Chatgpt-4.1 Has Potential to Triage & Diagnose Complications Following Common Shoulder Procedures When Appropriately Prompted

Charles J. Patterson[1], Ryan T. Lin[1], Tyler C. Williams[1], Sahil Dadoo[2], Tyler Paras[3], Jonathan D. Hughes[2], Albert Lin[2]
[1]Univeristy of Pittsburgh School of Medicine, Pittsburgh, PA, [2]University of Pittsburgh Medical Center Department of Orthopedic Surgery, University of California San Diego, San Diego, CA
Pattersoncj2@upmc.edu

Disclosures: N/A

INTRODUCTION: Recently, large language models (LLMs) have demonstrated impressive performances in triaging, diagnosing, and generating treatment plans for traumatic and soft tissue injuries, as well as improving education in many medical specialties, especially orthopedics. However, even though ChatGPT is the most widely used LLM, its performance involving post-operative shoulder complications remains unexamined. This study evaluates the ability of ChatGPT-4.1 to formulate accurate diagnoses and treatment options when prompted with triage questions and clinical vignettes. We hypothesized that ChatGPT-4.1 will provide clinically appropriate and accurate differential diagnoses and management recommendations in most scenarios.

METHODS: Twelve triage stems and twelve clinical vignettes with corresponding imaging were derived from common shoulder procedures. Prompts were input to ChatGPT-4.1 as either triage plus imaging with instructions to generate the three most likely diagnoses, or clinical vignette plus imaging with instructions to generate the most likely diagnosis and the best next step for treatment. Two fellowship-trained shoulder surgeons graded ChatGPT4.1 responses and provided their own responses. Accuracy represents the percentage of overlap between surgeons and ChatGPT-4.1 diagnoses. Appropriateness was rated by surgeons on a scale from 0-2: 0 = inappropriate; 1 = somewhat appropriate; 2 = appropriate. This study was approved by IRB.

RESULTS SECTION: Triage plus imaging demonstrated 65% accuracy and a 1.45 appropriateness score. ChatGPT4.1 matched the surgeons' top diagnosis in 58% of triage scenarios and listed it first or second in 75% of triage scenarios. Clinical vignette plus imaging demonstrated 79% accuracy and an appropriateness score of 1.54.

DISCUSSION: ChatGPT4.1 demonstrated the potential to triage, diagnose, and provide next steps for common shoulder surgery complications. The increased detail of the clinical vignettes compared to the triage stems may account for the higher accuracy and appropriateness of ChatGPT4.1's clinical vignette responses. While ChatGPT4.1 is a valuable tool for patients, it can fall short of identifying the intricacies of imaging and patient history if not appropriately prompted.

SIGNIFICANCE/CLINICAL RELEVANCE: (1-2 sentences): This study demonstrates that ChatGPT-4.1 can provide clinically appropriate support in diagnosing and managing postoperative shoulder complications, offering a potential decision-support tool for orthopedic surgeons while also improving patient access to timely guidance and communication in their care.

REFERENCES:
1. Chatterjee, S., Bhattacharya, M., Pal, S., Lee, S. S., & Chakraborty, C. (2023). ChatGPT and large language models in orthopedics: from education and surgery to research. *Journal of Experimental Orthopaedics*, 10(1), 128.
2. Kunze, K. N., Varady, N. H., Mazzucco, M., Lu, A. Z., Chahla, J., Martin, R. K., ... & Williams Iii, R. J. (2025). The large language model ChatGPT-4 exhibits excellent triage capabilities and diagnostic performance for patients presenting with various causes of knee pain. Arthroscopy: *The Journal of Arthroscopic & Related Surgery*, 41(5), 1438-1447.
3. Picazo-Sanchez, P., & Ortiz-Martin, L. (2024). Analysing the impact of ChatGPT in research. *Applied Intelligence*, 54(5), 4172-4188.

ACKNOWLEDGEMENTS: The authors would like to acknowledge UPMC Sports Medicine and the Pittsburgh Shoulder Institute for assistance with abstract preparation.