Clinical Acceptability of a Generative AI System for Automated Chest X-ray Reporting: A Multi-Institutional Non-inferiority Study

Moon Young Kim[1], Ye Ra Choi[1], Kwang Nam Jin[1], Soo-youn Ham[2]

[1]Department of Radiology, SMG-SNU Boramae Medical Center, Seoul National University College of Medicine, Seoul, Republic of Korea
[2]Department of Radiology, Kangbuk Samsung Hospital, Sungkyunkwan University School of Medicine, Seoul, Republic of Korea

1. BACKGROUND

- Recent advances in generative artificial intelligence and vision-language models have enabled the automated generation of comprehensive radiology reports directly from chest radiographs. Although early studies have demonstrated the potential clinical value of these systems, their generalizability remains restricted by single-institution cohorts and narrowly defined diagnostic tasks.
- Furthermore, standardized frameworks for evaluating the real-world clinical acceptability of AI-generated reports are limited, often lacking structured peer-review metrics like RADPEER combined with objective reference standards. Institutional variations in disease prevalence and case complexity can also substantially influence the relative performance of AI systems compared to human readers.
- Therefore, this multicenter study aimed to evaluate the diagnostic acceptability and performance of AI-generated chest radiograph reports across heterogeneous clinical settings

2. MATERIALS AND METHOD

- This retrospective, multicenter, blinded crossover study included 1,455 adult chest radiographs from a tertiary referral hospital (n=925) and a secondary care hospital (n=530).
- For each radiograph, two report versions were generated: an AI-generated report using a generative vision-language model (M4CXR) and a radiologist-generated report produced by a board-certified general radiologist.
- To assess diagnostic adequacy, four blinded board-certified radiologists at each institution evaluated the reports using the RADPEER scoring system, with scores dichotomized into acceptable or unacceptable categories.
- A prespecified non-inferiority margin of a 3.3% absolute difference in unacceptable diagnosis rates was established to compare AI and human readers.
- Additionally, diagnostic performance, including sensitivity, specificity, and accuracy across nine major diagnostic categories, was evaluated against a consensus reference standard.
- This reference standard was meticulously established by two experienced thoracic radiologists using chest CT, electronic medical records, and short-term follow-up imaging.

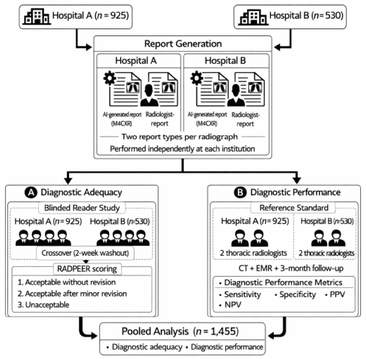
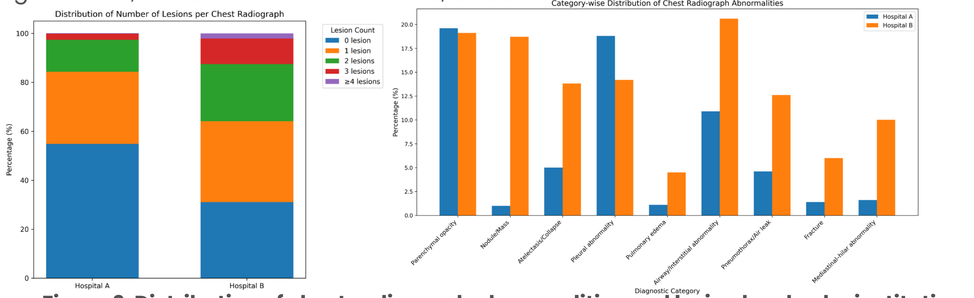


Figure 1. Study design and workflow.



- **Figure 2. Distribution of chest radiograph abnormalities and lesion burden by institution.**
  (A) Overall CXR status and number of lesions per radiograph in Hospital A and Hospital B.
  (B) Category-wise distribution of chest radiograph abnormalities by institution.

3. RESULTS

- **Table 1. Characteristics of the study population**

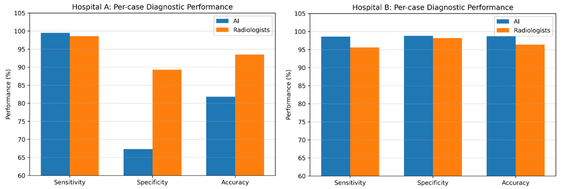
Characteristic	Hospital A (n = 925)	Hospital B (n = 530)	Overall (n = 1,455)	p-value
Age, years (mean ± SD)	60.3 ± 19.4	63.5 ± 16.5	61.5 ± 18.4	0.001
Male sex, n	486 (52.5%)	280 (52.8%)	766 (52.6%)	0.959
Clinical setting, n				<0.001
Outpatient	119 (12.9%)	318 (60.0%)	437 (30.0%)	
Emergency	156 (16.9%)	79 (14.9%)	235 (16.2%)	
Screening	375 (40.5%)	19 (3.6%)	394 (27.1%)	
Inpatient	275 (29.7%)	114 (21.5%)	389 (26.7%)	

Note. Data are presented as mean ± standard deviation or number (%). p -values were calculated using the independent t-test for continuous variables and the chi-square test for categorical variables.

- **Table 3. Unacceptable Diagnosis Rates and Non-inferiority Analysis**

Institution	Group	Unacceptable diagnoses, n/N	Difference (AI –Radiologist)	One-sided 97.5% CI upper bound	Non-inferiority margin
Hospital A	AI	16/925 (1.73%)	–0.11%	1.15	3.3
	Radiologist	17/925 (1.84%)			
Hospital B	AI	7/530 (1.32%)	–0.19%	1.35	3.3
	Radiologist	8/530 (1.51%)			
Overall	AI	23/1455 (1.58%)	–0.14%	0.82	3.3
	Radiologist	25/1455 (1.72%)			

Note. Unacceptable diagnosis rates were compared between AI and radiologists using paired analyses. Non-inferiority was concluded when the upper bound of the one-sided 97.5% confidence interval for the difference (AI – Radiologist) was below the prespecified non-inferiority margin of 3.3%.



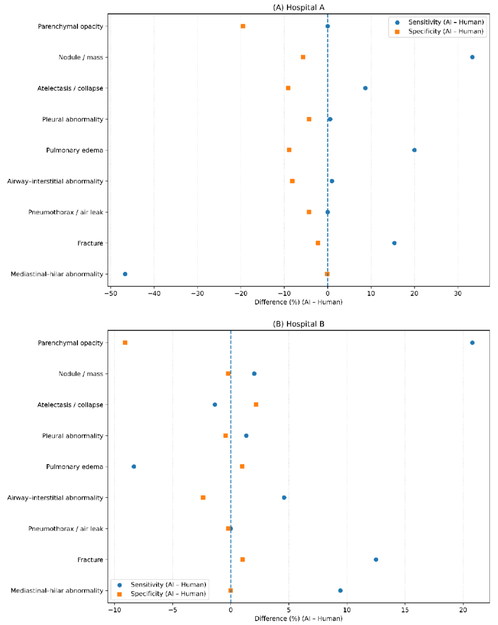
- **Figure 3. Per-case diagnostic performance of AI and radiologists.**

- **Limitations**
  1. A retrospective study
  2. Subjective scoring for diagnostic acceptability
  3. Limited generalizability
     - Selective case sampling
     - Evaluation using single generative AI system across only two institutions without external validation

- **Table 2. Distribution of Chest Radiograph Abnormalities and Lesion Burden**

Characteristic	Hospital A (n = 925)	Hospital B (n = 530)	Overall (n = 1,455)	p-value
Overall CXR status				<0.001
Normal CXR (0 lesion)	507 (54.8%)	165 (31.1%)	672 (46.2%)	
Abnormal CXR (≥1 lesion)	418 (45.2%)	365 (68.9%)	783 (53.8%)	
Number of lesions per CXR				<0.001
1 lesion	272 (29.5%)	175 (33.0%)	447 (30.7%)	
2 lesions	121 (13.3%)	124 (23.4%)	245 (16.8%)	
3 lesions	23 (2.5%)	55 (10.4%)	78 (5.4%)	
≥4 lesions	2 (0.2%)	11 (2.1%)	13 (0.9%)	
Category-wise number of lesions*				
Parenchymal opacity	181 (19.6%)	101 (19.1%)	282 (19.4%)	0.760
Nodule / mass	9 (1.0%)	99 (18.7%)	108 (7.4%)	<0.001
Atelectasis / collapse	46 (5.0%)	73 (13.8%)	119 (8.2%)	<0.001
Pleural abnormality	174 (18.8%)	75 (14.2%)	249 (17.1%)	0.020
Pulmonary edema	10 (1.1%)	24 (4.5%)	34 (2.3%)	<0.001
Airway / interstitial abnormality	101 (10.9%)	109 (20.6%)	210 (14.4%)	<0.001
Pneumothorax / air leak	43 (4.6%)	67 (12.6%)	110 (7.6%)	<0.001
Fracture	13 (1.4%)	32 (6.0%)	45 (3.1%)	<0.001
Mediastinal–hilar abnormality	15 (1.6%)	53 (10.0%)	68 (4.7%)	<0.001

Note. * Percentages for category-wise abnormalities were calculated using the total number of chest radiographs at each hospital as the denominator; individual radiographs could contribute to more than one diagnostic category; therefore, percentages may exceed 100%. p-values were calculated using the chi-square test.



- **Figure 4. Category-wise differences in sensitivity and specificity between AI and radiologists.** The plot shows differences in sensitivity and specificity (AI – Radiologist) for each diagnostic category in Hospital A/B. Positive values indicate higher performance of AI, whereas negative values indicate higher performance of radiologists.

4. CONCLUSION

Although AI-generated reporting systems demonstrate non-inferior diagnostic acceptability, their institution-dependent performance highlights the importance of deploying them as context-aware decision-support tools in clinical practice.