JB&JS OPEN ACCESS

ChatGPT Is Moderately Accurate in Providing a General Overview of Orthopaedic Conditions

Chandler A. Sparks, MS, Sydney M. Fasulo, MD, Jordan T. Windsor, BS, Vita Bankauskas, BA, Edward V. Contrada, BS, Matthew J. Kraeutler, MD, and Anthony J. Scillia, MD

Investigation performed at St. Joseph's University Medical Center, Paterson, New Jersey

Background: ChatGPT is an artificial intelligence chatbot capable of providing human-like responses for virtually every possible inquiry. This advancement has provoked public interest regarding the use of ChatGPT, including in health care. The purpose of the present study was to investigate the quantity and accuracy of ChatGPT outputs for general patient-focused inquiries regarding 40 orthopaedic conditions.

Methods: For each of the 40 conditions, ChatGPT (GPT-3.5) was prompted with the text "I have been diagnosed with [condition]. Can you tell me more about it?" The numbers of treatment options, risk factors, and symptoms given for each condition were compared with the number in the corresponding American Academy of Orthopaedic Surgeons (AAOS) OrthoInfo website article for information quantity assessment. For accuracy assessment, an attending orthopaedic surgeon ranked the outputs in the categories of <50%, 50% to 74%, 75% to 99%, and 100% accurate. An orthopaedics sports medicine fellow also independently ranked output accuracy.

Results: Compared with the AAOS OrthoInfo website, ChatGPT provided significantly fewer treatment options (mean difference, −2.5; $p < 0.001$) and risk factors (mean difference, −1.1; $p = 0.02$) but did not differ in the number of symptoms given (mean difference, −0.5; $p = 0.31$). The surgical treatment options given by ChatGPT were often non-descript (n = 20 outputs), such as "surgery" as the only operative treatment option. Regarding accuracy, most conditions (26 of 40; 65%) were ranked as mostly (75% to 99%) accurate, with the others (14 of 40; 35%) ranked as moderately (50% to 74%) accurate, by an attending surgeon. Neither surgeon ranked any condition as mostly inaccurate (<50% accurate). Interobserver agreement between accuracy ratings was poor ($\kappa = 0.03$; $p = 0.30$).

Conclusions: ChatGPT provides at least moderately accurate outputs for general inquiries of orthopaedic conditions but is lacking in the quantity of information it provides for risk factors and treatment options. Professional organizations, such as the AAOS, are the preferred source of musculoskeletal information when compared with ChatGPT.

Clinical Relevance: ChatGPT is an emerging technology with potential roles and limitations in patient education that are still being explored.

The use of online resources to obtain health-related medical information has risen considerably during the previous 2 decades and is now common practice. Among patients receiving care at an outpatient orthopaedic clinic, 84.9% report access to the Internet and 64.7% report the use of online resources to obtain orthopaedic information[1]. Such practices have the potential to allow patients to participate in decision-making regarding treatment of their condition[2-4]. In surgical fields, patient access to accurate medical information can align patient expectations, improve satisfaction, and improve outcomes[5]. The quality of online medical information, however, is variable and it is often inaccurate, of low quality, and/or of low pertinence[5-8].

The use of artificial intelligence (AI) chatbots to provide patients with accurate medical advice and knowledge regarding their conditions dates back to the 1990s, but interest in the use of such technologies has risen as the technological capabilities and their availability increases[9,10]. AI chatbots have previously shown some success in delivering accurate medical advice to patients[9]. It has been suggested that the use of

Disclosure: No external funding was received for this work. The Article Processing Charge for open access publication was funded by St. Joseph's University Medical Center. The **Disclosure of Potential Conflicts of Interest** forms are provided with the online version of the article (http://links.lww.com/JBJSOA/A636).

chatbots may be able to help save patients with minor conditions from needing to visit the hospital, help patients avoid consultations that are not required, empower patients, and improve adherence[9,11]. It has been noted, however, that if the quality of these resources is not closely assessed, they could be dangerous and put patients at risk[9,11]. Recent discussion regarding the role of AI chatbots in patient education stems from the release of Chat Generative Pre-Trained Transformer (ChatGPT), an advanced chatbot capable of creating human-like responses for unlimited inquiries to anyone with internet access[12].

ChatGPT is a large language model (LLM) recently developed by OpenAI[12]. LLMs, such as ChatGPT, are pre-trained on vast resources (e.g., internet articles, books, etc.) and analyze the patterns in these texts to provide relevant answers to questions. ChatGPT is trained on resources up to September 2021. LLMs generate outputs by predicting a sequence of words according to the patterns and relationships it has learned from the words that came before it. Multiple prompts can be made in a single chat session with ChatGPT, with the chatbot generating responses in light of the entire session history, allowing for a simulated conversation.

Despite only recently being made publicly available, there has been a considerable amount of discussion regarding the applications of ChatGPT, including education, literature, scientific writing, customer service, journalism, misinformation detection, public health, and medicine, to name a few[13]. In the field of medicine, research investigating the potential uses and pitfalls of ChatGPT is rapidly occurring, with some results showing that the technology may be able to reason through medical information[14]. It is, however, unclear how ChatGPT may act as a patient interface for accessing medical information. The purpose of the present study was to investigate the quantity and accuracy of ChatGPT outputs for general patient-focused inquiries regarding orthopaedic conditions. We hypothesized that ChatGPT would provide a lower quantity of information than American Academy of Orthopaedic Surgeons (AAOS) OrthoInfo, a peer-reviewed website for patient information provided by the AAOS. We also hypothesized that the information would be of low to moderate quality on average.

Materials and Methods

Study Design

During the week of February 26 through March 4, 2023, we prompted ChatGPT (GPT-3.5) with the text "I have been diagnosed with [condition]. Can you tell me more about it?" for a total of 40 orthopaedic conditions, then compared these outputs to information in the corresponding OrthoInfo articles over the same time frame. Conditions were grouped into 1 of 5 categories (Sports, Joints, Spine, Trauma, or Hand), with 8 conditions selected per group (Table I). All inquiries were made in separate chats to ensure that ChatGPT would not draw information from previous inquiries in the same chat.

Assessment of Outputs

To assess the quantity of information given by ChatGPT, the numbers of symptoms, risk factors, and treatment options in the output for each condition were recorded. The same parameters were recorded from a corresponding AAOS OrthoInfo article. For both resources, all individual symptoms, risk factors, and treatment options were recorded (e.g., "older female" counted as 2 individual risk factors, age and female sex). When a treatment option such as "surgery" or "pain medications" was listed, these were each recorded as a single treatment option; however, when multiple surgical and medical options were given, each surgical procedure and individual medicine were recorded as individual treatment options. Only information directly pertinent to the input was recorded from each AAOS OrthoInfo article, even when symptoms, risk factors, and treatment options for other or related conditions were given that did not directly apply to the input.

To assess the accuracy of information provided by ChatGPT, a sports medicine fellowship-trained attending orthopaedic surgeon (A.J.S.) reviewed the outputs and ranked the accuracy of information into 4 categories, <50%, 50% to 74%, 75% to 99%, or 100% accurate, with categories representing accuracy estimates. Rationale for rank and additional comments were also recorded. An orthopaedic surgery sports medicine fellow (M.J.K) also independently ranked the accuracy of the outputs and recorded rationale and other comments.

TABLE I Conditions Evaluated and Their Respective Categories

Category	Conditions
Sports	Meniscus tear, anterior cruciate ligament tear, lateral epicondylitis, biceps tendinitis, sports hernia, rotator cuff tear, patellofemoral pain syndrome, shoulder bursitis
Joints	Knee osteoarthritis, hip dysplasia, femoral head osteonecrosis, knee rheumatoid arthritis, slipped capital femoral epiphysis, knee osteochondritis dissecans, gout, infection of joint prosthesis
Spine	Herniated disc in the lumbar spine, kyphosis, spinal stenosis, spondylolisthesis, sciatica, lumbar vertebral compression fracture, cervical radiculopathy, cervical spondylosis
Trauma	Clavicle fracture, ankle fracture, femoral nonunion, hip fracture, shoulder dislocation, femoral head osteomyelitis, toe fracture, compartment syndrome
Hand	Carpal tunnel syndrome, ganglion cyst, scaphoid fracture, Dupuytren contracture, De Quervain tenosynovitis, stenosing tenosynovitis, boxer's fracture, boutonnière deformity

JBJS Open Access • 2024:e23.00129.

Statistical Analysis

All statistical analyses were performed with use of RStudio (version 2021.09.2; Posit PBC). For paired data (i.e., comparisons of ChatGPT versus AAOS OrthoInfo parameters), a paired t test was utilized. To compare mean differences in parameters across categories (i.e., Sports, Joints, Spine, Trauma, and Hand), a 1-way analysis of variance (ANOVA) was utilized. To assess the inter-rater agreement for assessments of accuracy, the Cohen weighted kappa was calculated.

Results

ChatGPT Output Characteristics

In response to the prompt utilized in the preset study, ChatGPT provided responses with a mean (and standard error of the mean [SEM]) of 258 ± 8 words (range, 177 to 376 words). ChatGPT provided symptoms and treatment options for all 40 conditions and risk factors for 26 of the 40 conditions. The outputs were all primarily in sentence format; however, the chatbot provided bullet points for individual symptoms in 5 outputs, risk factors in 4 outputs, and treatment options in 4 outputs, as opposed to providing them within a sentence or short paragraph. Follow-up with a medical professional was recommended in 38 of the 40 outputs. An example ChatGPT output is provided in Figure 1.

ChatGPT Versus AAOS OrthoInfo Information Quantity

Compared with AAOS OrthoInfo articles, ChatGPT provided significantly fewer treatment options (mean difference, −2.5; $p < 0.001$) (Fig. 2-A). ChatGPT provided an average of 4.4 treatment options (including mean of 1.3 surgical options, 1.4 medical options, and 0.9 physical therapy options), whereas OrthoInfo gave an average of 6.9 treatment options (including a mean of 2.3 surgical options, 2.4 medical options, and 1.0 physical therapy options). On average, 60.4% of the risk factors given by ChatGPT were similar to treatment options given in the corresponding OrthoInfo article.

Of note, ChatGPT often did not give specific surgical options, instead suggesting a nondescript option (e.g., "surgery"). Specifically, 20 of the outputs were nondescript, 11 provided a single specific procedure and/or description of the surgery, 8 listed and/or described various surgical options or techniques, and 1 did not give any surgical option. However, ChatGPT did not provide

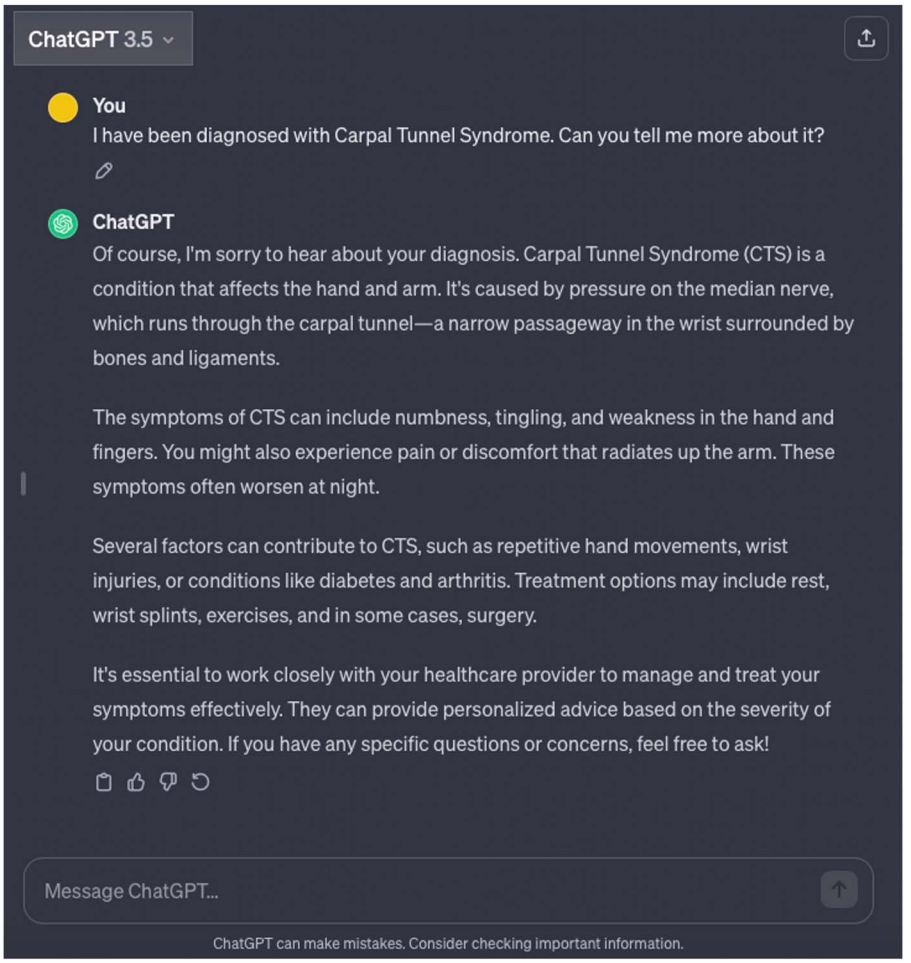

Fig. 1

An example of a ChatGPT output. This is an example output and does not represent an output that was used in the analysis.

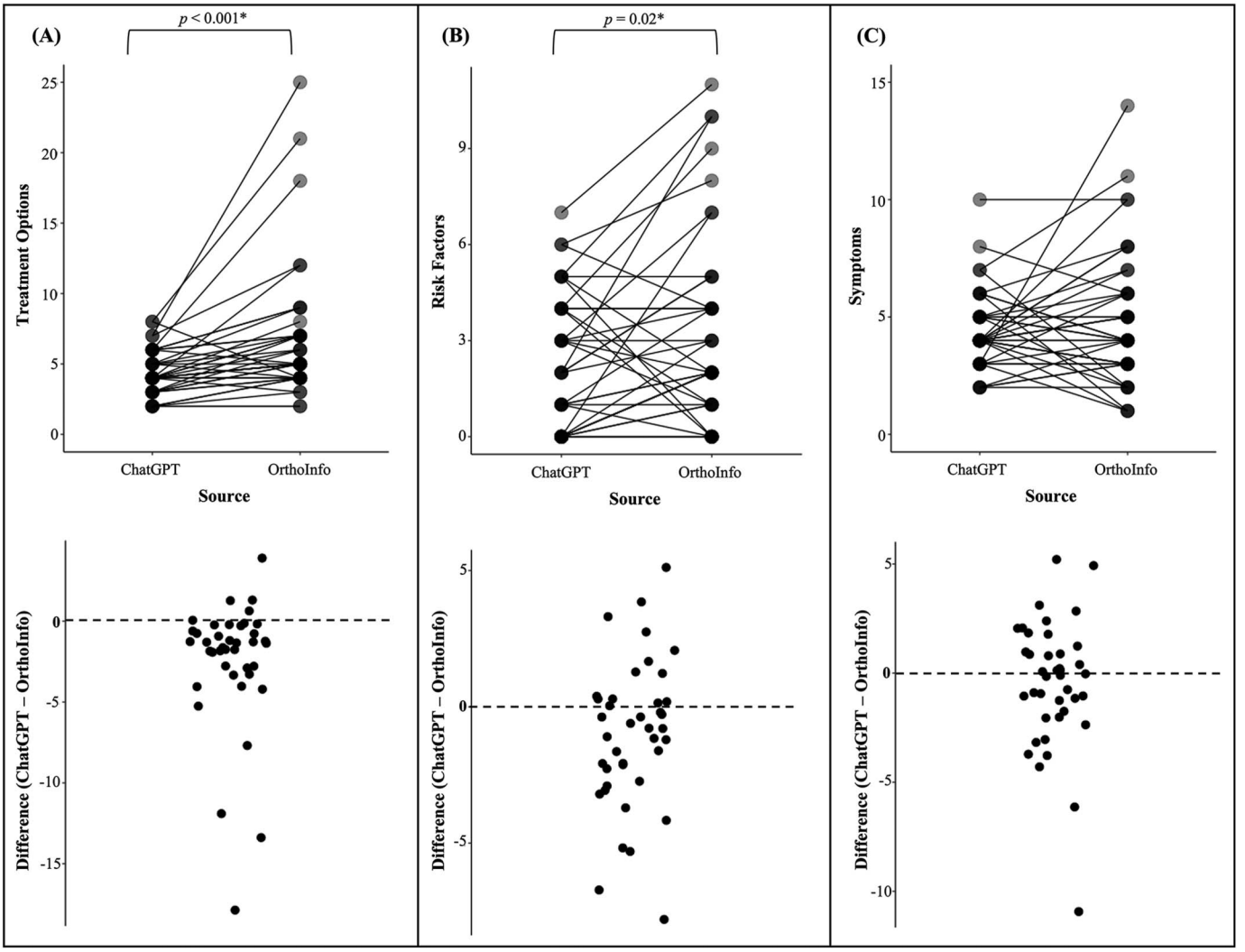

Fig. 2

ChatGPT versus AAOS OrthoInfo parameters. Figures represent raw paired data between ChatGPT and OrthoInfo (top) and the difference (ChatGPT – AAOS OrthoInfo; bottom) for treatment options (**Fig. 2-A**), risk factors (**Fig. 2-B**), and symptoms (**Fig. 2-C**). *Significance was set at 0.05 and determined with a paired t test.

discussion of the indications and/or evidence-based outcomes when discussing different surgical treatment options or techniques. Conversely, OrthoInfo articles described multiple surgical treatment options or techniques for 26 conditions, with additional discussion of the indications and/or evidence-based outcomes for 20 of those conditions (e.g., stating that the outcomes of open and endoscopic carpal tunnel release are similar). Of the 14 conditions without multiple treatment options, 9 listed and/or described a single specific procedure and 5 gave a nondescript surgical option (e.g., "surgery"). However, 3 of the 5 OrthoInfo articles that gave a nondescript surgical option provided links to additional resources discussing the various surgical options for the condition.

ChatGPT also gave significantly fewer risk factors compared with the corresponding OrthoInfo articles (mean difference, −1.1; p = 0.02) (Fig. 2-B). In cases in which both ChatGPT and the corresponding OrthoInfo article gave risk factors (23 of the 40 conditions), an average of 51.3% of the risk factors were similar those given by to the trusted source per condition. The number of symptoms given by ChatGPT did not significantly differ from corresponding articles on OrthoInfo (mean difference, −0.5; p = 0.31) (Fig. 2-C). ChatGPT and OrthoInfo articles did often give differing symptoms. On average, the symptoms given by ChatGPT had 60.4% similarity to those on the corresponding AAOS OrthoInfo article per condition.

Information Quantity Among Condition Categories

The difference (ChatGPT minus AAOS) in treatment options given by ChatGPT compared with AAOS OrthoInfo for corresponding conditions significantly differed by category (Sports, Joints, Spine, Trauma, and Hand) (p = 0.02) (Fig. 3-A). Per an unplanned (Tukey-Kramer) post hoc analysis, this difference

JBJS Open Access • 2024:e23.00129. **openaccess.jbjs.org** 5

was the result of ChatGPT giving fewer treatment options in the category of Joints (mean difference, −6.4) compared with Hand (mean difference, −0.6; p = 0.02) and Trauma (mean difference, −1.0; p = 0.04). There was no significant difference in the number of risk factors (p = 0.62) (Fig. 3-B) or symptoms (p = 0.97) (Fig. 3-C) given by ChatGPT compared with AAOS OrthoInfo across the different categories.

Information Accuracy

As ranked by an attending orthopaedic surgeon, 26 (65%) of the 40 outputs were 75% to 99% accurate. The other 14 outputs (35%) were 50% to 74% accurate. No outputs were ranked as 100% accurate or <50% accurate by the attending surgeon. Inter-rater agreement was found to be poor, with κ = 0.03 (95% confidence interval, −0.03 to 0.09) and p = 0.30. Of note, most areas of disagreement were between the ranking of an output as 100% and as 75% to 99% accurate (58.8%), with fewer being the result of disagreement between 100% and 50% to 74% accurate (20.6%) or between 75% to 99% and 50% to 74% accurate (20.6%). Although the attending surgeon ranked most conditions as 75% to 99% accurate, the sports medicine fellow ranked most as 100% accurate. Neither surgeon ranked any output as <50% accurate.

Discussion

The most important findings of the present study were that, in response to a general inquiry, ChatGPT may provide less information to patients regarding treatment options and risk factors than established patient-information websites and may occasionally provide misinformation. Inter-rater agreement regarding the accuracy of the outputs was poor between 2 practicing orthopaedic surgeons, with the attending surgeon being more critical of the outputs than the sports medicine fellow. However, disagreement was mostly regarding the information being mostly accurate (75% to 99%) versus completely accurate (100%), a nuanced and objective characterization. This reflects a need to better determine the accuracy of ChatGPT as a tool for finding accurate medical information.

Future studies might investigate a less subjective assessment of ChatGPT's medical information accuracy, such as comparing outputs to current guidelines. Traditional scored assessments of online medical information quality (e.g., the Quality Evaluation Scoring Tool and the DISCERN Tool) are not equipped to evaluate AI chatbots because of the lack of authorship and citation of sources from ChatGPT[15,16]. Regardless, most outputs were ranked as either mostly (75% to 99%) or entirely (100%) accurate by either orthopaedic surgeon, with no output being ranked as mostly inaccurate (<50%) by either surgeon. These results indicate that the accuracy of medical information by ChatGPT for musculoskeletal conditions is at least of moderate accuracy.

When both surgeons found deficiencies in the accuracy of the outputs, it was frequently because of misinformation. Often, misinformation regarded the condition description or causes. For instance, ChatGPT falsely reported that hip dysplasia occurs as a result of repetitive stress being placed on the hip joint. As another example, ChatGPT suggested that a femoral fracture may be a fracture of the femur or pelvis, not only the femur. In other cases, ChatGPT provided misinformation regarding treatment options for certain conditions, such as recommending chiropractic manipulation as a treatment option for spondylosis of the cervical spine or stating that surgery "may be required" for periprosthetic joint infection.

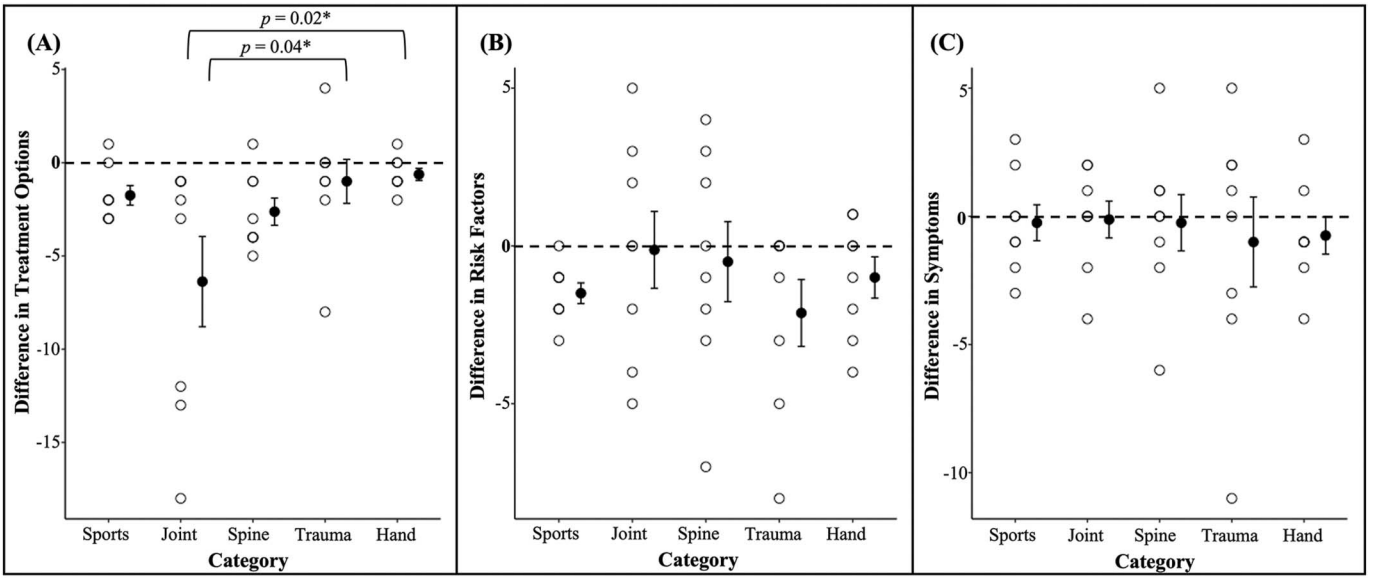

Fig. 3

ChatGPT versus AAOS OrthoInfo parameters by category of condition. Figures represent differences in the number of treatment options (**Fig. 3-A**), risk factors (**Fig. 3-B**), and symptoms (**Fig. 3-C**) (ChatGPT – AAOS OrthoInfo) among categories of conditions. Unfilled points represent individual conditions, and filled points represent the mean ± standard error. *Significance was set at 0.05 and determined with a Tukey-Kramer post hoc analysis.

ChatGPT may be limited by its knowledge cutoff of September 2021, but this does not explain all of the deficiencies found in its outputs[12]. The chatbot may also be limited by its inability to filter accurate from inaccurate information. ChatGPT acquires knowledge by searching the vast online resources it was trained on, which may contain contrasting or unreliable information[17].

In a related study focusing on the information provided by ChatGPT for shoulder impingement syndrome, it was found that the chatbot provided mostly accurate and complete information but sometimes provided controversial risk factors that have yet to be validated, failed to provide important differential diagnoses based on symptoms, was missing information about tests for the condition, and had missing and/or inaccurate information regarding treatment options[10]. In a study assessing ChatGPT responses to queries regarding hip osteoarthritis, the authors found that outputs were largely accurate but were superficial and occasionally missing important references and up-to-date research[18]. Both studies concluded that the use of ChatGPT was limited to providing general and basic information to patients who were using it as a tool to investigate musculoskeletal conditions, and that the chatbot could not provide personalized information or recommendations[10,18]. These findings are consistent with those of the present study, given that we found that the AI chatbot was capable of providing at least moderately accurate information, but that it occasionally provided misinformation, lacked critical information, and lacked in the amount of treatment options and risk factors it reported for musculoskeletal conditions.

Given these findings, the role of ChatGPT for patient education in musculoskeletal health care is unclear. Seth et al.[18] noted that the general knowledge provided in the outputs may be useful for providing patients preoperative information, but that the occasional inaccuracy of the chatbot may limit its use and that such outputs are unlikely to be superior to the current preoperative patient handouts. Under the review and editing of an orthopaedic surgeon, the technology may be of use for enhancing the surgeon's workflow. For instance, it has been proposed that ChatGPT may provide physicians and surgeons a means of creating more personalized information packets and/or discharge summaries than the current templates while also not reducing the time available for direct patient care[19]. Without the direct review and editing of an orthopaedic surgeon, however, the outputs may fail to adequately inform patients or, worse, may misinform patients.

There were several limitations to this study, including the use of only 1 resource (AAOS OrthoInfo) to compare against the ChatGPT outputs. Furthermore, it should be noted that although the AAOS OrthoInfo website served as the control in this study, it may not always be the first resource that patients see when using a search engine. Secondly, we only prompted ChatGPT with 1 inquiry; however, multiple inquiries can be made in the same chat to have a more extensive conversation and to obtain more information. We also only analyzed a limited number of parameters and conditions in the present study. Finally, the poor interobserver agreement limits the conclusions that can be drawn from this study. However, the poor interobserver agreement also reflects the need for the development of a better tool for measuring the accuracy of information given by AI chatbots. Future studies should investigate ChatGPT outputs compared with the information provided by other patient-education websites and should investigate requests other than those utilized in the present study (i.e., different inquiries). Furthermore, the accuracy of ChatGPT outputs should be investigated in future studies.

Conclusions

ChatGPT provides at least moderately accurate outputs for general inquiries regarding orthopaedic conditions but lacks in the quantity of information it provides for risk factors and treatment options. Professional organizations such as the AAOS remain the preferred source of musculoskeletal information when compared with ChatGPT. ■

Chandler A. Sparks, MS[1]
Sydney M. Fasulo, MD[2]
Jordan T. Windsor, BS[1]
Vita Bankauskas, BA[1]
Edward V. Contrada, BS[1]
Matthew J. Kraeutler, MD[3]
Anthony J. Scillia, MD[2]

[1]Hackensack Meridian School of Medicine, Nutley, New Jersey

[2]Department of Orthopedic Surgery, St. Joseph's University Medical Center, Paterson, New Jersey

[3]Department of Orthopedics, University of Colorado Anschutz Medical Campus, Aurora, Colorado

Email for corresponding author: chandler.sparks@hmhn.org

References

1. Burrus MT, Werner BC, Starman JS, Kurkis GM, Pierre JM, Diduch DR, Hart JM. Patient Perceptions and Current Trends in Internet Use by Orthopedic Outpatients. HSS J. 2017 Oct;13(3):271-5.

2. Clarke MA, Moore JL, Steege LM, Koopman RJ, Belden JL, Canfield SM, Meadows SE, Elliott SG, Kim MS. Health information needs, sources, and barriers of primary care patients to achieve patient-centered care: A literature review. Health Informatics J. 2016 Dec;22(4):992-1016.

3. Xie B, Wang M, Feldman R, Zhou L. Internet use frequency and patient-centered care: measuring patient preferences for participation using the health information wants questionnaire. J Med Internet Res. 2013 Jul 1;15(7):e132.

4. Lee CJ, Gray SW, Lewis N. Internet use leads cancer patients to be active health care consumers. Patient Educ Couns. 2010 Dec;81(0)(Suppl):S63-9.

5. Truumees D, Duncan A, Mayer EK, Geck M, Singh D, Truumees E. Cross sectional analysis of scoliosis-specific information on the internet: potential for patient confusion and misinformation. Spine Deform. 2020 Dec;8(6):1159-67.

6. O'Neill SC, Baker JF, Fitzgerald C, Fleming C, Rowan F, Byrne D, Synnott K; O'Neill SC. Cauda equina syndrome: assessing the readability and quality of patient information on the Internet. Spine (Phila Pa 1976). 2014 May 1;39(10):E645-9.

7. Corcelles R, Daigle CR, Talamas HR, Brethauer SA, Schauer PR. Assessment of the quality of Internet information on sleeve gastrectomy. Surg Obes Relat Dis. 2015 May-Jun;11(3):539-44.

8. Truumees D, Duncan A, Kunj A, Rajagopalan D, Geck M, Singh D, Stokes J, Truumees E. Quality and Accuracy of Cervical Radiculopathy-specific Information on

the Internet: A Cross-sectional Analysis. Spine (Phila Pa 1976). 2022 May 1;47(9): E399-406.

9. Bibault JE, Chaix B, Guillemassé A, Cousin S, Escande A, Perrin M, Pienkowski A, Delamon G, Nectoux P, Brouard B. A Chatbot Versus Physicians to Provide Information for Patients With Breast Cancer: Blind, Randomized Controlled Noninferiority Trial. J Med Internet Res. 2019 Nov 27;21(11): e15787.

10. Gwak GT, Hwang UJ, Jung SH, Kim JH. Search for Medical Information and Treatment Options for Musculoskeletal Disorders through an Artificial Intelligence Chatbot: Focusing on Shoulder Impingement Syndrome. J Musculoskelet Sci Technol. 2023 Jun 30;7(1):8-16.

11. Chaix B, Guillemassé A, Nectoux P, Delamon G, Brouard B. Vik: A Chatbot to Support Patients with Chronic Diseases. Health (London). 2020;12(7):804-10.

12 OpenAI. ChatGPT-Optimizing Language Models for Dialogue. 2022. Accessed 2023 Jul 28. https://chat.openai.com.

13. Shahriar S, Hayawi K. Let's have a chat! A Conversation with ChatGPT: Technology, Applications, and Limitations. Artificial Intelligence and Applications. 2024 Jun 2;2(1):11-20.

14. Kung TH, Cheatham M, Medenilla A, Sillos C, De Leon L, Elepaño C, Madriaga M, Aggabao R, Diaz-Candido G, Maningo J, Tseng V. Performance of ChatGPT on USMLE: Potential for AI-assisted medical education using large language models. PLOS Digit Health. 2023 Feb 9;2(2):e0000198.

15. Robillard JM, Jun JH, Lai JA, Feng TL. The QUEST for quality online health information: validation of a short quantitative tool. BMC Med Inform Decis Mak. 2018 Oct 19;18(1):87.

16. 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 Feb;53(2):105-11.

17. Kung JE, Marshall C, Gauthier C, Gonzalez TA, Jackson JB 3rd. Evaluating ChatGPT Performance on the Orthopaedic In-Training Examination. JB JS Open Access. 2023 Sep 8;8(3):e23.00056.

18. Seth I, Rodwell A, Tso R, Valles J, Bulloch G, Seth N. A Conversation with an Open Artificial Intelligence Platform on Osteoarthritis of the Hip and Treatment. J Orthop Sports Med. 2023.

19. Patel SB, Lam K. ChatGPT: the future of discharge summaries? Lancet Digit Health. 2023 Mar;5(3):e107-8.