Radiology

ORIGINAL RESEARCH • EMERGENCY RADIOLOGY

Deep Learning for Chest Radiograph Diagnosis in the Emergency Department

Eui Jin Hwang, MD • Ju Gang Nam, MD • Woo Hyeon Lim, MD • Sae Jin Park, MD • Yun Soo Jeong, MD • Ji Hee Kang, MD • Eun Kyoung Hong, MD • Taek Min Kim, MD • Jin Mo Goo, MD, PhD • Sunggyun Park, PhD • Ki Hwan Kim, MD, PhD • Chang Min Park, MD, PhD

From the Department of Radiology, Seoul National University College of Medicine, 101 Daehak-ro, Jongno-gu, Seoul 03080, Korea (E.J.H., J.G.N., W.H.L., S.J.P., Y.S.J., J.H.K., E.K.H., T.M.K., J.M.G., C.M.P.); and Lunit, Seoul, Korea (S.P., K.H.K.). Received May 30, 2019; revision requested July 29; revision received August 9; accepted September 5. **Address correspondence to** C.M.P. (e-mail: *cmpark.morphius@gmail.com*).

Supported by the Seoul National University Hospital research fund (grants 06-2016-3000 and 03-2019-0190) and the Seoul Research & Business Development Program (grant FI170002).

Conflicts of interest are listed at the end of this article.

See also the editorial by Munera and Infante in this issue.

Radiology 2019; 00:1–8 • https://doi.org/10.1148/radiol.2019191225 • Content codes: ER IN

Background: The performance of a deep learning (DL) algorithm should be validated in actual clinical situations, before its clinical implementation.

Purpose: To evaluate the performance of a DL algorithm for identifying chest radiographs with clinically relevant abnormalities in the emergency department (ED) setting.

Materials and Methods: This single-center retrospective study included consecutive patients who visited the ED and underwent initial chest radiography between January 1 and March 31, 2017. Chest radiographs were analyzed with a commercially available DL algorithm. The performance of the algorithm was evaluated by determining the area under the receiver operating characteristic curve (AUC), sensitivity, and specificity at predefined operating cutoffs (high-sensitivity and high-specificity cutoffs). The sensitivities and specificities of the algorithm were compared with those of the on-call radiology residents who interpreted the chest radiographs in the actual practice by using McNemar tests. If there were discordant findings between the algorithm and resident, the residents reinterpreted the chest radiographs by using the algorithm's output.

Results: A total of 1135 patients (mean age, 53 years ± 18; 582 men) were evaluated. In the identification of abnormal chest radiographs, the algorithm showed an AUC of 0.95 (95% confidence interval [CI]: 0.93, 0.96), a sensitivity of 88.7% (227 of 256 radiographs; 95% CI: 84.1%, 92.3%), and a specificity of 69.6% (612 of 879 radiographs; 95% CI: 66.5%, 72.7%) at the high-sensitivity cutoff and a sensitivity of 81.6% (209 of 256 radiographs; 95% CI: 76.3%, 86.2%) and specificity of 90.3% (794 of 879 radiographs; 95% CI: 88.2%, 92.2%) at the high-specificity cutoff. Radiology residents showed lower sensitivity (65.6% [168 of 256 radiographs; 95% CI: 59.5%, 71.4%], $P < .001$) and higher specificity (98.1% [862 of 879 radiographs; 95% CI: 96.9%, 98.9%], $P < .001$) compared with the algorithm. After reinterpretation of chest radiographs with use of the algorithm's outputs, the sensitivity of the residents improved (73.4% [188 of 256 radiographs; 95% CI: 68.0%, 78.8%], $P = .003$), whereas specificity was reduced (94.3% [829 of 879 radiographs; 95% CI: 92.8%, 95.8%], $P < .001$).

Conclusion: A deep learning algorithm used with emergency department chest radiographs showed diagnostic performance for identifying clinically relevant abnormalities and helped improve the sensitivity of radiology residents' evaluation.

Online supplemental material is available for this article.

In 2015, approximately 137 million patients presented to the emergency department (ED) in the United States (43.3 visits per 100 persons) (1). Respiratory diseases were the second most common primary diagnosis in these patients, accounting for 9.8% of all visits (1). Chest radiography is the first-line examination for the evaluation of various thoracic diseases (2–8). The number of chest radiographs per ED visit increased by 81% between 1994 and 2014, suggesting an increasing dependency on chest radiographs (9).

The interpretation of chest radiographs is a challenging task, requiring experience and expertise. Previous studies have reported suboptimal performance in the interpretation of chest radiographs by ED physicians compared with expert radiologists (10–13). In addition, the American College of Radiology recommends that qualified radiologists be available to interpret all radiographs obtained in the ED (14). However, there is a practical limitation with regard to the full-time availability of expert radiologists, especially for after-hours coverage. In a 2014 survey (15), 73% of the academic radiology departments in the United States did not provide overnight coverage by faculty. Thus, for after-hours ED coverage, a computer-aided detection system for clinically relevant findings on chest radiographs may help improve the quality of radiographic interpretation and overall turnaround time.

Recently, deep learning (DL) algorithms with medical image analysis systems have been evaluated for retinal fundus photographs (16,17), pathologic images (18), and chest

Abbreviations

AUC = area under the receiver operating characteristic curve, CI = confidence interval, DL = deep learning, ED = emergency department, NPV = negative predictive value, PPV = positive predictive value

Summary

Use of a commercially available deep learning algorithm for evaluating chest radiographs resulted in identification of clinically relevant abnormalities on emergency department radiographs and improved the sensitivity of radiology residents' interpretations.

Key Results

- In 1135 consecutive patients who presented to the emergency department, a deep learning (DL) algorithm showed an area under the receiver operating characteristic curve of 0.95 in the identification of chest radiographs with clinically relevant abnormalities.
- For clinically relevant abnormalities, the DL algorithm showed a sensitivity of 88.7% and specificity of 69.6% at the high-sensitivity cutoff and a sensitivity of 81.6% and specificity of 90.3% at the high-specificity cutoff.
- After use of the DL algorithm, the sensitivity of the radiology residents showed a modest improvement (from 65.6% to 73.4%, $P = .003$); however, this was accompanied by a small reduction in specificity (from 98.1% to 94.3%, $P < .001$).

radiographs (19,20). Most of those studies evaluated the efficacy of these algorithms in enriched data sets, which differ from real-world findings in terms of disease prevalence, spectrum of presentation, and population diversity. For DL algorithms to be clinically useful in medical imaging, their performance should be validated in a study sample that reflects clinical applications of this new technology (21).

Thus, the purpose of our study was to evaluate the performance of a DL algorithm in the identification of chest radiographs with clinically relevant abnormalities in the ED setting.

Materials and Methods

Lunit (Seoul, Korea) provided technical support for analyzing chest radiographs with a DL algorithm and obtaining outputs from the algorithm. However, Lunit did not have any role in study design, data collection, statistical analysis, data interpretation, or manuscript preparation. Two authors (S.P. and K.H.K.) are employees of Lunit; however, all data and information were controlled by another author (E.J.H.) without any conflict of interest.

The study was approved by the institutional review board of Seoul National University Hospital, and the requirement to obtain written informed consent was waived.

Patients

We retrospectively included consecutive patients who presented to the ED of a tertiary academic institution between January 1 and March 31, 2017, and underwent chest radiography. Among them, patients for whom previously obtained chest radiographs were available were excluded; patients whose initial chest radiographs were obtained in the ED were included in the study.

DL Algorithm

We used a previously reported, commercially available DL algorithm (Lunit INSIGHT for Chest Radiography, version 4.7.2; Lunit [accessible at *https://insight.lunit.io*]) (22). The algorithm was designed to classify chest radiographs of patients with four major thoracic diseases, including pulmonary malignancy, active pulmonary tuberculosis, pneumonia, and pneumothorax, and was developed with use of 54 221 normal chest radiographs and 35 613 chest radiographs in patients with major thoracic diseases (prevalence, 39.6%). Given an input chest radiograph, the algorithm provided a probability score between 0 and 1 for the presence of any of the target diseases, and a heat map was overlaid on each input radiograph to show the location of abnormalities. For binary classification of positive and negative results, we used two different cut-off values of the probability score that were developed in a prior study (22). The high-sensitivity cutoff was defined as a probability score of 0.16, where the algorithm showed a sensitivity of 95% and a specificity of 75% in the held-out validation data set; the high-specificity cutoff was defined as a probability score of 0.46, where the algorithm showed a sensitivity of 92% and a specificity of 95% in the held-out validation data set (22).

Data Collection

We included both posteroanterior and bedside anteroposterior chest radiographs. All posteroanterior radiographs were obtained with a single dedicated radiography unit (Multix FD; Siemens Healthineers, Erlangen, Germany), and all anteroposterior radiographs were obtained with a single bedside unit (DRX-Revolution; Carestream Health, Rochester, NY). In addition to the radiographic image, the following parameters were recorded: patient age, sex, chief complaint for the ED visit, clinical diagnosis, radiology report made by on-call radiology residents in the actual practice, and time between radiograph acquisition and the final report (Table 1).

Definition of Reference Standard and Categorization of Abnormalities

The reference standard for the presence of a clinically relevant abnormality was defined retrospectively and apart from the actual practice. First, a junior and senior thoracic radiologist (E.J.H. and C.M.P., with 8 and 20 years of experience in the interpretation of chest radiographs, respectively) independently reviewed radiographs, medical records (including follow-up records), and results of laboratory and any additional radiologic examinations (eg, chest CT) to determine whether the radiograph showed any abnormality necessitating further diagnostic evaluation or treatment. The relevant abnormalities included not only abnormalities in the lungs but also those in the mediastinum, pleural space, bones, and upper abdomen. In the event of discordant findings between the two radiologists, they reanalyzed the radiographs with all the available clinical information and provided a final reference standard in consensus (Table E1 [online]). The time between initial and consensus review was 4 months for the junior thoracic radiologist and 1 month for the senior thoracic radiologist. Hereafter, the term *abnormal radiographs* indicates radiographs with clinically relevant abnormalities, whereas the term *normal radiographs* indicates those without clinically relevant abnormalities.

The abnormal radiographs were categorized into the following five classes by the junior thoracic radiologist who defined the

Table 1: Summary of Demographic Information

Parameter	All Patients (n = 1135)	Patients with Normal Radiographs (n = 879)	Patients with Abnormal Radiographs (n = 256)	P Value
Age (y)*	55 (28)	53 (28)	64 (25)	<.001†
Men	582 (51)	425 (48)	157 (61)	<.001‡
Posteroanterior radiographs	951 (84)	766 (87)	185 (72)	<.001‡
Presence of respiratory symptoms	160 (14)	55 (6)	105 (41)	<.001‡
Acquisition of chest CT	126 (11)	47 (5)	79 (31)	<.001‡
Time to report (min)*	88 (161.5)	81 (156)	114 (175.25)	.02†
Discrepant findings between thoracic radiologists	88 (7.8)	17 (2)	71 (28)	<.001‡

Note.—Unless otherwise specified, data are numbers of patients, with percentages in parentheses. P values are for comparison of normal and abnormal radiographs.

* Data are medians, with interquartile range in parentheses.

† Obtained with the Mann-Whitney U test.

‡ Obtained with the χ^2 test.

reference standards (E.J.H.): *(a)* focal lung abnormality, *(b)* diffuse lung abnormality, *(c)* mediastinal abnormality, *(d)* pleural abnormality, and *(e)* other abnormality (abnormalities in the bones or upper abdomen). In addition, abnormal radiographs were classified as depicting the algorithm's target diseases (ie, pulmonary malignancy, active pulmonary tuberculosis, pneumonia, and pneumothorax) or not depicting the algorithm's target diseases.

Assessment of Algorithm Performance

After analysis of radiographs with use of the algorithm, the area under the receiver operating characteristic curve (AUC) was obtained on the basis of the output probability scores and predefined reference standards. Sensitivity, specificity, positive predictive value (PPV), and negative predictive value (NPV) were calculated at the two predefined operating cutoffs. Subsequently, the junior thoracic radiologist (E.J.H.) reviewed heat maps provided by the algorithm and determined whether the abnormality had been correctly localized. Finally, the sensitivities were recalculated; positively classified radiographs with incorrect localization were categorized as false-negative findings. Sensitivities calculated from the probability scores alone were defined as "crude sensitivity," and sensitivities calculated after review of the heat maps were defined as "corrected sensitivity" (Fig E1 [online]).

Radiology Report Evaluation and Reinterpretation

The radiology reports for each radiograph interpreted by an on-call radiology resident in the ED (one of seven radiology residents: J.G.N., W.H.L., S.J.P., Y.S.J., J.H.K., E.K.H., and T.M.K., in their 3rd year of training) during the actual practice were reviewed by a junior thoracic radiologist (E.J.H.) and classified as indicating the presence or absence of any clinically relevant abnormality. The sensitivity, specificity, PPV, and NPV of the radiology residents' evaluations were assessed based on the predefined reference standards.

In the case of discordant interpretation between the algorithm and radiology reports, either at the high-sensitivity or the high-specificity cutoff, the residents who initially interpreted each radiograph reinterpreted it, apart from the actual practice, retrospectively. During the repeat interpretation, residents were provided with the radiographic image, the algorithm's output, and a brief description of the patient (age, sex, chief complaint) and the indication for radiography. However, they were blinded to their previous reports. After the repeat interpretation, residents were asked to classify radiographs as normal or abnormal. Finally, the sensitivity, specificity, PPV, and NPV after repeat interpretation were determined.

Performance Assessment in Different Subgroups

To evaluate the consistency of the algorithm's performance in different subgroups, patients were classified according to the following criteria: *(a)* patient age of 55 years or younger versus age older than 55 years, *(b)* men versus women, *(c)* posteroanterior versus anteroposterior radiographs, *(d)* patients with versus patients without respiratory symptoms, and *(e)* radiographs with concordant versus discordant classification by two thoracic radiologists in the definition of reference standards. The performances of the algorithm and of the radiology residents were evaluated for each subgroup.

Statistical Analysis

Software (R version 3.5.1; R Foundation for Statistical Computing, Vienna, Austria) was used for statistical analyses. The AUCs and their 95% confidence intervals (CIs) were calculated with the nonparametric method suggested by DeLong et al (23). Sensitivities and specificities were compared with use of the McNemar test, and comparisons of PPVs and NPVs were performed with the method suggested by Leisenring et al (24). $P < .05$ was considered indicative of a statistically significant difference.

Results

Patient Characteristics

A total of 1135 patients (one radiograph per patient; 582 men, 553 women; mean age ± standard deviation, 53 years ± 18) were included in the study (Table 1); 3116 patients with available previous radiographs were excluded from the study. According to the reference standard, 256 of the 1135 radiographs (22.6%) were classified as abnormal (Table 2).

Table 2: Clinically Relevant Abnormalities on Abnormal Radiographs

Abnormality	No. of Patients (n = 256)	No. of Patients with CT Examinations (n = 79)
Pulmonary parenchymal diseases		
Pneumonia	69 (27)	36
Pulmonary edema	32 (12)	9
Parenchymal infiltration with indeterminate nature	30 (12)	6
Pulmonary nodule or mass with indeterminate nature	18 (7)	0
Pulmonary tuberculosis suspected from radiographs	15 (6)	2
Interstitial lung disease	9 (3)	3
Primary lung cancer	9 (4)	6
Pulmonary metastasis	8 (4)	3
Bacteriologically proven pulmonary tuberculosis	7 (3)	2
Giant bulla	1 (0.4)	0
Pleural diseases		
Pleural effusion without parenchymal abnormality	34 (13)	7
Pneumothorax	7 (3)	2
Mediastinal diseases		
Clinically significant cardiomegaly	4 (2)	0
Acute aortic syndrome	4 (2)	2
Mediastinal mass	2 (1)	1
Other diseases		
Rib fracture without other abnormality	5 (2)	2
Small bowel obstruction	1 (0.4)	0
Scoliosis	1 (0.4)	0

Note.—Numbers in parentheses are percentages.

Comparison of Algorithm and Resident Performance

The algorithm had an AUC of 0.95 (95% CI: 0.93, 0.96) in the identification of abnormal radiographs. At the high-sensitivity cutoff, the crude sensitivity, corrected sensitivity, and specificity were 95.7% (245 of 256 radiographs; 95% CI: 92.4%, 97.8%), 88.7% (227 of 256 radiographs; 95% CI: 84.1%, 92.3%), and 69.6% (612 of 879 radiographs; 95% CI: 66.5%, 72.7%), respectively. PPV and NPV were 47.9% (245 of 512 radiographs; 95% CI: 43.5%, 52.3%) and 98.2% (612 of 623 radiographs; 95% CI: 96.9%, 99.1%), respectively. In the high-specificity cutoff, the crude sensitivity, corrected sensitivity, and specificity were 85.9% (220 of 256 radiographs; 95% CI: 81.1%, 90.0%), 81.6% (209 of 256 radiographs; 95% CI: 73.3%, 86.2%), and 90.3% (794 of 879 radiographs; 95% CI: 88.2%, 92.2%), respectively. PPV and NPV were 72.1% (220 of 305 radiographs; 95% CI: 66.7%, 77.1%) and 95.7% (794 of 830 radiographs; 95% CI: 94.0%, 96.9%), respectively (Fig 1; Tables 3, E2 [online]). The residents' radiology reports showed lower sensitivity (65.6% [168 of 256 reports; 95% CI: 59.5%, 71.4%]) and NPV (90.7% [862 of 950 reports; 95% CI: 88.7%, 92.5%]) but higher specificity (98.1% [862 of 879 reports; 95% CI: 96.9%, 98.9%]) and PPV (90.8% [168 of 185 reports; 95% CI: 85.7%, 94.6%]) compared with the algorithm at both cutoffs ($P < .001$ for all; Tables 3, E3 [online]; Figs 1–5). The median time between radiograph acquisition and the radiology resident's report was 88 minutes. The interval was longer for abnormal radiographs than for normal radiographs (median interval, 114 vs 81 minutes, respectively; $P = .02$, Mann-Whitney U test).

Resident Performance after Reviewing Algorithm's Output

After reinterpretation of the radiographs with the algorithm's outputs, the radiology residents' sensitivity (73.4% [188 of 256 radiographs; 95% CI: 68.0%, 78.8%]; $P = .003$) and NPV (92.4% [829 of 897 radiographs; 95% CI: 90.7%, 94.2%]; $P = .01$) improved when compared with that of initial reports. Conversely, the specificity (94.3% [829 of 879 radiographs; 95% CI: 92.8%, 95.8%]; $P < .001$) and PPV (79.0% [188 of 238 radiographs; 95% CI: 73.8%, 84.2%]; $P < .001$) were reduced after repeat interpretation (Tables 3, E4 [online]).

Sensitivities Varied according to Pathologic Conditions

The sensitivities of the algorithm and radiology residents' evaluations varied for each category of abnormal radiographs (Table 4). Among the different types of abnormalities, corrected sensitivities of the algorithm for focal lung abnormalities (90.4% [85 of 94 radiographs; 95% CI: 82.6%, 95.5%; $P < .001$; high-sensitivity cutoff] and 81.9% [77 of 94 radiographs; 95% CI: 72.6%, 89.1%; $P < .001$; high-specificity cutoff] vs 47.9% [45 of 94 radiographs; 95% CI: 72.6%, 89.1%; radiology residents]) and diffuse lung abnormalities (96.9% [95 of 98 radiographs; 95% CI: 91.3%, 99.4%; $P < .001$; high-sensitivity cutoff] and 88.8% [87 of 98 radiographs; 95% CI: 80.8%, 94.3%; $P = .03$; high-specificity cutoff] vs 76.5% [75 of 98 radiographs; 95% CI: 66.9%, 84.5%; radiology residents]) were higher than those of radiology residents. No differences were observed for pleural abnormalities ($P = .75$ for both high-sensitivity and high-specificity cutoffs) and mediastinal abnormalities ($P = .63$ and $P > .99$ for high-sensitivity and high-specificity cutoffs, re-

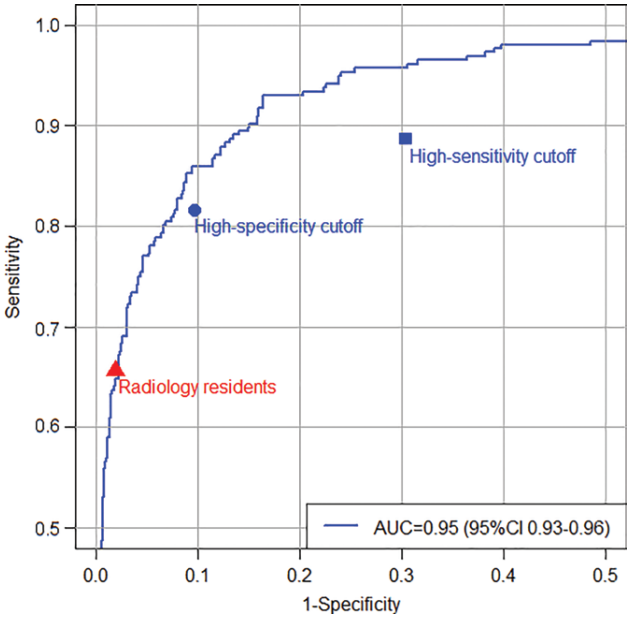


Figure 1: Graph shows performance of algorithm and radiology residents for all consecutive patients. Algorithm showed an area under the receiver operating characteristic curve (AUC) of 0.95 in the classification of chest radiographs with clinically relevant abnormalities. Blue square and circle indicate corrected performance of algorithm at high-sensitivity (corrected sensitivity, 88.7%; specificity, 69.6%) and high-specificity (corrected sensitivity, 81.6%; specificity, 90.3%) cutoffs, respectively, thus reflecting corrected sensitivity. Radiology residents who initially interpreted radiographs in actual clinical practice had a sensitivity of 65.6% and specificity of 98.1%. CI = confidence interval.

spectively) between the algorithm and the radiology residents. Meanwhile, the algorithm showed higher corrected sensitivities in both radiographs with target diseases (93.6% [132 of 141 radiographs; 95% CI: 89.6%, 97.7%; $P < .001$; high-sensitivity cutoff] and 87.9% [124 of 141 radiographs; 95% CI: 82.6%, 93.3%; $P < .001$; high-specificity cutoff] vs 71.6% [101 of 141 radiographs; 95% CI: 64.2%, 79.1%, radiology residents]) and nontarget diseases (82.6% [95 of 115 radiographs; 95% CI: 75.7%, 89.5%; $P < .001$, high-sensitivity cutoff] and 73.9% [85 of 115 radiographs; 95% CI: 65.9%, 81.9%; $P = .01$; high specificity cutoff] vs 58.3% [67 of 115 radiographs; 95% CI: 49.2%, 67.3%; radiology residents]) compared with radiology residents.

Performances in Different Subgroups

The algorithm showed a higher AUC in patients with respiratory symptoms compared with those without respiratory symptoms (AUC, 0.99 [95% CI: 0.97, 1.00] vs 0.93 [95% CI: 0.90, 0.95], respectively; $P < .001$); it also showed a higher AUC in radiographs with concordant classification by thoracic radiologists compared with those with discordant classification (AUC, 0.97 [95% CI: 0.96, 0.98] vs 0.67 [95% CI: 0.52, 0.81], respectively; $P < .001$), without overlapping 95% CIs. In terms of patient age ($P = .08$) and sex ($P = .76$) and radiographic projection ($P = .73$), the algorithm showed no differences in AUC (Fig E2, Tables E5–E9 [online]).

Discussion

In our study, we validated the performance of a commercialized deep learning (DL) algorithm for the classification of chest radiographs with clinically relevant abnormalities in consecutive patients in an emergency department (ED). In the identification of abnormal chest radiographs, the algorithm showed an area under the receiver operating characteristic curve (AUC) of 0.95 (95% confidence interval [CI]: 0.93, 0.96), sensitivity of 88.7% (95% CI: 84.1%, 92.3%), and specificity of 69.6% (95% CI: 66.5%, 72.7%) at the high-sensitivity cutoff and a sensitivity of 81.6% (95% CI: 76.3%, 86.2%) and specificity of 90.3% (95% CI: 88.2%, 92.2%) at the high-specificity cutoff. Radiology residents showed lower sensitivity (65.6% [95% CI: 59.5%, 71.4%], $P < .001$) and higher specificity (98.1% [95% CI: 96.9%, 98.9%], $P < .001$) than the algorithm. After the residents reinterpreted radiographs with discordant classifications by using the algorithm's output, their sensitivity improved (73.4% [95% CI: 68.0%, 78.8%], $P = .003$); however, their specificity was slightly reduced (94.3% [95% CI: 92.8%, 95.8%], $P < .001$).

The most important advantage of our study over previous studies that evaluated the performance of DL algorithms (16,18–20,22) was its application in a clinical setting. Previously, this algorithm exhibited excellent and consistent performance in an enriched data set of normal and abnormal radiographs (22). To determine whether any DL algorithm can be used in clinical practice, its performance should be validated clinically (21). The algorithm showed high efficacy in the classification of radiographs with clinically relevant abnormalities from the ED in this ad hoc retrospective review. This suggests that this DL algorithm is ready for further testing in a controlled real-time ED setting.

In comparison with reports provided by on-call radiology residents, the algorithm showed a different diagnostic performance. In addition, after the reinterpretation of radiographs with discordant classifications between the initial radiology resident's report and the algorithm's classification, the sensitivity and NPV of the residents improved, whereas the specificity and PPV were reduced. Although there was a trade-off between sensitivities and specificities, considering that chest radiographs serve as a screening examination in various acute thoracic diseases (2–8), sensitivity may be a more important measure of performance than specificity—especially in the ED. Therefore, if this algorithm is used as a computer-aided diagnosis tool, we believe it has the potential to improve the interpretation of radiographs in the ED by reducing the number of false-negative interpretations.

Another future application of the algorithm includes use as a screening or triage tool. During the study period, the interval between image acquisition and reporting was paradoxically longer in radiographs with relevant abnormalities. In this regard, the algorithm may improve clinical workflow in the ED by screening radiographs before interpretation by ED physicians and radiologists. The algorithm can inform physicians and radiologists if there is a high probability of relevant disease necessitating timely diagnosis and management.

Among the different types of abnormalities, the algorithm showed an excellent sensitivity for pulmonary parenchymal abnormalities; however, the sensitivity for mediastinal and skeletal abnormalities remained suboptimal. This may be attributed to the algorithm's training, which was limited to the detection of parenchymal abnormalities (pulmonary malignancy, tuberculosis, and pneumonia) and pneumothorax. Considering the low proportion

Table 3: Performances of Algorithm and Radiology Residents

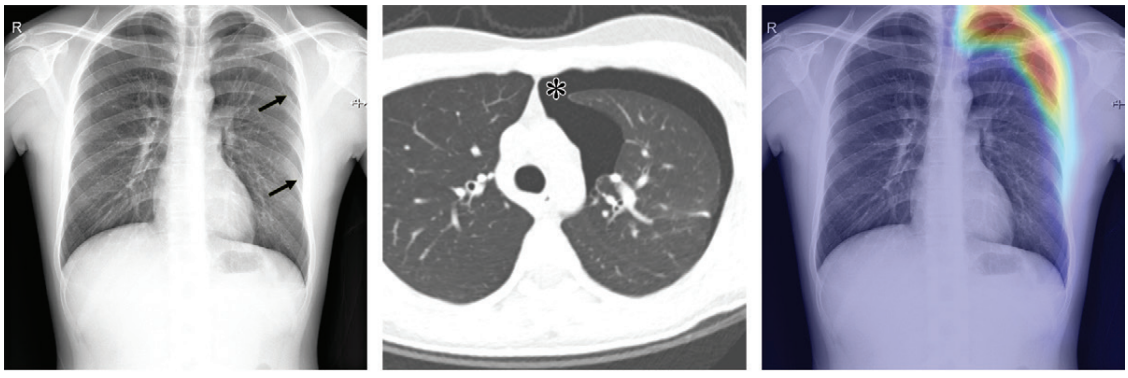
Classifier	Crude Sensitivity (%)	*P* Value	Corrected Sensitivity (%)	*P* Value	Specificity (%)	*P* Value	PPV (%)	*P* Value	NPV (%)	*P* Value
Radiology residents (initial report)	65.6 (168/256) [59.5, 71.4]	NA	NA	NA	98.1 (862/879) [96.9, 98.9]	NA	90.8 (168/185) [85.7, 94.6]	NA	90.7 (862/950) [88.7, 92.5]	NA
Algorithm										
High-sensitivity cutoff	95.7 (245/256) [92.4, 97.8]	<.001	88.7 (227/256) [84.1, 92.3]	<.001*	69.6 (612/879) [66.5, 72.7]	<.001	47.9 (245/512) [43.5, 52.3]	<.001	98.2 (612/623) [96.9, 99.1]	<.001
High-specificity cutoff	85.9 (220/256) [81.1, 90.0]	<.001	81.6 (209/256) [76.3, 86.2]	<.001*	90.3 (794/879) [88.2, 92.2]	<.001	72.1 (220/305) [66.7, 77.1]	<.001	95.7 (794/830) [94.0, 96.9]	<.001
Radiology residents (after reinterpretation)	73.4 (188/256) [68.0, 78.8]	.003	NA	NA	94.3 (829/879) [92.8, 95.8]	<.001	79.0 (188/238) [73.8, 84.2]	<.001	92.4 (829/897) [90.7, 94.2]	.01

Note.—Numbers in parentheses are the raw data (numbers of radiographs). Numbers in brackets are 95% confidence intervals. All *P* values indicate results of comparison with initial reports from radiology residents. NA = not applicable, NPV = negative predictive value, PPV = positive predictive value.

* Comparison with crude sensitivity of initial reports from radiology residents.

a. b. c.

Figure 2: Images in 65-year-old man with pneumonia who presented to the emergency department with fever and cough. **(a)** Chest radiograph shows focal area of increased opacity at juxtaphrenic right basal lung (arrows). Radiograph was initially misinterpreted as normal by the on-call radiology resident. **(b)** Corresponding chest CT scan shows patchy consolidation in right lower lobe of lung (arrows), which is suggestive of pneumonia. **(c)** Heat map from algorithm overlaid on chest radiograph shows that algorithm successfully detected the lesion, with probability score of 0.862. After reviewing the radiograph with the algorithm's output, the radiology resident was able to detect the lesion.



a. b. c.

Figure 3: Images in 19-year-old man with pneumothorax who presented to the emergency department with pleuritic chest pain and dyspnea. **(a)** Chest radiograph shows pneumothorax in left hemithorax (arrows). Radiograph was initially misinterpreted as normal by the on-call radiology resident. **(b)** Corresponding chest CT scan shows left pneumothorax (*). **(c)** Heat map from algorithm overlaid on chest radiograph shows that algorithm successfully detected the pneumothorax, with probability score of 0.974. With the algorithm's output, the radiology resident detected the left pneumothorax.

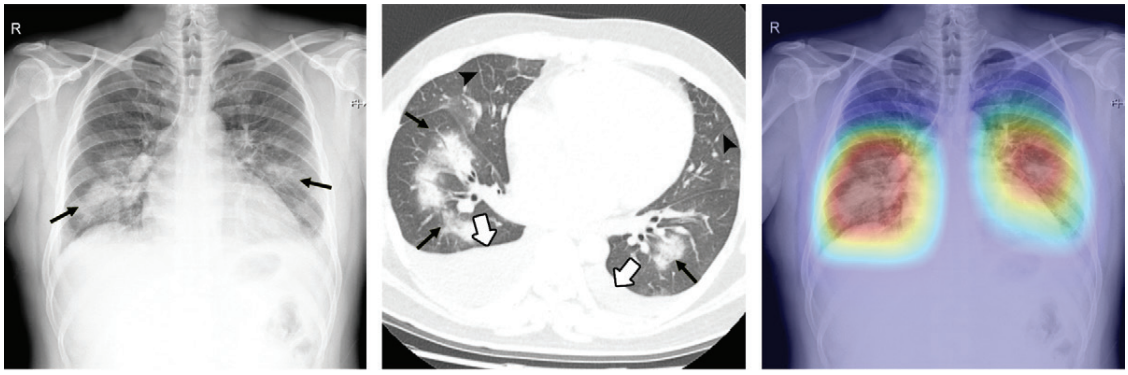


Figure 4: Images in 37-year-old man with pulmonary edema who presented to the emergency department with dyspnea. **(a)** Chest radiograph shows bilateral patchy increased area of consolidation (arrows). **(b)** Corresponding chest CT scan shows bilateral consolidations (black arrows), interlobular septal line thickening (arrowheads), and bilateral pleural effusion (white arrows). **(c)** Heat map from algorithm overlaid on chest radiograph shows that algorithm correctly classified the radiograph as abnormal with both cutoffs, with probability score of 0.999, although pulmonary edema was not included in target diseases of the algorithm.

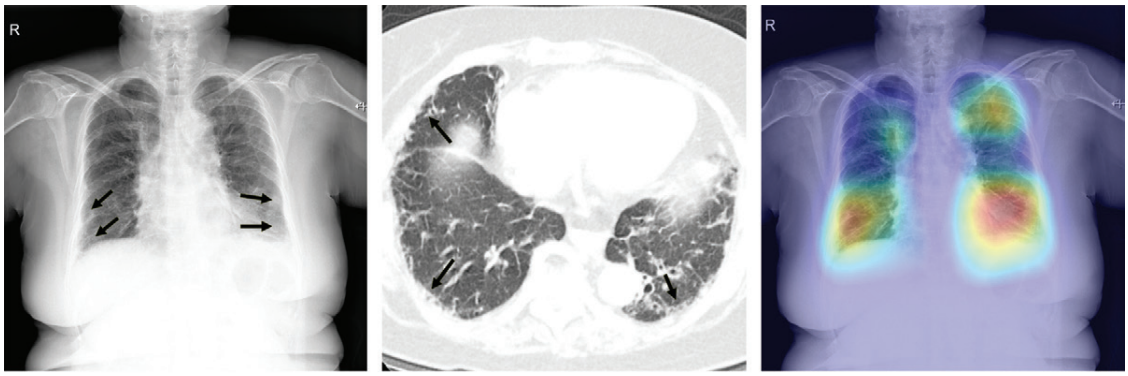


Figure 5: Images in 76-year-old woman with pulmonary fibrosis who presented to the emergency department with cough and dyspnea. **(a)** Chest radiograph shows reticular opacities of both basal and peripheral lungs (arrows). **(b)** Corresponding chest CT scan shows bilateral subpleural reticular opacities (arrows), a finding suggestive of interstitial fibrosis. **(c)** Heat map from algorithm overlaid on chest radiograph shows that algorithm correctly classified the radiograph as abnormal with both cutoffs, with probability score of 0.958, although pulmonary fibrosis was not included in target diseases of the algorithm.

Table 4: Sensitivities of Algorithm and Radiology Residents according to Category of Abnormal Radiographs

		Algorithm: High-Sensitivity Cutoff				Algorithm: High-Specificity Cutoff			
Category	Radiology Residents	Crude Sensitivity	*P* Value	Corrected Sensitivity	*P* Value	Crude Sensitivity	*P* Value	Corrected Sensitivity	*P* Value
Focal lung abnormality	47.9 (45/94) [72.6, 89.1]	95.7 (90/94) [89.5, 98.8]	<.001	90.4 (85/94) [82.6, 95.5]	<.001	86.2 (81/94) [77.5, 92.4]	<.001	81.9 (77/94) [72.6, 89.1]	<.001
Diffuse lung abnormality	76.5 (75/98) [66.9, 84.5]	99.0 (97/98) [94.4, 100]	<.001	96.9 (95/98) [91.3, 99.4]	<.001	90.8 (89/98) [83.3, 95.7]	.01	88.8 (87/98) [80.8, 94.3]	.03
Pleural abnormality	85.5 (47/55) [73.3, 93.5]	92.7 (51/55) [82.4, 98.0]	.29	81.8 (45/55) [69.1, 90.9]	.75	85.5 (47/55) [73.3, 93.5]	>.99	81.8 (45/55) [69.1, 90.9]	.75
Mediastinal abnormality	53.8 (7/13) [25.1, 80.8]	92.3 (12/13) [64.0, 99.8]	.13	69.2 (9/13) [38.6, 90.9]	.63	61.5 (8/13) [31.6, 86.1]	>.99	53.8 (7/13) [25.1, 80.8]	>.99
Other abnormality	40.0 (4/10) [12.2, 73.8]	80.0 (8/10) [44.4, 97.5]	.38	0 (0/10) [0, 30.8]	NA	70.0 (7/10) [34.8, 93.3]	.69	0 (0/10) [0, 30.8]	NA
Target disease	71.6 (101/141) [64.2, 79.1]	97.2 (137/141) [94.4, 99.9]	<.001	93.6 (132/141) [89.6, 97.7]	<.001	90.8 (128/141) [86.0, 95.6]	<.001	87.9 (124/141) [82.6, 93.3]	<.001
Nontarget disease	58.3 (67/115) [49.2, 67.3]	93.9 (108/115) [89.5, 98.3]	<.001	82.6 (95/115) [75.7, 89.5]	<.001	80.0 (92/115) [72.7, 87.3]	<.001	73.9 (85/115) [65.9, 81.9]	.008

Note.—Unless otherwise specified, data are percentages. Numbers in parentheses are raw data (numbers of radiographs). Numbers in brackets are 95% confidence intervals. All *P* values indicate results of comparison with radiology residents. NA = not applicable.

of mediastinal and skeletal abnormalities among abnormal radiographs, these lower sensitivities may not impede the clinical utilization of the algorithm.

In actual clinical situations, the algorithm may encounter various types of diseases, including diseases that the algorithm did not specifically target. The wider disease spectrum could be an important cause of low performance in the clinical setting and a major obstacle to its clinical implementation (21). Although the algorithm in our study targeted four specific diseases, it also showed excellent sensitivities in nontarget diseases and outperformed residents. We believe that the algorithm's sensitivity in nontarget diseases is due to the considerable overlap in their radiographic findings with target diseases (ie, increased opacities in the lung fields). This result suggests the algorithm's usefulness in actual clinical practice, where various target and nontarget diseases may be present.

In the subgroup analyses, the algorithm showed consistent performances regardless of patient age, sex, and radiograph projection, thereby indicating its robustness. The algorithm showed slightly higher performance in patients with respiratory symptoms compared to patients without respiratory symptoms. This may be attributed to the presence of more obvious abnormal findings on the radiographs of patients with respiratory symptoms. In addition, the algorithm showed a lower performance for radiographs in which thoracic radiologists had discordant findings. Abnormalities in these radiographs may have been very subtle and debatable even between expert radiologists. This difficulty in detection would also have affected the performance of the algorithm.

This study has several limitations. First, our study was performed at a single institution and thus it is unknown whether the performance of the algorithm is reproducible in different institutions. Second, because of the retrospective nature of the study design, the effect of the algorithm on a real-time clinical workflow in the ED was not evaluated. Third, we compared the performance of the algorithm with that of on-call radiology residents, rather than experienced radiologists, because primary interpretation by radiology residents is routine practice in our institution. Finally, we evaluated a single frontal chest radiograph per patient in this study. The inability to evaluate lateral radiographs or a series of radiographs in comparison is a weakness of the algorithm.

In conclusion, we tested a deep learning algorithm in emergency department patients during their first visit for the identification of chest radiographs with clinically relevant abnormalities. We found that this algorithm improved the sensitivity of radiology resident trainee interpretations. Further prospective studies are necessary to confirm whether the use of the algorithm can improve clinical workflow and patient outcomes.

Author contributions: Guarantors of integrity of entire study, E.J.H., K.H.K., C.M.P.; study concepts/study design or data acquisition or data analysis/interpretation, all authors; manuscript drafting or manuscript revision for important intellectual content, all authors; approval of final version of submitted manuscript, all authors; agrees to ensure any questions related to the work are appropriately resolved, all authors; literature research, E.J.H., T.M.K., C.M.P.; clinical studies, E.J.H., J.G.N., W.H.L., S.J.P., Y.S.J., E.K.H., T.M.K., J.M.G., K.H.K., C.M.P.; experimental studies, K.H.K.; statistical analysis, E.J.H., S.P.; and manuscript editing, E.J.H., J.G.N., J.M.G., C.M.P.

Disclosures of Conflicts of Interest: **E.J.H.** disclosed no relevant relationships. **J.G.N.** disclosed no relevant relationships. **W.H.L.** disclosed no relevant relationships. **S.J.P.** disclosed no relevant relationships. **Y.S.J.** disclosed no relevant relationships. **J.H.K.** disclosed no relevant relationships. **E.K.H.** disclosed no relevant relationships. **T.M.K.** disclosed no relevant relationships. **J.M.G.** Activities related to the present article: disclosed no relevant relationships. Activities not related to the present article: received research grants from Lunit. Other relationships: disclosed no relevant relationships. **S.P.** Activities related to the present article: is employed by Lunit. Activities not related to the present article: is employed by Lunit; has stock/stock options in Lunit. Other relationships: disclosed no relevant relationships. **K.H.K.** Activities related to the present article: is employed by Lunit. Activities not related to the present article: is employed by Lunit; has stock/stock options in Lunit. Other relationships: disclosed no relevant relationships. **C.M.P.** Activities related to the present article: disclosed no relevant relationships. Activities not related to the present article: received research grants from Lunit. Other relationships: disclosed no relevant relationships.

References

1. Rui P, Kang K. National Hospital Ambulatory Medical Care Survey: 2015 Emergency Department Summary Tables. https://www.cdc.gov/nchs/data/nhamcs/web_tables/2015_ed_web_tables.pdf. Accessed February 19, 2019.
2. Chung JH, Cox CW, Mohammed TL, et al. ACR Appropriateness Criteria® Blunt Chest Trauma. J Am Coll Radiol 2014;11(4):345–351.
3. Expert Panel on Thoracic Imaging, Jokerst C, Chung JH, et al. ACR Appropriateness Criteria® Acute Respiratory Illness in Immunocompetent Patients. J Am Coll Radiol 2018;15(11S):S240–S251.
4. Expert Panel on Thoracic Imaging, McComb BL, Chung JH, et al. ACR Appropriateness Criteria Routine Chest Radiography. J Thorac Imaging 2016;31(2):W13–W15.
5. Expert Panels on Cardiac and Thoracic Imaging, Kirsch J, Brown RKJ, et al. ACR Appropriateness Criteria Acute Chest Pain-Suspected Pulmonary Embolism. J Am Coll Radiol 2017;14(5S):S2–S12.
6. Heitkamp DE, Albin MM, Chung JH, et al. ACR Appropriateness Criteria® Acute Respiratory Illness in Immunocompromised Patients. J Thorac Imaging 2015;30(3):W2–W5.
7. Hoffmann U, Akers SR, Brown RK, et al. ACR Appropriateness Criteria Acute Nonspecific Chest Pain-Low Probability of Coronary Artery Disease. J Am Coll Radiol 2015;12(12 Pt A):1266–1271 [Published correction appears in J Am Coll Radiol 2016;13(2):231.].
8. Ketai LH, Mohammed TL, Kirsch J, et al. ACR Appropriateness Criteria Hemoptysis. J Thorac Imaging 2014;29(3):W19–W22.
9. Chung JH, Duszak R Jr, Hemingway J, Hughes DR, Rosenkrantz AB. Increasing utilization of chest imaging in US emergency departments from 1994 to 2015. J Am Coll Radiol 2019;16(5):674–682.
10. Al Aseri Z. Accuracy of chest radiograph interpretation by emergency physicians. Emerg Radiol 2009;16(2):111–114.
11. Eng J, Mysko WK, Weller GE, et al. Interpretation of emergency department radiographs: a comparison of emergency medicine physicians with radiologists, residents with faculty, and film with digital display. AJR Am J Roentgenol 2000;175(5):1233–1238.
12. Gatt ME, Spectre G, Paltiel O, Hiller N, Stalnikowicz R. Chest radiographs in the emergency department: is the radiologist really necessary? Postgrad Med J 2003;79(930):214–217.
13. Petinaux B, Bhat R, Boniface K, Aristizabal J. Accuracy of radiographic readings in the emergency department. Am J Emerg Med 2011;29(1):18–25.
14. American College of Radiology. ACR practice parameter for radiologist coverage of imaging performed in hospital emergency departments. https://www.acr.org/-/media/ACR/Files/Practice-Parameters/HospER.pdf?la=en. Published 2018. Updated 2018. Accessed February 19, 2019.
15. Sellers A, Hillman BJ, Wintermark M. Survey of after-hours coverage of emergency department imaging studies by US academic radiology departments. J Am Coll Radiol 2014;11(7):725–730.
16. Gulshan V, Peng L, Coram M, et al. Development and validation of a deep learning algorithm for detection of diabetic retinopathy in retinal fundus photographs. JAMA 2016;316(22):2402–2410.
17. Ting DSW, Cheung CY, Lim G, et al. Development and validation of a deep learning system for diabetic retinopathy and related eye diseases using retinal images from multiethnic populations with diabetes. JAMA 2017;318(22):2211–2223.
18. Ehteshami Bejnordi B, Veta M, Johannes van Diest P, et al. Diagnostic assessment of deep learning algorithms for detection of lymph node metastases in women with breast cancer. JAMA 2017;318(22):2199–2210.
19. Hwang EJ, Park S, Jin KN, 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(5):739–747.
20. 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(1):218–228.
21. Park SH, Han K. Methodologic guide for evaluating clinical performance and effect of artificial intelligence technology for medical diagnosis and prediction. Radiology 2018;286(3):800–809.
22. Hwang EJ, Park S, Jin KN, 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(3):e191095.
23. DeLong ER, DeLong DM, Clarke-Pearson DL. Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach. Biometrics 1988;44(3):837–845.
24. Leisenring W, Alonzo T, Pepe MS. Comparisons of predictive values of binary medical diagnostic tests for paired designs. Biometrics 2000;56(2):345–351.