Large Language Models and Recommendations for Rotator Cuff Injuries: Risks of Discordance with AAOS Guidelines

Benjamin R Caruso, Suzanna Marie Ohlsen, Jaewon (Freddy) Yang, Albert Ooguen Gee

INTRODUCTION: Artificial intelligence (AI) is increasingly accessible to the public, with large language models (LLMs) such as ChatGPT and Gemini commonly used by patients to obtain medical information. These tools have the potential to influence patient understanding, guide decision-making, and affect clinical outcomes. However, the accuracy of the medical guidance provided by LLMs remains uncertain. This study aims to evaluate and compare the responses of ChatGPT and Gemini to the American Academy of Orthopaedic Surgeons (AAOS) Clinical Practice Guidelines (CPG) for the management of rotator cuff injuries.

METHODS: Both ChatGPT and Gemini were queried regarding 21 treatment recommendations of strong or moderate strength evidence from the 2022 AAOS CPG “Management of Rotator Cuff Injuries.” Two blinded reviewers independently evaluated each response and classified it as “Concordant,” “Discordant,” or “Neutral” with respect to the CPG recommendation. Discrepancies were resolved by a third blinded reviewer. Cohen’s Kappa was used to assess inter-rater agreement, and chi-square analysis was planned to compare overall agreement rates between models.

RESULTS: Of the 21 AAOS CPG recommendations evaluated, ChatGPT was concordant with 12 (57.1%), discordant with 2 (9.5%), and neutral on 7 (33.3%). ChatGPT cited 43 PubMed-indexed studies across its responses, of which 31 (72.1%) were deemed relevant and supportive of its claims.

Gemini was concordant with 14 of 21 recommendations (66.7%), discordant with 2 (9.5%), and neutral on 5 (23.8%). Gemini also cited 43 PubMed-indexed studies, 35 (81.4%) of which were judged relevant. The remaining references were either confabulated, inaccurately cited or did not support the claims made by ChatGPT or Gemini.

Inter-rater reliability was substantial, with a Cohen’s Kappa coefficient of 0.7685. Chi-square analysis revealed no significant difference between models ($p = 0.7838$).

DISCUSSION AND CONCLUSION:

This study compared two large language models, ChatGPT and Gemini, in their adherence to AAOS Clinical Practice Guidelines for rotator cuff injuries. While both demonstrated a majority of concordant responses (57.1% for ChatGPT and 66.7% for Gemini), a substantial proportion of responses either failed to align with guideline-based recommendations or were classified as neutral. This level of non-concordance may pose risks in clinical contexts, particularly when LLMs are used by patients to inform medical decisions.

Although both models referenced the same number of PubMed-indexed studies (43), the proportion of studies deemed relevant differed: 72.1% for ChatGPT and 81.4% for Gemini. Gemini also demonstrated slightly greater alignment with the AAOS guidelines. These findings suggest that while both models are capable of generating medically grounded responses, variability remains in both the accuracy of recommendations and the relevance of supporting citations.

These results reinforce the need for caution when using LLMs in patient-facing roles. Despite their potential as educational aids, their inconsistent alignment with evidence-based guidelines underscores the importance of clinical oversight. Future development of LLMs should prioritize guideline concordance, citation transparency, and the implementation of safeguards to prevent the dissemination of misleading or vague medical information.