

Research article

Diagnostic agreement between a commercial AI system and radiologists in interpreting checkup chest radiographs

Nattawut Jiraaram*, Piyaporn Kitpraphasiri, Panida Wongsansuk,
Teeraphat Pataraphitakpong, Nithipat Phinklao

*Diagnostic Imaging and Image-Guided Therapy Centre for the People,
Chulabhorn Hospital, Chulabhorn Royal Academy*

*Corresponding author, e-mail: nattawut.jir@cra.ac.th

Received: 11 December 2022; Revised: 22 April 2024; Accepted: 29 April 2024

Abstract

Background: Chest radiography is widely used to screen for thoracic diseases. Recent artificial intelligence (AI) systems have demonstrated outstanding standalone performance in diagnostic tasks concerning chest radiographs (CXRs), often comparable to that of radiologists. However, few studies have evaluated AI performance in real-world clinical practice settings, especially with regard to checkup examinations and diagnostic agreement with humans. **Objective:** To evaluate the diagnostic agreement between AI and radiologists in evaluating checkup CXRs. **Methods:** An AI system and radiologists independently evaluated 500 checkup CXRs from a retrospective review period. We then quantified their diagnostic agreement. **Results:** When our analysis was restricted to lesions identified by the AI (AI-target lesions), we found fair agreement. If we regard the diagnoses of the radiologists as ground truth, the AI produced false-negative and false-positive rates of 8.0% and 8.7%, respectively. When extended to include lesions that had not been identified by the AI (both AI-target and AI-non-target lesions), our analysis showed a reduced agreement (76.0%) and an increased false-negative rate (16.7%). The AI demonstrated low sensitivity (26.5–36.8) but high specificity (90.1–90.5), with a significant number of AI-non-target lesions. **Conclusions:** Our results demonstrate fair to slight agreement between automatic AI diagnoses and the assessments of radiologists for checkup CXRs. Radiologists often reported numerous AI-non-target lesions. Despite certain limitations, our findings suggest that AI may play a valuable role as a diagnostic aid to enhance radiological evaluations. Further research is warranted to comprehensively assess AI performance in broader clinical contexts.

Keywords: artificial intelligence (AI) system, checkup chest radiograph (CXR), diagnostic agreement

Introduction

Chest radiography is a widely available and cost-effective imaging method characterized by low radiation exposure and reasonable diagnostic efficacy.¹ It serves as a primary screening tool for various thoracic pathologies, including lung cancer, owing to its accessibility. A recent study examining annual health checkups in Japan reported detection rates of 1.8–5.3 and 9.1–24.4 per 100,000 persons for tuberculosis and lung cancer, respectively.² However, unlike the findings of the National Lung Screening Trial, the efficacy of chest radiography in detecting early-stage lung cancer or reducing mortality remains uncertain. This uncertainty has led to ongoing debate regarding the utility of chest radiography for lung

cancer screening.¹ Nevertheless, chest radiographs (CXR) are routinely incorporated into annual health checkups by many healthcare providers. At the Diagnostic Imaging and Image-Guided Therapy Centre for the People, over 6,000 CXRs are performed annually as part of routine checkup examinations.

Artificial intelligence (AI) systems represent a domain of computer science aimed at emulating human cognitive functions, offering the capacity to analyze and address complex problems. Radiological applications of AI, and more specifically machine learning, have become a reality in current clinical practice. These advancements have extended to thoracic imaging, encompassing tasks such as lung nodule evaluation, tuberculosis detection, pneumonia diagnosis, and the assessment of diffuse lung diseases.³ Recent developments have yielded several AI systems that demonstrate standalone performance comparable to that of radiologists across various clinical scenarios.²

The AI-Rad Companion Chest X-ray (Siemens Healthcare GmbH, Erlangen, Germany⁴) is a diagnostic aid for radiologists that identifies and highlights prespecified radiographic findings using AI. It is intended to be used in conjunction with interpretations of original images by radiologists before arriving at a final diagnosis. Overall, the performance of the AI-Rad Companion Chest X-ray aligns closely with the average performance of radiologists participating in controlled studies.⁴

Given the extensive production of CXRs coupled with a scarcity of qualified readers, the timely assessment of radiographs poses challenges in both developing and developed nations.⁵ AI integration holds significant promise for enhancing workflow efficiency, particularly in environments where reporting delays are prevalent owing to a pronounced shortage of radiologists relative to the substantial volume of diagnostic images.²

Nevertheless, the performance of AI applications in real-world clinical practice, particularly concerning routine checkup examinations and diagnostic agreement with radiologist, has only been investigated in a limited number of studies.¹⁷ Differences in the study population and country may impact the

prevalence of thoracic diseases and the variability in abnormalities, which can affect the diagnostic performance of AI. To address this gap in understanding, our study assesses the agreement between AI and radiologists in diagnosing and evaluating checkup CXRs obtained from a single-center health screening cohort. Additionally, within this clinical context, we investigate the prevalence of lesions missed by the AI algorithm (non-target findings).

Methods

We aggregated data from a retroactive 3-month period at a single center in Thailand. The research methodology received approval from the Ethics Committee of the Chulabhorn Research Institute, Bangkok, Thailand. All data were anonymized to ensure privacy and confidentiality.

Study population:

The study included all patients who went through the checkup chest CXRs reporting system and the Picture Archiving and Communication System (PACS). From the pool of 1,292 eligible patients, a random sample of 500 patients was selected. Patients whose CXRs were not interpretable by the AI were excluded. Complete clinical and demographic information, including age, sex, and underlying diseases, was successfully retrieved for all included patients.

Radiographic assessment by radiologists:

Three board-certified radiologists, one certified by the subspecialty board of Advanced Diagnostic Body Imaging, another by the board of Pediatric Diagnostic Imaging, and the third by the board of Diagnostic Radiology, with 4 years, 4 years, and 1 year of experience, respectively, conducted independent evaluations of all abnormalities detected in the given images. Clinical data, comparison images, and AI results were blinded to ensure unbiased assessment. Institutional equipment, including reporting stations and workstations, was utilized to carry out the evaluations. Consensus readings were conducted by two out of the three radiologists for each CXR. In cases of disagreement, a majority vote from all three radiologists was employed to

determine the presence or absence of a given abnormality.

Radiographic interpretation was categorized as follows:

- Findings from AI-target lesions: specifying category and location
- Findings from AI-non-target lesions: specifying lesion and location
- Diagnosis based on AI-target lesions only: determining whether the lesion was normal or abnormal
- Diagnosis based on both AI-target and AI-non-target lesions: determining whether the lesion was normal or abnormal

The AI-target lesions were categorized into 5 distinct groups: pulmonary lesions, atelectasis, pneumothorax, consolidation, and pleural effusion. Lesions identified outside the scope of AI-target areas were deemed AI-non-target lesions and further divided into two subgroups: 1) lesions categorized as normal lesions, variants, or post-operative changes and 2) lesions classified as abnormal or indicative of potential disease.

Abnormal lesions or those suggesting possible disease within AI-non-target areas, as well as all findings within AI-target areas, were diagnosed as abnormal. Conversely, lesions classified as normal, variations, or post-operative changes within AI-non-target areas were diagnosed as normal.

Lesion locations were classified into six distinct categories: right/left upper-lung zone (RUL/LUL), right/left mid-lung zone (RML/LML), and right/left lower-lung zone (RLL/LLL).

Radiographic assessment via AI:

We employed a commercially available AI-Rad Companion Chest X-ray VA30B to analyze CXRs from the study population.

Radiographic interpretation was categorized as follows:

- Findings from AI-target lesions: specifying category and location
- Diagnosis: determining if the lesion was normal or abnormal

Because, by definition, AI-non-target lesions cannot be assessed by the AI algorithm, the AI diagnosis was the same whether it was based on AI-target

lesions only or on all lesions (both AI-target and AI-non-target lesions).

Lesion locations were categorized using the same conventions as those adopted by radiologists (described previously).

Statistical analysis:

Data preparation and analysis were conducted using PASW Statistics 18.0 (SPSS Inc., Chicago IL, USA). Continuous data are presented as medians with interquartile ranges, while categorical data are expressed as proportions and percentages.

To assess diagnostic agreement, we calculated Cohen's kappa and utilized the following classification system: poor (<0.0), slight (0.01-0.20), fair (0.21-0.40), moderate (0.41-0.60), substantial (0.61-0.80), and almost perfect (0.81-1.00).⁶

Statistical significance was determined using a p-value threshold of 0.05.

Results

Baseline characteristics, findings, and lesion types:

The AI platform was unable to interpret 51 out of the initially selected 500 checkup radiographs. Consequently, these cases were excluded, resulting in a total of 449 radiographs included in the study. The median age of the patients in the study was 46 years (interquartile range: 36–55), with females accounting for 62.8% of the sample. Among the included patients, 60.4% had no known underlying diseases.

Table 1: Demographic data

Sample size	n = 449
Age (Median (IQR))	46 (36–55)
Sex (n (%)) • Female • Male	309 (62.8) 140 (31.2)
Underlying disease (n (%)) • No • Yes	271 (60.4) 178 (39.6)

Table 2 presents the different diagnoses, findings, and lesions identified by AI and radiologists. The majority of the diagnoses were normal, comprising 77.3% to 87.3% of the total number. Pulmonary lesions were the most frequently reported AI-target lesions by both AI (10.0%) and radiologists (16.7%).

AI-non-target lesions, detectable solely by radiologists, exhibited abnormal findings in 52 cases (11.6%) and normal findings in 228 cases (50.8%). Further details of AI-non-target lesions are provided in Table 3. The three most prevalent abnormal findings among AI-non-target lesions were cardiomegaly, scoliosis, and an enlarged pulmonary artery.

Table 2: Diagnoses, findings, and lesions identified by the AI and by radiologists

Diagnosis and findings	AI n (%)	Radiologist n (%)
Diagnosis (based on both AI-target and AI-non-target lesions)		
• Normal	389 (86.6)	347 (77.3)
• Abnormal	60 (13.4)	102 (22.7)
Diagnosis (based on AI-target lesions only)		
• Normal	389 (86.6)	392 (87.3)
• Abnormal	60 (13.4)	57 (12.7)
AI-target lesion category		
• Pulmonary lesion	45 (10.0)	66 (16.7)
• Consolidation	12 (2.7)	0 (0.0)
• Atelectasis	11 (2.5)	11 (2.5)
• Pleural effusion	2 (0.5)	4 (0.9)
• Pneumothorax	3 (0.7)	0 (0.0)
AI-non-target lesion		242 (53.9)
• Abnormal finding		52 (11.6)
• Normal finding		228 (50.8)

Table 3: AI-non-target lesions identified by radiologists

AI-non-target lesion	Radiologist n (%)
Abnormal lesion or possible disease of AI-non-target lesions	
• Cardiomegaly	44 (9.8)
• Scoliosis	11 (2.5)
• Enlarged pulmonary artery	4 (0.9)
• Elevated hemidiaphragm	3 (0.7)
• Acute fracture	2 (0.5)
• Vertebral compression fracture	2 (0.5)
• Enlarged thyroid gland	2 (0.5)
• Emphysema	2 (0.5)
• Hiatal hernia	2 (0.5)

AI-non-target lesion	Radiologist n (%)
Normal, variant, or possibly inactive lesions of AI-non-target lesions	
• Degenerative spine / Curvature spine • Scoliosis	197 (43.9)
• Pleural thickening	157 (35.0)
• Atherosclerotic change of aorta • Acute fracture	116 (25.8)
• Borderline heart size	37 (8.2)
• Bone lesion / Rib deformity / Old fracture / Osteopenia	37 (8.2)
• Nipple shadow	19 (4.2)
• Soft tissue summation	8 (1.8)
• Post-surgical change; Mastectomy / Lobectomy / Surgical material	6 (1.3)
• Gallstone	5 (1.1)
• Soft tissue calcification / Abdominal calcification	5 (1.1)
• Prominent lung marking	4 (0.9)
• Foreign body	2 (0.5)
• Pectus excavatum	1 (0.2)

Diagnostic agreement between AI and radiologists

In Table 4, the diagnostic agreement between the AI and radiologists is presented. Upon analysis of results solely from AI-target lesions, we observed fair agreement in diagnosis (83.30%, p-value < 0.001), with a kappa value of 0.263. However, upon considering both AI-target and AI-non-target lesions, we found only slight agreement (75.95%, p-value < 0.001, kappa = 0.199).

Table 4: Diagnostic agreement between AI and radiologists

	Agreement (%)	Expected agreement (%)	Kappa	P-value
Diagnosis (based on both AI-target and AI-non-target lesions)	76.0	70.0	0.199	<0.001
Diagnosis (based on AI-target lesions only)	83.3	77.3	0.263	<0.001
AI-target lesion category				
• Pulmonary lesion	86.4	81.8	0.254	<0.001
• Consolidation	97.3	97.3	-	-
• Atelectasis	97.8	95.8	0.489	<0.001
• Pleural effusion	98.9	99.0	-0.01	0.539
• Pneumothorax	100.0	100.0	0.00	0.500

Standalone diagnostic performance of AI using radiologist diagnoses as ground truth:

Treating the diagnosis provided by radiologists as the ground truth, we found that the AI exhibited false-negative and false-positive rates of 8.0% and 8.7%, respectively, based solely on AI-target lesions. However, after we incorporated both AI-target and AI-non-target lesions, the false-negative rate for the AI notably increased to 16.7%.

Table 5: AI false-negative and false-positive rates for diagnosis, evaluated using radiologist diagnoses as ground truth

	False negative	False positive	Sensitivity	Specificity
Diagnosis (based on both AI-target and AI-non-target lesions)	75 (16.7)	33 (7.4)	26.5	90.5
Diagnosis (based on AI-target lesions only)	36 (8.0)	39 (8.7)	36.8	90.1

Discussion

In our study, the radiologists diagnosed a higher proportion of abnormal cases (22.7%) compared with those in a prior study reporting 10.5% referable thoracic abnormalities⁷, and another study reporting 14.5% detected abnormalities.⁸ These variations can be attributed to the inclusion by our radiologists of subtle abnormalities such as fibrotic changes, calcified nodules, and small nonspecific nodules within abnormal pulmonary lesions. Furthermore, differences impacting the prevalence of thoracic diseases between study populations may have contributed to the variance in abnormality rates.

Kim et al.⁷ reported pulmonary nodules as the most frequent finding (3.1%) from checkup radiographs, consistent with our findings from AI-target lesions assessed by AI (prevalence of 10.0% for pulmonary lesions) and by radiologists (16.7%). The radiologists demonstrated the ability to identify a broader range of pulmonary lesions, including pulmonary fibrosis, compared with AI. This discrepancy may account for the higher rate of pulmonary lesion detection observed among radiologists compared with AI.

To our knowledge, there are only a few studies assessing the diagnostic agreement between AI applications and radiologists in the evaluation of checkup CXRs.¹⁷ We report fair agreement for diagnoses based solely on AI-target lesions identified by both AI and radiologists. However, this diminishes to slight agreement when both AI-target and AI-non-target lesions are considered.

Previous studies have demonstrated the remarkable performance of AIs, on par with that of radiologists, in diagnosing lung cancer, tuberculosis, and multiple abnormal findings.^{7,9,10} Despite achieving sensitivity values comparable to those of the radiologists when operating without human assistance, the AI exhibited substantially lower specificity values and a higher frequency of false-positive findings.⁵ However, our findings exhibit the opposite trend: when considering the diagnosis by radiologists as the ground truth, we observed low sensitivity (26.5–36.8) but high specificity (90.1–90.5) values for the AI. This discrepancy in sensitivity estimates could be attributed to the inclusion of subtle abnormalities by our radiologists, such as fibrotic changes, calcified nodules, and small nonspecific nodules. Additionally, the radiologists identified numerous abnormalities within regions discarded by the AI algorithm (AI-non-target lesions).

In addition, the performance of AI diagnostic systems can be influenced by the characteristics of the training population dataset, potentially accounting for interstudy differences. This AI relied on datasets primarily from American and German populations, which may exhibit significant variations in disease prevalence.

Regarding the higher AI specificity observed in our study compared with previous results, this disparity can be attributed to the large number of negative/normal cases in our checkup population, leading to a high number of true-negative cases and elevated specificity. These hypotheses warrant further investigation using larger cohorts comprising classified, clinically significant findings, as well as using more recent AI applications.

Agreement of individual AI-target lesions ranged from negative (pleural effusion and pneumothorax) to fair (pulmonary lesions) to moderate (atelectasis). Furthermore, we observed notable disagreement among the abnormal cases diagnosed by the AI, with 33–39 cases showing disagreement out of 60 abnormal

cases. These findings suggest that abnormal cases diagnosed by AI may require additional peer review and diagnosis.

In a previous study, an AI-based triaging system effectively removed a significant proportion of normal CXRs from the worklist, enhancing the reading efficiency of radiologists, albeit with an erroneous removal rate of 1.8%–3.5%.² In contrast, our AI exhibited higher false-negative rates of 8.0% (36 cases) and 16.7% (75 cases), compared with our measurements. Notably, most false-negative cases in our study involved cardiomegaly (32 cases) from AI-non-target lesions and pulmonary lesions (30 cases) from AI-target lesions.

Consistent with prior research from Kim et al.⁷, our results showed that the AI identified a substantial number of AI-non-target lesions in both the abnormal and normal groups. The presence of these abnormal AI-non-target lesions notably elevated the AI false-negative rate from 8.0% to 16.7%. However, the clinical significance of these lesions extends beyond the scope of our study.

Our study has several limitations. First, it relied on a retrospective single-center cohort, which may limit its generalizability to populations with diverse ethnic compositions and clinical profiles. In addition, the absence of a gold standard diagnostic tool, such as CT chest scans, made confirming the presence of abnormalities more difficult. Furthermore, the inclusion of subtle abnormalities such as fibrotic changes, calcified nodules, small nonspecific nodules, and AI-non-target lesions identified by radiologists, may have led to an underestimation of the true benefit of AI, as some of these lesions may have limited clinical significance. Future studies should classify the degree of clinical significance of findings to evaluate and assess the potential of AI applications more accurately.

Conclusion

Our study revealed fair to slight agreement between a standalone AI and radiologists in diagnosing checkup CXRs from a single-center health-screening cohort. We also observed a significant number of AI-non-target lesions.

Compared with previous studies, our results demonstrated a lower AI sensitivity and a higher false-negative rate, particularly when including AI-non-target lesions. The significant disagreement for abnormal cases diagnosed by AI underscores the need for further scrutiny before clinical decisions are made. AI systems can serve as valuable diagnostic aids for improving radiologists' reading efficiency. Future research should focus on evaluating the performance of standalone AI systems in broader clinical applications.

Acknowledgements

We extend our sincere gratitude to Associate Prof. Dr. Chanisa Chotipanich, our esteemed advisor, for her invaluable guidance, support, and instruction throughout our studies. Additionally, we express our gratitude to the Data Management Unit at the HRH Princess Chulabhorn College of Medical Science and the Chulabhorn Royal Academy for their assistance with the research methodology and statistical analysis. We thank Edanz (www.edanz.com/ac) for editing a draft of this manuscript.

Declaration of Conflicting Interests

We declare no conflict of interest regarding the use of a commercial AI system routinely employed in our hospital for research purposes. Our study utilized this AI system without additional support or benefit from the owning company. As such, there are no financial or personal interests that could potentially bias the interpretation or outcomes of this research.

References

1. Lee JH, Sun HY, Park S, 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(3):687-696. doi:10.1148/radiol.2020201240
2. Yoo H, Kim EY, Kim H, et al. Artificial Intelligence-Based Identification of Normal Chest Radiographs: A Simulation Study in a Multicenter Health Screening Cohort. *Korean J Radiol*. 2022;23(10):1009-1018. doi:10.3348/kjr.2022.0189
3. Gampala S, Vankeshwaram V, Gadula SSP. Is Artificial Intelligence the New Friend for Radiologists? A Review Article. *Cureus*. 2020; 12(10):e11137. Accessed October 24, 2020. doi:10.7759/cureus.11137
4. Siemens Healthcare GmbH. Operating Characteristics and Algorithm Performance AI-Rad Companion Chest X-ray Version 2.0 – 4th February 2021. Accessed January 1, 2022.
5. Ahn JS, Ebrahimian S, McDermott S, et al. Association of Artificial Intelligence-Aided Chest Radiograph Interpretation with Reader Performance and Efficiency. *JAMA Netw Open*. 2022;5(8):e2229289. Accessed August 1, 2022. doi:10.1001/jamanetworkopen.2022.29289
6. McHugh ML. Interrater reliability: the kappa statistic. *Biochem Med (Zagreb)*. 2012;22(3): 276-282.
7. Kim EY, Kim YJ, Choi WJ, 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(2):e0246472. Accessed February 19, 2021. doi:10.1371/journal.pone.0246472
8. Khosangruang O, Chitranonth C. Radiographic findings of screening chest films in Priest Hospital. *J Med Assoc Thai*. 2008;91 Suppl 1:S21-S23.
9. Thammarach P, Khaengthanyakan S, Vongsurakrai S, et al. AI Chest 4 All. *Annu Int Conf IEEE Eng Med Biol Soc*. 2020;2020: 1229-1233. doi:10.1109/EMBC44109.2020.9175862
10. Qin ZZ, Sander MS, Rai B, et al. Using artificial intelligence to read chest radio graphs for tuberculosis detection: A multi-site evaluation of the diagnostic accuracy of three deep learning systems. *Sci Rep*. 2019;9(1):15000. Accessed October 18, 2019. doi:10.1038/s41598-019-51503-3

การอ้างอิง

Jiraaram N., Kitpraphasiri P., Wongsansuk P., Pataraphitakpong T., Phinklao N. Diagnostic agreement between a commercial AI system and radiologists in interpreting checkup chest radiographs. *J Chulabhorn Royal Acad*. 2024; 6(2): 54-62. <https://he02.tci-thaijo.org/index.php/jcra/article/view/260322>

Online Access

<https://he02.tci-thaijo.org/index.php/jcra/article/view/260322>



Appendix

FIGURE 1:

- 1A. False-positive cases diagnosed by the AI (consolidation with an AI confidence score of 7), which were deemed normal by the radiologists.
- 1B. False-positive cases diagnosed by the AI (pneumothorax with an AI confidence score of 6), which were deemed normal by the radiologists.

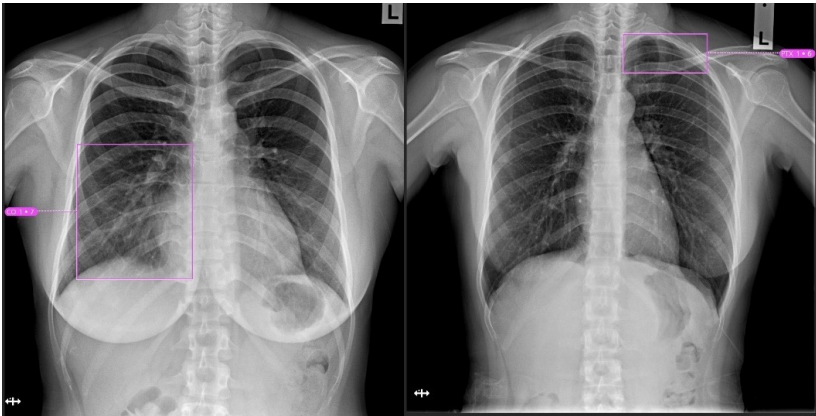


FIGURE 2: Lesions misclassified by the AI (pneumothorax, consolidation, and pleural effusion with AI confidence scores of 10, 6, and 9, respectively), which were identified as pleural thickening, mixed fibrosis, atelectasis, and consolidation by radiologists.

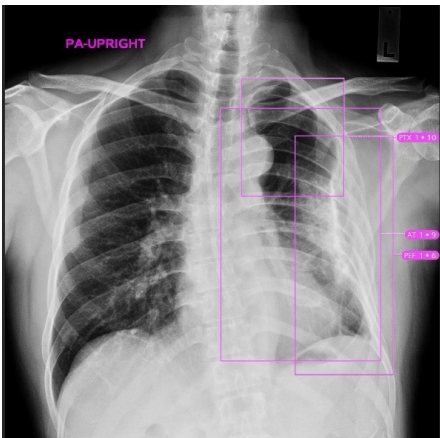


FIGURE 3: Cardiomegaly and scoliosis (AI-non-target lesions) were identified by the radiologists but were not diagnosed by the AI.

