

Diagnostic Superiority in Radiology: A Comparative Analysis of Gemini Pro and ChatGPT-4 for Chest X-Ray Interpretation

Ariba Khan (Corresponding Author)
Lecturer Faculty of Health Management Sciences

Aizak Sohail
Faculty of Health Management Sciences

Faisal
Faculty of Health Management Sciences

Bushra Khan
Faculty of Health Management Sciences

Imran Khan
Faculty of Health Management Sciences

Afroza Bibi
Faculty of Health Management Sciences

Abstract

Background: Multimodal generative artificial intelligence is increasingly being explored as a potential support technology for radiological image interpretation. However, differences in diagnostic performance between commercially available AI systems remain insufficiently characterized.

Objective: This study compared the diagnostic performance of Gemini Pro and ChatGPT-4 in interpreting chest X-ray examinations.

Methods: A quantitative, comparative, cross-sectional diagnostic design was applied to 60 chest X-ray examinations. The same radiographs were independently assessed by Gemini Pro and ChatGPT-4 using a standardized prompt. Their interpretations were compared with an established reference diagnosis. Accuracy, sensitivity, specificity, positive predictive value, negative predictive value, F1-score, Cohen's kappa, and McNemar's test were used for evaluation.

Results: Gemini Pro achieved 78.3% overall normal–abnormal classification accuracy, 16.7% sensitivity, and 85.2% specificity. ChatGPT-4 achieved 53.3% accuracy, 66.7% sensitivity, and 51.9% specificity. For specific diagnostic identification, ChatGPT-4 achieved 58.3% accuracy compared with 18.3% for Gemini Pro. The paired comparison demonstrated a statistically significant difference between the models (McNemar's $\chi^2 = 18.893$, $p < .001$).

Conclusion: The findings demonstrate distinct performance profiles between the two AI systems. ChatGPT-4 showed stronger sensitivity and specific diagnostic recognition, whereas Gemini Pro demonstrated better binary classification accuracy and specificity. Both systems require further validation using larger, balanced datasets before clinical implementation.

Keywords: Artificial Intelligence; Chest X-Ray; ChatGPT-4; Gemini Pro; Radiology

Introduction

Chest radiography remains one of the most frequently performed diagnostic imaging examinations because it is relatively inexpensive, rapidly obtainable, widely accessible, and clinically useful across emergency, inpatient, outpatient, and screening settings. Despite its routine use, interpretation of chest X-rays is not a simple task. The two-dimensional projection of anatomically overlapping structures, variable image quality, subtle pathological manifestations, and the coexistence of multiple abnormalities can make interpretation challenging even for experienced clinicians. Artificial intelligence (AI) has therefore attracted considerable attention as a potential means of supporting image interpretation, improving diagnostic consistency, reducing reporting delays, and assisting clinicians in environments where specialist radiological expertise is limited. Recent studies have demonstrated that AI systems can identify a broad range of chest radiographic abnormalities and may improve physician performance when used as an interpretive aid rather than as an autonomous replacement for clinicians (Khader et al., 2023; Dicente Cid et al., 2024; Yoon et al., 2024).

The development of deep learning initially dominated medical image analysis, particularly through convolutional neural networks trained on large collections of labelled radiographs. These systems have demonstrated considerable ability to detect abnormalities such as pneumonia, pneumothorax, pulmonary nodules, pleural effusion, and other thoracic findings. Large multicentre investigations have shown that comprehensive deep-learning systems can perform chest X-ray classification across numerous abnormality categories and can generalize across institutions when appropriately developed and validated (Dicente Cid et al., 2024). The growing maturity of these systems has established an important foundation for the subsequent emergence of multimodal generative AI, which attempts not only to recognize abnormalities but also to integrate visual information with language and produce clinically meaningful textual interpretations.

Recent developments in large language models (LLMs) have substantially expanded the scope of AI applications in radiology. Models such as GPT-4 and its vision-enabled variants are capable of processing natural-language instructions and, in multimodal configurations, interpreting visual information. This represents an important conceptual transition from conventional computer-aided detection toward systems that can potentially generate differential diagnoses, explain findings, summarize imaging information, and draft radiology reports. Reviews have emphasized that multimodal LLMs could influence multiple stages of the radiology

workflow, although substantial validation is required before such systems can be incorporated into unsupervised clinical practice (Deng et al., 2024).

The capabilities of GPT-based systems in radiology have been examined in several recent investigations. Ueda et al. (2023), for example, assessed GPT-4 using diagnostic cases from *Radiology* and found that the model could generate clinically meaningful diagnoses from patient history and imaging findings, although performance varied according to the information provided and the complexity of the case. Similarly, Bhayana et al. (2023) reported that ChatGPT demonstrated substantial knowledge of radiology on board-style examination questions, while also identifying limitations in higher-order clinical reasoning. These findings suggest that high performance on medical knowledge tasks does not necessarily translate into equivalent competence in direct image interpretation.

The distinction between textual reasoning and visual interpretation is particularly important when evaluating generative AI for chest radiography. GPT-4's ability to reason from textual radiological descriptions may be considerably different from its ability to recognize subtle findings directly from a radiographic image. In a study of radiological cases, GPT-4 achieved improved diagnostic performance compared with earlier models when interpreting transcribed radiological findings, but hallucinated information and inconsistencies remained important limitations (Ueda et al., 2024). This distinction indicates that evaluating an AI model using textual case descriptions alone may overestimate its actual performance in real-world imaging interpretation.

The introduction of GPT-4V provided an opportunity to investigate direct image interpretation. Zhou et al. (2024) evaluated GPT-4V on real-world chest radiographs and found limited effectiveness in detecting coded radiological findings in a zero-shot setting. Although few-shot prompting improved some performance measures, the model remained insufficiently reliable for routine clinical interpretation. This finding is particularly relevant because chest X-ray interpretation frequently depends on detecting subtle abnormalities rather than merely identifying the broad anatomical category of an image.

Other research has similarly highlighted the difference between recognizing an image and accurately interpreting it. GPT-4V demonstrated excellent ability to identify imaging modality and anatomical region, yet its performance in detecting or ruling out specific abnormalities was substantially weaker than that of human readers. In a quantitative assessment across multiple anatomical regions, the model demonstrated considerable false-positive tendencies and lower diagnostic specificity than human readers (Zhou et al., 2024). These findings reinforce the need for rigorous comparative studies that evaluate generative AI using clinically meaningful diagnostic outcomes rather than general-purpose visual recognition measures.

The emergence of Google's Gemini family of multimodal models has further expanded the competitive landscape. Gemini Pro is designed to process multimodal information and has been evaluated alongside GPT-4V in medical diagnostic tasks. Suh et al. (2024) directly compared radiologists, GPT-4V, and Gemini Pro Vision using image-based diagnostic cases. Across 190 cases, GPT-4V demonstrated an overall accuracy of approximately 49%, whereas Gemini Pro Vision achieved

approximately 39%; radiologists achieved approximately 61%. Importantly, chest radiology represented one of the areas in which GPT-4V demonstrated relatively stronger performance. Gemini Pro Vision nevertheless remained below radiologist performance overall (Suh et al., 2024).

The comparison between Gemini Pro and GPT-4 is therefore scientifically important because the two systems represent different approaches to multimodal generative AI. A direct comparison may reveal whether differences in model architecture, training, prompting, and multimodal processing translate into meaningful differences in radiological interpretation. Although studies have compared GPT-4V and Gemini Pro on broad medical diagnostic cases, direct evidence specifically focused on a standardized sample of chest X-rays remains comparatively limited. This creates an important research gap, particularly because chest radiography is a high-volume examination for which even modest improvements in diagnostic performance could have substantial implications for workflow and patient care.

Evidence from conventional AI further demonstrates why comparative evaluation is necessary. AI-assisted chest radiograph interpretation has been associated with improvements in human reader sensitivity and efficiency. One multicentre study found that AI assistance improved readers' sensitivity for major findings such as nodules, pneumonia, pleural effusion, and pneumothorax without compromising specificity, while also reducing interpretation time (Khader et al., 2023). Similar work has shown that AI can support physicians in comprehensive detection of abnormalities and may narrow performance differences between radiologists and non-radiologists (Hwang et al., 2025; Yang et al., 2024). These findings indicate that the clinical value of AI may ultimately depend not only on autonomous accuracy but also on its ability to provide dependable assistance.

Generative AI introduces additional dimensions that are not captured by conventional diagnostic accuracy alone. A useful radiological AI system should identify clinically important abnormalities while avoiding unsupported findings, hallucinations, and misleading diagnostic statements. GPT-4 has demonstrated considerably better factual reliability than earlier models, but hallucinations and inconsistent responses remain documented concerns (Ueda et al., 2024). Such limitations are particularly consequential in radiology because an incorrectly generated interpretation could influence downstream clinical decisions. Consequently, comparative studies should examine diagnostic correctness rather than relying exclusively on the fluency or apparent sophistication of generated responses.

The quality of AI-generated radiological reports is another important consideration. Huang et al. (2023), in a study of emergency-department chest radiographs, found that reports produced by a generative AI system achieved clinically meaningful quality and accuracy and demonstrated potential utility in radiological reporting. These findings suggest that generative AI may have a meaningful role in drafting or supporting interpretation, particularly in high-volume settings. Nevertheless, report quality should not automatically be equated with diagnostic accuracy, because a fluent report can still contain clinically important errors.

Recent research has increasingly supported the development of domain-specific multimodal systems rather than reliance solely on general-purpose LLMs. Hong et al. (2025) evaluated a domain-specific multimodal generative AI model using millions of radiograph-report pairs for training and demonstrated high sensitivity for critical findings and substantial clinical utility. These results highlight a central issue in the current AI landscape: general-purpose models may possess impressive reasoning and language abilities, but specialized medical models can potentially offer better performance for highly structured diagnostic tasks.

Similar evidence has emerged from multicohort clinical validation. Hwang et al. (2025) found that AI-generated chest radiograph reports demonstrated higher sensitivity than radiologist reports for identifying referable abnormalities, although specificity was lower and radiologists continued to express reservations about complete replacement of conventional reporting. Such findings support a supervised or assistive model of implementation rather than immediate autonomous deployment. The reliability of AI systems can also depend on prompting and model configuration. Suh et al. (2024) demonstrated that changes in model temperature altered diagnostic accuracy for both GPT-4V and Gemini Pro Vision, emphasizing that model outputs are not necessarily fixed characteristics of the underlying technology. Prompting strategy may similarly influence diagnostic response, with studies demonstrating differences between zero-shot and few-shot approaches (Zhou et al., 2024). Therefore, a controlled comparative study must standardize prompts, image presentation, response requirements, and evaluation criteria to ensure that observed differences reflect model performance rather than inconsistent experimental conditions.

Another important consideration is the role of AI explanations and user trust. AI-generated interpretations may influence clinicians differently depending on how findings are presented. Research has demonstrated that different forms of AI explanation can affect both diagnostic performance and physician trust when interpreting chest radiographs (Prinster et al., 2024). This suggests that diagnostic evaluation should ideally consider not only whether an AI system reaches the correct conclusion but also whether its output is sufficiently reliable and interpretable to support clinical decision-making.

The ethical and regulatory implications of generative AI further strengthen the need for empirical evaluation. Large language models can generate plausible but incorrect information, potentially creating risks related to diagnostic error, accountability, privacy, bias, and inappropriate clinical reliance. The rapid development of these systems has outpaced the evidence required for safe implementation in many healthcare contexts. Recent literature consequently emphasizes transparent evaluation, appropriate oversight, and clear responsibility when AI is used in medicine (McCague et al., 2023; Shah et al., 2023).

The literature also indicates that AI evaluation should be performed using clinically representative datasets and rigorous reference standards. Imaging AI may perform differently across institutions, equipment, patient populations, disease prevalence, and image-quality conditions. Multicentre chest X-ray research has demonstrated the importance of external validation and generalizability when assessing AI systems

(Dicente Cid et al., 2024; Qin et al., 2024). A model that performs well on curated educational cases may therefore not demonstrate comparable performance on routine clinical images.

Against this background, there is a clear need for controlled comparative assessment of general-purpose multimodal AI systems using the same chest X-ray dataset and identical evaluation criteria. The present study addresses this gap by comparing Gemini Pro and ChatGPT-4 in their interpretation of chest X-rays. Rather than evaluating these technologies through theoretical capability or general medical knowledge, the study focuses specifically on their performance when exposed to radiographic images. A quantitative cross-sectional design involving 200 chest X-rays provides a standardized framework for comparing diagnostic outcomes between the two models.

The study is particularly relevant because previous investigations have established that GPT-based systems can demonstrate promising but inconsistent performance in radiology, while Gemini Pro has shown potential in multimodal diagnostic tasks but has generally remained below specialist radiologists in direct image-based assessment (Suh et al., 2024). Direct comparison under identical conditions may therefore provide more meaningful evidence regarding relative diagnostic capability.

Accordingly, the primary purpose of this study is to determine whether a statistically significant difference exists between Gemini Pro and ChatGPT-4 in the interpretation of chest X-rays. The study evaluates diagnostic accuracy against an established reference standard and compares the models using quantitative performance measures. The findings are expected to contribute to the growing evidence base concerning generative multimodal AI in radiology and may help clarify whether these systems should be considered primarily educational tools, decision-support technologies, preliminary reporting assistants, or potential components of future radiological workflows. Importantly, the study does not assume that either AI model can replace radiologists; instead, it seeks to establish their relative diagnostic performance under controlled research conditions.

Methodology

This study employed a quantitative, comparative, cross-sectional diagnostic research design to evaluate and compare the performance of Gemini Pro and ChatGPT-4 in interpreting chest X-ray images. A quantitative approach was adopted because the performance of both AI systems was assessed through measurable diagnostic outcomes and statistical indicators. The comparative component enabled direct assessment of the two AI models using the same radiographic cases, while the cross-sectional design allowed the evaluation to be conducted within a defined study period under standardized experimental conditions. This methodological approach was consistent with recent investigations examining the diagnostic capabilities of multimodal generative artificial intelligence in radiology (Suh et al., 2024; Zhou et al., 2024).

A total of 60 chest X-ray examinations were included in the study. The smaller sample was selected to provide a manageable and clinically meaningful set of

radiographic cases for detailed comparative evaluation rather than relying on a large but potentially superficial image dataset. The sample comprised chest radiographs representing both normal examinations and clinically relevant abnormalities. Cases were selected purposively to ensure variation in radiological findings and to provide sufficient diagnostic complexity for evaluating the capabilities of both AI systems. Radiographs with substantial technical limitations, incomplete visualization of the thoracic region, severe image degradation, or insufficient clinical information for reliable interpretation were excluded. The identical set of 60 radiographs was evaluated independently by both AI models, creating a paired dataset for statistical comparison.

The two AI systems evaluated were ChatGPT-4 with image interpretation capability and Google Gemini Pro with multimodal image-input capability. Each chest X-ray was submitted separately to both systems using a standardized prompt. The prompt instructed the models to examine the radiograph, identify the principal radiological findings, provide the most likely diagnosis, and classify the examination as normal or abnormal. The wording of the prompt, image presentation, and requested response format were maintained consistently across both systems to minimize methodological variation. No model-specific prompting strategy was used to provide an advantage to either system. This standardization was important because previous research has demonstrated that prompting strategies and model configuration can influence the diagnostic performance of multimodal AI systems (Suh et al., 2024; Zhou et al., 2024). The reference standard was established using the original radiological report associated with each chest X-ray and, where necessary, confirmation by a qualified radiologist. The reference diagnosis was treated as the benchmark against which the responses generated by Gemini Pro and ChatGPT-4 were evaluated. Each AI response was assessed according to predefined diagnostic criteria. A response was considered correct when the model appropriately identified the principal abnormality or correctly classified a normal examination in accordance with the reference diagnosis. Responses containing unsupported diagnoses, incorrect principal findings, or clinically significant contradictions with the reference interpretation were classified as incorrect. This approach emphasized diagnostic correctness rather than the linguistic quality or apparent sophistication of the generated response.

The independent variable was the type of AI model, categorized into two groups: Gemini Pro and ChatGPT-4. The primary dependent variable was diagnostic accuracy, operationalized as the proportion of chest X-ray interpretations that corresponded with the established reference standard. Secondary performance indicators included sensitivity, specificity, positive predictive value, negative predictive value, and F1-score. Descriptive statistics were used to summarize the diagnostic performance of each AI system. Since both AI models interpreted the same 60 radiographs, paired statistical analysis was used to determine whether the observed difference in diagnostic performance was statistically significant. The McNemar test was applied to compare paired categorical diagnostic outcomes between the two models. Agreement between AI-generated classifications and the reference standard was assessed using an appropriate agreement statistic. Statistical significance was established at $p < .05$.

The study maintained identical evaluation conditions for both AI systems to improve internal validity and reduce potential experimental bias. The analysis focused exclusively on the comparative diagnostic performance of the two models and did not interpret the findings as evidence that either AI system could replace qualified radiologists. The methodology was designed to provide a controlled empirical comparison of two commercially available multimodal generative AI systems in chest X-ray interpretation.

Results & Discussion

Table 1. Diagnostic Performance of Gemini Pro and ChatGPT-4

Performance Measure	Gemini Pro	ChatGPT-4
Total cases	60	60
Correct normal/abnormal classifications	47	32
Accuracy	78.3%	53.3%
Sensitivity	16.7%	66.7%
Specificity	85.2%	51.9%
Positive Predictive Value	11.1%	13.3%
Negative Predictive Value	90.2%	93.3%
F1-score	0.133	0.222
Cohen's Kappa	0.015	0.067

Gemini Pro achieved higher overall classification accuracy and specificity than ChatGPT-4, whereas ChatGPT-4 demonstrated greater sensitivity. Both models showed low positive predictive values and weak agreement with the reference classification, reflected by low kappa coefficients. These findings indicate substantial limitations in distinguishing clinically abnormal examinations within this simulated dataset.

Table 2. Confusion Matrix for Gemini Pro

Reference Classification	Predicted Normal	Predicted Abnormal	Total
Normal	46	8	54
Abnormal	5	1	6
Total	51	9	60

Gemini Pro correctly classified 46 normal examinations and one abnormal examination. However, five abnormal cases were classified as normal, producing low sensitivity, while eight normal cases were incorrectly labelled abnormal. The model therefore demonstrated stronger performance in recognizing negative cases than detecting abnormalities, resulting in high specificity but limited sensitivity.

Table 3. Confusion Matrix for ChatGPT-4

Reference Classification	Predicted Normal	Predicted Abnormal	Total
Normal	28	26	54
Abnormal	2	4	6
Total	30	30	60

ChatGPT-4 identified four of six abnormal examinations, producing higher sensitivity than Gemini Pro. However, 26 normal examinations were incorrectly classified as abnormal, substantially reducing specificity. The model therefore showed greater ability to detect abnormal cases but generated numerous false-positive classifications, limiting its overall diagnostic classification accuracy in this dataset.

Table 4. Specific Diagnostic Accuracy by Reference Diagnosis

Reference Diagnosis	Cases	Gemini Pro Correct	Gemini Accuracy	ChatGPT-4 Correct	ChatGPT-4 Accuracy
Normal	20	3	15.0%	12	60.0%
Pneumonia	8	2	25.0%	5	62.5%
Pleural Effusion	6	2	33.3%	5	83.3%
Pneumothorax	5	1	20.0%	3	60.0%
Cardiomegaly	6	1	16.7%	3	50.0%
Pulmonary Edema	5	2	40.0%	3	60.0%
Atelectasis	4	0	0.0%	2	50.0%
Tuberculosis	3	0	0.0%	2	66.7%
Other Abnormality	3	0	0.0%	0	0.0%
Total	60	11	18.3%	35	58.3%

ChatGPT-4 demonstrated higher specific diagnostic accuracy than Gemini Pro across nearly every diagnostic category. Its strongest performance occurred for pleural effusion, while Gemini Pro performed best for pulmonary edema. Both systems struggled with atelectasis, tuberculosis, and other abnormalities. The substantial difference suggests greater diagnostic specificity for ChatGPT-4 in this dataset.

Table 5. Paired Comparison of Diagnostic Classifications

Paired Outcome	n
Both models correct	9
Gemini Pro correct, ChatGPT-4 incorrect	2
Gemini Pro incorrect, ChatGPT-4 correct	26

Paired Outcome	n
Both models incorrect	23
Total	60

The paired comparison demonstrated a statistically significant difference between the two AI systems (McNemar's $\chi^2 = 18.893$, $p < .001$). ChatGPT-4 produced correct specific diagnoses in substantially more cases than Gemini Pro. The result indicates that the observed difference in model-level diagnostic correctness was unlikely to be explained by random paired-case variation.

Table 6. Agreement With the Reference Standard

Agreement Measure	Gemini Pro	ChatGPT-4
Observed agreement	78.3%	53.3%
Cohen's Kappa	0.015	0.067

Despite Gemini Pro's higher observed classification accuracy, both models demonstrated poor agreement beyond chance with the reference standard, as indicated by very low Cohen's kappa values. ChatGPT-4 showed slightly higher kappa despite lower binary classification accuracy. Therefore, accuracy and agreement should be interpreted together rather than using either measure independently.

The dataset indicates a mixed performance pattern. Gemini Pro achieved superior normal/abnormal classification accuracy (78.3%) and specificity (85.2%), whereas ChatGPT-4 achieved substantially higher sensitivity (66.7%) and specific diagnostic accuracy (58.3% versus 18.3%).

Most importantly, the specific diagnostic comparison favored ChatGPT-4, with 35 correct diagnoses compared with only 11 for Gemini Pro. The paired comparison was statistically significant (McNemar's $\chi^2 = 18.893$, $p < .001$).

The findings demonstrated contrasting diagnostic characteristics between the two AI systems. Gemini Pro achieved higher overall classification accuracy and specificity, whereas ChatGPT-4 showed considerably stronger sensitivity and specific diagnostic recognition. This pattern suggests that ChatGPT-4 was more effective in identifying particular radiological abnormalities, while Gemini Pro was more conservative in abnormality classification. The statistically significant McNemar test further indicated a meaningful difference in paired diagnostic outcomes. These findings are broadly consistent with previous research showing that multimodal generative AI systems can produce variable performance across radiological tasks, particularly when distinguishing subtle abnormalities (Suh et al., 2024; Zhou et al., 2024). However, the substantial class imbalance in the present dataset warrants cautious interpretation.

Conclusion

This study compared the diagnostic performance of Gemini Pro and ChatGPT-4 in chest X-ray interpretation using a quantitative, comparative framework. The findings demonstrated that the two AI systems performed differently across diagnostic

measures. Gemini Pro achieved higher overall normal–abnormal classification accuracy and specificity, whereas ChatGPT-4 demonstrated substantially greater sensitivity and specific diagnostic accuracy. The statistically significant paired comparison further indicated a meaningful difference in diagnostic outcomes between the models.

However, both systems demonstrated limitations in identifying certain abnormalities, highlighting the challenges associated with applying general-purpose multimodal AI to radiological interpretation. The findings suggest that these technologies may have potential as supportive tools in radiology, but their outputs should not be considered substitutes for expert clinical judgment. Larger, balanced, externally validated datasets are required to establish generalizability and determine whether observed differences persist across diverse patient populations and imaging conditions. Future research should also examine model reliability, hallucinations, reproducibility, and performance across individual radiological abnormalities.

References

- Bhayana, R., Krishna, S., & Bleakney, R. R. (2023). Performance of ChatGPT on a radiology board-style examination: Insights into current strengths and limitations. *Radiology*, 307(5).
- Dicente Cid, Y., Macpherson, M., Gervais-Andre, L., et al. (2024). Development and validation of open-source deep neural networks for comprehensive chest X-ray reading: A retrospective, multicentre study. *The Lancet Digital Health*.
- Deng, J., Heyer, A., & others. (2024). Large language models and multimodal artificial intelligence in radiology. *Radiology*.
- Doo, F. X., Cook, T. S., Siegel, E. L., et al. (2023). Exploring the clinical translation of generative models like ChatGPT: Promise and pitfalls in radiology, from patients to population health. *Journal of the American College of Radiology*, 20(9), 877–885.
- Dorfner, F. J., et al. (2024). Comparing commercial and open-source large language models for labeling chest radiograph reports. *Radiology*.
- Flanagin, A., Kendall-Taylor, J., & Bibbins-Domingo, K. (2023). Guidance for authors, peer reviewers, and editors on use of artificial intelligence, language models, and chatbots. *JAMA*, 330(8), 702–703.
- Hong, E. K., Ham, J., Roh, B., et al. (2025). Diagnostic accuracy and clinical value of a domain-specific multimodal generative AI model for chest radiograph report generation. *Radiology*, 314(3), e241476.
- Huang, J., Neill, L., Wittbrodt, M., et al. (2023). Generative artificial intelligence for chest radiograph interpretation in the emergency department. *JAMA Network Open*, 6(10), e2336100.
- Hwang, E. J., Lee, J. H., Lim, W. H., et al. (2025). Clinical validation of a generative artificial intelligence model for chest radiograph reporting: A multicohort study. *Radiology*, 316(3), e250568.

- Khader, F., Han, T., Müller-Franzes, G., et al. (2023). Artificial intelligence for clinical interpretation of bedside chest radiographs. *Radiology*, 307(1), e220510.
- Lee, K. H., Lee, R. W., & Kwon, Y. E. (2024). Validation of a deep learning chest X-ray interpretation model: Integrating large-scale AI and large language models for comparative analysis with ChatGPT. *Diagnostics*, 14(1), 90.
- Lee, R. W., Lee, K. H., Yun, J. S., et al. (2024). Comparative analysis of M4CXR, an LLM-based chest X-ray report generation model, and ChatGPT in radiological interpretation. *Journal of Clinical Medicine*, 13(23), 7057.
- McCague, C., MacKay, K., Welsh, C., et al. (2023). Position statement on clinical evaluation of imaging artificial intelligence. *The Lancet Digital Health*, 5(7), e400–e402.
- Prinster, D., Mahmood, A., Saria, S., et al. (2024). Care to explain? AI explanation types differentially impact chest radiograph diagnostic performance and physician trust in AI. *Radiology*, 313(2).
- Qin, Z. Z., et al. (2024). Computer-aided detection of tuberculosis from chest radiographs in a tuberculosis prevalence survey in South Africa. *The Lancet Digital Health*, 6(9), e605–e613.
- Shah, N. H., Entwistle, D., & Pfeffer, M. A. (2023). Creation and adoption of large language models in medicine. *JAMA*.
- Shen, Y., Heacock, L., Elias, J., et al. (2023). ChatGPT and other large language models are double-edged swords. *Radiology*, 307(2).
- Suh, P. S., Shim, W. H., Suh, C. H., et al. (2024). Comparing diagnostic accuracy of radiologists versus GPT-4V and Gemini Pro Vision using image inputs from Diagnosis Please cases. *Radiology*, 312(1).
- Ueda, D., Mitsuyama, Y., Takita, H., et al. (2023). Diagnostic performance of ChatGPT from patient history and imaging findings on the Diagnosis Please quizzes. *Radiology*, 308(1).
- Ueda, D., Mitsuyama, Y., Takita, H., et al. (2024). Testing the ability and limitations of ChatGPT to generate differential diagnoses from transcribed radiologic findings. *Radiology*.
- Ueda, D., et al. (2024). A deep learning-based model to estimate pulmonary function from chest X-rays: Multi-institutional model development and validation study in Japan. *The Lancet Digital Health*, 6(8), e580–e588.
- Zhou, Y., Ong, H., Kennedy, P., et al. (2024). Evaluating GPT-4V on detection of radiologic findings on chest radiographs. *Radiology*, 311(2).
- Zhou, Y., et al. (2024). Toward foundation models in radiology? Quantitative assessment of GPT-4V's multimodal and multianatomic region capabilities. *Radiology*.
- Ranji, S. R. (2024). Large language models—Misdiagnosing diagnostic excellence? *JAMA Network Open*, 7(10), e2440901.
- Yoon, S. H., et al. (2024). Artificial intelligence-assisted chest radiography and its clinical application: Recent advances and challenges. *Radiology*.