

Evaluating the Reliability, Accuracy, and Readability of ChatGPT-3.5 and GPT-4 in Providing Patient Education Related to Glenohumeral Joint Osteoarthritis

MOHAMAD Y. FARES, MD; TARISHI PARMAR, BSc; PETER BOUFADEL, MD; MOHAMMAD DAHER, MD; JONATHAN BERG, BSc; ADAM Z. KHAN, MD; BRIAN W. HILL, MD; JOHN G. HORNEFF, MD; JOSEPH A. ABBOD, MD

ABSTRACT

OBJECTIVES: Chatbots have been increasingly recognized as modern tools that provide patients with reliable health-related information. This study aimed to evaluate and compare ChatGPT-3.5 and GPT-4's ability to answer glenohumeral osteoarthritis-related questions.

METHODS: Fifteen questions were derived from the 2020 AAOS Clinical Practice Guidelines for the Surgical Management of Glenohumeral Joint Osteoarthritis. Questions were categorized into three groups: risk factors, implant/intraoperative considerations, and pain/functional outcomes. ChatGPT-3.5 and GPT-4 were prompted with these questions, and responses were evaluated by four fellowship-trained shoulder and elbow surgeons. Each response was rated on a scale (scores:1-5) based on relevance, accuracy, clarity, completeness, and evidence-based support. Data was analyzed descriptively and statistically to compare the scores between ChatGPT-3.5 and GPT-4.

RESULTS: Average score for ChatGPT-3.5 was 19.7/25, with "Risk Factor" prompts achieving the highest mean score. GPT-4 averaged 18.7/25, with "Functional Outcomes" prompts scoring highest. However, there were no statistically significant differences between different prompt themes for GPT-3.5 and GPT-4. "Clarity" category received the highest score for GPT-3.5, while "Relevance" was highest for GPT-4. Both models scored lowest on "Evidence-based" prompts. On the Flesch-Kincaid scale, GPT-3.5 responses had a significantly higher score of 18.3 compared to GPT-4's 15.4, indicating a more difficult reading level in GPT-3.5's responses.

CONCLUSION: Both ChatGPT-3.5 and GPT-4 performed adequately in providing well-informed medical responses to patient queries about glenohumeral osteoarthritis. Future chatbot versions should focus on providing evidence-based content through systematic and reliable reviews of literature, in an accessible readable manner.

KEYWORDS: artificial intelligence; chatbot; patients; health information; Flesch-Kincaid

INTRODUCTION

One of the oldest goals of computer science: "to create a machine that imitates intelligent human behavior" is the definition of modern-day artificial intelligence (AI).^{1,2} Among the spectrum of generative AI models, ChatGPT has emerged as a revelation, owing to its sophisticated capabilities in handling intricately demanding tasks. Developed by OpenAI (San Francisco, California), ChatGPT has a multitude of uses, including promoting teaching and learning.³ On March 14, 2023, OpenAI also launched GPT-4, the latest version of ChatGPT: a multimodal model that accepts both images and text as input, thus furthering its usage.⁴ As expected, the benefits of ChatGPT do not come without the shortcomings. The potential drawbacks of ChatGPT, such as fabricated references, biased training data, outdated knowledge, factual inaccuracies, and inconsistency in quality, raise crucial concerns about its effectiveness in answering questions.^{5,6}

From providing medical facts to patients,⁷ to providing answers and explanations for examination questions to medical students,⁸ to helping doctors make diagnoses,⁹ ChatGPT is being increasingly used in the field of medicine. However, considering the limitations of ChatGPT, its use in healthcare has been a topic of skepticism.⁷⁻⁹ Consequently, numerous research projects have been undertaken to assess ChatGPT's reliability, efficiency, and accuracy in medical contexts.¹⁰⁻¹⁵ Studies assessing ChatGPT's capacity to address medical school-level questions have found it capable of answering at a level comparable to a third-year medical student.¹¹ Further ChatGPT's expertise has been put to test by researchers around the world in a range of medical subspecialties like plastic surgery,¹² ophthalmology,^{16,17} obstetrics/gynecology,¹⁸ neurosurgery,^{19,20} orthopedic surgery and others.^{13-15,21,22} Studies have yielded mixed and variable results highlighting the need for niche, subspecialty and topic specific evaluation of ChatGPT's skills.

Although research has examined ChatGPT's efficacy in responding to orthopedic questions,^{14,15,23-26} there has been limited attention directed specifically towards questions related to shoulder and elbow orthopedics. Furthermore, a majority of these evaluations have focused on inquiries typical of medical students or healthcare professionals, overlooking questions from the patient's viewpoint. Despite an exhaustive literature review, no studies have been identified

that investigate ChatGPT's responses to inquiries patients might pose about glenohumeral osteoarthritis or total shoulder arthroplasty (TSA) procedure. With an anticipated surge in shoulder arthroplasty procedures to approximately 350,558 in 2025, and the growing reliance of patients on ChatGPT for medical advice,²⁷ it is crucial to assess ChatGPT's capability to accurately and consistently address patient queries concerning the TSA procedure. Therefore, the purpose of this study was to evaluate and compare ChatGPT-3.5 and GPT-4's ability to answer questions adopted from the 2020 American Academy of Orthopaedic Surgeons (AAOS) Clinical Practice Guidelines for the Surgical Management of Glenohumeral Joint Osteoarthritis.²⁸

METHODS

The methodology was adopted, in part, from a similar study assessing ChatGPT's ability to answer orthopedic-related questions.¹⁴ The questions were derived from the 2020 AAOS Clinical Practice guidelines for the Surgical Management of Glenohumeral Joint Osteoarthritis.²⁸ Fifteen of the 22 reviewed recommendations were reformulated by

MYF and TP into question formats reflecting a patient's viewpoint, i.e., framed as patient-posed inquiries while preserving the original clinical content. The seven excluded recommendations were primarily technical or intraoperative in nature and could not be meaningfully reformulated as patient-posed questions. Consequently, the questions were organized into three primary categories according to the theme of the prompt: 1) risk factors: assessing the impact of preoperative risk factors on postoperative results; 2) implant/intraoperative: considerations pertaining to implants and intraoperative choices; and 3) pain/functional outcome: interventions influencing postoperative pain levels and functional outcomes [Table 1]. Attention was given to ensure incorporation of questions encompassing diverse levels of AAOS recommendations: 5 strong, 4 moderate, 1 limited, and 5 consensus.

Each prompt was provided to both ChatGPT-3.5 and ChatGPT-4 and the responses were recorded using Microsoft Excel (state of Washington). Each prompt included an explicit requirement that the information provided be supported by cited sources. All queries were conducted between May 25 and May 30, 2024 using the ChatGPT-web-based

Table 1. Questions derived from the 2020 AAOS Clinical Practice guidelines for the surgical management of glenohumeral joint osteoarthritis

	QUESTION		AAOS RECOMMENDATION	STRENGTH OF AAOS RECOMMENDATION
RISK FACTORS	1	Do obese patients with glenohumeral osteoarthritis experience an increase in rate of post-operative complications?	No	Strong
	2	Is gender/sex associated with better or worse post-operative outcomes?	No	Strong
	3	Do patients with glenohumeral joint osteoarthritis who have more comorbidities experience higher rates of early post-arthroplasty complications?	Yes	Strong
	4	Is smoking associated with inferior post-operative outcomes?	Yes	Moderate
	5	Will patients with higher pre-operative function experience less functional improvement than patients with lower pre-operative outcomes?	Yes	Moderate
	6	Is depression associated with inferior post-operative outcomes in patients with glenohumeral joint osteoarthritis undergoing arthroplasty?	Yes	Moderate
INTRAOPERATIVE	7	Can surgeons utilize subscapularis peel, lesser tuberosity osteotomy, or tenotomy when performing shoulder arthroplasty?	Yes	Moderate
	8	Can surgeons utilize pegged or keeled glenoid components in patients with glenohumeral joint osteoarthritis and a well-functioning rotator cuff?	Yes	Strong
	9	Can clinicians utilize stemmed, stemless or resurfacing prosthesis for patients with glenohumeral joint osteoarthritis undergoing total or hemi-arthroplasty?	Yes	Limited
	10	Would use of tranexamic acid during shoulder arthroplasty result in reduced blood loss and reduced risk of blood transfusion?	N/A	Consensus
FUNCTIONAL OUTCOME	11	Is there a benefit associated to the use of hyaluronic acid in the treatment of glenohumeral joint osteoarthritis?	No	Strong
	12	Should post-operative physical therapy be recommended to patients following shoulder arthroplasty?	N/A	Consensus
	13	Should opioids be prescribed as routine and long-term pain management of glenohumeral osteoarthritis?	N/A	Consensus
	14	Should anatomic total shoulder arthroplasty be performed on patients with small, isolated, repairable supraspinatus tears?	N/A	Consensus
	15	Is same-day discharge an option after shoulder arthroplasty in select patients?	N/A	Consensus

interface. Each prompt was entered individually in a new chat session to prevent contextual carryover between queries. Chat history was cleared after each prompt to further ensure independence of responses. Outputs were generated once per prompt without repeated runs. In order to evaluate the chatbots' responses to each question, a rating scale based on five categories was used: relevance, accuracy, clarity, completeness and evidence-based. Relevance referred to the correlation of the answer to the posted inquiry; accuracy referred to the alignment of the answer to the AAOS guidelines; clarity referred to the ease of comprehension; completeness referred to the comprehensiveness of the provided response and evidence-based referred to the authentication of claims with relevant citations provided by the bots. Four shoulder and elbow fellowship-trained surgeons were asked to evaluate both ChatGPT-3.5 and ChatGPT-4's responses to every question, and the surgeons were blinded as to which chatbot provided which response. Possible scores for every attribute ranged between 1 and 5, and as such, the overall score of the responses ranged between 5 and 25. The description of each attribute was mentioned at the top of the forms for the surgeons to increase convenience and minimize personal bias in interpreting the attributes. Mean scores were derived by calculating the average scores from every surgeon for every chatbot response. In addition, the Flesch-Kincaid grade level was calculated to evaluate the readability of each chatbot's response to each prompt.

Derived scores from the shoulder and elbow surgeons were recorded and analyzed. Descriptive statistics were used to describe the scores of the chatbots' responses to the derived prompts. Analysis of variance (ANOVA) testing with a post-hoc Tukey's test was then conducted to compare the scores between different question categories and different scoring categories for both ChatGPT-3.5 and GPT-4. An independent t-test was used to compare the overall response scores between ChatGPT-3.5 and GPT-4, as well as the overall readability between the two chatbots. A P-value less than 0.05 was considered significant. All statistical analyses were conducted using the statistical package for the social sciences for Windows software version 25.0 (IBM SPSS, 2017).

RESULTS

ChatGPT-3.5

The average score of a prompt response by ChatGPT-3.5 was 19.7/25, with question 10 achieving the highest mean score at 22.3, and question 13 achieving the lowest mean score at 15.5/25 [Table 2].

Mean score was highest for "Risk Factor" prompts scoring at 4.15/5. This was followed by "Intraoperative" prompts with a score of 3.9/5, and "Functional Outcomes" prompts with a score of 3.9/5. We have provided

additional details about the average scores of ChatGPT-3.5 for the different prompt themes in our study [Table 3]. There was no statistical significance when exploring the differences of scores between different prompt themes in ChatGPT-3.5 ($p=0.887$).

Overall, the average ChatGPT-3.5 response in our study received a mean score of 4.7/5 on relevance, 4.2/5 on accuracy, 4.8/5 on clarity, 3.9/5 on completeness, and 2.5/5 on being evidence-based. Generally, the average ChatGPT-3.5's response scored lower on being evidence-based than on relevance ($p=0.001$), accuracy ($p=0.01$), clarity ($p<0.001$), and completeness ($p=0.027$).

On average, ChatGPT-3.5 response received a Flesch-Kincaid grade level of 18.3 (range: 15.4-20.4), indicating a reading level pertaining to a college graduate (very difficult to read).

Table 2. Average Prompt Score for both ChatGPT and GPT-4

Questions	ChatGPT-3.5	GPT-4
1	19.5	16.25
2	20.25	12.75
3	19.75	19
4	21.5	19.25
5	21	18.25
6	22.5	19.5
7	16.75	14.25
8	17.75	19.5
9	20.25	18
10	22.25	21
11	21.5	21.25
12	20.75	19.25
13	15.5	20.75
14	16.75	19.5
15	19	21.75
Total Average	19.7	18.7

Table 3. Performance of ChatGPT-3.5 when responding to AAOS prompts according to type of question and rating category

ChatGPT-3.5	Risk Factors		Intraoperative		Functional Outcomes	
	Average	Standard Deviation	Average	Standard Deviation	Average	Standard Deviation
Relevance	4.88	0.16	4.56	0.31	4.40	0.28
Accuracy	4.45	0.67	3.75	0.79	4.19	0.47
Clarity	4.78	0.18	4.88	0.14	4.68	0.28
Completeness	3.92	1.24	3.81	1.23	4.04	0.86
Evidence-based	2.71	0.72	2.38	0.92	2.35	0.93
Average	4.15		3.88		3.93	

GPT-4

The average score for a prompt response by GPT-4 was 18.7/25, with Question 15 achieving the highest mean score at 21.8/25 and Question 2 achieving the lowest mean score at 12.8/25 [Table 2]. While the average prompt response score for ChatGPT-3.5 was higher than that of GPT-4, this was not deemed significant on independent t-test analysis ($p=0.239$; MD: 1.0233).

Mean score was highest for “Functional Outcomes” prompts scoring at 4.1/5. This was followed by “Intraoperative” prompts with a score of 3.7/5, and “Risk Factors” prompts with a score of 3.5/5. We have provided additional details about the average scores of GPT-4 for the different prompt themes in our study [Table 4]. There was no statistical significance when exploring the differences of scores between different prompt themes in GPT-4 ($p=0.485$). There was also no statistically significant differences in mean scores between ChatGPT-3.5 and GPT-4 when responding to “Risk Factors” prompts ($p=0.285$), “Intraoperative” prompts ($p=0.751$), and “Functional Outcomes” prompts ($p=0.759$).

Overall, the average GPT-4 response in our study received a mean score of 4.3/5 on relevance, 4.0/5 on accuracy, 4.0/5 on clarity, 3.1/5 on completeness, and 2.1/5 on being evidence-based. Generally, the average GPT-4’s response scored lower on being evidence-based than on relevance ($p=0.003$), accuracy ($p=0.007$), and clarity ($p=0.008$). In addition, there was also no statistically significant differences in mean scores between ChatGPT-3.5 and GPT-4’s responses according to relevance ($p=0.097$), accuracy ($p=0.326$), clarity ($p=0.078$), completeness ($p=0.307$), and being evidence-based ($p=0.312$).

On average, GPT-4’s response received a Flesch-Kincaid grade level of 15.4 (range: 12.7–17.6), indicating a reading level pertaining to a college graduate (very difficult to read). That being said, the Flesch-Kincaid grade level of GPT-3.5’s responses was significantly higher than that of GPT-4’s responses ($p<0.001$), indicating a more difficult reading level in GPT-3.5’s responses.

Table 4. Performance of ChatGPT-3.5 when responding to AAOS prompts according to type of question and rating category

ChatGPT-4	Risk Factors		Intraoperative		Functional Outcomes	
	Average	Standard Deviation	Average	Standard Deviation	Average	Standard Deviation
Relevance	4.25	0.62	4.25	0.54	4.50	0.53
Accuracy	4.04	0.37	3.85	0.27	4.45	0.50
Clarity	4.00	0.71	4.23	0.49	4.60	0.43
Completeness	3.13	0.70	3.56	0.31	4.15	0.72
Evidence-based	2.08	0.88	2.63	0.78	2.80	1.05
Average	3.50		3.70		4.10	

DISCUSSION

With the increasing reliance on ChatGPT by patients for medical advice across various medical specialties, it becomes essential to assess its credibility and efficiency, particularly in niche subspecialties and specific topics.²⁹ Moreover, as new models of ChatGPT gain popularity, it is important to compare their performance in addressing these questions, in order to highlight areas of improvement and address noteworthy limitations. Our study evaluated and compared the performance of ChatGPT-3.5 and GPT-4 in answering questions related to the surgical management of glenohumeral joint osteoarthritis from a patient’s perspective. The findings indicate that both models provide moderate-quality responses, with average scores of 19.7/25 for ChatGPT-3.5 and 18.7/25 for GPT-4. These scores reflected interesting perceptions regarding the relevance, accuracy, clarity, completeness, and evidence-based support of ChatGPT responses.

Interestingly, ChatGPT-3.5 performed slightly better overall, with better readability scores than their more advanced counterpart.

Both ChatGPT-3.5 and GPT-4 performed similarly well in our study, with no significant differences overall ($p=0.239$). It was seen that responses related to “Risk factors” questions had highest scores for ChatGPT-3.5, while those relating to “Functional Outcomes” had highest scores for GPT-4. Moreover, it was noted that both chatbots had significantly lower scores on being “evidence-based” when compared to other scoring categories. This has multiple implications on performance. Chatbots derive responses to scientific inquiries based on the available scientific literature online.^{30,31} Hence, limitations in responses reflect either weaknesses in the available literature or shortcomings in the processing methods by which chatbots screen and gather information from scientific sources.^{30, 31} In this setting, a new more specialized chatbot powered by advanced AI technology, termed Consensus has emerged, focusing particularly on enhancing the academic research experience, using improved search capabilities, and employing a research-oriented interface.³² Perhaps future studies would prove the superior efficacy and accuracy of this chatbot when compared to its predecessors.

When assessing readability of responses based on the Flesch Kincaid grading level, both chatbots showcased reading levels that may not be accessible to patients with lower health literacy. That being said, ChatGPT-4 demonstrated a lower Flesch-Kincaid grading level than GPT-3.5, suggesting more readable content, which may provide more accessible information for a broader audience. This highlights important areas of improvement for future chatbots, while improving accuracy

and reliability of information is essential, doing so in an accessible and more readable manner is of major importance.

Our study showed that Chat GPT-3.5 performed better on average than GPT-4, though the difference was not significant overall. This finding is notable because GPT-4 is reportedly significantly more capable of solving difficult problems with better accuracy due to improved reasoning and a much larger base of learned information.³³ This result also falls in contrast with previous studies comparing the performance of both GPT-models in responding to medical questions. GPT-4 was found to answer more accurately on orthopaedic assessment examinations than ChatGPT-3.5.³⁴ GPT-4 was also found to answer Orthopedic In-Training Examination (OITE) questions with significantly better accuracy than ChatGPT-3.5.³⁵ For other medical specialties too, GPT-4 has performed significantly better than GPT-3.5.³⁶⁻³⁸

To our knowledge, this is the first study to assess ChatGPT-responses to patient-related queries targeting glenohumeral osteoarthritis. Nevertheless, several limitations exist. We asked shoulder surgery experts to rate ChatGPT's answers, introducing subjectivity as a potential limitation of our study. To address this, we provided detailed criteria for each characteristic and score to ensure a consistent evaluation. Additionally, we minimized bias between GPT-3.5 and GPT-4 by blinding the surgeons to which responses corresponded to which model. The experts were also provided with the AAOS recommendations and their strengths for each question to guide their assessments.

Beyond evaluator subjectivity, this study is subject to several limitations inherent to large language model-based analyses. Model outputs may vary over time due to temporal drift as underlying models are updated, which may limit reproducibility. Responses may also differ across repeated runs for the same prompt. Since our methods did not include repeated sampling, within-prompt variability could not be quantified. Additionally, LLM outputs are sensitive to prompt phrasing, and alternative wording may have yielded different responses despite similar intent. These factors should be considered when interpreting the findings and when comparing results across studies or time points.

CONCLUSION

While both ChatGPT-3.5 and GPT-4 demonstrate substantial potential in providing well-informed medical responses to patient queries about glenohumeral osteoarthritis, enhancements remain necessary and essential to ensure the reliability and accessibility of their answers. Future versions should focus on incorporating more evidence-based content and more accessible language to cater to diverse patient populations with varying levels of health literacy. Additionally, patients should be mindful of the limitations of current chatbots when considering the medical advice and information provided by it. Further research should explore future iterations of ChatGPT and assess their ability to respond to queries related to other niche medical subspecialties, in an aim to evaluate improvement and ensure continuous growth.

References

1. Rawassizadeh R, Sen T, Kim SJ, et al. Manifestation of virtual assistants and robots into daily life: vision and challenges. *CCF Transactions on Pervasive Computing and Interaction*. 2019;1:163-174. DOI: 10.1007/s42486-019-00014-1.
2. Feuerriegel S, Hartmann J, Janiesch C, Zschech P. Generative AI. *Business @amp; Information Systems Engineering*. 2023;66:111-126. DOI: 10.1007/s12599-023-00834-7.
3. BalDoo-Anu D, Owusu Ansah L. Education in the Era of Generative Artificial Intelligence (AI): Understanding the Potential Benefits of ChatGPT in Promoting Teaching and Learning. *Journal of AI*. 2023;7:52-62. DOI: 10.61969/jai.1337500.
4. Aydin Ö, Karaarslan E. Is ChatGPT Leading Generative AI? What is Beyond Expectations? *Academic Platform Journal of Engineering and Smart Systems*. 2023;11:118-134. DOI: 10.21541/apjess.1293702.
5. Ray PP. ChatGPT: A comprehensive review on background, applications, key challenges, bias, ethics, limitations and future scope. *Internet of Things and Cyber-Physical Systems*. 2023;3:121-154. DOI: 10.1016/j.iotcps.2023.04.003.
6. Tyson J. Shortcomings of ChatGPT. *Journal of Chemical Education*. 2023;100:3098-3101. DOI: 10.1021/acs.jchemed.3c00361.
7. Walker HL, Ghani S, Kuemmerli C, et al. Reliability of Medical Information Provided by ChatGPT: Assessment Against Clinical Guidelines and Patient Information Quality Instrument. *J Med Internet Res*. 2023;25: e47479-e47479. DOI: 10.2196/47479.
8. Jeyaraman M, K SP, Jeyaraman N, et al. ChatGPT-in Medical Education and Research: A Boon or a Bane? *Cureus*. 2023;15: e44316-e44316. DOI: 10.7759/cureus.44316.
9. Mehnen L, Gruarin S, Vasileva M, Knapp B. ChatGPT as a medical doctor? A diagnostic accuracy study on common and rare diseases. Cold Spring Harbor Laboratory, 2023.
10. IStrong E, DiGiammarino A, Weng Y, et al. Chatbot vs Medical Student Performance on Free-Response Clinical Reasoning Examinations. *JAMA Intern Med*. 2023;183:1028-1030. DOI: 10.1001/jamainternmed.2023.2909.
11. Gilson A, Safranek CW, Huang T, et al. How Does ChatGPT Perform on the United States Medical Licensing Examination (USMLE)? The Implications of Large Language Models for Medical Education and Knowledge Assessment. *JMIR Med Educ*. 2023;9:e45312-e45312. DOI: 10.2196/45312.
12. Gupta R, Herzog I, Park JB, et al. Performance of ChatGPT on the Plastic Surgery Inservice Training Examination. *Aesthetic Surgery Journal*. 2023;43:NP1078-NP1082. DOI: 10.1093/asj/sjad128.
13. Daher M, Koa J, Boufadel P, et al. Breaking barriers: can ChatGPT compete with a shoulder and elbow specialist in diagnosis and management? *JSES Int*. 2023;7:2534-2541. DOI: 10.1016/j.jseint.2023.07.018.
14. Magruder ML, Rodriguez AN, Wong JCJ, et al. Assessing Ability for ChatGPT to Answer Total Knee Arthroplasty-Related Questions. *The Journal of Arthroplasty*. 2024. DOI: 10.1016/j.arth.2024.02.023.
15. Lum ZC. Can Artificial Intelligence Pass the American Board of Orthopaedic Surgery Examination? Orthopaedic Residents Versus ChatGPT. *Clin Orthop Relat Res*. 2023;481:1623-1630. 2023/05/23. DOI: 10.1097/CORR.0000000000002704.
16. Antaki F, Touma S, Milad D, et al. Evaluating the Performance of ChatGPT in Ophthalmology: An Analysis of Its Successes and Shortcomings. *Ophthalmol Sci*. 2023;3: 100324-100324. DOI: 10.1016/j.xops.2023.100324.
17. Balas M, Ing EB. Conversational AI Models for ophthalmic diagnosis: Comparison of ChatGPT and the Isabel Pro Differential Diagnosis Generator. *JFO Open Ophthalmology*. 2023;1:100005. DOI: 10.1016/j.jfop.2023.100005.
18. Grünebaum A, Chervenak J, Pollet SL, et al. The exciting potential for ChatGPT in obstetrics and gynecology. *American Journal of Obstetrics and Gynecology*. 2023;228:696-705. DOI: 10.1016/j.ajog.2023.03.009.
19. Kuang Y-R, Zou M-X, Niu H-Q, et al. ChatGPT encounters multiple opportunities and challenges in neurosurgery. *Int J Surg*. 2023;109:2886-2891. DOI: 10.1097/JS9.0000000000000571.

20. Hopkins BS, Nguyen VN, Dallas J, et al. ChatGPT versus the neurosurgical written boards: a comparative analysis of artificial intelligence/machine learning performance on neurosurgical board-style questions. *Journal of Neurosurgery*. 2023;139:904-911. DOI: 10.3171/2023.2.jns23419.
21. Kasthuri VS, Glueck J, Pham H, et al. Assessing the Accuracy and Reliability of AI-Generated Responses to Patient Questions Regarding Spine Surgery. *JBJS*. 2021;10:2106.
22. Chalhoub R, Mouawad A, Aoun M, et al. Will ChatGPT be able to replace a spine surgeon in the clinical setting? *World Neurosurgery*. 2024;185:e648-e652.
23. Fayed AM, Mansur NSB, de Carvalho KA, et al. Artificial intelligence and ChatGPT in Orthopaedics and sports medicine. *J Exp Orthop*. 2023;10:74-74. DOI: 10.1186/s40634-023-00642-8.
24. Giorgino R, Alessandri-Bonetti M, Luca A, et al. ChatGPT in orthopedics: a narrative review exploring the potential of artificial intelligence in orthopedic practice. *Front Surg*. 2023; 10: 1284015-1284015. DOI: 10.3389/fsurg.2023.1284015.
25. Kaarre J, Feldt R, Keeling LE, et al. Exploring the potential of ChatGPT as a supplementary tool for providing orthopaedic information. *Knee Surg Sports Traumatol Arthrosc*. 2023;31:5190-5198. DOI: 10.1007/s00167-023-07529-2.
26. Kuroiwa T, Sarcon A, Ibara T, et al. The Potential of ChatGPT as a Self-Diagnostic Tool in Common Orthopedic Diseases: Exploratory Study. *J Med Internet Res*. 2023;25:e47621-e47621. DOI: 10.2196/47621.
27. Wagner ER, Farley KX, Higgins I, et al. The incidence of shoulder arthroplasty: rise and future projections compared with hip and knee arthroplasty. *Journal of Shoulder and Elbow Surgery*. 2020;29:2601-2609. DOI: 10.1016/j.jse.2020.03.049.
28. Khazzam M, Gee AO, Pearl M. Management of Glenohumeral Joint Osteoarthritis. *Journal of the American Academy of Orthopaedic Surgeons*. 2020;28:781-789. DOI: 10.5435/jaaos-d-20-00404.
29. Clark M, Bailey S. Chatbots in Health Care: Connecting Patients to Information. *Canadian Journal of Health Technologies*. 2024;4.
30. Aleedy M, Shaiba H, Bezbradica M. Generating and analyzing chatbot responses using natural language processing. *International Journal of Advanced Computer Science and Applications*. 2019;10.
31. Liu J, Wang C, Liu S. Utility of ChatGPT in Clinical Practice. *J Med Internet Res*. 2023; 25:e48568-e48568. DOI: 10.2196/48568.
32. Consensus. AI Search Engine for Research, <https://consensus.app/> (2024).
33. Koubaa A. GPT-4 vs. GPT-3.5: A concise showdown. 2023.
34. Massey PA, Montgomery C, Zhang AS. Comparison of Chat GPT-3.5, ChatGPT-4, and Orthopaedic Resident Performance on Orthopaedic Assessment Examinations. *The Journal of the American Academy of Orthopaedic Surgeons*. 2023;31:1173-1179. 2023/09/04. DOI: 10.5435/JAAOS-D-23-00396.
35. Rizzo MG, Cai N, Constantinescu D. The performance of ChatGPT on orthopaedic in-service training exams: A comparative study of the GPT-3.5 turbo and GPT-4 models in orthopaedic education. *J Orthop*. 2024;50:70-75.
36. Lewandowski M, Łukowicz P, Świetlik D, Barańska-Rybak W. ChatGPT-3.5 and ChatGPT-4 dermatological knowledge level based on the Specialty Certificate Examination in Dermatology. *Clinical and Experimental Dermatology*. 2023. DOI: 10.1093/ced/llad255.
37. Lee GU, Hong DY, Kim SY, et al. Comparison of the problem-solving performance of ChatGPT-3.5, ChatGPT-4, Bing Chat, and Bard for the Korean emergency medicine board examination question bank. *Medicine (Baltimore)*. 2024;103:e37325-e37325. DOI: 10.1097/MD.00000000000037325.
38. Lee TJ, Rao AK, Campbell DJ, et al. Evaluating ChatGPT-3.5 and ChatGPT-4.0 Responses on Hyperlipidemia for Patient Education. *Cureus*. 2024;16: e61067-e61067. DOI: 10.7759/cureus.61067.

Authors

Mohamad Y. Fares, MD, Division of Shoulder and Elbow Surgery, Rothman Orthopaedic Institute, Philadelphia, PA.
 Tarishi Parmar, BSc, Penn State College of Medicine, Centre County, PA.
 Peter Boufadel, MD, Division of Shoulder and Elbow Surgery, Rothman Orthopaedic Institute, Philadelphia, PA.
 Mohammad Daher, MD, Brown University, Providence, RI.
 Jonathan Berg, BSc, Sidney Kimmel Medical College, Thomas Jefferson University, Philadelphia, PA.
 Adam Z. Khan, MD, Southern Permanente Medical Group, CA.
 Brian W. Hill, MD, Palm Beach Orthopaedic Institute, West Palm Beach, FL.
 John G. Horneff, MD, Division of Shoulder and Elbow Surgery, Department of Orthopaedics, University of Pennsylvania, Philadelphia, PA.
 Joseph A. Abboud, MD, Division of Shoulder and Elbow Surgery, Rothman Orthopaedic Institute, Philadelphia, PA.

Disclosures

Study Location: Research was performed at the Rothman Orthopaedic Institute, Philadelphia, PA.

Declaration of Conflicts of Interests: JAA would like to disclose: Royalties from a company or supplier; Disclosures; Osteocentric Technologies, Enovis, Zimmer-Biomet, Stryker, Globus Medical, Inc. Stocks in: Shoulder Jam, Aevumed, Oberd, Ots Medical, Orthobullets, Atreon, Restore 3D. Research support from a company or supplier as a PI; Disclosures; Enovis, Arthrex. Royalties, financial or material support from publishers; Disclosures; Wolters Kluwer, Slack Orthopaedics, Elsevier. Board member/committee appointments for a society; Disclosures; American Shoulder And Elbow Society, Mid Atlantic Shoulder And Elbow Society, Shoulder 360, Pacira.

Declaration of Funding: None

Declaration of Ethical Approval: This study does not involve any patients or patient-related information and as such, does not require Institutional Board Review approval.

Declaration of Informed Consent: Not applicable.

Correspondence

Mohamad Y. Fares, MD
 Division of Shoulder and Elbow Surgery
 Rothman Orthopaedic Institute
 925 Chestnut St, Philadelphia, PA 19107
mohamadfaresmd@gmail.com