

Research Article

Concordance of ChatGPT with a Fellowship-Trained Orthopaedic Surgeon in Detecting Periprosthetic Osteolysis Following Total Shoulder Arthroplasty

Colton C. Mowers, MD¹, Justin T. Childers, MD², Alexander L. Hornung, MD¹, Amelia Hummel, MD¹, Arden Shen, BS¹, Gregory P. Nicholson, MD¹, Grant E. Garrigues¹

¹ Orthopedic Surgery, Rush University Medical Center, ² Orthopedic Surgery, Florida Atlantic University

Keywords: chatgpt, periprosthetic osteolysis, total shoulder arthroplasty, Gruen zone scoring, artificial intelligence, radiographic analysis

<https://doi.org/10.60118/001c.163588>

Journal of Orthopaedic Experience & Innovation

Vol. 7, Issue 2, 2026

Purpose

This study aimed to evaluate the agreement between ChatGPT 4.0 and a fellowship-trained orthopaedic surgeon in detecting periprosthetic osteolysis on radiographs following anatomic total shoulder arthroplasty (aTSA), using the Gruen and Modified Gruen zone scoring systems. Agreement with a fellowship-trained orthopaedic surgeon was evaluated.

Methods

A total of 73 radiographs from patients with failed aTSAs were retrospectively analyzed. Each radiograph was assessed for osteolysis across predefined zones using both traditional and modified Gruen zone scoring. ChatGPT-4.0 received zone definitions and reference images before evaluating each zone as positive (1) or negative (0) for osteolysis. A stem was labeled "loose" if ≥ 2 zones showed osteolysis. A fellowship-trained orthopaedic surgeon independently scored the same radiographs. Agreement was evaluated using Cohen's kappa, and agreement metrics relative to the surgeon were calculated.

Results

Agreement metrics relative to the surgeon demonstrated variable sensitivity across zones, ranging from 33% to 100%, with consistently high specificity (55% to 100%). The highest agreement was observed in Modified Gruen M2 ($\kappa = 1.000$), followed by Modified Gruen Loose ($\kappa = 0.812$), Gruen 7 and Gruen 8 ($\kappa = 0.754$), Gruen 1 ($\kappa = 0.734$), Gruen 6 ($\kappa = 0.686$), Overall Loose ($\kappa = 0.718$), and Gruen Loose ($\kappa = 0.670$). Moderate agreement was observed in Gruen 3 ($\kappa = 0.456$) and Modified Gruen L2 ($\kappa = 0.437$). Fair agreement was observed in Modified Gruen US ($\kappa = 0.395$) and Gruen 5 ($\kappa = 0.333$), while slight to poor agreement was observed in Gruen 4 ($\kappa = 0.183$), Modified Gruen M1 ($\kappa = 0.161$), and Modified Gruen L1 ($\kappa = -0.067$). Overall, substantial or greater agreement was observed in 5 of 8 Gruen zones (62.5%) and 1 of 5 Modified Gruen zones (20%). Agreement relative to the surgeon ranged from 50% to 100%, with most zones exceeding 85%.

Conclusion

ChatGPT 4.0 demonstrates variable agreement with a fellowship-trained orthopaedic surgeon in detecting periprosthetic osteolysis, particularly in zones with well-defined criteria. However, agreement remained inconsistent across anatomical zones. Since the cohort consisted exclusively of failed anatomic total shoulder arthroplasties from an implant retrieval database, these findings may overrepresent the prevalence of osteolysis and limit generalizability.

Level of Evidence

Level III, Retrospective Comparative Study.

INTRODUCTION

ChatGPT (Chat Generative Pre-trained Transformer) is an publicly available artificial intelligence (AI) language model (Kunze et al. 2022). It employs deep-learning techniques and has been trained on a diverse corpus of text, enabling it to generate human-like responses across a wide range of conversational contexts. Since its debut, ChatGPT has undergone significant advancements, evolving from earlier versions with relatively fewer parameters to more sophisticated iterations. Recently, a new feature was added, namely the GPT-4V(vision) system card, enabling it to interpret image-based queries and generate corresponding responses (Zhu et al. 2024). This innovative feature holds promise as it may aid in diagnosing and managing common orthopaedic fractures by integrating patient data, radiological findings, and medical history to generate informed diagnostic suggestions and treatment strategies (Saad et al. 2025). Recent studies, such as Hiredesai et al (Hiredesai et al. 2024)., demonstrated variable agreement in identifying common upper extremity bony pathologies (e.g. distal radius fracture, metacarpal fracture, carpometacarpal osteoarthritis, etc.) with agreement ranging from 0% to 55%. Hiredesai's study, among others, demonstrate the potential for ChatGPT; however, additional studies are needed to validate its consistency and reliability across a broader spectrum of clinical scenarios.

Osteolysis is defined as the progressive loss of bone resulting from an inflammatory response, typically initiated by wear particles and other biological mediators that stimulate osteoclast activity (Algan et al. 1996; Blaine et al. 1996; Teitelbaum 2000). In the context of periprosthetic osteolysis, this inflammatory cascade leads to the resorption of bone adjacent to a prosthetic implant, potentially resulting in implant loosening and eventual failure (Goodman and Gallo 2019). Radiographically, periprosthetic osteolysis is characterized by the appearance of radiolucent lines or cavity defects at the interface of bone with either cement or the implant itself. These findings are integral for early diagnosis and ongoing monitoring of implant integrity as osteolysis may lead to the positional shift of the implant observed on successive radiographs without infection (Hornung et al. 2022; Maloney and Smith 1996). To date, only a single study has evaluated radiolucent lesions in field of dentomaxillofacial radiology with ChatGPT, albeit with reportedly low accuracy (Silva et al. 2024). The application of ChatGPT in orthopaedic radiology, particularly in detecting periprosthetic radiolucency, may provide a valuable tool for enhancing diagnostic accuracy and assisting surgeons in decision-making. As the number of shoulder arthroplasty procedures continues to rise in the US, the importance of recognizing osteolysis post-arthroplasty grows (Kunze et al. 2022; Farley et al. 2021).

The purpose of this study is to investigate the agreement between ChatGPT and a fellowship-trained orthopaedic surgeon in detecting periprosthetic radiolucency on radiographs. Specifically, we hypothesized that although ChatGPT would successfully detect osteolysis on postoperative

radiographs, its overall agreement with a fellowship-trained shoulder and elbow surgeon would be inferior.

METHODS

PATIENT SELECTION

A total of 130 consecutive cases of failed anatomic total shoulder arthroplasty (aTSA) from the years 2010 to 2024 were retrospectively reviewed from a prospectively maintained implant retrieval database to determine eligibility for inclusion. Clinical data, including patient age, implant duration (time in situ), laterality, and sex, were abstracted from the electronic medical records. Radiographic categorization of stem length was performed, classifying implants as short, standard, or long. Inclusion criteria comprised adult patients with a documented history of anatomic total shoulder arthroplasty and the availability of at least one pre-revision radiograph. Both primary and revision aTSA cases were considered eligible.

Exclusion criteria included the absence of pre-revision radiographic imaging and cases involving non-anatomic arthroplasty types, including stemless total shoulder arthroplasties, reverse total shoulder arthroplasties (rT-SAs), and hemiarthroplasties. Of the 130 cases reviewed, 73 met the inclusion criteria and were selected for analysis.

Lastly, each patient was assigned a PT-ID, and the patient information was removed. A blinded researcher was then provided with the radiographs to put into ChatGPT, unaware of the scoring by the orthopaedic surgeon. This ensured that the assessment remained unbiased, allowing for a more standardized comparison of ChatGPT's agreement with the surgeon. This retrospective study received institutional review board (IRB) exemption. All radiographs and patient data were deidentified prior to analysis, and no protected health information was entered into ChatGPT during the study.

COMPARISON WITH ORTHOPAEDIC SURGEON ASSESSMENT

To benchmark agreement with the surgeon, a board-certified, fellowship-trained orthopaedic surgeon specializing in shoulder and elbow surgery independently evaluated the same set of radiographs. These assessments served as the comparator for agreement analysis and not as an independent diagnostic gold standard.

RADIOGRAPHIC ANALYSIS BY CHATGPT

Periprosthetic radiolucency was evaluated using a large language model (ChatGPT) based on a standardized input prompt. The model was prompted to apply the Gruen and Modified Gruen scoring systems corresponding to long and short stem implants, respectively (Gruen et al. 1979). It was provided with detailed zone definitions, annotated reference radiographs, and representative examples. We instructed ChatGPT to analyze periprosthetic osteolysis around the implant in the provided radiographs using these

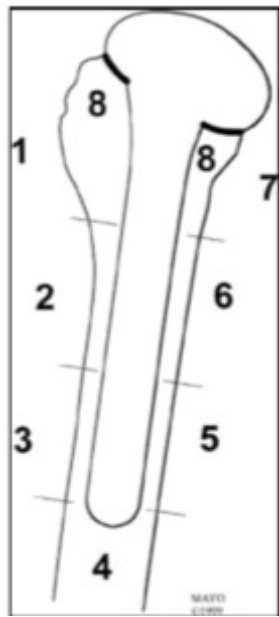


Figure 1. Diagram illustrating Gruen zone scoring for aTSA long stems, highlighting the eight zones used to assess radiolucencies and implant stability.

scoring systems. Instructions were consistent for all short stems but utilized the Modified Gruen Zone Scoring system.

To support standardized AI interpretation, we provided ChatGPT with detailed definitions, including the term ‘osteolysis’, first defined by Jacobs et al (Jacobs et al. 2001), in 2001 as the bone loss or deterioration observed around an implant, impacting implant stability and longevity. Additionally, we supplied example radiographs demonstrating both the presence and absence of osteolysis in all zones. These examples were provided as contextual reference examples to guide ChatGPT with both scoring systems and what to look for in each type of radiograph to standardize interpretation across different stem types.

Gruen zone scoring, traditionally used to assess femoral stem loosening in total hip arthroplasty, has been adapted to evaluate the humeral stem in anatomic total shoulder arthroplasty (aTSA). For standard and long humeral stems, the scoring evaluates radiolucencies in 8 distinct zones (Figures 1, 2, 3), using a binary system where 1 indicates observed osteolysis and 0 indicates no osteolysis, with a maximum possible score of 8 if osteolysis is present in all zones (Sperling et al. 2000). Agreement categories were defined based on concordance between ChatGPT and the surgeon as follows: “Agree +” indicates both identified osteolysis in a given zone; “1+” indicates ChatGPT identified osteolysis while the surgeon did not; “1–” indicates the surgeon identified osteolysis while ChatGPT did not; and “Agree –” indicates that neither identified osteolysis.

For short humeral stems, a Modified Gruen Zone scoring system is utilized, which specifically adapts the original framework to accommodate the shorter stem length (Figure 4, 5) (Schnetzke et al. 2016). This system also employs a binary scoring mechanism across 5 designated zones (L1, L2, US, M1, M2), with each zone scored as either 1 (ob-



Figures 2 and 3. Radiographs of the same shoulder showing osteolysis around the humeral stem and glenoid component in different views, highlighting potential implant loosening.

Figure 2.

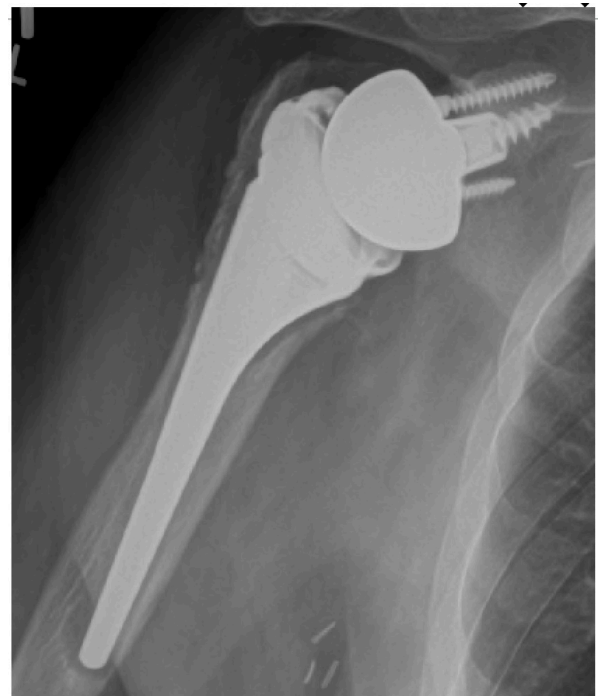


Figure 3.

served osteolysis) or 0 (no osteolysis), culminating in a maximum score of 5.

The orthopaedic surgeon graded this short stem osteolysis as a 5, meaning osteolysis in all zones; ChatGPT gave it

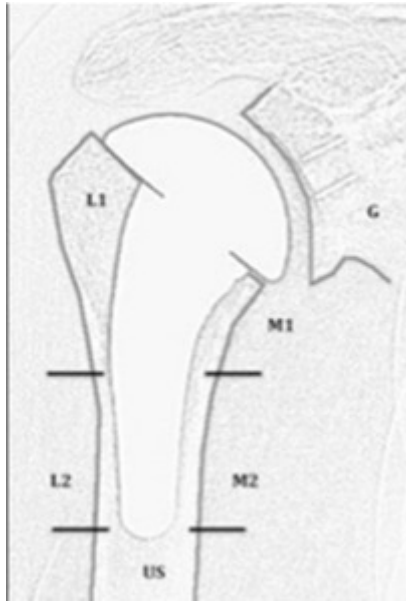


Figure 4. Diagram illustrating the Modified Gruen zone scoring system for short humeral stems in anatomic total shoulder arthroplasty (aTSA).

Legend: M1, medial 1; M2, medial 2; L1, lateral 1; L2, lateral 2; US, under stem; G, glenoid.



Figure 5. Radiographic evidence of severe bone loss in a short stem, showing radiolucencies (indicated by arrows) around the bone-implant interface

a score of 4. We applied this same methodology for all radiographs.

STATISTICAL ANALYSIS

STATISTICAL ANALYSIS

Descriptive statistics were computed, including means and standard deviations (SD) for continuous variables. Group comparisons were performed using two-sample independent t-tests (Student's t-test) and non-parametric Mann-Whitney U tests, as appropriate. Inter-rater agreement be-

tween the AI model and the human evaluator was determined using Cohen's kappa statistic. Agreement with the surgeon was quantified using standard metrics: agreement, sensitivity, specificity, positive predictive value (PPV), negative predictive value (NPV), and false positive rate (FPR). These metrics reflect agreement relative to the surgeon evaluator rather than true diagnostic accuracy. A p-value of less than 0.05 was considered indicative of statistical significance. All statistical analyses were conducted using RStudio (RStudio, PBC, Boston, MA). Subgroup analyses were performed to evaluate the impact of stem type on the detection of periprosthetic osteolysis by ChatGPT.

RESULTS

BASELINE CHARACTERISTICS

73 patients fit the inclusion criteria, with the mean age of the cohort 66.2 (48-72) years. The average time *in situ* was 88.1 (12-302) months. The majority of the patients were men (n=43, 58.9%), and right-sided arthroplasty was more common (n=37). The cohort consisted of 39 patients with long humeral stems, 24 with short stems, and 10 with standard stems.

COMPARISON WITH FELLOWSHIP-TRAINED ORTHOPAEDIC SURGEON

Inter-rater reliability, as assessed by Cohen's kappa (κ), demonstrated a wide range of agreement across scoring models (Table 1). Kappa values ranged from -0.067 to 1.000. The highest agreement was observed in Modified Gruen M2 ($\kappa = 1.000$) and Modified Gruen Loose ($\kappa = 0.812$), followed by Gruen 7 and Gruen 8 ($\kappa = 0.754$), Gruen 1 ($\kappa = 0.734$), Gruen 6 ($\kappa = 0.686$), and Gruen Loose ($\kappa = 0.670$). Moderate agreement was observed in Gruen 3 ($\kappa = 0.456$), Modified Gruen L2 ($\kappa = 0.437$), and Modified Gruen US ($\kappa = 0.441$). Fair agreement was seen in Gruen 5 ($\kappa = 0.333$), while slight to poor agreement was observed in Gruen 4 ($\kappa = 0.183$), Modified Gruen M1 ($\kappa = 0.161$), and Modified Gruen L1 ($\kappa = -0.067$). The composite Overall Loose category demonstrated substantial agreement ($\kappa = 0.718$).

AGREEMENT METRICS RELATIVE TO THE SURGEON

Agreement metrics relative to the surgeon revealed notable variation across models (Table 2). Agreement ranged from 50% in Modified Gruen L1 to 100% in Modified Gruen M2, with most zones exceeding 85%. Sensitivity demonstrated the greatest variability, ranging from 33% in Modified Gruen L1 to 100% in multiple zones, including Gruen 7, Gruen 8, Modified Gruen M2, and Modified Gruen Loose. Specificity remained consistently high, ranging from 55% to 100%. The false positive rate (FPR) was generally low, with the highest values observed in Modified Gruen M1 (0.45), while most zones demonstrated low false positive rates.

Agreement metrics differed across individual zones. The highest false positive rates were observed in Modified Gruen M1 (0.45), Modified Gruen L1 (0.40), and Gruen

Table 1. Summary of inter-rater agreement data comparing ChatGPT to a Fellowship-Trained Orthopedic Surgeon

Zone	Agree +	1+	1-	Agree -	Total	Kappa (95% CI)	Agreement
Gruen (n=49)							
Gruen 1	5	1	2	41	49	0.734 (0.449 to 1.000)	Substantial
Gruen 2	7	3	3	36	49	0.623 (0.349 to 0.897)	Substantial
Gruen 3	2	2	2	43	49	0.456 (0.006 to 0.905)	Moderate
Gruen 4	1	3	3	42	49	0.183 (-0.235 to 0.602)	None to Slight
Gruen 5	3	4	4	38	49	0.333 (-0.027 to 0.694)	Fair
Gruen 6	10	3	3	33	49	0.686 (0.453 to 0.918)	Substantial
Gruen 7	20	3	3	23	49	0.754 (0.570 to 0.938)	Substantial
Gruen 8	20	3	3	23	49	0.754 (0.570 to 0.938)	Substantial
Modified Gruen (n=24)							
L1	3	6	6	9	24	-0.067 (-0.462 to 0.329)	None to Slight
L2	5	3	3	13	24	0.437 (0.057 to 0.818)	Moderate
M1	8	5	5	6	24	0.161 (-0.235 to 0.556)	None to Slight
M2	3	0	0	21	24	1.000 (1.000 to 1.000)	Perfect
US	4	3	3	14	24	0.395 (-0.006 to 0.796)	Fair
Composite							
Gruen Loose (n=49)	30	4	3	12	49	0.670 (0.446 to 0.894)	Substantial
Modified Loose (n=24)	15	1	1	7	24	0.812 (0.564 to 1.000)	Substantial
Overall Loose (n=73)	45	5	4	19	73	0.718 (0.546 to 0.889)	Substantial

Legend: Agree + = both positive; 1+ = ChatGPT positive, surgeon negative; 1- = ChatGPT negative, surgeon positive; Agree - = both negative

Loose (0.25). Lower specificity was observed in Modified Gruen M1 (55%) and Modified Gruen L1 (60%) compared with other zones. No significant difference in agreement with the surgeon or inter-rater agreement was observed based on stem type ($p > 0.05$). Subgroup analysis revealed that ChatGPT's agreement with the surgeon remained consistent across long, short, and standard stem types.

DISCUSSION

This study demonstrates that ChatGPT was able to detect osteolysis in some cases, but its agreement varied across zones, with sensitivity ranging from 33% to 100%. When compared to a fellowship-trained orthopaedic surgeon, ChatGPT demonstrated substantial or greater agreement in 5/8 (62.5%) of Gruen zones and 1/5 (20%) of Modified Gruen zones. The highest agreement was observed in Modified Gruen M2 ($\kappa = 1.000$), followed by Modified Gruen Loose ($\kappa = 0.812$), Gruen 7 and 8 ($\kappa = 0.754$), Gruen 1 ($\kappa = 0.734$), Gruen 6 ($\kappa = 0.686$), and Gruen Loose ($\kappa = 0.670$). Moderate agreement was observed in Gruen 3 ($\kappa = 0.456$) and Modified Gruen L2 ($\kappa = 0.437$). Fair agreement was observed in Modified Gruen US ($\kappa = 0.395$) and Gruen 5 ($\kappa = 0.333$), while slight to poor agreement was observed in Gruen 4 ($\kappa = 0.183$), Modified Gruen M1 ($\kappa = 0.161$), and Modified

Gruen L1 ($\kappa = -0.067$). These findings indicate that ChatGPT's consistency varies significantly by anatomical zone.

Agreement ranged from 50% to 100%, with most zones exceeding 85%. Sensitivity demonstrated the greatest variability, ranging from 33% to 100%, while specificity remained consistently high (55% to 100%). False positive rates were generally low but were higher in certain zones, including Modified Gruen M1 and Gruen 7. While these agreement-based metrics suggest that ChatGPT may be a useful adjunct to surgeon interpretation, the variability in sensitivity and agreement across zones indicates it is not yet a reliable standalone tool. Importantly, these findings should be interpreted within the context of the study cohort, which consisted exclusively of failed aTSA's from an implant retrieval database. This selected population likely has a higher prevalence of osteolysis and implant loosening than a routine postoperative surveillance cohort, which may influence agreement-based metrics and limit extrapolation of these findings to well-functioning shoulder arthroplasties.

Both scoring systems revealed high overall agreement (50% to 100%) and specificity (55% to 100%), but large variability in sensitivity (33% to 100%). Still, the broader composite categories demonstrated substantial agreement, with Gruen Loose ($\kappa = 0.670$), Modified Gruen Loose ($\kappa = 0.812$), and Overall Loose ($\kappa = 0.718$), supporting Chat-

Table 2. Summary of agreement metrics (agreement, specificity, sensitivity, positive predictive value (PPV), negative predictive value (NPV), and false positive rate (FPR)) comparing ChatGPT to a fellowship-trained orthopaedic surgeon

Zone	Agreement	Specificity	Sensitivity	PPV	NPV	FPR
Gruen (n=49)						
Gruen 1	94%	98%	71%	0.83	0.95	0.02
Gruen 2	88%	92%	70%	0.7	0.92	0.08
Gruen 3	92%	96%	50%	0.5	0.96	0.04
Gruen 4	88%	93%	25%	0.25	0.93	0.07
Gruen 5	84%	90%	43%	0.43	0.9	0.1
Gruen 6	88%	92%	77%	0.77	0.92	0.08
Gruen 7	88%	88%	87%	0.87	0.88	0.12
Gruen 8	88%	88%	87%	0.87	0.88	0.12
Modified Gruen (n=24)						
L1	50%	60%	33%	0.33	0.6	0.4
L2	75%	81%	63%	0.63	0.81	0.19
M1	58%	55%	62%	0.62	0.55	0.45
M2	100%	100%	100%	1	1	0
US	75%	82%	57%	0.57	0.82	0.18
Composite						
Gruen Loose	86%	75%	91%	0.88	0.8	0.25
Modified Loose	92%	88%	94%	0.94	0.88	0.12
Overall Loose	88%	79%	92%	0.9	0.83	0.21

GPT's potential role as a supplemental reviewer in detecting periprosthetic osteolysis. However, further validation is necessary to improve agreement and reliability across all zones.

The distribution of stem lengths and types was fairly balanced, with long humeral stems being the most common. While it is possible that stem type could be associated with the radiographic presentation of osteolysis, our analysis revealed no significant effect on ChatGPT's agreement across different stem types. The Modified Gruen system, used specifically for short stems, also showed no significant differences in agreement with the surgeon, suggesting that ChatGPT's agreement was not notably affected by the stem length or type.

A major obstacle in osteolysis research is the lack of standardized definitions and grading systems, making comparisons between studies challenging. Hornung et al (Hornung et al. 2022). identified significant heterogeneity in radiographic assessment methods for glenoid osteolysis, with most relying on subjective, categorical grading and few using quantitative measurements. Our findings similarly highlight the importance of consistent, objective criteria to assess and define osteolysis.

To date, much of the orthopaedic literature involving ChatGPT and other AI tools has centered on patient education (Eng et al. 2025; Lack et al. 2024). The current study is among the first to examine ChatGPT's utility in radiographic evaluation of orthopaedic pathology. A systematic review by Keshavarz et al (Keshavarz et al. 2024). found that

37/44 studies (84.1%) reported high agreement for ChatGPT in radiology, yet none supported unsupervised use. Of the 24 studies that assessed model agreement, median agreement was 70.5% and agreement 83.6%. Our results align with this trend, with comparable agreement relative to the surgeon (50% to 100%) but slightly lower agreement (62.5% and 20% for Gruen and Modified Gruen, respectively).

One included study from the Keshavarz review assessed ChatGPT's ability to generate fracture reports using structured templates and found high agreement with expert reports (Bosbach et al. 2024). However, criticisms included lack of medical precision and verbosity, limitations that may extend to its application in identifying osteolysis, a less-defined radiographic finding.

Current evidence suggests that while ChatGPT performs reasonably well when given structured inputs like classification systems, it is more consistent with distinct radiographic pathologies (e.g., fractures) than more subjective findings such as osteolysis (Saad et al. 2025; Bosbach et al. 2024). While it is not yet capable of replacing human expertise, ChatGPT may serve as a useful secondary reviewer, similar to how radiology reports are used to guide interpretation by orthopaedic surgeons. Discrepancies between radiologists and surgeons are common, and ChatGPT may offer additional context in borderline or uncertain cases. Ultimately, it remains the surgeon's responsibility to determine clinical relevance.

Interestingly, studies in neuroradiology have shown that the standard (text-only) ChatGPT 4.0 outperformed the vi-

sion-enabled 4.0V model in image-based tasks (Hiredesai et al. 2024; Horiuchi et al. 2024). This may reflect the complexity of medical image interpretation and differences in training data. It underscores the need for standardized inputs, especially in uncontrolled clinical environments.

Prompt engineering and pretraining are crucial to maximize AI utility (Meskó 2023). Minor changes in input wording can significantly alter outputs. Tools like RadLex, developed by the RSNA, offer standardized terminology that may improve AI performance by minimizing ambiguity and aligning output with radiologic conventions (Keshavarz et al. 2024). Until such standardization is widespread, variability will continue to hinder consistent clinical application.

Lastly, the false negative rate in our study illustrates a major concern: AI may overlook pathology, leading to missed diagnoses and poor outcomes. While AI offers speed and scalability, it must be supervised, especially when variability in sensitivity ranges from 33% to 100%. Notably, ChatGPT did not undergo iterative learning during this study; all analyses were performed using a fixed prompt with standardized definitions and reference examples, and agreement did not change over time. While representative contextual examples of osteolytic and non-osteolytic radiographs were provided during prompt development, false positives and false negatives were still observed, highlighting limitations in pattern recognition despite structured input.

A key limitation of ChatGPT identified in our study was variability in interpreting glenohumeral osteolysis following aTSA. While some reads were consistent with expert interpretation, others were poor in quality. This inconsistency limits its immediate clinical utility, particularly for general orthopaedic surgeons less specialized in shoulder arthroplasty. However, agreement has improved with successive ChatGPT iterations, suggesting future potential as a decision-support tool (Temperley et al. 2024). Addition-

ally, this cohort consisted exclusively of failed anatomic total shoulder arthroplasties undergoing pre-revision evaluation, which may overrepresent the prevalence of osteolysis and limit the generalizability of PPV, NPV, and overall agreement metrics to routine surveillance or well-functioning arthroplasty populations. The single-center, retrospective design may further restrict external applicability. Furthermore, the use of a single fellowship-trained orthopaedic surgeon as both the comparator and reference standard limits the robustness of the findings and precludes assessment of interobserver variability. In the absence of a true independent gold standard, the reported sensitivity, specificity, PPV, and NPV should be interpreted as measures of agreement with the surgeon rather than true diagnostic agreement. Future studies incorporating multiple surgeons and standardized reference criteria are needed to strengthen the validity and generalizability of these findings.

CONCLUSION

This study found that ChatGPT demonstrated variable agreement with a fellowship-trained orthopaedic surgeon in identifying periprosthetic osteolysis, with lower agreement observed in more complex scenarios. The cohort consisted exclusively of failed anatomic total shoulder arthroplasties from an implant retrieval database, which may overrepresent the prevalence of osteolysis and limit the generalizability of PPV, NPV, and overall agreement metrics. These results highlight the need for ongoing improvements in AI technologies and careful integration into clinical workflows to ensure they enhance rather than replace the expertise of healthcare professionals.

Submitted: March 10, 2026 EDT. Accepted: June 18, 2026 EDT.
Published: October 16, 2026 EDT.



This is an open-access article distributed under the terms of the Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License (CCBY-NC-ND-4.0). View this license's legal deed at <https://creativecommons.org/licenses/by-nc-nd/4.0> and legal code at <https://creativecommons.org/licenses/by-nc-nd/4.0/legalcode> for more information.

REFERENCES

- Algan, S. M., M. Purdon, and S. M. Horowitz. 1996. "Role of Tumor Necrosis Factor Alpha in Particulate-Induced Bone Resorption." *J Orthop Res* 14 (1): 30–35. <https://doi.org/10.1002/jor.1100140107>.
- Blaine, T. A., R. N. Rosier, J. E. Puzas, et al. 1996. "Increased Levels of Tumor Necrosis Factor-Alpha and Interleukin-6 Protein and Messenger RNA in Human Peripheral Blood Monocytes Due to Titanium Particles." *J Bone Joint Surg Am* 78 (8): 1181–92. <https://doi.org/10.1002/jor.1100140107>.
- Bosbach, W. A., J. F. Senge, B. Nemeth, et al. 2024. "Ability of ChatGPT to Generate Competent Radiology Reports for Distal Radius Fracture by Use of RSNA Template Items and Integrated AO Classifier." *Curr Probl Diagn Radiol* 53 (1): 102–10. <https://doi.org/10.1067/j.cpradiol.2023.04.001>.
- Eng, E., C. Mowers, D. Sachdev, et al. 2025. "Chat Generative Pre-Trained Transformer (ChatGPT) - 3.5 Responses Require Advanced Readability for the General Population and May Not Effectively Supplement Patient-Related Information Provided by the Treating Surgeon Regarding Common Questions About Rotator Cuff Repair." *Arthroscopy* 41 (1): 42–52. <https://doi.org/10.1016/j.arthro.2024.05.009>.
- Farley, K. X., J. M. Wilson, A. Kumar, et al. 2021. "Prevalence of Shoulder Arthroplasty in the United States and the Increasing Burden of Revision Shoulder Arthroplasty." *JB JS Open Access* 6 (3): e20.00156. <https://doi.org/10.2106/JBJS.OA.20.00156>.
- Goodman, S. B., and J. Gallo. 2019. "Periprosthetic Osteolysis: Mechanisms, Prevention and Treatment." *J Clin Med* 8 (12): 2091. <https://doi.org/10.3390/jcm8122091>.
- Gruen, T. A., G. M. Mcneice, and H. C. Amstutz. 1979. "'Modes of Failure' of Cemented Stem-Type Femoral Components: A Radiographic Analysis of Loosening." *Clinical Orthopaedics and Related Research* (1976-2007) 141: 17. <https://doi.org/10.1097/00003086-197906000-00002>.
- Hireesai, A. N., C. J. Martinez, M. L. Anderson, C. P. Howlett, K. D. Unadkat, and S. S. Noland. 2024. "Is Artificial Intelligence the Future of Radiology? Accuracy of ChatGPT in Radiologic Diagnosis of Upper Extremity Bony Pathology." *Hand (N Y)*, December 6, 15589447241298982. <https://doi.org/10.1177/15589447241298982>.
- Horiuchi, D., H. Tatekawa, T. Oura, et al. 2024. "Comparing the Diagnostic Performance of GPT-4-Based ChatGPT, GPT-4V-Based ChatGPT, and Radiologists in Challenging Neuroradiology Cases." *Clin Neuroradiol* 34 (4): 779–87. <https://doi.org/10.1007/s00062-024-01426-y>.
- Hornung, A. L., M. R. Cohn, N. Mehta, et al. 2022. "The Definition of Periprosthetic Osteolysis in Shoulder Arthroplasty: A Systematic Review of Grading Schemes and Criteria." *JBJS Rev* 10 (5). <https://doi.org/10.2106/JBJS.RVW.22.00002>.
- Jacobs, J. J., K. A. Roebuck, M. Archibeck, N. J. Hallab, and T. T. Glant. 2001. "Osteolysis: Basic Science." *Clinical Orthopaedics and Related Research*® 393: 71. <https://doi.org/10.1097/00003086-200112000-00008>.
- Keshavarz, P., S. Bagherieh, S. A. Nabipoorashrafi, et al. 2024. "ChatGPT in Radiology: A Systematic Review of Performance, Pitfalls, and Future Perspectives." *Diagn Interv Imaging* 105 (7–8): 251–65. <https://doi.org/10.1016/j.diii.2024.04.003>.
- Kunze, K. N., L. M. Krivicich, C. Brusalis, et al. 2022. "Pathogenesis, Evaluation, and Management of Osteolysis After Total Shoulder Arthroplasty." *Clin Shoulder Elb* 25 (3): 244–54. <https://doi.org/10.5397/cise.2021.00738>.
- Lack, B. T., E. Mouhawasse, J. T. Childers, et al. 2024. "Can ChatGPT Answer Patient Questions Regarding Reverse Shoulder Arthroplasty?" *J ISAKOS* 9 (6): 100323. <https://doi.org/10.1016/j.jisako.2024.100323>.
- Maloney, W. J., and R. L. Smith. 1996. "Periprosthetic Osteolysis in Total Hip Arthroplasty: The Role of Particulate Wear Debris." *Instr Course Lect* 45: 171–82. <https://doi.org/10.2106/00004623-199509000-00022>.
- Meskó, B. 2023. "Prompt Engineering as an Important Emerging Skill for Medical Professionals: Tutorial." *J Med Internet Res* 25: e50638. <https://doi.org/10.2196/50638>.
- Saad, A., H. Rahij, B. K. Durgaprasad, K. P. Iyengar, and R. Botchu. 2025. "Assessing Chat Generative Pre-Training Transformer's Proficiency in Identifying, Diagnosing, and Managing Orthopedic Fractures." *Journal of Arthroscopy and Joint Surgery* 12 (1): 27. https://doi.org/10.4103/jajs.jajs_32_24.
- Schnetzke, M., S. Coda, P. Raiss, G. Walch, and M. Loew. 2016. "Radiologic Bone Adaptations on a Cementless Short-Stem Shoulder Prosthesis." *J Shoulder Elbow Surg* 25 (4): 650–57. <https://doi.org/10.1016/j.jse.2015.08.044>.
- Silva, T. P., M. F. S. Andrade-Bortoletto, T. S. C. Ocampo, et al. 2024. "Performance of a Commercially Available Generative Pre-Trained Transformer (GPT) in Describing Radiolucent Lesions in Panoramic Radiographs and Establishing Differential Diagnoses." *Clin Oral Investig* 28 (3): 204. <https://doi.org/10.1007/s00784-024-05587-5>.
- Sperling, J. W., R. H. Cofield, S. W. O'Driscoll, M. E. Torchia, and C. M. Rowland. 2000. "Radiographic Assessment of Ingrowth Total Shoulder Arthroplasty." *J Shoulder Elbow Surg* 9 (6): 507–13. <https://doi.org/10.1067/MSE.2000.109384>.
- Teitelbaum, S. L. 2000. "Bone Resorption by Osteoclasts." *Science* 289 (5484): 1504–8. <https://doi.org/10.1126/science.289.5484.1504>.
- Temperley, H. C., N. J. O'Sullivan, B. M. Mac Curtain, et al. 2024. "Current Applications and Future Potential of ChatGPT in Radiology: A Systematic Review." *J Med Imaging Radiat Oncol* 68 (3): 257–64. <https://doi.org/10.1111/1754-9485.13621>.

Zhu, L., W. Mou, Y. Lai, et al. 2024. "Step into the Era of Large Multimodal Models: A Pilot Study on ChatGPT-4V(Vision)'s Ability to Interpret Radiological Images." *Int J Surg* 110 (7): 4096–102. <https://doi.org/10.1097/JS9.0000000000001359>.