, Li Q, Cruz C, Do S, Dreyer K, et al. Artificial intelligence and machine learning in radiology: opportunities, challenges, pitfalls, and criteria for success. *J Am Coll Radiol* 2018;15(3 Pt B):504-508
19. Annarumma M, Withey SJ, Bakewell RJ, Pesce E, Goh V, Montana G. Automated triaging of adult chest radiographs with deep artificial neural networks. *Radiology* 2019;291:272
20. Park SH. Diagnostic case-control versus diagnostic cohort studies for clinical validation of artificial intelligence algorithm performance. *Radiology* 2019;290:272-273
21. Kim EY, Kim YJ, Choi WJ, Lee GP, Choi YR, Jin KN, et al. Performance of a deep-learning algorithm for referable thoracic abnormalities on chest radiographs: a multicenter study of a health screening cohort. *PLoS One* 2021;16:e0246472
22. Hansell DM, Bankier AA, MacMahon H, McLoud TC, Müller NL, Remy J. Fleischner society: glossary of terms for thoracic imaging. *Radiology* 2008;246:697-722
23. Johnson AEW, Pollard TJ, Berkowitz SJ, Greenbaum NR, Lungren MP, Deng CY, et al. MIMIC-CXR, a de-identified publicly available database of chest radiographs with free-text reports. *Sci Data* 2019;6:317
24. Wang X, Peng Y, Lu L, Lu Z, Bagheri M, Summers R. ChestX-ray8: hospital-scale chest X-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR); 2017 Jul 21-26; Honolulu, HI, USA: IEEE; 2017. p.2097-2106
25. Brämer GR. International statistical classification of diseases and related health problems. Tenth revision. *World Health Stat Q* 1988;41:32-36
26. Dembrower K, Wåhlin E, Liu Y, Salim M, Smith K, Lindholm P, et al. Effect of artificial intelligence-based triaging of breast cancer screening mammograms on cancer detection and radiologist workload: a retrospective simulation study. *Lancet Digit Health* 2020;2:e468-e474
27. Lång K, Dustler M, Dahlblom V, Åkesson A, Andersson I, Zackrisson S. Identifying normal mammograms in a large screening population using artificial intelligence. *Eur Radiol* 2021;31:1687-1692
28. Yala A, Schuster T, Miles R, Barzilay R, Lehman C. A deep learning model to triage screening mammograms: a simulation study. *Radiology* 2019;293:38-46
29. Youk JH, Kim EK. Research highlight: artificial intelligence for ruling out negative examinations in screening breast MRI. *Korean J Radiol* 2022;23:153-155

https://doi.org/10.3348/kjr.2022.0189