Can artificial intelligence replace the role of the radiologist in the reporting of wrist radiographs?

Abstract

PurposeArtificial intelligence (AI) has demonstrated improved accuracy and efficiency in several areas of medical imaging; however, its role in musculoskeletal radiograph interpretation remains unclear. This study aimed to evaluate the diagnostic accuracy of freely accessible AI platforms in interpreting wrist radiographs and to determine their ability to detect abnormalities and provide correct diagnoses compared with an expert reference standard.

MethodsA retrospective observational study was conducted using 100 anonymised adult wrist radiographs sourced from publicly available medical imaging repositories. The dataset included 50 normal and 50 abnormal images, encompassing traumatic, congenital, growth-related, and post-operative conditions. All reference diagnoses were independently verified by a board-certified orthopaedic specialist and served as the gold standard. Each radiograph was uploaded to two AI platforms (Grok and CT-read) using a standardised prompt requesting a holistic report. AI outputs were assessed for abnormality detection and diagnostic correctness. Sensitivity, specificity, accuracy, positive predictive value (PPV), negative predictive value (NPV), and Youden's index were calculated and compared against chance performance.

ResultsBoth AI platforms demonstrated limited diagnostic performance. Grok showed sensitivity above chance for fracture detection but poor specificity, frequently over-identifying abnormalities, with an overall accuracy of 41% and a negative Youden's index. For abnormality detection, Grok achieved moderate accuracy (56%) but limited discriminatory ability. CT-read demonstrated higher specificity than sensitivity, performing better at identifying normal radiographs than detecting abnormalities. Its overall diagnostic accuracy was 55%, with PPV of 0.53 and NPV of 0.64. Across both platforms, performance metrics were near chance levels.

ConclusionFreely accessible AI platforms showed limited reliability in interpreting wrist radiographs and are not yet suitable for independent clinical decision-making. Further model refinement and training are required before such tools can be safely integrated into musculoskeletal imaging practice.

KeywordsArtificial Intelligence, Orthopaedics, Radiology, Wrist fractures, Wrist abnormalities

Introduction

The use of artificial intelligence (Al) is increasingly common in medicine. The quoted advantages of Al generated radiograph reporting include improved detection and highlighting of potential abnormalities on radiographs, guiding the clinician's attention to areas that might otherwise be overlooked, improved sensitivity, increased efficiency with reduction in the time it takes for a radiologist to interpret an image, increased overall productivity, reducing diagnostic errors and lastly a potential source for support and teaching for trainees. Furthermore, in many clinics and hospitals worldwide, an instantly available radiology report is not always possible especially when the radiographic examination is performed outside normal working hours. Typically, the radiologist report is available within a few days of the examination, and the report is then sent back to the referring clinician. The report is then checked and if there has been a missed diagnosis, the patient is contacted and brought back for further treatment. This can adversely affect the patient's experience. Bone fractures stand out as a critical area where instantly available reports can greatly help the referring clinician in making the correct diagnosis and instigating early accurate treatment.

The use of Al in chest radiography reporting is already widespread, and studies have shown that there is an absolute increase in accuracy and reduction in reading times for radiologists of all levels of experience. However, little is known regarding the accuracy of Al in interpreting musculoskeletal radiographs. We chose to study wrist radiographs only in this study as this is one of the commonest radiographs performed and it covers both traumatic and non-traumatic conditions (congenital abnormalities, degenerative conditions, post-surgical complications.

The aim is to evaluate the diagnostic accuracy (sensitivity, specificity, positive predictive value, negative predictive value) of Al in interpreting wrist radiographs. The study would also aim to identify the type of conditions (e.g. bone fractures) which some studies claim Al has high accuracy, as well as some conditions (e.g.post-surgical conditions) in which models and algorithms have not yet been established.

Method

A total of 100 anonymised wrist radiographs were included in this study. The images weresourced from publicly available medical imaging repositories, including the NationalInstitutes of Health (NIH) and Kaggle Medical Imaging Collections. The dataset comprised50 normal radiographs and 50 abnormal radiographs, with the abnormal images furthercategorised as trauma-related, growth-related, congenital, or post-operative. Inclusion criteria required that radiographs be posterior–anterior (PA) or lateral views of adult wrists, accompanied by confirmed diagnostic metadata indicating either a normalfinding or an abnormality within the categories described above. Radiographs of poor imagequality or lacking metadata were excluded from the analysis.All reference diagnoses were independently reviewed and verified by a board-certifiedorthopaedic specialist registered with the Hong Kong Medical Council, who holds fullaccreditation in musculoskeletal imaging and orthopaedic trauma care. These expertevaluations served as the gold standard for diagnostic comparison.

Materials and Design

This study employed a retrospective observational design to evaluate the diagnostic accuracyof freely accessible artificial intelligence (AI) platforms in detecting distal radius fractures onwrist radiographs.

Each radiograph was individually uploaded to the selected AI platforms. To maintainconsistency, the following standardised prompt was used for all platforms:“Can you provide a holistic report for the uploaded x-ray?”The AI-generated responses wererecorded verbatim. Each AI report was analysed todetermine:

1. Whether the AI correctly identified the presence or absence of a abnormality, and

2. Whether the AI recognised have correctly recognise the diagnosis of the x-ray .

The diagnostic output for each image was then categorised using two binary variables:

• Abnormality present: Yes or No

• Diagnosis correct: Yes or No

The gold standard orthopaedic diagnosis provided by the specialist was used as the referenceto determine whether each AI classification was accurate.Diagnostic performance measures—including sensitivity, specificity, and overall accuracy—were subsequently calculated bycomparing AI results with the specialist's referencediagnoses.

Procedure

All eligible radiographs were collected and anonymised prior to analysis. Each image wasreviewed and confirmed to meet inclusion criteria before being uploaded to the AI platforms(Grok and CT-read). The uploads were performed sequentially to ensure identical handlingacross platforms.

For each radiograph, the AI was prompted using the standardised question, "Can you providea holistic report for the uploaded x-ray?", and its full narrative interpretation was recorded.The AI-generated reports were then independently reviewed and coded according topredefineddiagnostic categories, with categories defined as "abnormality present" versus "absent" and"diagnosis correct" versus "incorrect." For data processing, these categoricalassessments weresubsequently converted into binary yes/no values. The board-certified orthopaedic specialist, blinded to all AI outputs, independently reviewedall radiographs and provided the diagnosis for each x-ray, e.g., fracture of little fingermetacarpal. These classifications served as the gold standard reference against which AIoutputs were compared.

Following data collection, diagnostic performance metrics were computed to evaluate the agreement between AI-generated interpretations and the expert reference standard.

Results

All analyses were conducted using Jamovi v2.4.12.0

The main analysis assessed the diagnostic performance of Grok and CT-read on wrist radiographs. Metrics evaluated included sensitivity, specificity, PPV, NPV, accuracy, and Youden's index (J). Observed values were compared against chance ($H_0 = 0.50$) to determine each platform's ability to detect abnormalities. Results are presented in Tables 1–8.

Table 1:

Sensitivity and specificity for diagnostic ability of Grok

Support: Diagnostic statistics

	Value	Difference	S	Param	G	df	p
H_0 vs Sensitivity	0.5000	0.3163	-10.0952	1, 2	21.1904	1	< .001
H_0 vs Specificity	0.5000	-0.1275	-1.1753	1, 2	3.3506	1	0.067

Note. S uses Occam's Bonus correction for parameters (Param).

The results showed that sensitivity was significantly higher than chance ($p < .001$), indicating that the AI platforms could reliably detect distal radius fractures when present. However, specificity did not differ significantly from chance ($p = 0.067$), suggesting limited accuracy in identifying normal radiographs

Table 2:

Accuracy and NPV and PPV for diagnostic ability of Grok

Diagnostic statistics

Statistic	LR	Neg LR	Accuracy	Odds Ratio	Prevalence	PPV	NPV	Youden's index J
Value	0.4930	1.3010	0.4100	0.3789	0.4900	0.3214	0.4444	-0.1889

The results showed an overall diagnostic accuracy of 41%, with a positive predictive value (PPV) of 0.32 and a negative predictive value (NPV) of 0.44. The negative Youden's index (J = –0.19) indicated performance below chance level, suggesting poor diagnostic reliability for fracture detection.

Table 3:

Sensitivity and specificity results for abnormality detection of Grok

Support: Diagnostic statistics

	Value	Difference	S	Param	G	df	p
H_0 vs Sensitivity	0.5000	0.0556	0.2216	1, 2	0.5567	1	0.456
H_0 vs Specificity	0.5000	-0.1545	-2.1708	1, 2	5.3416	1	0.021

Note. S uses Occam's Bonus correction for parameters (Param).

The results showed that specificity was significantly below chance (p = 0.021), while sensitivity did not differ significantly from chance (p = 0.456). These findings suggest that the model tended to over-identify abnormalities and lacked consistent discrimination between normal and abnormal images.

Table 4:

Accuracy and NPV and PPV for abnormality detection of Grok

Diagnostic statistics

Statistic	LR	Neg LR	Accuracy	Odds Ratio	Prevalence	PPV	NPV	Youden's index J
Value	1.2865	0.8488	0.5600	1.5158	0.4500	0.5128	0.5902	0.0990

The results showed a moderate overall accuracy of 56%, with PPV = 0.51 and NPV = 0.59. The slightly positive Youden's index (J = 0.10) suggested marginal improvement over chance performance but still reflected limited diagnostic accuracy for detecting abnormalities.

Table 5:

Sensitivity and specificity for diagnostic ability of CT-read

Support: Diagnostic statistics

	Value	Difference	S	Param	G	df	p
H_0 vs Sensitivity	0.5000	-0.3367	-11.6570	1, 2	24.3140	1	< .001
H_0 vs Specificity	0.5000	0.2255	-4.8782	1, 2	10.7563	1	0.001

Note. S uses Occam's Bonus correction for parameters (Param).

The results showed significant differences from chance for both sensitivity ($p < .001$) and specificity ($p = 0.001$). Sensitivity was lower than 0.5, whereas specificity exceeded chance levels, indicating that the platform was more effective at identifying normal images than detecting abnormal findings.

Table 6:

Accuracy and NPV and PPV for diagnostic ability of CT-read

Diagnostic statistics

Statistic	LR	Neg LR	Accuracy	Odds Ratio	Prevalence	PPV	NPV	Youden's index J
Value	1.1533	0.5948	0.5500	1.9392	0.4900	0.5256	0.6364	0.1112

The results showed an overall diagnostic accuracy of 55%, with PPV = 0.53 and NPV = 0.64. The Youden's index (J = 0.11) suggested a slight improvement over chance, reflecting moderate diagnostic capability for wrist radiograph interpretation.

Table 7:

Sensitivity and specificity results for abnormality detection of CT-read

Support: Diagnostic statistics

	Value	Difference	S	Param	G	df	p
H_0 vs Sensitivity	0.5000	-0.1889	-2.7922	1, 2	6.5844	1	0.010
H_0 vs Specificity	0.5000	0.2455	-6.4230	1, 2	13.8461	1	< .001

Note. S uses Occam's Bonus correction for parameters (Param).

The results showed statistically significant sensitivity ($p = 0.010$) and specificity ($p < .001$). The higher specificity values indicated that the system performed better at correctly identifying normal wrist radiographs than detecting abnormal cases.

Table 8:

Accuracy and NPV and PPV for abnormality detection of CT-read

Diagnostic statistics

Statistic	LR	Neg LR	Accuracy	Odds Ratio	Prevalence	PPV	NPV	Youden's index J
Value	0.9241	1.2222	0.4500	0.7561	0.4500	0.4306	0.5000	-0.0566

The results showed an overall accuracy of 45%, with PPV = 0.43 and NPV = 0.50. The slightly negative Youden's index (J = –0.06) suggested near-chance performance, indicating limited reliability in differentiating between normal and abnormal wrist radiographs.

Overall, the results showed that both Grok and CT-read demonstrated limited diagnostic performance on wrist radiographs. Sensitivity and specificity varied across tasks, with specificity generally outperforming sensitivity, indicating that the platforms were more reliable at identifying normal images than detecting abnormalities.

Accuracy, positive predictive value (PPV), and negative predictive value (NPV) were moderate too low for both systems, and Youden's indices were near zero or slightly negative. These findings suggest that, while the platforms performed slightly better than chance in some measures, their overall reliability for interpreting wrist radiographs was limited. Collectively, these results highlight the challenges of using freely accessible AI tools for clinical imaging interpretation and set the stage for further discussion regarding their potential limitations and areas for improvement.

Discussion

Artificial intelligence has been widely promoted as a tool that can enhance clinical efficiency, highlight subtle abnormalities, and support clinicians when immediate radiology input is not available[1]. This is particularly relevant in musculoskeletal imaging, where delayed or missed fracture detection can negatively impact patient outcomes. Although AI has demonstrated strong performance in other areas such as chest radiography, evidence shows that AI systems can detect up to 90% of pulmonary nodules and can improve radiologist sensitivity by approximately 9 to 10 percent[2]. In contrast, much less is known about its performance in interpreting wrist radiographs across both traumatic and non-traumatic presentations. Recent studies have focused mainly on fracture identification[3], and there remains a lack of research exploring broader musculoskeletal applications. Our study aimed to address this gap by evaluating the diagnostic accuracy of two AI systems, Grok and CT-read.

Although this study contributes to the limited literature on AI interpretation of wrist radiographs, an important limitation is that it relies on non–medical-grade AI applications (Grok and CT-scan). These systems differ substantially from the clinically validated AI tools typically available in large hospitals and academic centers. However, this limitation is also reflective of real-world practice: the clinicians most likely to depend on AI for immediate radiograph interpretation are often those working in settings without access to medical-grade platforms. Understanding how readily available, publicly accessible AI models perform is therefore essential, even if their diagnostic capabilities are not directly comparable to specialised clinical systems.

The results showed that neither system currently performs at a level that would reliably support clinical decision making. Grok demonstrated inconsistent diagnostic behaviour and frequently identified abnormalities where none were present. Although its sensitivity for distal radius fractures exceeded chance, the low specificity and below chance overall performance suggest that it may generate unnecessary follow up and patient anxiety. Its difficulty distinguishing normal from abnormal wrist radiographs, particularly in the abnormality detection condition, indicates that it is not yet suitable for routine musculoskeletal interpretation.

CT-read performed relatively better, especially in recognising normal radiographs, which is important for reducing unnecessary referrals. However, its low sensitivity to fractures means that clinically important injuries could still be missed. Since early and accurate fracture identification is one of the main reasons for incorporating AI into frontline practice, this limitation significantly restricts its usefulness.

Overall, these findings suggest that although AI has the potential to improve workflow and support clinicians who do not have immediate access to radiology reporting, current musculoskeletal models are not yet ready for clinical deployment. More extensive training, greater dataset diversity, and continued refinement are needed before these tools can reliably assist in the interpretation of wrist radiographs.

Bibliography:

[1] R. Najjar, "Redefining Radiology: A Review of Artificial Intelligence Integration in Medical Imaging," *Diagnostics*, vol. 13, no. 17, p. 2760, Aug. 2023, doi: 10.3390/diagnostics13172760.

[2] M. Meetschen *et al.*, "AI-Assisted X-ray Fracture Detection in Residency Training: Evaluation in Pediatric and Adult Trauma Patients," *Diagnostics*, vol. 14, no. 6, p. 596, Mar. 2024, doi: 10.3390/diagnostics14060596.

[3] M. E. Adam Essa, "Diagnostic accuracy of AI in chest radiography for pneumonia and lung cancer: A meta-analysis," *Eur. J. Radiol. Open*, vol. 15, p. 100701, Dec. 2025, doi: 10.1016/j.ejro.2025.100701.