Do Currently Available Large Language Models Provide Musculoskeletal Treatment Recommendations That Are Concordant with Evidence-Based Clinical Practice Guidelines?

Benedict U Nwachukwu[1], Nathan Varady, Answorth Anthony Allen[2], David W Altchek[2], Joshua S Dines[1], Riley Joseph Williams[3], Kyle N Kunze[4]
[1]Hospital For Special Surgery, [2]Hosp for Special Surgery, [3]Hospital For Special Surgery - Weill Cornell Med, [4]Hospital for Special Surgery

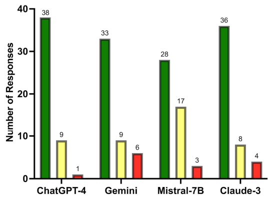
INTRODUCTION: An important application of large language models (LLMs) consists of the feasibility for providing evidenced-based guidelines for medical providers, as this may help guide patient management including injury triage and necessity of referral to musculoskeletal specialists, and therefore have implications for deployment within healthcare settings. The purpose of this study was to determine whether several leading, commercially-available LLMs provide treatment recommendations concordant with evidenced-based clinical practice guidelines (CPGs) developed by the American Academy of Orthopedic Surgeons (AAOS).

METHODS: All evidence-based CPGs concerning the management of rotator cuff tears (n=33) and anterior cruciate ligament (ACL) injuries (n=15) were extracted from the AAOS (**Table 1**). Utilizing a structured framework for query development, information from each AAOS guideline was transformed into a question-based format to prompt each LLM to provide a recommendation concerning the topic. Subsequently, treatment recommendations from four contemporary LLMs (Chat-generative pretrained transformer version-4 [ChatGPT-4; OpenAI], Gemini (Google), Mistral-7B (Mistral AI), and Claude-3 (Anthropic) were obtained. Each question was systematically queried three times utilizing each LLM, and additional follow-up dialogue was not permitted after the initial query. After the three prompts, responses were categorized into a single unique recommendation for analysis by two blinded physicians based on majority vote (576 total responses were therefore combined into 192 unique recommendations for analysis) as being “concordant,” “discordant,” or “indeterminate” (i.e., neutral response without definitive recommendation) with respect to AAOS CPGs. Prior to each new query, the previous dialogue was deleted and history cleared in order to avoid the propagation of bias from stored memory stored. The overall concordance between LLM and AAOS recommendations were quantified, while the comparative overall concordance of recommendations amongst the four LLMs was evaluated through the Fischer’s-exact test. Inter-rater reliability of response concordance was assessed utilizing Cohen’s Kappa coefficient.

RESULTS:

A total of 192 responses were elicited. Overall, 135 (70.3%) responses were concordant, 43 (22.4%) were indeterminate, and 14 (7.3%) were discordant (**Table 2**). Inter-rater reliability for classification of concordance was deemed as excellent (Kappa=0.92). Concordance with AAOS CPGs was most frequently observed with ChatGPT-4 (n=38, 79.2%), and least frequently with Mistral-7B (n=28,58.3%). Indeterminate recommendations were most frequently observed with Mistral-7B (n=17,35.4%) and least frequently with Claude-3 (n=8, 6.7%). Discordant recommendations were most frequently observed with Gemini (n=6,12.5%) and least frequently with ChatGPT-4 (n=1,2.1%). Overall, no statistically significant differences in concordant recommendations was observed across LLMs (p=0.12; **Figure 1**). Only 20 (10.4%) of all recommendations were transparent and provided references with full bibliographic details or links to specific peer-reviewed content to support recommendations (**Figure 2**) ChatGPT-4 was the most transparent LLM evaluated, providing 11 (22.9%) responses with complete and transparent citations. Mistral-7B was the least transparent LLM, with 45 (93.4%) of responses not providing any reference or source to support the recommendation.

DISCUSSION AND CONCLUSION: More than one-in-four recommendations provided by leading commercially-available LLMs concerning the evaluation and management of rotator cuff and ACL injuries are not concordant with current evidenced-based CPGs. Although ChatGPT-4 demonstrated the highest performance, clinically significant rates of recommendations without concordance or supporting evidence exist across all the evaluated LLMs. Only 10% of responses by LLMs were transparent and provided complete, evidence-based resources, precluding users from fully understanding and interpreting the sources from which recommendations were provided. Future academic research should include analyses of multiple LLMs beyond ChatGPT given that performance may vary; in the interim, while all leading LLMs generally provide recommendations concordant with CPGs, the substantial proportion of recommendations that do not align with these guidelines suggest that LLMs are not adequate clinical support tools.



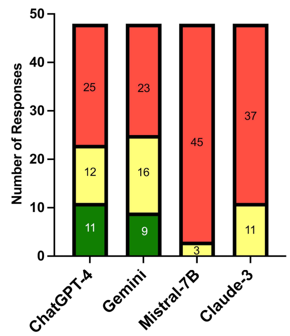


Table IA. American Academy of Orthopaedic Surgeons (AAOS) Clinical Practice Guidelines (CPGs) for the Management of Rotator Cuff Injuries.

Subject	AAOS Recommendation	Strength of Recommendation
Management of Small to Medium Full-Thickness Tears*	Strong evidence supports that both physical therapy and operative treatment result in significant improvement in patient-reported outcomes for patients with symptomatic small to medium full-thickness rotator cuff tears.	Strong
Management of High-Grade Partial Thickness Tears*	Strong evidence supports the use of either conversion to full-thickness or transtendinous/in-situ repair in patients that failed conservative management with high-grade partial thickness rotator cuff tears.	Strong
Long-term Non-operative Management*	Strong evidence supports that patient reported outcomes improve with physical therapy in symptomatic patients with full thickness rotator cuff tears. However, the rotator cuff tear size, muscle atrophy, and fatty infiltration may progress over 5 to 10 years with non-operative management.	Strong
Diagnosis (Clinical Examination)*	Strong evidence supports that clinical examination can be useful to diagnose or stratify patients with rotator cuff tears; however, combination of tests will increase diagnostic accuracy.	Strong
Diagnosis (Imaging)*	Strong evidence supports that MRI, MRA, and ultrasound are useful adjuncts to a clinical exam for identifying rotator cuff tears.	Strong
Post-operative Mobilization Timing*	Strong evidence suggests similar postoperative clinical and patient-reported outcomes for small to medium sized full-thickness rotator cuff tears between early mobilization and delayed mobilization up to 8 weeks for patients who have undergone arthroscopic rotator cuff repair.	Strong
Prognostic Factors (Age)*	Strong evidence supports that older age is associated with higher failure rates and poorer patient reported outcomes after rotator cuff repair.	Strong
Prognostic Factors (Worker's Compensation)*	Strong evidence supports the presence of a worker's compensation claim is associated with poorer patient reported outcomes after rotator cuff repair.	Strong
Biologic Augmentation with Platelet Derived Products*	Strong evidence does not support biological augmentation of rotator cuff repair with platelet-derived products on improving patient reported outcomes; however, limited evidence supports the use of liquid platelet rich plasma (PRP) in the context of decreasing re-tear rates.	Strong
Single-Row vs. Double-Row Repair*	Strong evidence does not support double row rotator cuff repair constructs on improving patient-reported outcomes compared to single row vertical mattress repair constructs.	Strong
Single-Row vs. Double-Row Repairs for Re-Tears*	Strong evidence supports lower re-tear rates after double row repair compared to single row vertical mattress repair when evaluating for both partial and full thickness retears after primary repair; however, when evaluating the data for only full thickness retears, limited evidence does not support lower re-tear rates after double row primary repair.	Strong
Open vs. Arthroscopic Repair*	Strong evidence supports no difference in long-term (> 1 year) patient-reported outcomes or cuff healing rates between open and arthroscopic repairs; however, arthroscopic-only technique	Strong

Table 2. Large language model text-query inputs and grading of concordance with American Academy of Orthopedic Surgeons (AAOS) clinical practice guidelines. + indicates peer-reviewed reference cited (completely or incompletely); - indicates no reference cited. C, concordance; I, indeterminate; D, discordant.

AAOS Recommendation	Query	ChatGPT-4	Gemini	Mistral-7B	Claude-3
Rotator Cuff Injuries					
Management of Small to Medium Full-Thickness Tears	Are both physical therapy and operative treatment recommended for small to medium full-thickness rotator cuff tears?	C/-	C/+	C/-	C/-
Management of High-Grade Partial Thickness Tears	Is conversion to a full-thickness tear recommended for patients with high-grade partial rotator cuff tears?	C/+	D/-	I/-	I/-
Long-term Non-operative Management	Is long-term nonoperative management recommended for patients that have full-thickness rotator cuff tears?	C/-	C/-	I/-	D/-
Diagnosis (Clinical Examination)	Is a clinical exam recommended to diagnose patients with rotator cuff tears?	C/-	C/-	C/-	C/-
Diagnosis (Imaging)	Is imaging, such as MRI, MRA, and ultrasound, recommended to diagnose patients with rotator cuff tears?	C/+	C/-	C/-	C/-
Post-operative Mobilization Timing	Are both early and delayed mobilization recommended for patients after rotator cuff repair?	C/+	D/+	I/-	D/-
Prognostic Factors (Age)	Is older age associated with higher re-tear rates after rotator cuff repair?	C/+	C/-	C/-	C/+
Prognostic Factors (Worker's Compensation)	Is worker's compensation associated with higher re-tear rates after rotator cuff repair?	D/+	I/+	I/-	I/+
Biologic Augmentation with Platelet Derived Products	Is biological augmentation of rotator cuff repairs with platelet-derived products recommended for improving patient reported outcomes?	I/+	I/+	I/+	I/+
Single-Row vs. Double-Row Repair	Are both the single-row and double-row repair techniques recommended for improving patient reported outcomes after rotator cuff repair?	C/+	C/-	C/-	C/+
Single-Row vs. Double-Row Repairs for Re-Tears	Are both the single-row and double-row repair techniques recommended for preventing re-tears after rotator cuff repair?	I/+	C/+	I/-	I/+
Open vs. Arthroscopic Repair	Are both the open and arthroscopic rotator cuff repair techniques recommended for improving patient reported outcomes, range of motion, or pain?	C/-	C+	C/-	C/+
Operative Management	Is operative management of rotator cuff tears recommended to improve outcomes compared to physical therapy and unhealed rotator cuff repairs?	I/-	C/-	I/-	I/-
Acromioplasty and Rotator Cuff Repair	Is the routine use of acromioplasty during arthroscopic rotator cuff repair recommended for patients with small to medium sized rotator cuff tears?	I/+	C/+	C/+	C/+