Evaluating the Accuracy of ChatGPT as a Patient Education Resource on Osteosarcoma

Brigitte Alexis Lieu[1], Deaquan J Nichols, Teja Yeramosu, Logan K Laubach, John William Krumme[2], Gregory F Domson[3]
[1]Orthopaedic Surgery, Virginia Commonwealth University School of Medicine, [2]Dickson and Diveley Orthopaedics, [3]VCU Medical Center

INTRODUCTION:

As artificial intelligence (AI) continues to advance, it is likely to gain popularity among patients as an educational resource. In anticipation of this change in healthcare, it is crucial to understand whether AI is a reliable information source on orthopaedic conditions. Prior studies investigate the utility of ChatGPT in providing evidence-based information for patients considering common orthopaedic surgeries for sports, pediatric orthopaedic, and joint-related injuries. However, studies have yet to assess ChatGPT's responses for more rare and complex diseases. This study evaluates the accuracy and comprehensibility of responses produced by ChatGPT to answer common patient questions about osteosarcoma.

METHODS:

Frequently asked questions (FAQs) regarding osteosarcoma were compiled through a literature review and national society patient FAQ pages. The questions reflected those that patients routinely ask in clinics regarding the indications and management of osteosarcoma. ChatGPT (Version 3.5) was subsequently utilized to answer these questions. For each of the 10 responses, a detailed description was written based on relevant literature supporting or refuting the chatbot's claims. Responses were analyzed for accuracy and clarity using a previously validated scoring system for ChatGPT response accuracy and a modified DISCERN score. In accordance with the former scoring system, a numerical score of 1 to 4 was assigned to the responses based on their accuracy, with 1 representing the highest response accuracy and 4 representing the lowest response accuracy. The responses were independently reviewed by three authors, and scores were averaged as a crowd-sourced scoring strategy. The DISCERN instrument is a validated tool to help contextualize the quality of written health information; in the present study, DISCERN scores were assigned to the ChatGPT responses by two orthopaedic oncology surgeons. Readability was assessed using several published educational-level indices, namely the Flesch-Kincaid grade level, Simple Measure of Gobbledygook (SMOG) index, Coleman-Liau index, Gunning Fog index, and Automated Readability index. FAQ compilation and scoring were completed in collaboration with two fellowship-trained orthopaedic oncology surgeons.

RESULTS: ChatGPT's responses generally required moderate clarification, with a mean accuracy score of 3 (satisfactory but requiring moderate clarification). One response received a mean rating of 2 (satisfactory requiring minimal clarification), five responses received a rating of 2.5, and four responses received a rating of 3. Zero responses received a rating of 1 (excellent response not requiring clarification) or 4 (unsatisfactory requiring substantial clarification). The 10 responses received an average mean DISCERN score of 36 (classified as poor, 28-38). The interrater reliability between the two orthopaedic oncology surgeons for the DISCERN criteria was 0.601, qualifying as moderate agreement. Readability level ranged from college graduate to 7th grade, higher than is recommended for patient educational materials. The individual education indices were as follows: Flesch-Kincaid Grade Level 16.01 (SD 12.04), Gunning Fog Index 20.07 (2.79), Coleman-Liau Index 18.21 (1.71), Simple Measure of Gobbledygook Index 14.24 (1.78), Automated Readability Index 15.81 (1.94).

DISCUSSION AND CONCLUSION: We hypothesized that ChatGPT would offer high-quality information regarding osteosarcoma with regard to accuracy, clarity, and readability. Similar to prior literature on the use of ChatGPT for common orthopaedic conditions, most responses regarding osteosarcoma were moderately accurate but required further clarification and were written in an inaccessible reading level. ChatGPT can therefore be considered a starting point for patient education on osteosarcoma to supplement traditional patient education strategies, but it should not replace professional medical advice. One major limitation of AI that was apparent in the present study is its inability to provide personalized recommendations. The chatbot does not account for the unique clinical presentations that might affect an individual patient's counseling and treatment options. Furthermore, ChatGPT lacks the ability to gauge a patient's verbal competency and scientific literacy and adjust its language accordingly, as is the case in face-to-face consultations with healthcare professionals. Future research should apply similar methodologies to other AI platforms to more comprehensively investigate the breadth and accuracy of online resources available to patients.

Table 1: Scoring Systems

ChatGPT Response Accuracy Score	Response Accuracy Description
1	Excellent response not requiring clarification
2	Satisfactory requiring minimal clarification
3	Satisfactory requiring moderate clarification
4	Unsatisfactory requiring substantial clarification
DISCERN Score	
63-75	Classified as excellent
51-62	Classified as good
39-50	Classified as average
28-38	Classified as poor
<27	Classified as very poor

Adapted from Mika et al, Charnock et al, and Weil et al.

Mika AP, Martin JR, Engstrom SM, Polkowski GG, Wilson JM. Assessing ChatGPT Responses to Common Patient Questions Regarding Total Hip Arthroplasty. J Bone Jt Surg. Published online July 17, 2023. doi:10.2106/JBJS.23.00209

Charnock D, Shepperd S, Needham G, Gann R. DISCERN: an instrument for judging the quality of written consumer health information on treatment choices. J Epidemiol Community Health. 1999;53(2):105-111. doi:10.1136/jech.53.2.105

Weil AG, Bojanowski MW, Jamart J, Gustin T, Lévêque M. Evaluation of the Quality of Information on the Internet Available to Patients Undergoing Cervical Spine Surgery. World Neurosurg. 2014;82(1-2):e31-e39. doi:10.1016/j.wneu.2012.11.003

Table 2: Mean Scores of ChatGPT Responses

Question	Mean ChatGPT Response Accuracy Score	Mean DISCERN Score
1. What is osteosarcoma?	2.5	37
2. What is the cause of osteosarcoma?	2.5	37
3. What is the survival rate of osteosarcoma?	3	35
4. What are the risk factors for osteosarcoma?	2	29
5. What are the symptoms of osteosarcoma?	2.5	36
6. What are the treatments for osteosarcoma?	2.5	43
7. How long is recovery after surgery for osteosarcoma?	2.5	34
8. What does surgery for osteosarcoma entail?	3	37
9. How do you monitor the recurrence of osteosarcoma?	3	31
10. What is the outcome after surgery for osteosarcoma?	3	37
Averaged Means	3	36

Adapted from Mika et al, Charnock et al, and Weil et al.

Mika et al: 1 (Excellent response not requiring clarification), 2 (Satisfactory requiring minimal clarification), 3 (Satisfactory requiring moderate clarification), 4 (Unsatisfactory requiring substantial clarification)

DISCERN: 63-75 (excellent), 51-62 (good), 39-50 (average), 28-38 (poor), <27 (very poor)

Table 3: Readability Indices

Readability Formula	Average Score (Std)	Grade Level
Flesch-Kincaid Grade Level	16.01 (12.04)	College Graduate
Gunning Fog Index	20.07 (2.79)	College Graduate
Coleman-Liau Index	18.21 (1.71)	Professional
Simple Measure of Gobbledygook Index	14.24 (1.78)	7th grade
Automated Readability Index	15.81 (1.94)	7th grade