

Is ChatGPT a trusted source of information for total hip and knee arthroplasty patients?

Cite this article:
Bone Jt Open 2024;5(2):
139–146.

DOI: 10.1302/2633-1462.
52.BJO-2023-0113.R1

Correspondence should be
sent to B. M. Wright
benjaminwright@usf.edu

B. M. Wright,¹ M. S. Bodnar,¹ A. D. Moore,² M. C. Maseda,² M. P. Kucharik,² C. C. Diaz,² C. M. Schmidt,² H. R. Mir³

¹Morsani College of Medicine, University of South Florida, Tampa, Florida, USA

²Department of Orthopaedic Surgery, University of South Florida, Tampa, Florida, USA

³Orthopaedic Trauma Service, Florida Orthopedic Institute, Tampa, Florida, USA

Aims

While internet search engines have been the primary information source for patients' questions, artificial intelligence large language models like ChatGPT are trending towards becoming the new primary source. The purpose of this study was to determine if ChatGPT can answer patient questions about total hip (THA) and knee arthroplasty (TKA) with consistent accuracy, comprehensiveness, and easy readability.

Methods

We posed the 20 most Google-searched questions about THA and TKA, plus ten additional postoperative questions, to ChatGPT. Each question was asked twice to evaluate for consistency in quality. Following each response, we responded with, "Please explain so it is easier to understand," to evaluate ChatGPT's ability to reduce response reading grade level, measured as Flesch-Kincaid Grade Level (FKGL). Five resident physicians rated the 120 responses on 1 to 5 accuracy and comprehensiveness scales. Additionally, they answered a "yes" or "no" question regarding acceptability. Mean scores were calculated for each question, and responses were deemed acceptable if \geq four raters answered "yes."

Results

The mean accuracy and comprehensiveness scores were 4.26 (95% confidence interval (CI) 4.19 to 4.33) and 3.79 (95% CI 3.69 to 3.89), respectively. Out of all the responses, 59.2% (71/120; 95% CI 50.0% to 67.7%) were acceptable. ChatGPT was consistent when asked the same question twice, giving no significant difference in accuracy ($t = 0.821$; $p = 0.415$), comprehensiveness ($t = 1.387$; $p = 0.171$), acceptability ($\chi^2 = 1.832$; $p = 0.176$), and FKGL ($t = 0.264$; $p = 0.793$). There was a significantly lower FKGL ($t = 2.204$; $p = 0.029$) for easier responses (11.14; 95% CI 10.57 to 11.71) than original responses (12.15; 95% CI 11.45 to 12.85).

Conclusion

ChatGPT answered THA and TKA patient questions with accuracy comparable to previous reports of websites, with adequate comprehensiveness, but with limited acceptability as the sole information source. ChatGPT has potential for answering patient questions about THA and TKA, but needs improvement.

Take home message

- With the rising popularity of artificial intelligence chatbots such as ChatGPT, patients will increasingly use them as a source for answering medical questions.
- ChatGPT answered total hip (THA) and knee arthroplasty (TKA) patient questions with accuracy comparable to previous reports of websites, with adequate

comprehensiveness, but limited acceptability as the sole information source.

- ChatGPT has potential for answering patient questions about THA and TKA, but needs improvement.

Introduction

ChatGPT is an artificial intelligence (AI) large language model (LLM) developed by OpenAI

(USA). It was pre-trained to understand the structure and patterns of language using text from a wide variety of sources, then fine-tuned using human feedback to produce optimal responses.¹ It utilizes a complex neural network, consisting of more than 175 billion parameters, to predict the next best text option based on user input,² giving it the ability to engage in sensible dialogue with the user. Having begun as a free research preview on 30 November 2022, it has since evolved and garnered unprecedented interest. As of January 2023, it had over 100 million monthly users.³ For over 70% of American adults, the internet is the first source they consult for medical information,⁴ and search engines, such as Bing and Google, have begun testing to incorporate ChatGPT and other LLMs into their browsers.⁵ As the use of LLMs and other forms of AI enter the mainstream, patients will use these new sources to answer medical enquiries. Thus far, ChatGPT has shown impressive capability in medicine-related tasks. This includes passing the USA Medical Licensing Examinations,⁶ and generating diagnostically accurate differential diagnoses from clinical cases.⁷ A few recent papers have assessed ChatGPT's ability to answer medical questions,⁸⁻¹⁰ and one compared physicians' responses to patient questions on Reddit's "r/AskDocs" with ChatGPT's responses.¹¹

In a previous study of orthopaedic patients, 94.1% reported access to the internet, and 55.8% reported using the internet for information on their injuries. Overall, 64.5% of these patients used at least one unreliable source. The most common sources used were WebMD, their treating institution's website, Google, YouTube, Wikipedia, Mayo Clinic, and Facebook.¹² In other studies, in which 65% of patients used the internet to find information on their orthopaedic condition,^{13,14} Google, WebMD, and the treating institution's website were the most common sources.¹⁴ Other than the internet, telephone calls seem to be a common way surgical patients address medical inquiries, especially postoperatively. Hadeed et al¹⁵ found that 29% of orthopaedic trauma patients initiated a telephone call within 14 days of discharge.

Previous studies have demonstrated that patients both have access to the internet and use it to find medical information, but are prone to finding unreliable sites that potentially contain inaccurate information.¹²⁻¹⁴ Furthermore, the average American reads at between a sixth to ninth grade level,^{16,17} although most patient educational material is written at a ninth to eleventh grade level.¹⁸⁻²⁰ Previous studies have emphasized the importance of patient education websites to not only be accurate and comprehensive, but also easily readable.^{19,21-27} With the rising popularity of ChatGPT, we wanted to assess the accuracy, comprehensiveness, and readability of ChatGPT when answering the most common patient questions surrounding total hip arthroplasty (THA) and total knee arthroplasty (TKA), two of the most common surgeries performed in the USA, with approximately 866,410 THAs and 1,223,299 TKAs annually.²⁸

Methods

In March 2023, we posed the 20 most Google-searched questions about THA and TKA, as reported by Shen et al,²⁹ to ChatGPT (Feb 13 Free Research Preview Version; OpenAI). Ethical approval was not applicable for this study design. If the questions said, "hip surgery" or "knee surgery", they were modified to say, "hip arthroplasty surgery" or "knee

arthroplasty surgery". Using the style of the first 20 questions as a guide, we added ten questions that a patient might ask following TKA or THA. These questions were based on the most common categories of post-discharge phone call reasons reported by Hällfors et al.³⁰ Since ChatGPT will give similar (but not the exactly the same) responses to questions every time, we asked each question twice. Each time a question was asked, it was done in a "New Chat", because ChatGPT learns from previous questions within a conversation. We then replied to each response with, "Please explain so it is easier to understand". Example questions and responses are shown in Table I. All questions and responses are shown in Supplementary Table i. All responses were copied into Word (Microsoft, USA) to calculate the Flesch-Kincaid Grade Level (FKGL), a formula that measures the approximate level of education required to understand a text based on the number of words per sentence and syllables per word.³¹ In order to avoid underestimating the text difficulty, we removed paragraph breaks, bullet points, one- or two-word headings before bullet points, colons, semicolons, and numbers from responses before calculating the score, as recommended by Friedman and Hoffman-Goetz³¹ and Eltorai et al.¹⁸ We also replaced colons and semicolons with periods.

Next, a spreadsheet containing the responses to each question was sent to five orthopaedic surgery residents (ADM, MCM, MPK, CCD, CMS) for evaluation of the responses. They scored each response for accuracy using a scale implemented in previous investigations of medical information accuracy obtained from online sources.²¹⁻²⁶ If they agreed with 0% to 25%, 26% to 50%, 51% to 75%, 75% to 99%, or 100%, scores of 1, 2, 3, 4, or 5 were given, respectively.

In addition to rating the accuracy, the responses were rated on a 1 to 5 comprehensiveness scale, shown in Table II, to ensure the responses were not only accurate, but also thorough. Lastly, to gauge acceptability of responses, each reviewer responded "yes" or "no" to the question: "Would you be comfortable if this was the only information your patient received for their question?" The purpose of this question was to determine if essential information was excluded from the answer, or if patient harm could result from the answer.

Statistical analysis

Due to the subjectivity of rating accuracy and comprehensiveness, we employed an ensemble (or crowd sourcing) scoring strategy,^{32,33} by averaging the ratings of the five reviewers for each response. This is comparable to a panel of judges averaging their scores for a performance. Therefore, the mean rating scores represent the consensus among the reviewers and the confidence intervals are influenced by reviewer agreement.¹¹ A response was deemed acceptable if \geq four reviewers answered "yes" to the acceptability question. A chi-squared test of independence was used to assess differences in ratings of acceptability between original and easier responses, and between the first and second time the questions were asked. Lastly, we found the mean FKGL and calculated 95% confidence intervals (CIs). We performed independent-samples *t*-tests comparing accuracy, comprehensiveness, and FKGL between the responses to the first and second time questions were asked. We also used independent-samples *t*-tests to compare these measures between original and easier responses. A two-tailed α of $p < 0.05$ was used to

Table 1. Examples of ChatGPT responses to patient questions.

Question	Response 1	Response 1, "Please explain so it is easier to understand."
How long does it take for bone to grow into hip replacement?	<p>The process of bone growth into a hip replacement, also known as osseointegration, can take several months to a year or more. The exact timeline can vary depending on factors such as the type of implant used, the patient's age and overall health, and the quality of the patient's bone.</p> <p>In general, the first few weeks after surgery are critical for the initial stability of the implant. During this time, the bone will begin to grow around the implant and integrate with it. Over the following months, the bone will continue to grow and remodel around the implant, gradually increasing its strength and stability.</p> <p>The surgeon will typically monitor the patient's progress with regular follow-up visits and X-rays or other imaging studies to assess the bone growth and ensure that the implant is properly integrated. The patient may also need to undergo physical therapy or rehabilitation to help regain strength and mobility in the hip joint.</p> <p>It's important for patients to follow their surgeon's postoperative instructions closely to maximize the chances of a successful outcome and minimize the risk of complications. Ultimately, the time it takes for bone to grow into a hip replacement will depend on several factors and can vary from patient to patient.</p>	<p>After hip replacement surgery, the bone around the new implant gradually grows and attaches to the implant. This process is called osseointegration and can take several months to a year or more. During the first few weeks after surgery, the bone starts to grow around the implant and integrate with it. Over time, the bone continues to grow and strengthen around the implant.</p> <p>The surgeon will monitor the patient's progress with follow-up visits and imaging studies. The patient may also need physical therapy to help regain strength and mobility in the hip joint. It's important for patients to follow their surgeon's instructions carefully to reduce the risk of complications and increase the chances of a successful outcome.</p> <p>Overall, the time it takes for bone to grow into a hip replacement varies depending on factors such as the type of implant used, the patient's age and health, and the quality of the bone.</p>
How do I know if I have a blood clot after joint replacement?	<p>If you have had joint replacement surgery and are concerned about the possibility of a blood clot, there are several signs and symptoms to look out for. These may include:</p> <p>Swelling: If the affected area is significantly swollen, it could be a sign of a blood clot. Pay attention to whether the swelling is increasing, staying the same, or decreasing over time.</p> <p>Redness: If the skin around the affected area is red or warm to the touch, it could be a sign of a blood clot.</p> <p>Pain: Pain is a common symptom after joint replacement surgery, but if the pain is getting worse instead of better or is accompanied by other symptoms, it could be a sign of a blood clot.</p> <p>Difficulty walking or standing: If you are having difficulty putting weight on the affected leg or standing up, it could be a sign of a blood clot.</p> <p>Shortness of breath: If you are experiencing shortness of breath, chest pain, or coughing, it could be a sign of a blood clot that has traveled to your lungs.</p> <p>If you experience any of these symptoms, it is important to contact your doctor immediately. They may order diagnostic tests such as ultrasound or blood tests to confirm whether or not you have a blood clot. Prompt medical attention is crucial to prevent serious complications from blood clots.</p>	<p>After joint replacement surgery, it's possible to develop a blood clot in your leg. This can be a serious problem because the clot can break off and travel to your lungs, which can be life-threatening.</p> <p>Here are some signs that you might have a blood clot:</p> <p>Your leg is swollen, red, or warm to the touch.</p> <p>You have pain that's getting worse, not better, or you're having trouble walking or standing.</p> <p>You feel short of breath, have chest pain, or are coughing.</p> <p>If you have any of these symptoms, it's important to contact your doctor right away. They may do some tests to see if you have a blood clot. The sooner you get treatment, the better your chances of avoiding serious complications.</p>

determine significance. All analyses were conducted with SPSS v. 28 (IBM, USA).

Results

The mean accuracy score of all responses was 4.26 (95% CI 4.19 to 4.33), corresponding to agreement with > 75% of information. The mean comprehensiveness score of all responses was 3.79 (95% CI 3.69 to 3.89), corresponding to above the 'adequate' level. The mean FKGL of all responses was 11.65 (95% CI 11.19 to 12.10). Out of all the responses, ≥ four reviewers rated 59.2% (71.0/120, 95% CI 50.0% to 67.7%) as acceptable.

To determine if ChatGPT answers the same question with similar accuracy, comprehensiveness, acceptability, and FKGL, each time the question was asked, we grouped the responses by the first and second time the questions were

asked. There was no statistically significant difference in accuracy ($t = 0.821$; $p = 0.415$) between the first question responses (4.33; 95% CI 4.21 to 4.46) and the second question responses (4.25; 95% CI 4.11 to 4.40). There was no statistically significant difference in comprehensiveness ($t = 1.387$; $p = 0.171$) between the first question responses (3.99; 95% CI 3.85 to 4.14) and the second question responses (3.82; 95% CI 3.62 to 4.02). There was no statistically significant difference in acceptability ($\chi^2 = 1.832$; $p = 0.176$) between the first question responses (21.9/30, 73.3%; 95% CI 54.8% to 86.2%) and the second question responses (17.0/30, 56.7%; 95% CI 38.6% to 73.1%). There was no statistically significant difference in FKGL ($t = 0.264$; $p = 0.793$) between the first question responses (12.25; 95% CI 11.29 to 13.20) and the second question responses (12.06; 95% CI 11.02 to 13.10).

Table II. Comprehensiveness scoring.

Score	Description
1	Incomplete The answer only addresses a small portion of the question and leaves out significant details or important information.
2	Partial The answer gives some relevant information, but it is not comprehensive and is missing key elements.
3	Adequate The answer covers the main points of the question and provides enough information to understand the topic, but lacks depth or detail.
4	Thorough The answer is comprehensive and provides detailed information that addresses all aspects of the question.
5	Exhaustive The answer is extremely detailed and thorough, covering all aspects of the question and providing a deep understanding of the topic.

Table III. ChatGPT compared to previous studies that assessed accuracy and/or readability of orthopaedic patient information websites.

Study	Topic	Number of websites assessed	Raters	Mean accuracy score (SD)	Percent of maximum accuracy score	Mean FKGL (SD)
Wang et al ²⁷	Articular cartilage defects management	53	3	11.7 (0.6) (maximum 12)	97.5	13.4 (8.0)
Crozier-Shaw et al ²⁶	Metal-on-metal hip arthroplasty	61	2	6.6 (1.2) (maximum 8)	82.5	11.7 (0.88)
Fabricant et al ²⁵	Developmental hip dysplasia	63	3	10.7 (1.8) (maximum 12)	89.2	10.5 (2.4)
Badarudeen et al ¹⁹	Paediatric orthopaedics patient education	101	-	-	-	9.0 (2.7)
Mathur et al ²⁴	Scoliosis	50	3	5.9 (0.64) (maximum 12)	49.2	-
Dy et al ²³	Distal radius fracture/broken wrist	70	3	11.1 (2.2) (maximum 12)	92.5	10 (2.3)
Zade et al ²²	Bone tumours	48	3	10.3 (1.7) (maximum 12)	85.8	10.5 (1.4)
Garcia et al ²¹	Shoulder instability	82	3	8.6 (2.6) (maximum 12)	71.7	10.96 (2.5)
ChatGPT	THA and TKA	120 responses evaluated	5	4.25 (maximum 5)	85.0	11.65

FKGL, Flesch-Kincaid Grade Level; SD, standard deviation; THA, total hip arthroplasty; TKA, total knee arthroplasty.

Next, we wanted to determine if ChatGPT could decrease the FKGL when asked to, "Please explain so it is easier to understand," while maintaining comparable accuracy, comprehensiveness, and acceptability. There was a statistically significant lower FKGL ($t = 2.204$; $p = 0.029$) for the easier responses (11.14; 95% CI 10.57 to 11.71) than the original question responses (12.15; 95% CI 11.45 to 12.85). There was no statistically significant difference in accuracy ($t = 0.975$; $p = 0.332$) between the original question responses (4.29; 95% CI 4.20 to 4.39) and the easier responses (4.23; 95% CI 3.28 to 5.17), nor were there differences in acceptability ($\chi^2 = 1.690$; $p = 0.194$) between the original (39/60, 65%; 95% CI 52% to 76.1%) and the easier responses (31.9/60, 53.3%; 95% CI 40.6% to 65.7%). There was, however, a statistically significant lower comprehensiveness ($t = 2.246$; $p = 0.027$) for the easier responses (3.68; 95% CI 3.53 to 3.83) than the original question responses (3.91; 95% CI 3.78 to 4.03).

Discussion

Our results show that ChatGPT can consistently answer patient questions about THA and TKA with accuracy, comprehensiveness, and in a readable manner comparable to other online sources. ChatGPT was consistent in accuracy, comprehensiveness, acceptability, and FKGL when asked the same

question multiple times. Between the original responses and the easier responses, there was no significant change in accuracy or acceptability, but there was a significant reduction in comprehensiveness and reading level. We speculate that easier responses excluded critical information at the expense of providing a more accessible reading level.

Previous studies have analyzed various orthopaedic patient informational websites. For the purpose of comparing accuracy and FKGL of ChatGPT to previous papers that have reported accuracy as the sum of rater scores, we converted ChatGPT's mean score of 4.26 out of a maximum 5, to 85.2% of the maximum score, and converted previous papers' scores into a percentage of the maximum. ChatGPT's score was better than websites on metal-metal-hip arthroplasty,²⁶ scoliosis,²⁴ and shoulder instability,²¹ but was slightly lower than websites on developmental hip dysplasia,²⁵ distal radius fractures,²³ and bone tumours,²² and substantially lower than websites on articular cartilage defects management.²⁷ Table III shows a summary of these papers.

ChatGPT's mean comprehensiveness score of 3.91 for the original responses was above the 'adequate' level and just below 'thorough'. Previous papers evaluating comprehensiveness of orthopaedic condition patient informational websites used pre-determined content that they believed a website

Table IV. Example of reduction in acceptability from original response to easier response.

Question	Response 2	Response 2, "Please explain so it is easier to understand."
How do I know if I have an infection after joint replacement?	<p>If you have had joint replacement surgery, it is important to monitor the site for signs of infection. Some signs of infection after joint replacement surgery may include:</p> <p>Pain: Pain around the joint that does not go away with rest or medication may be a sign of infection.</p> <p>Swelling: Swelling around the joint that does not go away with ice or elevation may be a sign of infection.</p> <p>Redness: Redness around the joint may be a sign of infection.</p> <p>Fever: A fever may indicate an infection.</p> <p>Drainage: Pus or fluid draining from the surgical incision may indicate an infection.</p> <p>Difficulty moving: If you are having difficulty moving the joint or experiencing stiffness, this could be a sign of infection.</p> <p>If you experience any of these symptoms, it is important to contact your surgeon or healthcare provider immediately, as they may need to evaluate you for a possible infection. Early detection and treatment of an infection after joint replacement surgery is critical to prevent serious complications.</p>	<p>After joint replacement surgery, it's important to keep an eye out for signs of infection. Infection can cause pain, swelling, redness, fever, fluid draining from the incision site, or difficulty moving the joint. If you experience any of these symptoms, it's important to contact your doctor right away so they can check if you have an infection. Catching an infection early and treating it promptly is important to prevent serious complications.</p>

covering a topic should include to be fully comprehensive. This method worked for webpages discussing an entire topic, but was not practical for our approach of posing numerous, specific questions about a variety of topics.

Recent papers have assessed the accuracy and comprehensiveness of ChatGPT's responses to medical questions. Johnson et al¹⁰ evaluated 284 questions pertaining to 17 specialties. For accuracy, a six-point Likert scale was used. For comprehensiveness, a three-point Likert scale was used. Mean accuracy score was 4.4 with 39.4% (70.9/180) of responses scored at the highest level of accuracy. The mean comprehensiveness score was 2.4. Yeo et al⁸ assessed the ability of ChatGPT to answer questions obtained from FAQ sections on professional society websites and Facebook support group pages regarding liver cirrhosis and hepatocellular carcinoma. They used a grading scale that combined comprehensiveness and accuracy. Overall, 74% (54.0/73) of responses were given the highest score, rated as comprehensive and correct. Potapenko et al⁹ asked four questions, five times each, about five common retinal diseases, and used a five-point Likert scale (different than the one used here) to evaluate accuracy. In total, 45% (45/100) of responses were rated as very good/no inaccuracies, and 26% (26/100) were rated as containing minor, non-harmful inaccuracies. Our study is the first to focus on ChatGPT's capacity to answer orthopaedic questions, and it did so with similar accuracy and comprehensiveness compared to the aforementioned studies.

Although not measuring accuracy or comprehensiveness, Ayers et al¹¹ drew 195 questions from Reddit's r/AskDocs and had five reviewers rate physicians' responses on Reddit and ChatGPT's responses on five-point Likert scales for "the quality of information provided" and "the empathy or bedside manner provided." The ChatGPT responses had statistically significant higher quality and empathy scores than the physician responses. Additionally, 78.6% (459.8/585) of responses by ChatGPT were rated as "better" than the physician responses on Reddit.

No previous studies have used an acceptability metric. Only 59.2% (71.0/120) of all responses were graded as acceptable by at least four raters in our study. This implies that

ChatGPT may have been leaving out key elements of answers that a surgeon would include when responding to patients, and therefore may not be reliable as the sole source of information for patient questions. Table IV shows an example of a response that 0 out of 5 raters deemed acceptable. This was an easier response that the original question received a 5 out of 5, which is also shown in the table. While raters did not write reasons for marking a response unacceptable, we speculate that in this example, the lack of detail regarding signs of infection was responsible for the drop in acceptability.

Thus far, making direct comparisons to studies that assessed ChatGPT's accuracy and comprehensiveness to medical questions is difficult due to varying methodologies and a low volume of research. Other studies have shown ChatGPT to be both accurate and comprehensive, and our paper's results are congruent with this theme.

The mean FKGL of ChatGPT's answers, before and after asking for easier wording, was 12.15 and 11.14, respectively. These are above previously reported FKGLs of 9.2 for American Academy of Orthopaedic Surgeons (AAOS) patient information articles,¹⁸ 10.5 for websites on developmental hip dysplasia,²⁵ 8.9 for AAOS and Paediatric Orthopaedic Society of North America (POSNA) articles on paediatric orthopaedics,¹⁹ 10.0 for websites on distal radius fractures,²³ 10.5 for websites on bone tumours,²² and 10.96 for websites on shoulder instability.²¹ On the other hand, ChatGPT's responses was similar to the FKGL of 11.7 for websites describing metal-on-metal hip arthroplasty,²⁶ and lower than 13.4 for websites on the management of articular cartilage defects.²⁷ Although the mean reduction in FKGL was only 1.01, this is a statistically significant, unique feature of ChatGPT to respond to requests for easier readability.

An advantage we see of ChatGPT is ease of use. Finding an answer to a question by using an internet search engine requires one to select a website from a list of multiple results, assess the credibility, then locate the relevant information within the website. Locating health information through an internet search engine is done with variable accuracy ranging from 16% to 100%,^{34,35} and can take more than five minutes to find some answers.³⁵ Alternatively, ChatGPT provides a

concise, conversational response without relying on patient judgement, and can provide further clarification upon request. Future research should evaluate whether real patients actually prefer the conversational style of ChatGPT to an internet search engine.

Another advantage over an internet search is that one does not need to know medical terminology. In an analysis of websites on distal radius fractures by Dy et al,²³ the accuracy of website results was reduced when “broken wrist” was searched compared to “distal radius fracture”. ChatGPT does not return results based on searched keywords contained on a webpage, so it can respond to either “knee replacement” or the more technical medical language “total knee arthroplasty”. Future research should determine if differences in question terminology affect the accuracy of ChatGPT’s answers.

While other healthcare chatbots exist, ChatGPT is more advanced. Most available health-focused apps with chatbot integration rely on rule-based approaches and finite-state dialogue management. They direct the user through a predetermined path to a response, despite claims of machine learning and natural language processing.³⁶ On the other hand, ChatGPT and other LLMs adapt organically to each input, providing greater flexibility and personalization.

There are, however, several limitations to using ChatGPT for answering patient questions. The first is that in the future, ChatGPT may not be free for patients to access; OpenAI may decide to offer ChatGPT only as a paid subscription. Second, ChatGPT was trained on text containing information up to 2021.³⁷ As medical knowledge changes, its responses may be outdated. However, ChatGPT will have continued improvement, with the announcement of GPT-4 and GPT-4 Turbo, OpenAI’s more advanced, paid versions of ChatGPT, which has information up to April 2023.^{38,39}

A limitation of our study was that review of responses was done by resident-level physicians. Previous studies have used residents to evaluate orthopaedic patient information websites,^{21–23,27} but it is possible that lack of experience prevented the reviewers from accurately assessing responses. This limitation was mitigated by the fact that the questions asked were patient-generated, rudimentary, and would be responded to without attending physician guidance in real-world scenarios. Ideally, a reference standard would have been used. However, this was impractical given the large quantity of responses and diversity of question topics.

Another limitation of our study is that not all raters were blinded to the source of the responses. Two of the raters (ADM, MCM) were unable to be blinded, since they were involved in planning the study. This could have caused bias in their assessment of the responses.

Lastly, while comparisons can be made to previous studies on websites for patient education, our study lacked a comparative group, such as website results for internet-searched questions, or physician answers to the questions. External validity is also limited by the dynamic nature of the software, updates from OpenAI, and upgrades to ChatGPT including release of GPT-4. Future responses to identical questions may be drastically different.

The rapid growth in popularity of ChatGPT means that orthopaedic patients will increasingly use it to answer questions concerning surgery. ChatGPT must be properly evaluated so

that clinicians can advise their patients on whether it is trustworthy. Additional potential applications of LLMs like ChatGPT include clinical decision-making support, clinical documentation assistance, administrative work, and medical education.^{40,41}

To conclude, we found that ChatGPT answered common patient questions about THA and TKA with accuracy comparable to the average orthopaedic patient informational website, with adequate comprehensiveness, and with readability that can be improved upon request. However, improvements are needed before ChatGPT can be trusted to give safe, fully comprehensive information to patients. Further research is needed, including studies that directly compare ChatGPT to other sources of information for patient questions.

Supplementary material

All questions, responses, and Flesch-Kincaid Grade Levels.

References

1. **Ouyang L, Wu J, Jiang X, et al.** Training language models to follow instructions with human feedback. *arXiv*. 2022.
2. **Brown TB, Mann B, Ryder N, et al.** Language models are few-shot learners. *arXiv*. 2020.
3. **Hu K.** ChatGPT sets record for fastest-growing user base - analyst note. Reuters. February 2, 2023. <https://www.reuters.com/technology/chatgpt-sets-record-fastest-growing-user-base-analyst-note-2023-02-01/> (date last accessed 30 January 2024).
4. **No authors listed.** National Cancer Institute: Health Information National Trends Survey 5 Cycle 3. National Institutes of Health. 2019. https://hints.cancer.gov/view-questions/question-detail.aspx?PK_Cycle=11&qid=688 (date last accessed 30 January 2024).
5. **Shakir U.** From ChatGPT to Google Bard: how AI is rewriting the internet. The Verge. <https://www.theverge.com/23610427/chatbots-chatgpt-new-bing-google-bard-conversational-ai> (date last accessed 30 January 2024).
6. **Kung TH, Cheatham M, Medenilla A, et al.** Performance of ChatGPT on USMLE: potential for AI-assisted medical education using large language models. *PLOS Digit Health*. 2023;2(2):e0000198.
7. **Hirosawa T, Harada Y, Yokose M, Sakamoto T, Kawamura R, Shimizu T.** Diagnostic accuracy of differential-diagnosis lists generated by generative pretrained transformer 3 chatbot for clinical vignettes with common chief complaints: a pilot study. *Int J Environ Res Public Health*. 2023;20(4):3378.
8. **Yeo YH, Samaan JS, Ng WH, et al.** Assessing the performance of ChatGPT in answering questions regarding cirrhosis and hepatocellular carcinoma. *Clin Mol Hepatol*. 2023;29(3):721–732.
9. **Potapenko I, Boberg-Ans LC, Stormly Hansen M, Klefter ON, van Dijk EHC, Subhi Y.** Artificial intelligence-based chatbot patient information on common retinal diseases using ChatGPT. *Acta Ophthalmol*. 2023;101(7):829–831.
10. **Johnson D, Goodman R, Patrinely J, et al.** Assessing the accuracy and reliability of AI-generated medical responses: an evaluation of the ChatGPT model. *Res Sq*. 2023;rs.3.rs-2566942.
11. **Ayers JW, Poliak A, Dredze M, et al.** Comparing physician and artificial intelligence chatbot responses to patient questions posted to a public social media forum. *JAMA Intern Med*. 2023;183(6):589–596.
12. **Hautala GS, Comadoll SM, Raffetto ML, et al.** Most orthopaedic trauma patients are using the internet, but do you know where they’re going? *Injury*. 2021;52(11):3299–3303.
13. **Fralav A, Ming Chong Y, Holcdorf D, Plunkett V, Tran P.** Internet use by orthopaedic outpatients - current trends and practices. *Australas Med J*. 2012;5(12):633–638.
14. **Burrus MT, Werner BC, Starman JS, et al.** Patient perceptions and current trends in internet use by orthopedic outpatients. *HSS J*. 2017;13(3):271–275.

15. Hadeed MM, Kandil A, Patel V, Morrison A, Novicoff WM, Yarboro SR. Factors associated with patient-initiated telephone calls after orthopaedic trauma surgery. *J Orthop Trauma*. 2017;31(3):e96–e100.
16. Spandorfer JM, Karras DJ, Hughes LA, Caputo C. Comprehension of discharge instructions by patients in an urban emergency department. *Ann Emerg Med*. 1995;25(1):71–74.
17. Doak CC, Doak LG, Root JH. *Teaching Patients with Low Literacy Skills*. Morton PG, ed. *American Journal of Nursing*, 1996.
18. Eltorai AEM, Sharma P, Wang J, Daniels AH. Most American Academy of Orthopaedic Surgeons' online patient education material exceeds average patient reading level. *Clin Orthop Relat Res*. 2015;473(4):1181–1186.
19. Badarudeen S, Sabharwal S. Readability of patient education materials from the American Academy of Orthopaedic Surgeons and Pediatric Orthopaedic Society of North America web sites. *J Bone Joint Surg Am*. 2008;90-A(1):199–204.
20. Kher A, Johnson S, Griffith R. Readability assessment of online patient education material on congestive heart failure. *Adv Prev Med*. 2017;2017:9780317.
21. Garcia GH, Taylor SA, Dy CJ, Christ A, Patel RM, Dines JS. Online resources for shoulder instability: what are patients reading? *J Bone Joint Surg Am*. 2014;96-A(20):e177.
22. Zade RT, Tartaglione JP, Chisena E, Adams CT, DiCaprio MR. The quality of online orthopaedic oncology information. *J Am Acad Orthop Surg Glob Res Rev*. 2020;4(3):e19.00181.
23. Dy CJ, Taylor SA, Patel RM, Kitay A, Roberts TR, Daluiski A. The effect of search term on the quality and accuracy of online information regarding distal radius fractures. *J Hand Surg Am*. 2012;37(9):1881–1887.
24. Mathur S, Shanti N, Brkaric M, et al. Surfing for scoliosis: the quality of information available on the Internet. *Spine (Phila Pa 1976)*. 2005;30(23):2695–2700.
25. Fabricant PD, Dy CJ, Patel RM, Blanco JS, Doyle SM. Internet search term affects the quality and accuracy of online information about developmental hip dysplasia. *J Pediatr Orthop*. 2013;33(4):361–365.
26. Crozier-Shaw G, Queally JM, Quinlan JF. Metal-on-metal total hip arthroplasty: quality of online patient information. *Orthopedics*. 2017;40(2):e262–e268.
27. Wang D, Jayakar RG, Leong NL, Leathers MP, Williams RJ, Jones KJ. Evaluation of the quality, accuracy, and readability of online patient resources for the management of articular cartilage defects. *Cartilage*. 2017;8(2):112–118.
28. Siddiqi A, Levine BR, Springer BD. Highlights of the 2021 American Joint Replacement Registry Annual Report. *Arthroplast Today*. 2022;13:205–207.
29. Shen TS, Driscoll DA, Islam W, Bovonratwet P, Haas SB, Su EP. Modern internet search analytics and total joint arthroplasty: what are patients asking and reading online? *J Arthroplasty*. 2021;36(4):1224–1231.
30. Hällfors E, Saku SA, Mäkinen TJ, Madanat R. A consultation phone service for patients with total joint arthroplasty may reduce unnecessary emergency department visits. *J Arthroplasty*. 2018;33(3):650–654.
31. Friedman DB, Hoffman-Goetz L. A systematic review of readability and comprehension instruments used for print and web-based cancer information. *Health Educ Behav*. 2006;33(3):352–373.
32. Whalen S, Pandey OP, Pandey G. Predicting protein function and other biomedical characteristics with heterogeneous ensembles. *Methods*. 2016;93:92–102.
33. Chang N, Lee-Goldman R, Tseng M. Linguistic wisdom from the crowd. Proceedings of the Third AAAI Conference on Human Computation and Crowdsourcing; 2016. <https://ojs.aaai.org/index.php/HCOMP/article/view/13266/13114> (date last accessed 30 January 2024).
34. Agree EM, King AC, Castro CM, Wiley A, Borzekowski DLG. "It's got to be on this page": age and cognitive style in a study of online health information seeking. *J Med Internet Res*. 2015;17(3):e79.
35. Chevalier A, Dommes A, Marquié J-C. Strategy and accuracy during information search on the Web: effects of age and complexity of the search questions. *Computers in Human Behavior*. 2015;53:305–315.
36. Parmar P, Ryu J, Pandya S, Sedoc J, Agarwal S. Health-focused conversational agents in person-centered care: a review of apps. *NPJ Digit Med*. 2022;5(1):21.
37. No authors listed. ChatGPT General FAQ. OpenAI. <https://help.openai.com/en/articles/6783457-chatgpt-general-faq> (date last accessed 30 January 2024).
38. Achiam J, Adler S, Agarwal S, et al. OpenAI: GPT-4 Technical Report, arXiv. 2023.
39. No authors listed. New models and developer products announced at DevDay. OpenAI. n.d. <https://openai.com/blog/new-models-and-developer-products-announced-at-devday>
40. Kunze KN, Jang SJ, Fullerton MA, Vigdorichik JM, Haddad FS. What's all the chatter about? *Bone Joint J*. 2023;105-B(6):587–589.
41. Polisetty TS, Jain S, Pang M, et al. Concerns surrounding application of artificial intelligence in hip and knee arthroplasty: a review of literature and recommendations for meaningful adoption. *Bone Joint J*. 2022;104-B(12):1292–1303.

Author information

B. M. Wright, BS, Medical Student
 M. S. Bodnar, BS, Medical Student
 Morsani College of Medicine, University of South Florida, Tampa, Florida, USA.

A. D. Moore, MD, Orthopaedic Surgery Resident
 M. C. Maseda, MD, Orthopaedic Surgery Resident
 M. P. Kucharik, MD, Orthopaedic Surgery Resident
 C. C. Diaz, MD, Orthopaedic Surgery Resident
 C. M. Schmidt, MD, Orthopaedic Surgery Resident
 Department of Orthopaedic Surgery, University of South Florida, Tampa, Florida, USA.

H. R. Mir, MD, MBA, Orthopaedic Surgeon, Orthopaedic Trauma Service, Florida Orthopedic Institute, Tampa, Florida, USA.

Author contributions

B. M. Wright: Conceptualization, Investigation, Data curation, Formal analysis, Methodology, Writing – original draft, Writing – review & editing.

M. S. Bodnar: Conceptualization, Data curation, Methodology, Visualization, Writing – original draft, Writing – review & editing.

A. D. Moore: Conceptualization, Data curation, Methodology, Project administration, Writing – review & editing.

M. C. Maseda: Data curation, Writing – review & editing.

M. P. Kucharik: Data curation, Writing – review & editing.
 C. C. Diaz: Data curation, Writing – review & editing.
 C. M. Schmidt: Data curation, Writing – review & editing.
 H. R. Mir: Conceptualization, Methodology, Project administration, Supervision, Writing – review & editing.

Funding statement

The authors received no financial or material support for the research, authorship, and/or publication of this article.

ICMJE COI statement

No authors have any relevant disclosures or conflicts of interest. H. Mir reports consulting fees from Smith & Nephew and Synthes DePuy, unrelated to this article. H. Mir also reports a role on the board of directors for the Orthopaedic Trauma Association and the Center for Orthopaedics.

Data sharing

The data that support the findings for this study are available to other researchers from the corresponding author upon reasonable request.

Acknowledgements

Thank you to Krista J. Howard, Emily Coughlin, and Diep Nguyen for help with statistical analyses.

Ethical review statement

No ethics board approval was necessary since neither patient information nor participants were used.

Open access funding

The open access fee for this article was self-funded.

© 2024 Wright et al. This is an open-access article distributed under the terms of the Creative Commons Attribution Non-Commercial No Derivatives (CC BY-NC-ND 4.0) licence, which permits the copying and redistribution of the work only, and provided the original author and source are credited. See <https://creativecommons.org/licenses/by-nc-nd/4.0/>