

The Concordance Between Radiologists and AI in Identifying Nipple Shadows on Chest X-rays During Health Checkups at a Private Hospital in Thailand

Cathaleeyakorn Khorana¹; Kewalin Rungsinaporn¹; Pornchalit Suksuputh¹; Warut Chiwong²; Vimolwan Ratanatharathorn¹



Cathaleeyakorn Khorana

Abstract

OBJECTIVES: To evaluate the diagnostic performance of an AI system in differentiating nipple shadows from non-nipple abnormalities, identify imaging characteristics associated with AI misclassification, and compare its performance according to nipple-marker status among annual health examination participants with discordant initial AI–radiologist interpretations of chest radiographs.

MATERIALS AND METHODS: This retrospective study included adults undergoing annual health examinations who had discordant initial chest radiograph interpretations between the AI system and radiologists. Of 54,809 screening examinations, 1,564 showed discordant interpretations and underwent repeat chest radiography with nipple markers followed by final radiologist review. A simple random sample of 902 discordant examinations was analyzed. The final radiologist diagnosis served as the reference standard for distinguishing nipple shadows from non-nipple abnormalities. Within this discordant subset, diagnostic performance was assessed using sensitivity, specificity, overall accuracy, and the area under the receiver operating characteristic curve (AUC), each with a 95% confidence interval, and was stratified by nipple-marker status on the initial radiograph.

RESULTS: Among 54,809 screening chest radiographs, initial AI and radiologist interpretations were concordant in 53,245 examinations (97.1%) and discordant in 1,564 (2.9%). Of the 902 randomly selected discordant examinations, 418 (46.3%) had nipple markers on the initial radiograph, whereas 484 (53.7%) did not. Within this discordant subset, AI sensitivity, specificity, accuracy, and AUC were 66.3%, 51.1%, 52.7%, and 0.587, respectively. Compared with examinations performed with nipple markers, those performed without markers showed higher specificity (74.2% vs. 25.2%), accuracy (72.1% vs. 29.9%), and AUC (0.658 vs. 0.538), but lower sensitivity (57.4% vs. 82.4%).

CONCLUSION: AI demonstrated high initial concordance with radiologist interpretation in a large health-screening population, supporting its potential use as an adjunct to chest radiograph screening and workflow prioritization. These findings suggest that routine nipple-marker placement may not provide additional benefit for AI-based interpretation. A selective approach, reserving repeat radiography with nipple markers for cases of persistent radiologic uncertainty requiring radiologist confirmation, may be more appropriate.

Keywords: artificial intelligence, chest radiography, nipple shadow, nipple markers, diagnostic performance, health screening

¹ Health Design Center, Bangkok Hospital Headquarters, Bangkok, Thailand

² BDMS Health Research Center, Bangkok Dusit Medical Services PCL., Bangkok, Thailand

* Address Correspondence to author:
Vimolwan Ratanatharathorn, MD
Health Design Center
Bangkok Hospital Headquarters,
2 Soi Soonvijai 7, New Petchburi Rd.,
Huaykwang, Bangkok 10310 Thailand
email: rvimolwan@gmail.com

Chest radiography (CXR) is one of the most widely used imaging modalities for screening and diagnosing thoracic diseases because of its accessibility, low cost, and relatively low radiation exposure.¹ In recent years, artificial intelligence (AI), particularly deep learning algorithms, has been increasingly integrated into CXR interpretation to improve diagnostic accuracy, reduce reporting time, and assist radiologists in identifying subtle abnormalities.²

Nipple shadows are common normal anatomical findings on chest radiographs and may mimic solitary pulmonary nodules.³ They are

observed in approximately 10% of chest radiographs⁴ and may lead to unnecessary additional investigations or repeat imaging when they cannot be confidently differentiated from true pulmonary lesions.⁵ When uncertainty exists, repeat CXR with nipple markers is commonly performed to confirm the diagnosis.

Although AI-assisted chest radiograph interpretation has demonstrated promising performance in detecting thoracic abnormalities, its ability to reliably distinguish nipple shadows from true pulmonary abnormalities remains insufficiently evaluated in routine clinical practice. Misclassification of nipple shadows may contribute to false-positive findings, potentially increasing unnecessary follow-up examinations and affecting radiologists' diagnostic confidence.⁶

Therefore, this retrospective study aimed to evaluate the diagnostic performance of an AI system in differentiating nipple shadows from non-nipple abnormalities among annual health examination participants with discordant initial AI–radiologist interpretations. We also examined imaging characteristics associated with AI misclassification and assessed performance according to nipple-marker status on the initial radiograph. These findings may inform the selective use of nipple markers and the role of AI as an adjunct in chest radiograph screening.

Materials and Methods

This study was approved by the Bangkok Hospital Headquarters–Institutional Review Board (Certificate of Approval no. COA.2025-22). The requirement for informed consent was waived because of the retrospective study design and the use of de-identified data. Adults aged 20–75 years undergoing annual health examinations at the health checkup department who had discordant initial chest radiograph interpretations between the AI system and a radiologist were eligible for inclusion. The AI system used was AI Spectra version 3.5.1. Participants with respiratory symptoms, a history of pulmonary nodules or lung tumours, previous lung surgery, inadequate chest radiographs, or other conditions that could interfere with image interpretation were excluded.

During the study period, chest radiographs were performed in two groups with different routine practices: one group routinely applied nipple markers, whereas the other did not. Therefore, nipple-marker status on the initial radiograph was determined by departmental practice rather than by the patient's clinical presentation or radiographic findings. All discordant examinations underwent repeat CXR with nipple markers and subsequent final radiologist review. Of the 1,564 discordant examinations, a simple random sample of 902 was selected using R software and included in the diagnostic performance and image characteristic analyses. The final radiologist diagnosis following repeat CXR with nipple markers was used as the reference standard for differentiating nipple shadows from non-nipple abnormalities. **The screening and participant-selection workflow is shown in Figure 1.**

Statistical analysis

Statistical analyses were performed using R version 4.5.3 (R Foundation for Statistical Computing, Vienna, Austria) in RStudio version 2026.06.0. The unit of analysis was the individual chest radiograph. Continuous variables were summarized using the mean and standard deviation, whereas categorical variables were presented as frequencies and percentages. For each examination, the AI classification on the initial chest radiograph was compared with the final radiologist diagnosis after repeat CXR with nipple markers, which served as the reference standard. A final diagnosis of a non-nipple abnormality was defined as the target condition, and an initial AI classification of a non-nipple abnormality was considered a positive result.

Diagnostic performance was summarized using sensitivity, specificity, overall accuracy, and the area under the receiver operating characteristic curve (AUC), each with a 95% confidence interval. Because the AI output used in this analysis was binary, the AUC represented discrimination at the observed classification threshold. Diagnostic performance was assessed overall and stratified according to nipple-marker status on the initial radiograph. Comparisons between marker-status groups were descriptive; no formal hypothesis

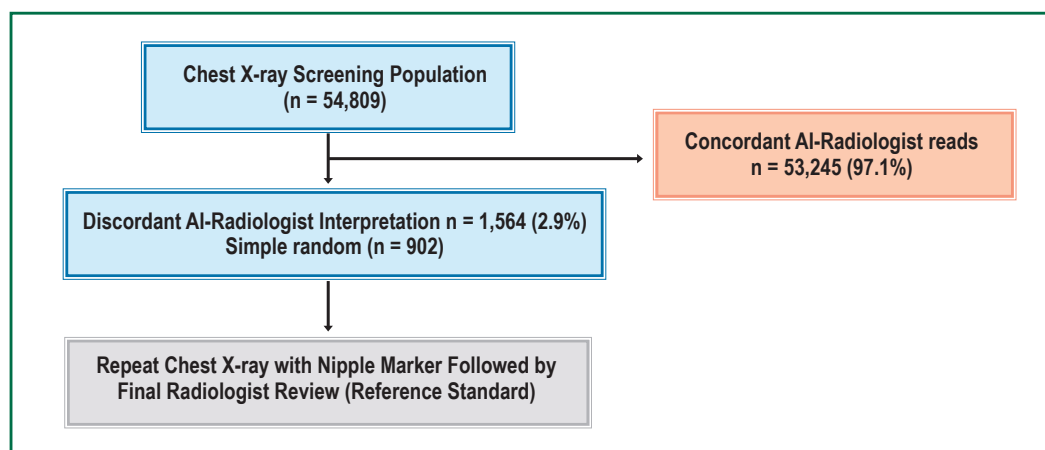


Figure 1: Screening workflow diagram.

testing was performed. Image characteristics associated with AI classification outcomes were summarized as frequencies and percentages.

Results

A total of 54,809 chest radiographs were screened, of which 15,299 (27.9%) were obtained with nipple markers and 39,510 (72.1%) were obtained without nipple markers. Initial AI and radiologist interpretations were concordant in 53,245 examinations (97.1%), whereas 1,564 examinations (2.9%) had discordant interpretations requiring further diagnostic clarification. All discordant examinations underwent repeat chest radiography with nipple markers, followed by final radiologist review. A simple random sample of 902 discordant examinations was included in the diagnostic-performance analysis. Mean age was 51.1 years. On the initial radiographs, nipple markers were present in 418 examinations (46.3%) and absent in 484

(53.7%). The screening workflow and characteristics of the discordant examinations are summarized in Table 1.

Image characteristics according to AI classification outcomes are presented in Table 2. Among correctly classified examinations, 36 non-nipple abnormalities (57.1%) and 57 nipple shadows (13.8%) had no identified image characteristics. Among correctly classified nipple shadows, asymmetry was the most frequently observed characteristic (49.0%), followed by superimposed structures (28.2%) and large nipple shadows measuring >1 cm (18.4%). Among false-positive AI classifications of non-nipple abnormalities, the most frequently observed characteristics were the presence or appearance of nipple markers (28.4%), large nipple shadows measuring >1 cm (26.1%), asymmetry (25.6%), and superimposed structures (22.8%) (Table 2).

Table 1: Screening Workflow and Characteristics of Discordant Examinations

Variable	n (%)
Screening population	n = 54,809
Initial radiographs with nipple markers	15,299 (27.9)
Initial radiographs without nipple markers	39,510 (72.1)
AI–radiologist concordant interpretations	53,245 (97.1)
AI–radiologist discordant interpretations	1,564 (2.9)
Diagnostic performance analysis cohort	n = 902
Mean age, years	51.1
Initial radiographs with nipple markers	418 (46.3)
Initial radiographs without nipple markers	484 (53.7)

Table 2. Distribution of Selected Image Characteristics According to AI Classification Outcome (n = 902)

Image characteristic	Correct non-NS (n = 63), n (%)	Correct NS (n = 412), n (%)	False non-NS (n = 395), n (%)	False NS (n = 32), n (%)
(1) Asymmetry	8 (12.7)	202 (49.0)	101 (25.6)	20 (62.5)
(2) Position mismatch	2 (3.2)	37 (9.0)	76 (19.2)	1 (3.1)
(3) Large nipple shadow (>1 cm)	1 (1.6)	76 (18.4)	103 (26.1)	3 (9.4)
(4) Superimposed structures	12 (19.0)	116 (28.2)	90 (22.8)	6 (18.8)
(5) Non-standard nipple position	3 (4.8)	33 (8.0)	51 (12.9)	12 (37.5)
(6) Presence or appearance of a nipple marker	7 (11.1)	2 (0.5)	112 (28.4)	1 (3.1)
(7) No identified image characteristic	36 (57.1)	57 (13.8)	0 (0.0)	0 (0.0)

Abbreviation: NS, nipple shadow. Correct non-NS indicates that both the AI and the reference radiologist classified the finding as a non-nipple abnormality. Correct NS indicates that both classified the finding as a nipple shadow. False non-NS indicates that the AI classified the finding as a non-nipple abnormality, whereas the reference radiologist classified it as a nipple shadow. False NS indicates that the AI classified the finding as a nipple shadow, whereas the reference radiologist classified it as a non-nipple abnormality. Data are presented as n (%), with percentages calculated using the total number of examinations in each classification group as the denominator. Because more than one image characteristic could be recorded for each examination, the percentages within a column may exceed 100%.

Reasons for AI misclassification were defined as follows (1) **Asymmetry**, differences in nipple density between the two sides, incomplete visualization of the nipple shadow, or unilateral nipple shadow; (2) **Position mismatch**, left and right nipple shadows located at different vertical or horizontal levels; (3) **Large nipple** (>1 cm), nipple shadow measuring >1 cm in diameter; (4) **Superimposed structures**, nipple shadow obscured by adjacent anatomical structures (e.g., soft tissue, ribs, or blood vessels); (5) **Non-standard nipple position**, nipple shadow located outside the expected anatomical position (e.g., excessively high, low, lateral, or off-center); and (6) ; **Presence of nipple marker**, misclassification related to the presence or appearance of a nipple marker; (7) **No error-associated characteristic**, no identifiable feature explaining AI misclassification.

The Concordance Between Radiologists and AI in Identifying Nipple Shadows on Chest X-rays During Health Checkups at a Private Hospital in Thailand

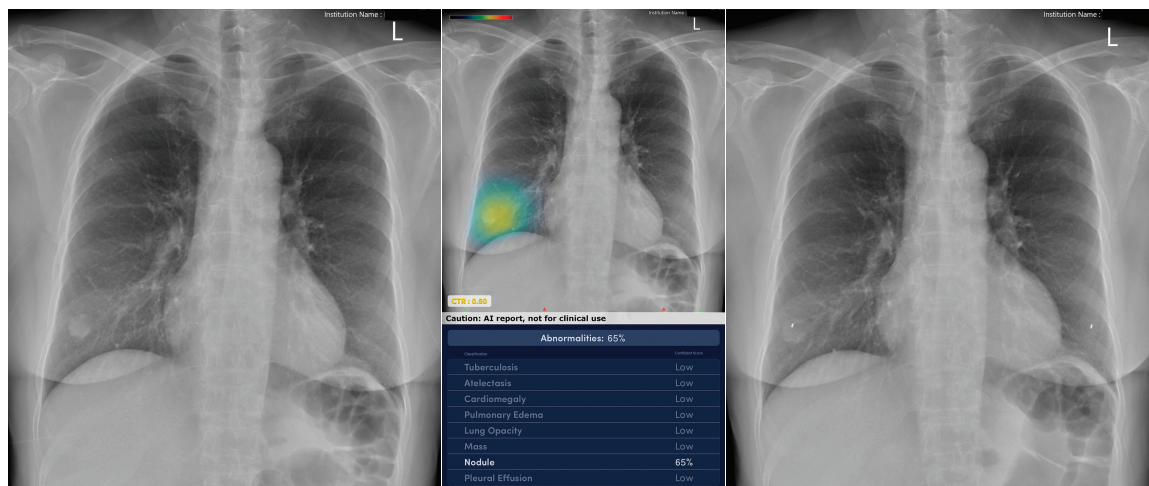


Figure 2: Chest radiographs of a 66-year-old woman. **(Left)** Initial chest radiograph showing an AI-positive nodule in the right lower lung, with a confidence score of 65%. **(Right)** Repeat chest radiograph with nipple markers confirming that the AI-detected nodule corresponded to a nipple shadow, with asymmetric nipple density measuring more than 1 cm.

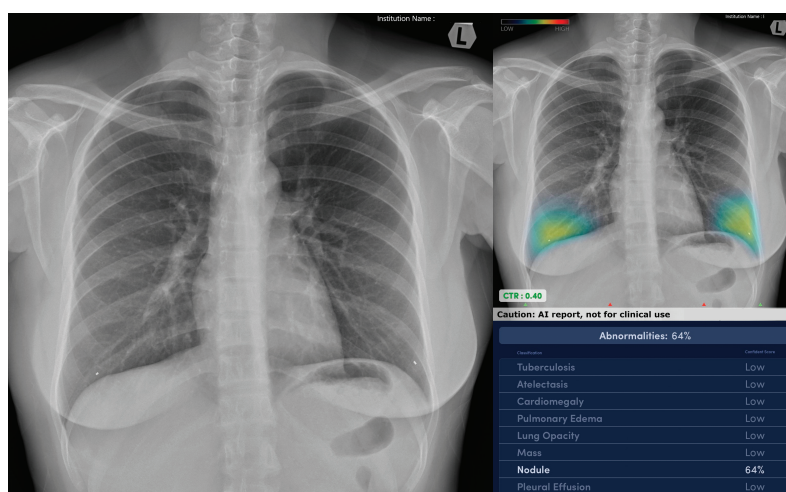


Figure 3: Chest radiograph of a 27-year-old woman showing AI-positive findings in both lower lung zones, with a confidence score of 64%, related to the presence of nipple markers.

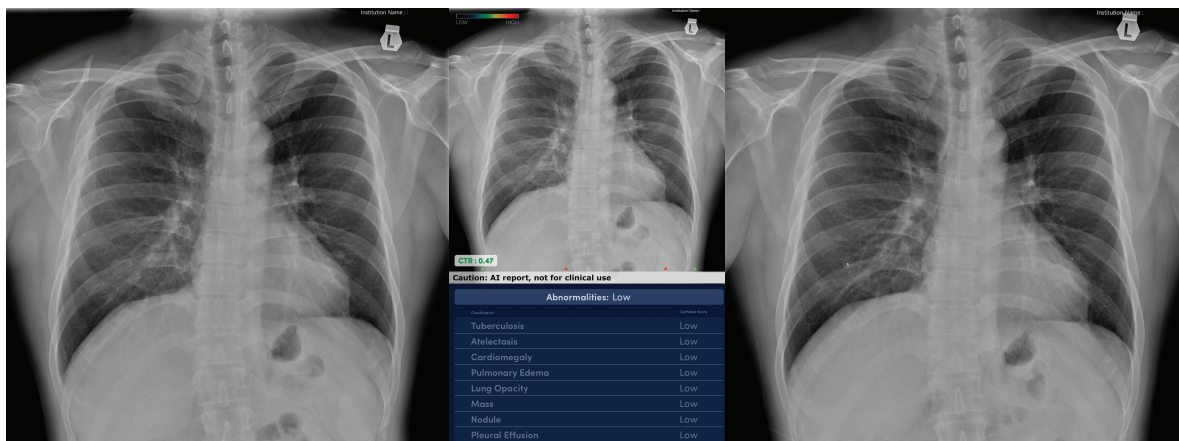


Figure 4: Chest radiographs of a 34-year-old man showing discordant interpretations between the radiologist and AI. **(Left)** Initial chest radiograph showing a questionable nodule-like opacity in the left lower lung identified by the radiologist, while the AI result was negative. **(Right)** Repeat chest radiograph with a nipple marker confirming that the suspected opacity corresponded to a nipple shadow.

Diagnostic performance stratified by nipple-marker status on the initial radiograph is shown in Table 3. Within the discordant subset, examinations without nipple markers demonstrated higher specificity (74.2% vs. 25.2%), overall accuracy (72.1% vs. 29.9%), and AUC (0.658 vs. 0.538) than examinations with nipple markers. However, sensitivity for detecting non-nipple abnormalities was lower in the group

without nipple markers (57.4% vs. 82.4%). Of the 458 examinations initially classified by AI as non-nipple abnormalities, 63 (13.8%) were confirmed as non-nipple abnormalities by the reference radiologist. Conversely, of the 444 examinations initially classified by AI as nipple shadows, 412 (92.8%) were confirmed as nipple shadows (Table 4).

Table 3: Diagnostic Performance of AI in Differentiating Nipple Shadows from Non-Nipple Abnormalities, Stratified by Initial Nipple-Marker Status (n = 902)

Performance metric	Overall (n = 902)	Marker = Yes (n = 418)	Marker = No (n = 484)
Sensitivity	0.663 (0.559 – 0.757)	0.824 (0.655 – 0.932)	0.574 (0.441 – 0.700)
Specificity	0.511 (0.475 – 0.546)	0.252 (0.209 – 0.299)	0.742 (0.697 – 0.783)
Overall accuracy	0.527 (0.493 – 0.560)	0.299 (0.255 – 0.345)	0.721 (0.679 – 0.760)
Area under the ROC curve (AUC)	0.587 (0.536 – 0.638)	0.538 (0.469 – 0.606)	0.658 (0.592 – 0.724)

Table 4: Initial AI Classification Compared with the Final Radiologist Diagnosis in a Random Sample of Discordant AI–Radiologist Examinations (n = 902)

Initial AI classification	Final nipple shadow	Final non-nipple abnormality	Total
Nipple shadow	412	32	444
Non-nipple abnormality	395	63	458
Total	807	95	902

Discussion

The principal finding of this study is the high concordance between initial AI and radiologist interpretations in a large health screening population. AI and radiologists reached the same initial classification in 97.1% of chest radiographs, including examinations performed with and without nipple markers, leaving only 2.9% of cases requiring further diagnostic clarification. This high level of agreement supports the potential role of AI as an adjunct to radiologist interpretation in routine chest radiograph screening. Previous studies have similarly suggested that AI may assist with the identification of normal examinations and the triage of chest radiographs, potentially improving workflow efficiency in screening and outpatient settings.^{7,8}

In this study, AI correctly classified 63 of the 95 reference-standard non-nipple abnormalities (66.3%). These findings indicate that AI retained meaningful discriminative ability in cases that required further diagnostic clarification. The misclassification of nipple shadows is not unexpected, as nipple shadows are a well-recognized and common source of false-positive findings on chest radiographs, including in previous AI studies.⁹ Therefore, initial AI–radiologist discordance in this context should not be interpreted simply as AI failure, but rather as a signal for targeted review and further adjudication. Although AI cannot independently resolve all equivocal findings, it may still provide useful complementary information when interpreted alongside radiologist assessment.

In subgroup analyses, AI performed more favorably in examinations obtained without nipple markers, with higher specificity (74.2% vs. 25.2%), overall accuracy (72.1% vs. 29.9%), and AUC (0.658 vs. 0.538). These findings suggest that AI more effectively distinguished nipple shadows from non-nipple abnormalities when marker-related image features were absent. Although nipple markers are traditionally used to assist radiologists in confirming nipple shadows, their routine application to all chest radiographs may have limited value because recall rates are low and marker placement is frequently inaccurate or noncontributory to interpretation.¹⁰ Their presence may not confer the same benefit to AI in this study. Examinations obtained without nipple markers showed better AI discrimination, raising the possibility that marker-related appearances may alter the visual characteristics of the nipple region or introduce non-pathological image features that affect AI classification. Such features may have contributed to false-positive AI classifications in the present study.

From a clinical perspective, false-positive AI findings may trigger unnecessary follow-up investigations, including repeat chest radiographs or chest computed tomography, thereby increasing radiologists' workload, healthcare costs, and patient anxiety.¹¹ These consequences are particularly relevant in routine health screening, where most individuals are asymptomatic and the prevalence of true pulmonary abnormalities is low; even a modest increase in the false-positive rate may therefore translate into a substantial number of unnecessary investigations at the population level.¹²

Our findings suggest that routine nipple-marker placement may not provide additional benefit for AI-based interpretation. However, in difficult or discordant examinations requiring further clarification, nipple markers remain a valuable tool for radiologists to differentiate nipple shadows from true pulmonary lesions.

Analysis of image characteristics showed that more than half of the correctly classified non-nipple abnormalities (57.1%) had no identified image characteristic, suggesting that AI performed well in relatively straightforward cases. False-positive cases were more likely to be associated with recognizable anatomical features, including large nipple shadows, nipple asymmetry, superimposed anatomical structures, and the presence of nipple markers on initial imaging. These findings suggest that the AI is particularly vulnerable to normal anatomical variants that resemble pulmonary abnormalities on chest radiographs, leading to unnecessary positive classifications. Similar observations have been reported in previous studies, which identified normal anatomical variants, anatomical overlap, and imaging artefacts as common causes of false-positive AI findings in chest radiograph interpretation.^{13,14}

The findings of this study are consistent with previous reports showing that, although AI-assisted chest radiograph interpretation has achieved high overall performance in screening for thoracic abnormalities, its diagnostic accuracy may decline in more challenging cases, such as examinations with multiple abnormalities, subtle lesions, anatomical overlap, or normal anatomical variants that mimic pulmonary disease.^{9,15,16}

Taken together, the findings support a selective nipple-marker strategy in routine health screening: initial chest radiography without universal nipple-marker placement, AI-assisted and radiologist interpretation, followed by repeat nipple-marker radiography only when diagnostic uncertainty or AI–radiologist discordance arises. Such an approach may simplify the initial imaging workflow while preserving a mechanism for diagnostic clarification in challenging cases.

Limitations

This study was limited by its retrospective, single-center design in a private health-screening population, which may limit generalizability to other clinical settings. In addition, nipple-marker use was based on routine departmental practice

rather than random allocation; therefore, differences in workflow or case mix between groups may have influenced the observed performance differences between examinations with and without nipple markers.

Image interpretation was performed as part of routine clinical practice by approximately 30–40 board-certified radiologists, reflecting real-world practice. However, data regarding individual radiologists' years of experience, subspecialty expertise, and interpretive thresholds were not available; therefore, interobserver variability could not be assessed.

Not all CXRs represented patients' first examinations, and prior images may have been available to radiologists during routine clinical interpretation. In contrast, the AI system analyzed each CXR independently without access to prior examinations. Because the availability and use of prior images were not systematically recorded, their potential influence on radiologist–AI agreement could not be assessed.

Conclusion

AI showed high initial concordance with radiologist interpretation in a large health-screening population, supporting its potential role as an adjunctive tool for screening workflow and prioritization of chest radiographs. These findings suggest that routine nipple-marker placement solely for AI interpretation may not provide additional benefit. A selective workflow, in which repeat radiography with nipple markers is considered for cases of persistent clinical or radiologic uncertainty, may be appropriate; however, prospective validation is required before this approach is implemented in routine practice.

Funding

This research received no external funding.

Conflict of Interest

The authors declare no financial or non-financial conflicts of interest related to the AI software or its manufacturer. The authors had no financial relationship, employment, consultancy, sponsorship, or other affiliation with the company that developed the AI system used in this study.

References

1. Broder J. Imaging the chest: the chest radiograph. In: Broder J, editor. *Diagnostic imaging for the emergency physician*. Philadelphia: Elsevier Saunders; 2011. p. 185-296. doi: 10.1016/B978-1-4160-6113-7.10005-5.
2. Jones CM, Buchlak QD, Oakden-Rayner L, et al. Chest radiographs and machine learning—past, present and future. *J Med Imaging Radiat Oncol*. 2021;65(5):538-44. doi: 10.1111/1754-9485.13274.
3. Ferris RA, White AF. The round nipple shadow. *Radiology*. 1976;121(2):293-4. doi: 10.1148/121.2.293.
4. Knipe H. Nipple shadows. *Radiopaedia.org* [Internet]. [cited 2026 Jul 25]. Available from: <https://radiopaedia.org/articles/nipple-shadows> doi: 10.5334/rID-29788.
5. Miller WT, Aronchick JM, Epstein DM, et al. The troublesome nipple shadow. *AJR Am J Roentgenol*. 1985;145(3):521-3. doi: 10.2214/ajr.145.3.521.

6. Huang HY, Huang YH, Lin CH, et al. AI-assisted chest radiograph interpretation enhances diagnostic confidence and standardizes diagnostic accuracy across radiologists: a multi-reader study. *Clin Imaging*. 2026;130:110694. doi: 10.1016/j.clinimag.2025.110694.
7. 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):1009-18. doi: 10.3348/kjr.2022.0189.
8. Yoon SH, Park S, Jang S, et al. Use of artificial intelligence in triaging of chest radiographs to reduce radiologists' workload. *Eur Radiol*. 2024;34(2):1094-103. doi: 10.1007/s00330-023-10124-1.
9. Plesner LL, Müller FC, Brejnbøl MW, et al. Using AI to identify unremarkable chest radiographs for automatic reporting. *Radiology*. 2024;312(2):e240272. doi: 10.1148/radiol.240272.
10. Lee E, Sayyough M, Aslam A, et al. Utility of nipple markers in the era of digital imaging. *J Thorac Imaging*. 2023;38(1): 4-9. doi: 10.1097/RTI.0000000000000628.
11. Hunink MGM, Gazelle GS. CT screening: a trade-off of risks, benefits, and costs. *J Clin Invest*. 2003;111(11):1612-9. doi: 10.1172/JCI18842.
12. Yoshida K, Takamatsu A, Tanaka R, et al. Can AI substitute the first reader in chest radiograph screening? A retrospective non-inferiority evaluation. *Jpn J Radiol*. 2026;44(7):1168-76. doi: 10.1007/s11604-026-01973-z.
13. Tugwell-Allsup JR, England A, Owen BW. Educational insights of artificial intelligence-assisted chest radiograph interpretation in lung cancer detection—a pictorial review. *BJR Artif Intell*. 2025;2(1):ubaf002. doi: 10.1093/bjrai/ubaf002.
14. Vasilev YA, Vladzmyrskyy AV, Arzamasov KM, et al. Limitations of using artificial intelligence services to analyze chest X-ray imaging. *Digit Diagn*. 2024;5(3):407-20. doi: 10.17816/DD626310.
15. Plesner LL, Müller FC, Brejnbøl MW, et al. Commercially available chest radiograph AI tools for detecting airspace disease, pneumothorax, and pleural effusion. *Radiology*. 2023;308(3):e231236. doi: 10.1148/radiol.231236.
16. Yanagawa M, Tomiyama N. Clinical performance of current-generation AI tools for chest radiographs. *Radiology*. 2023;308(3):e232139. doi: 10.1148/radiol.232139.