Comparing AAOS Appropriate Use Criteria with ChatGPT-4.5 and Gemini 2.0 Pro Recommendations for Hip Osteoarthritis

Charlene W Cai, Alexander Yu, Aneesh P Reddy, Jonathan J Huang, Prabhjot Singh, Samuel Kang-Wook Cho

INTRODUCTION: Hip osteoarthritis presents a complex clinical challenge, with treatment decisions influenced by a multitude of patient-specific factors, such as age, pain severity, radiographic findings, range of motion, and risk of adverse outcomes. The American Academy of Orthopaedic Surgeons (AAOS) developed appropriate use criteria (AUC) to guide treatment decisions for hip osteoarthritis based on expert consensus. This study evaluates the ability of two advanced AI (Artificial Intelligence) models, Chat Generative Pre-trained Transformer-4.5 (ChatGPT-4.5) and Gemini 2.0 Pro, to generate treatment recommendations for hip osteoarthritis by comparing their appropriateness scores for hip osteoarthritis treatment with those from the AUC.

METHODS: The AAOS AUC, ChatGPT-4.5, and Gemini 2.0 Pro were utilized to assess treatment suggestions for hip osteoarthritis. The 270 patient scenarios were assessed using a 1 to 9 scale to rate the appropriateness of various treatment options, including conservative options (risk factor assessment, activity modification, assistive devices, oral medications, intraarticular steroids, and physical therapy) and surgical interventions (arthroplasty, hip preservation, and arthrodesis). ChatGPT-4.5 and Gemini 2.0 Pro were queried to rate the treatments on the identical scale. Mean absolute and squared errors were calculated to assess the models' predictive accuracy compared to AAOS guidelines. Spearman correlations and paired t-tests ($\alpha < .05$) were conducted to evaluate consensus, with heatmaps to visualize findings.

RESULTS: Both ChatGPT-4.5 and Gemini 2.0 Pro provided hip osteoarthritis treatment recommendations that differed significantly from the AAOS guidelines ($p < 0.001$) (Tables 1 and 3). For most conservative treatments, Gemini demonstrated smaller mean absolute errors (e.g., assistive devices: +1.10 vs +2.30) compared to ChatGPT-4.5 (Tables 1 and 3). Relative to Gemini, ChatGPT-4.5 predictions showed greater standard deviations and variability in response to nuanced patient-specific factors (Tables 1 and 3; Figures 1 and 2). Though, this increased sensitivity to individual traits may also introduce potential inconsistencies in clinical decision-making. Conversely, Gemini's more conservative bias, reflected by its tighter clustering (Figures 1 and 2), may demonstrate limited responsiveness to patient-specific variation and inability to differentiate complex patient scenarios compared to ChatGPT-4.5 (Tables 1 and 3). Both models consistently underestimated the appropriateness of surgical interventions. However, the AI models demonstrated stronger rank correlations with AAOS guidelines for surgical interventions (e.g., GPT-4.5: $\rho = 0.648$ for arthroplasty; Gemini: $\rho = 0.522$) than for conservative treatments (e.g., GPT-4.5: $\rho = -0.00813$ for risk factor assessment; Gemini: $\rho = -0.129$) (Tables 2 and 4).

DISCUSSION AND CONCLUSION: Both ChatGPT-4.5 and Gemini 2.0 Pro significantly differed from AAOS recommendations in treating hip osteoarthritis ($p < 0.001$ across all treatments). While there was moderate concordance in certain treatments, the general trend towards favoring more conservative approaches and underrecommending surgical interventions raises concern about the reliability of AI-generated recommendations for clinical decision-making. If employed without human oversight, the AI models' underestimation of surgical appropriateness could lead to the undertreatment of hip osteoarthritis. Additionally, consistently over-/underestimating conservative treatments may lead to suboptimal early management decisions. As such, future development of AI models for orthopedic decision support should incorporate specialized training on clinical guidelines and real-world patient outcomes to improve both accuracy and applicability in practice. Currently, these models are not suited to generate recommendations for patients independently, but rather assist decision-making based on validated training data.

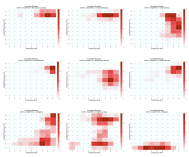

Figure 1. Heatmaps comparing ChatGPT-4.5 and AAOS ratings of the appropriateness of risk factor assessment (a), activity modification (b), assistive devices (c), oral medications (d), intraarticular steroids (e), physical therapy (f), arthroplasty (g), hip preservation (h), and arthrodesis (i) for 270 different patient scenarios.

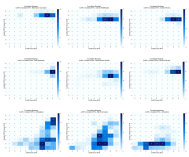

Figure 2. Heatmaps comparing Gemini 2.0 Pro and AAOS ratings of the appropriateness of risk factor assessment (a), activity modification (b), assistive devices (c), oral medications (d), intraarticular steroids (e), physical therapy (f), arthroplasty (g), hip preservation (h), and arthrodesis (i) for 270 different patient scenarios.

Table 1. Mean error, mean absolute error, mean squared error, and paired t-test results between AAOS and ChatGPT-4.5 scores for hip osteoarthritis treatment recommendations.

Treatment	Mean Error	Mean Absolute Error	Mean Squared Error	P-Value
Risk Factor Assessment	0.470 ± 1.23	0.670 ± 0.992	1.76 ± 4.17	<0.001
Activity Modification	0.570 ± 1.22	0.993 ± 0.913	1.82 ± 3.15	<0.001
Assistive Devices	-2.01 ± 2.06	2.30 ± 1.72	8.23 ± 9.17	<0.001
Oral Medication	-1.02 ± 0.813	1.13 ± 0.662	1.70 ± 1.55	<0.001
Intraarticular Steroids	-2.26 ± 1.72	2.32 ± 1.38	8.06 ± 6.85	<0.001
Physical Therapy	-0.044 ± 0.932	1.10 ± 0.752	1.56 ± 1.96	<0.001
Arthroplasty	-2.83 ± 2.11	2.89 ± 2.03	12.3 ± 12.3	<0.001
Hip Preservation	-0.956 ± 2.72	2.08 ± 1.73	8.29 ± 9.98	<0.001
Arthrodesis	-3.56 ± 1.64	3.56 ± 1.64	15.3 ± 12.1	<0.001

Table 2. Spearman correlation coefficients and p-values comparing AAOS and ChatGPT-4.5 scores for hip osteoarthritis treatment recommendations.

Treatment	Spearman's ρ	P-Value
Risk Factor Assessment	-0.00813	>0.001
Activity Modification	0.0918	>0.001
Assistive Devices	-0.771	<0.001
Oral Medication	-0.129	>0.001
Intraarticular Steroids	-0.302	<0.001
Physical Therapy	0.329	>0.001
Arthroplasty	0.648	>0.001
Hip Preservation	0.286	>0.001
Arthrodesis	0.169	>0.001

Table 3. Mean error, mean absolute error, mean squared error, and paired t-test results between AAOS and Gemini 2.0 Pro scores for hip osteoarthritis treatment recommendations.

Treatment	Mean Error	Mean Absolute Error	Mean Squared Error	P-Value
Risk Factor Assessment	0.244 ± 1.28	0.815 ± 0.918	1.86 ± 3.79	<0.001
Activity Modification	-0.244 ± 1.25	0.982 ± 0.808	1.61 ± 2.25	<0.001
Assistive Devices	1.01 ± 0.964	1.10 ± 0.703	1.83 ± 1.64	<0.001
Oral Medication	-1.73 ± 0.782	1.81 ± 0.632	3.68 ± 2.26	<0.001
Intraarticular Steroids	-1.44 ± 1.32	1.86 ± 0.909	4.42 ± 4.01	<0.001
Physical Therapy	-1.18 ± 1.08	1.43 ± 0.717	2.56 ± 2.17	<0.001
Arthroplasty	-3.85 ± 2.08	3.85 ± 2.08	19.1 ± 16.8	<0.001
Hip Preservation	-2.12 ± 1.41	3.16 ± 1.35	12.3 ± 9.81	<0.001
Arthrodesis	-2.68 ± 1.35	2.69 ± 1.33	9.57 ± 9.05	<0.001

Table 4. Spearman correlation coefficients and p-values comparing AAOS and Gemini 2.0 Pro scores for hip osteoarthritis treatment recommendations.

Treatment	Spearman's ρ	P-Value
Risk Factor Assessment	-0.129	>0.001
Activity Modification	0.0122	0.001
Assistive Devices	-0.0447	>0.001
Oral Medication	-0.0573	>0.001
Intraarticular Steroids	-0.123	>0.001
Physical Therapy	0.0466	>0.001
Arthroplasty	0.522	>0.001
Hip Preservation	0.501	>0.001
Arthrodesis	0.418	>0.001