Understanding How ChatGPT May Become a Clinical Administrative Tool Through an Investigation on the Ability to Answer Common Patient Questions Concerning Ulnar Collateral Ligament Injuries

Nathan H. Varady,* MD, MBA, Amy Z. Lu,† BS, Michael Mazzucco,† BS, Joshua S. Dines,* MD, David W. Altchek,† MD, Riley J. Williams III,* MD, and Kyle N. Kunze,*‡ MD

Investigation performed at the Hospital for Special Surgery, New York, New York, USA

Background: The consumer availability and automated response functions of chat generator pretrained transformer (ChatGPT-4), a large language model, poise this application to be utilized for patient health queries and may have a role in serving as an adjunct to minimize administrative and clinical burden.

Purpose: To evaluate the ability of ChatGPT-4 to respond to patient inquiries concerning ulnar collateral ligament (UCL) injuries and compare these results with the performance of Google.

Study Design: Cross-sectional study.

Methods: Google Web Search was used as a benchmark, as it is the most widely used search engine worldwide and the only search engine that generates frequently asked questions (FAQs) when prompted with a query, allowing comparisons through a systematic approach. The query "ulnar collateral ligament reconstruction" was entered into Google, and the top 10 FAQs, answers, and their sources were recorded. ChatGPT-4 was prompted to perform a Google search of FAQs with the same query and to record the sources of answers for comparison. This process was again replicated to obtain 10 new questions requiring numeric instead of open-ended responses. Finally, responses were graded independently for clinical accuracy (grade 0 = inaccurate, grade 1 = somewhat accurate, grade 2 = accurate) by 2 fellowship-trained sports medicine surgeons (D.W.A, J.S.D.) blinded to the search engine and answer source.

Results: ChatGPT-4 used a greater proportion of academic sources than Google to provide answers to the top 10 FAQs, although this was not statistically significant (90% vs 50%; $P = .14$). In terms of question overlap, 40% of the most common questions on Google and ChatGPT-4 were the same. When comparing FAQs with numeric responses, 20% of answers were completely overlapping, 30% demonstrated partial overlap, and the remaining 50% did not demonstrate any overlap. All sources used by ChatGPT-4 to answer these FAQs were academic, while only 20% of sources used by Google were academic ($P = .0007$). The remaining Google sources included social media (40%), medical practices (20%), single-surgeon websites (10%), and commercial websites (10%). The mean (± standard deviation) accuracy for answers given by ChatGPT-4 was significantly greater compared with Google for the top 10 FAQs (1.9 ± 0.2 vs 1.2 ± 0.6; $P = .001$) and top 10 questions with numeric answers (1.8 ± 0.4 vs 1 ± 0.8; $P = .013$).

Conclusion: ChatGPT-4 is capable of providing responses with clinically relevant content concerning UCL injuries and reconstruction. ChatGPT-4 utilized a greater proportion of academic websites to provide responses to FAQs representative of patient inquiries compared with Google Web Search and provided significantly more accurate answers. Moving forward, ChatGPT has the potential to be used as a clinical adjunct when answering queries about UCL injuries and reconstruction, but further validation is warranted before integrated or autonomous use in clinical settings.

Keywords: artificial intelligence; elbow; large language model; patient information; sports; technology

It is estimated that 70% to 90% of Americans search for health information online, with the vast majority of patients doing so before seeing a physician.[7] Data suggest that patients directly act on this information, and

The Orthopaedic Journal of Sports Medicine, 12(7), 23259671241257516
DOI: 10.1177/23259671241257516
© The Author(s) 2024

inaccurate or misleading information may lead to patient harm.[22,23] Therefore, it is critical as a medical community to understand what patients are seeing online, both to be informed when patients present to the office and to ensure quality control and initiate efforts to improve online information when needed.[32] As a result, considerable attention has been devoted to understanding what forums patients use to obtain health-related information and the quality of information provided by these resources.

Recently, there has been a dramatic advancement in the consumer availability of generative artificial intelligence (AI) large language models (LLMs) capable of responding to language commands and returning information based on pattern recognition and reinforcement learning.[15] In particular, these models have been popularized as open-source chatbots, such as chat generator pretrained transformer (GPT), with research suggesting ChatGPT (Open AI) is the fastest-growing consumer application in history.[10] The unsupervised and automated functions of ChatGPT are poised to provide information concerning a diverse range of subject matter, including health information. Recent studies have evaluated ChatGPT's performance on the United States Medical Licensing Examination,[9] compared its ability to post answers to an online health forum to human physicians,[1] and discussed its implications with respect to academic integrity.[16,19,26] More recently, several investigations have tested the ability of ChatGPT to triage and answer questions about various medical conditions.[8,28] However, there is a paucity of data on musculoskeletal conditions to date, which represents a substantial proportion of the health care burden nationally and globally.[2,24,31] Therefore, understanding its capabilities in this domain is clinically relevant.

Given the potential impact that an automated chatbot could confer on reducing administrative and clinical burden should it prove valid and accurate, the aim of the present study was to evaluate the ability of ChatGPT-4 to respond to patient inquiries concerning ulnar collateral ligament (UCL) injuries and compare these results with the performance of Google. We hypothesized that there would be no significant differences in the ability to compile answers to patients' frequently asked questions (FAQs) or in the accuracy of answers compared with Google.

METHODS

Search Engines and Rationale

This study did not examine human subjects; thus, it was exempt from institutional review board approval. This study evaluated the performance of ChatGPT-4 to deliver online health information with respect to simulated patient queries pertaining to both qualitative and quantitative information relating to UCL injuries. ChatGPT-4 is an AI-derived LLM that generates realistic human responses through a chatbot function. It is trained through supervised and reinforcement learning to optimize the accuracy, breadth, and relevance of responses to text prompts utilizing billions of modeling parameters and primarily information obtained from contemporary internet resources.[15]

UCL injuries were chosen as the ideal test case to understand how ChatGPT-4 may respond to patient queries and serve as a clinical adjunct for several reasons. First, rehabilitation protocols are graded, variable, and specific and require prolonged recovery periods in some cases.[14,17] Therefore, accurate information and expeditious treatment can be of great importance. Second, these injuries may be treated conservatively or operatively depending on several injury-related and patient-related factors.[3,6] Therefore, information pertaining to this decision may be highly searched. Similarly, there are multiple operative treatment options (ie, reconstruction versus repair), and 1 particular surgery is not necessarily appropriate for all patients (ie, in contrast to total joint arthroplasty for osteoarthritis). Finally, these injuries typically impact a young population familiar with new technologies who, therefore, may be early adopters of ChatGPT.

Google was chosen as the control case, as it is the most widely used search engine worldwide and the only search engine that generates FAQs when prompted with a query.[29] FAQs were specifically selected for investigation for the following reasons: (1) these are the questions that are the most commonly asked and, hence, of greatest interest to patients; (2) their use allows for objective assessment without bias from the authors in question generation; and (3) their use provides a systematic and reproducible method of question generation to compare between Google and ChatGPT-4.

Patient-replicated Query

A freshly installed browser was employed to minimize potential bias from previous search history, cookies, or cache. The first analysis was performed to identify the top FAQs on UCL injuries and reconstruction on Google and compare the relevance of the top questions as purported by ChatGPT-4 using Google as the historical gold

[‡]Address correspondence to Kyle N. Kunze, MD, Department of Orthopaedic Surgery, Hospital for Special Surgery, 535 East 70th Street, New York, NY 10021, USA (email: kylekunze7@gmail.com).

[*]Department of Orthopaedic Surgery, Hospital for Special Surgery, New York, New York, USA.

[†]Weill Cornell Medical College, New York, New York, USA.

Final revision submitted November 14, 2023; accepted January 10, 2024.

One or more of the authors has declared the following potential conflict of interest or source of funding: J.S.D. has received consulting fees from Arthrex and Trice Medical; royalties or license payments from Linvatec; nonconsulting fees from Arthrex; education payments from Gotham Surgical Solutions & Devices; and hospitality payments from Horizon Pharma. D.W.A. has received royalties or license payments from Stryker and hospitality payments from Arthrex. R.J.W. has received royalties or license payments from Arthrex; acquisitions from Smith+Nephew; nonconsulting fees from Arthrex; consulting fees from Arthrex; and hospitality payments from Joint Restoration Foundation. AOSSM checks author disclosures against the Open Payments Database (OPD). AOSSM has not conducted an independent investigation on the OPD and disclaims any liability or responsibility relating thereto.

Ethical approval was not sought for the present study.

TABLE 1
Modified Rothwell Criteria for Classification of Online Sources[13]

Website Categorization	
Commercial	Organizations that provide public health information, including medical device/manufacturing/pharmaceutical companies, and news outlets
Academic	Universities, academic medical centers, or academic societies
Medical practice	Local hospitals or medical groups without clear academic affiliation
Single surgeon practice	Personal websites maintained by individual surgeons
Government	Websites maintained by a national government
Social media	Blog, internet forms, support groups, and nonmedical organizations designed for information and video sharing

standard. On May 5, 2023, Google Web Search was first searched for the query “ulnar collateral ligament reconstruction.” The top 10 FAQs, answers, and their sources were recorded. FAQs were excluded if they were not relevant to UCL injuries of the elbow or if they represented a duplicate question. ChatGPT-4 was then queried as the experimental arm with the command “Perform a Google Search with the search term ‘ulnar collateral ligament reconstruction’ and record the most popular questions related to this search term,” in accordance with previously validated methods.[5] These additional 10 questions, answers, and their associated sources were also recorded for comparison. Concordance between the most commonly asked questions, as purported by Google and ChatGPT-4, was assessed. Question subjects—including cost, evaluation of surgery, indications and management, longevity, risks and complications, specific activities, technical details, and timeline of recovery using a modification of the Rothwell criteria, in accordance with previous methods[13]—were then assessed. Finally, we evaluated answers for the sources from which they were derived (ie, academic, commercial, group medical practice, single-surgeon practice, and social media) (Table 1).

A second independent search, after clearing all browsing history and memory, was then performed for 10 additional FAQs with answers necessitating a discrete, numeric value. By limiting the analysis to the most common questions with discrete answers, an objective comparison between search engines could be performed. For the numerical questions, answers were compared between Google and ChatGPT-4 for concordance, with answers rated as complete overlap, partial overlap (ie, if ranges provided), or completely discordant (or no overlap). All categorization was performed by the same 2 independent authors (A.Z.L., M.M.), with a third author called upon in the event of a discrepancy (K.N.K.).

Expert Comparison for Ground Truth Accuracy

Answers were assessed for accuracy based on the clinical judgment of 2 fellowship-trained sports medicine surgeons who have been in practice for >15 years and specialize in shoulder and elbow injuries (D.W.A., J.S.D.). Furthermore, these 2 surgeons assessed the accuracy and the clinical relevance of the top 10 general FAQs (Supplemental Table S1). Both graders were blinded to whether the answer was from Google or ChatGPT-4 and the source of information used to answer the question. The accuracy of each answer was rated on a 3-point response scale as inaccurate (0 points), somewhat accurate (1 point), or accurate (2 points). The clinical relevance of the top 10 FAQs was also graded on a 3-point scale as little to no clinical relevance (0 points), some clinical relevance (1 point), or very clinically relevant (2 points).

Statistical Analysis

Descripted statistics were computed and presented as frequencies with percentages. Comparisons of categorical data were made with chi-square test or Fisher exact test, as appropriate. Comparisons of continuous data were performed with Student t test. Statistical analyses were performed using Microsoft Excel (Microsoft) and GraphPad Prism Version 9.5.1 (GraphPad Software), with statistical significance being defined by using a $P < .05$ threshold in all circumstances.

RESULTS

Most Common Patient Queries for Google and ChatGPT-4

The top 10 FAQs concerning UCL injuries on both Google and ChatGPT-4 are presented in Table 2. Overall, question overlap was observed for 4/10 (40%) of the questions elicited by the respected queries. The most common questions for Google were related to evaluation of surgery (2/10, 20%), indications and management (2/10, 20%), and specific activities (2/10, 20%), while the most common questions for ChatGPT-4 were related to indications and management (3/10, 30%) and technical details (3/10, 30%; Figure 1). The 4 overlapping questions were related to 1 each (1/10, 10%) of technical details, timeline of recovery, evaluation of surgery, and longevity. Answers to these questions and their sources can be found in Supplemental Table S1. When graded by 2 experts blinded to search engine and answer source, the mean (± standard deviation) accuracy/correctness of answers given by ChatGPT-4 and Google were 1.9 ± 0.2 and 1.2 ± 0.6, respectively,

TABLE 2
Top 10 Open-Ended Internet FAQs for UCL Reconstruction per Google and ChatGPT-4[a]

Google	ChatGPT-4
1. How is UCL reconstruction done? (technical details)	1. What is UCL reconstruction? (technical details)
2. How long does it take to recover from UCL surgery? (timeline of recovery)	2. What is the purpose of UCL reconstruction surgery? (indications and management)
3. What age are patients that get Tommy John surgery? (indications and management)	3. How is UCL reconstruction surgery performed? (technical details)
4. What is the success rate of UCL reconstruction? (evaluation of surgery)	4. Who are the typical candidates for UCL reconstruction surgery? (indications and management)
5. Does Tommy John make you throw harder? (specific activities)	5. What is the success rate of UCL reconstruction surgery? (evaluation of surgery)
6. How long does it take to recover from a UCL injury? (timeline of recovery)	6. How long does it take to recover from UCL reconstruction surgery? (timeline of recovery)
7. Can you still pitch after Tommy John surgery? (specific activities)	7. What are the potential risks and complications of UCL reconstruction surgery? (risks and complications)
8. How long does UCL reconstruction last? (longevity)	8. What is the difference between UCL repair and UCL reconstruction? (technical details)
9. How much does Tommy John surgery cost? (cost)	9. How long does UCL reconstruction surgery last? (longevity)
10. What is Tommy John surgery for? (indications and management)	10. Can UCL reconstruction be performed more than once? (indications and management)

[a]FAQs are listed in order of appearance for each search engine. Overlapping questions are shaded gray. FAQs, frequently asked questions; GPT, generator pretrained transformer; QB, quarterback; UCL, ulnar collateral ligament.

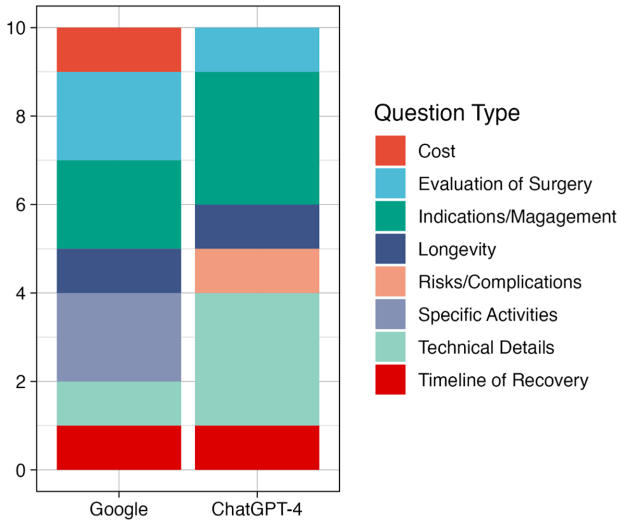

Figure 1. Question type for the 10 most commonly encountered questions relating to ulnar collateral ligament reconstruction for Google and ChatGPT-4. GPT, generator pretrained transformer.

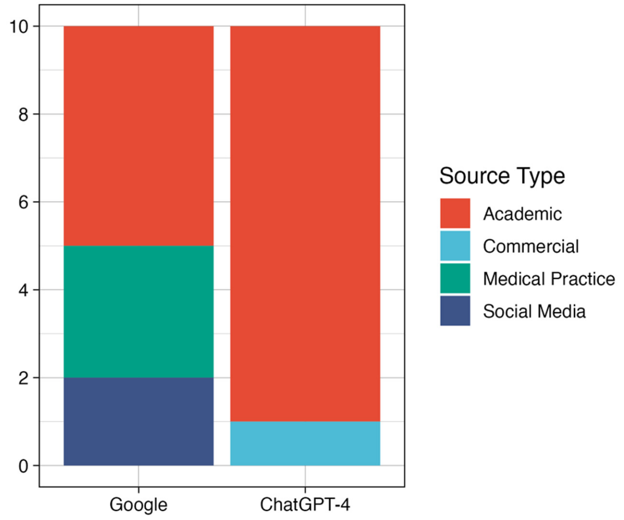

Figure 2. Sources of answers to the 10 most common questions for Google and ChatGPT-4. GPT, generator pretrained transformer.

representing a statistically significant difference in favor of ChatGPT-4 (P = .001).

In analyzing the source of answers to the 10 most common FAQs on the respective platforms, the sources utilized for Google answers to the top 10 FAQs were 50% academic, with the remaining 50% being answered from medical practices (3/10, 30%) and social media (2/10, 20%) (Figure 2). In contrast, the sources used by ChatGPT-4 were 90% academic, with the remaining FAQs being answered using a commercial source; however, although ChatGPT-4 used a greater proportion of academic resources, this difference failed to reach statistical significance (P = .14).

Most Common Patient Queries With Numeric Answers

The 10 most commonly asked numerical questions, with the respective answers from both Google and ChatGPT-4, are listed in Table 3. The most common question types

TABLE 3
Top 10 Questions and Answers (Sources) Comparison of Search Engines for Numerical Questions for UCL Reconstruction[a]

Question	Google Answer	Grade	ChatGPT-4 Answer	Grade
What is the success rate of Tommy John surgery? (evaluation of surgery)	80%-90% (academic)	2	The success rate of Tommy John surgery, also known as ulnar collateral ligament (UCL) reconstruction, is generally high, with approximately 80% to 90% of patients returning to their previous level of sports participation. (academic)	2
How long are you out with Tommy John surgery? (timeline of recovery)	1 Year (social media)	1	Generally, a pitcher can expect to return to competitive pitching between 12 and 18 months after the surgery. However, some players may take longer to regain their previous level of performance. (academic)	2
How much does Tommy John surgery cost? (cost)	$10,000 to $26,000 (medical practice)	2	The average cost of surgery in the United States ranges from $20,000 to $30,000, but this does not include additional expenses such as anesthesia, imaging, and physical therapy. (academic)	2
How long does UCL reconstruction last? (longevity)	8 to 9 months (medical practice)	0	Many years (academic)	1
When can you pitch after Tommy John surgery? (specific activity	18 months (academic)	1	12 to 18 months (academic)	2
How long after Tommy John can you move your arms? (timeline of recovery)	2 to 4 months (single surgeon practice)	0	First week or two (academic)	2
How many pitchers have had more than one Tommy John surgery? (indications/management)	Well over 1000 professional pitchers (social media)	1	While the exact number of pitchers who have undergone multiple Tommy John surgeries is not readily available, it is estimated that around 15% to 20% of professional baseball players. (academic)	1
How much harder do you throw after Tommy John? (specific activities)	Pitching velocity goes up 1.5 mph for every year of age and 1.2 mph for every inch of height (social media)	0	There is no guarantee that a pitcher will throw harder after Tommy John surgery. The primary goal of the surgery is to restore stability and function to the elbow joint. While some pitchers may experience an increase in velocity after surgery, this is not a consistent outcome. (academic)	2
How long are you in a sling after Tommy John? (restrictions)	7 to 10 days (Commercial)	2	7 to 10 days (Academic)	2
What age is Tommy John surgery for? (indications/management)	60% of Tommy John surgeries are for patients aged 15 to 19 (social media)	1	While Tommy John surgery can be performed on individuals of various ages, it's most common among high school, college-aged athletes, and professional baseball players. (academic)	2

[a] Answers with complete overlap are shaded green, answers with partial overlap are shaded yellow, and answers with no overlap are shaded red. Expert grading scale: grade 0, inaccurate; grade 1, somewhat accurate; and grade 2: accurate. UCL, ulnar collateral ligament.

were related to the timeline of recovery (2/10, 20%), indications and management (2/10, 20%), and specific activities (2/10, 20%). Overall, only 2 of 10 (20%) answers were the same, 3 of 10 (30%) demonstrated partial overlap, and 5 of 10 (50%) were completely discordant (Figure 3). All sources used by ChatGPT-4 to answer these FAQs were academic, while only 20% of sources used by Google were academic (*P* = .0007; Figure 4). The remaining Google sources included social media (40%), medical practices (20%), single-surgeon websites (10%), and commercial websites (10%).

When comparing the answers to these questions based on a blinded assessment by the 2 experts, the mean accuracy/correctness of answers was 1.8 ± 0.4 for answers provided by ChatGPT-4 versus 1 ± 0.82 for answers provided by Google, representing a statistically significant difference in favor of ChatGPT-4 (*P* = .013). Furthermore, there were no answers graded as inaccurate by ChatGPT-4, whereas Google had 3 answers graded as inaccurate. Across all 20 answers to both qualitative and quantitative questions, 16 of 20 (80%) were accurate for ChatGPT-4 compared with 5 of 20 (25%) for Google (*P* = .0005). When considering the clinical relevance of answers to FAQs, the mean clinical relevance score was 1.8 ± 4.2 for ChatGPT-4 versus 1.6 ± 0.7 for Google, which was

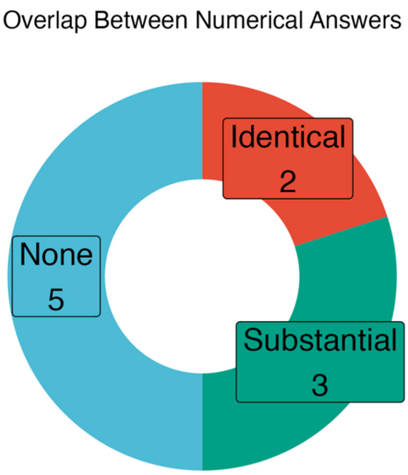

Figure 3. Overlap in answers to the 10 most common questions with numerical answers between Google and ChatGPT-4. GPT, generator pretrained transformer.

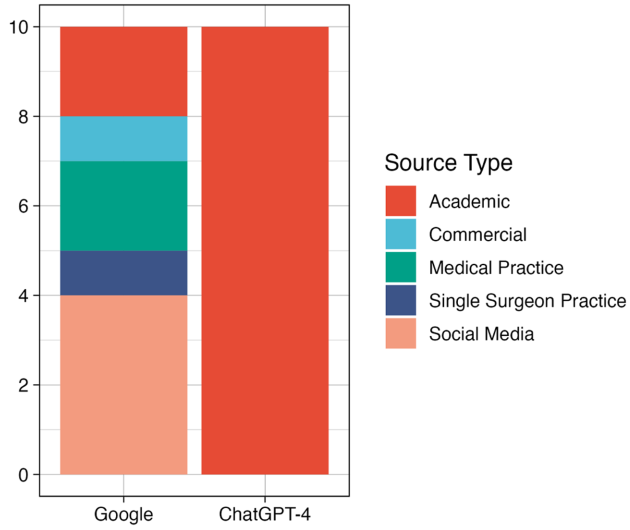

Figure 4. Sources of answers to the 10 most common questions with numerical answers between Google and ChatGPT-4. GPT, generator pretrained transformer.

not statistically different (P = .44). Only Google provided a FAQ graded as little to no clinical relevance, while all FAQs provided by ChatGPT-4 were graded as having some clinical relevance or being very clinically relevant.

DISCUSSION

The main findings of the present study were as follows: (1) ChatGPT-4 provided a comprehensive range of clinically relevant questions and answers across various topics related to UCL injuries using primarily academic resources; (2) queries yielded relatively little (40%) content overlap between Google and ChatGPT-4 in terms of the most FAQs; (3) ChatGPT-4 used a greater proportion of academic resources to answer questions compared with Google; and (4) ChatGPT-4 provided answers that were significantly more accurate compared with Google based on grading by a blinded expert. These results have several implications for both sports medicine patients and physicians.

As the adoption of ChatGPT-4 continues to grow at a breakneck pace,[10] more sports medicine patients are likely to turn to this service for online health information. The current results demonstrate that patients with UCL injuries using ChatGPT-4 are more likely to obtain information from academic sources compared with patients using Google. These findings are encouraging, as patients are likely to receive information that is both more accurate and less biased from academic sources compared with other sources (eg, commercial, social media, etc), which were frequently present in the Google results. In addition, an anecdotal analysis of the sources for the ChatGPT-4 answers demonstrated information was consistently derived from a mix of both reputable institutions, such as the American Academy of Orthopaedic Surgeons and the Mayo Clinic, as well as the primary medical literature, while Google used information from reputable sources in only few circumstances. This is also a positive finding, as it suggests ChatGPT-4 may source all available information for the most relevant answer, as opposed to deriving most of its answers from a single or few sources, which could also induce bias, even if from an academic resource.

In line with these expectations, and perhaps most importantly, we found that ChatGPT-4 answers were significantly more accurate than Google Web Search answers when assessed by 2 blinded fellowship-trained sports medicine surgeons. Moreover, no answers from ChatCPT-4 were judged to be completely inaccurate, compared with 3 answers from Google, while 16 of 20 (80%) answers were completely accurate for ChatGPT-4 compared with only 5 of 20 (25%) from Google. The high accuracy of ChatGPT-generated answers is consistent with and expands upon what has been observed in other fields. For instance, Samaan et al[25] assessed the accuracy of ChatGPT for answering questions pertaining to bariatric surgery where responses were graded by board-certified bariatric surgeons. The authors reported that ChatGPT provided comprehensive and accurate answers to 86.8% of the questions but did not compare the performance to current gold standard sources of information such as Google. In the cancer literature,[11] the accuracy of answers given by ChatGPT has been reported to exceed 95%. Taken together, the findings from the present study suggest that information on UCL injuries and reconstruction that patients obtain online from ChatGPT-4 is from high-quality sources and unlikely to be worse than when obtained from Google. Should patients inquire to search for more information about their condition online, physicians may choose to direct them to ChatGPT-4 as opposed to traditional search engines such as Google. However, future studies are necessary to continue to validate the accuracy and reliability of the information provided by ChatGPT-4 across a range of conditions.

Interestingly, there was relatively little overlap in the most commonly asked questions between Google and ChatGPT-4. The purpose of this analysis was to systematically analyze the relevance of FAQs generated by ChatGPT-4, compared with the most frequently used search engine in the world.[1] If Google is assumed to be the gold standard of relevant questions, then 1 interpretation of this finding could be that when nonspecifically prompted, ChatGPT-4 may not return the most common questions patients have. This finding could be due to the relative recency of ChatGPT-4 as a platform, and ongoing calibration through additional searches may observe results becoming more concordant over time. On the other hand, it is possible that the most commonly asked questions from Google over decades no longer represent the most common questions about UCL injuries and reconstruction that patients have today, and, as such, the ChatGPT-4 results could be more relevant and contemporary. In either event, patients can always ask ChatGPT-4 the specific questions they have, and when asked the same questions in the current study, ChatGPT-4 was more likely to respond with academically sourced and accurate answers.

The present study builds on previous work in other fields, suggesting that ChatGPT-4 may be an improved and reliable patient resource. Pertaining to musculoskeletal conditions, Dubin et al[5] prompted ChatGPT-3 with 10 questions on total knee arthroplasty and total hip arthroplasty and utilized Google Web Search to assess the utility of ChatGPT in total joint arthroplasty. They reported that ChatGPT provided a heterogeneous range of questions and answers compared with Google, with a high proportion of responses being informed by PubMed, which is in accordance with the finding in this study that ChatGPT has a predilection for using academic resources. Ayers et al[1] recently demonstrated that ChatGPT responded to patient questions on an online medical forum with significantly greater quality and empathy compared with verified physicians. On the other hand, ChatGPT has failed to exhibit the higher-level thinking required to pass the Fellowship of the Royal College of Surgeons (Trauma & Orthopaedics) examination, scoring only 35.8% (30% lower than the Fellowship of the Royal College of Surgeons pass rate)[4] and only in the 40th percentile for the Orthopaedic In-Training Examination.[20] Thus, while ChatGPT continues to improve and is likely to be an increasingly valuable clinical adjunct moving forward, it is far from a panacea that can replace clinical judgment. Therefore, expectations for its ability to answer medical questions should be measured. Regardless, with all the attention surrounding this technology, it is going to become increasingly common for sports medicine physicians to encounter patients who have obtained information from ChatGPT; therefore, it is critical for clinicians to be aware of this resource and the quality of information it provides. LLMs such as ChatGPT may eventually play an increasing role in other aspects of clinical practice besides serving as an information resource, with emerging literature demonstrating the potential to decrease administrative burden,[18] respond to patient inbox messages that are received, and effectively triage patients.[18,30] With increased validity and confidence in responses and information, ChatGPT may be incorporated into medical records and either send real-time responses or draft responses for providers to review, edit, and send, and screening systems could be implemented before new appointment bookings to ensure patients are seen by the appropriate provider (ie, new sports medicine patients with operative conditions routed to a sports medicine surgeon and those with nonoperative conditions routed to nonoperative sports medicine physicians). Although results from the present study are promising, continued validation and refinement of the ChatGPT-4 and its subsequent versions will be imperative before clinical use.

Limitations

Although this study has many strengths, including its novelty and systematic design to allow for objective comparison between Google and ChatGPT-4, it is not without limitations. First, both Google and ChatGPT-4 are dynamic resources that evolve over time. As such, the current results may not always hold in the future. Second, the study is limited to UCL injuries and reconstruction, and the results may not be generalizable to other conditions. Although previous literature concerning the use of ChatGPT has also restricted its search to a single topic involving a limited number of questions, continued research into the efficacy of ChatGPT as a source of online health information is likely to be beneficial. Third, the performance of a limited number of questions was assessed, which does not represent all possible questions or concerns a patient may have about this topic. However, previous literature has demonstrated samples of questions as low as 6 to provide a valid and representative sample of what patients may ask,[27] while other studies have reported on a more limited number of questions than investigated in the present study to demonstrate appropriate performance and reproducibility.[5,12,21] Fourth, although some statistically significant differences were observed between ChatGPT-4 and Google (ie, the mean medical accuracy of answers to general FAQs; Supplemental Table S1), it is unclear whether these differences are clinically significant. Fifth, patients may combine information from multiple resources (ie, both Google and ChatGPT-4), which was not assessed in this study. Last, although a methodological strength of this study is utilizing a new browser and erasing history with each iterative search, it is also a limitation because it may ultimately affect the generalizability of results as others may not achieve similar results if they do not refresh their browsers.

CONCLUSION

ChatGPT-4 is capable of providing responses with clinically relevant content concerning UCL injuries and reconstruction. ChatGPT-4 utilized a significantly greater proportion of academic websites to provide responses to FAQs representative of patient inquiries compared with

Google Web Search and provided significantly more accurate answers. Moving forward, ChatGPT can potentially be used as a clinical adjunct when answering queries about UCL injuries and reconstruction, but further validation is warranted before integrated or autonomous use in clinical settings.

Supplemental Material for this article is available at https://journals.sagepub.com/doi/full/10.1177/23259671241257516#supplementary-materials

REFERENCES

1. 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.
2. Blyth FM, Briggs AM, Schneider CH, Hoy DG, March LM. The global burden of musculoskeletal pain—where to from here? *Am J Public Health*. 2019;109(1):35-40.
3. Carr JB II, Camp CL, Dines JS. Elbow ulnar collateral ligament injuries: indications, management, and outcomes. *Arthroscopy*. 2020;36(5):1221-1222.
4. Cuthbert R, Simpson AI. Artificial intelligence in orthopaedics: can chat generative pre-trained Transformer (ChatGPT) pass Section 1 of the fellowship of the Royal College of Surgeons (Trauma & Orthopaedics) examination? *Postgrad Med J*. 2023;99(1176):1110-1114.
5. Dubin JA, Bains SS, Chen Z, et al. Using a Google web search analysis to assess the utility of ChatGPT in total joint arthroplasty. *J Arthroplasty*. 2023;38(7):1195-1202.
6. Erickson BJ, Harris JD, Chalmers PN, et al. Ulnar collateral ligament reconstruction: anatomy, indications, techniques, and outcomes. *Sports Health*. 2015;7(6):511-517.
7. Fox S, Duggan M. Health online 2013. Pew Research Center. Published January 15, 2013. Accessed August 1, 2023. https://www.pewresearch.org/internet/2013/01/15/health-online-2013/.2023.
8. Gebrael G, Sahu KK, Chigarira B, et al. Enhancing triage efficiency and accuracy in emergency rooms for patients with metastatic prostate cancer: a retrospective analysis of artificial intelligence-assisted triage using ChatGPT 4.0. *Cancers (Basel)*. 2023;15(14):3717.
9. Gilson A, Safranek CW, Huang T, et al. How does ChatGPT perform on the United States Medical Licensing Examination? The implications of large language models for medical education and knowledge assessment. *JMIR Med Educ*. 2023;9:e45312.
10. Gordon C. ChatGPT is the fastest growing App in the history of web applications. Forbes. Published February 2, 2023. Accessed August 1, 2023. https://www.forbes.com/sites/cindygordon/2023/02/02/chatgpt-is-the-fastest-growing-ap-in-the-history-of-web-applications/?sh=1bc69781678c
11. Johnson SB, King AJ, Warner EL, et al. Using ChatGPT to evaluate cancer myths and misconceptions: artificial intelligence and cancer information. *JNCI Cancer Spectr*. 2023;7(2):pkad015.
12. 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(11):5190-5198.
13. Kanthawala S, Vermeesch A, Given B, Huh J. Answers to health questions: internet search results versus online health community responses. *J Med Internet Res*. 2016;18(4):e95.
14. Kemler BR, Rao S, Willier DP III, et al. Rehabilitation and return to sport criteria following ulnar collateral ligament reconstruction: a systematic review. *Am J Sports Med*. 2022;50(11):3112-3120.
15. Kunze KN, Jang SJ, Fullerton MA, Vigdorchik JM, Haddad FS. What's all the chatter about? *Bone Joint J*. 2023;105-B(6):587-589.
16. Leopold SS, Haddad FS, Sandell LJ, Swiontkowski M. Artificial intelligence applications and scholarly publication in orthopaedic surgery. *J Bone Joint Surg Am*. 2023.
17. Lightsey HM, Trofa DP, Sonnenfeld JJ, et al. Rehabilitation variability after elbow ulnar collateral ligament reconstruction. *Orthop J Sports Med*. 2019;7(3):2325967119833363.
18. Liu S, McCoy AB, Wright AP, et al. Leveraging large language models for generating responses to patient messages. *medRxiv*. 2023:2023.07.14.23292669.
19. Lubowitz JH. ChatGPT, an artificial intelligence chatbot, is impacting medical literature. *Arthroscopy*. 2023;39(5):1121-1122.
20. Lum ZC. Can artificial intelligence pass the American Board of Orthopaedic Surgery Examination? Orthopaedic residents versus ChatGPT. *Clin Orthop Relat Res*. 2023;481(8):1623-1630.
21. Mika AP, Martin JR, Engstrom SM, Polkowski GG, Wilson JM. Assessing ChatGPT responses to common patient questions regarding total hip arthroplasty. *J Bone Joint Surg Am*. 2023;105(19):1519-1526.
22. Murray E, Lo B, Pollack L, et al. The impact of health information on the Internet on health care and the physician-patient relationship: national U.S. survey among 1.050 U.S. Physicians. *J Med Internet Res*. 2003;5(3):e17.
23. Olatunde Oduoye M, Javed B, Gupta N, Valentina Sih CM. Algorithmic bias and research integrity; the role of non-human authors in shaping scientific knowledge with respect to artificial intelligence (AI); a perspective. *Int J Surg*. 2023;109(10):2987-2990.
24. Safiri S, Kolahi AA, Cross M, et al. Prevalence, deaths, and disability-adjusted life years due to musculoskeletal disorders for 195 countries and territories 1990-2017. *Arthritis Rheumatol*. 2021;73(4):702-714.
25. Samaan JS, Yeo YH, Rajeev N, et al. Assessing the accuracy of responses by the language model ChatGPT to questions regarding bariatric surgery. *Obes Surg*. 2023;33(6):1790-1796.
26. Seth I, Bulloch G, Lee CHA. Redefining academic integrity, authorship, and innovation: the impact of ChatGPT on surgical research. *Ann Surg Oncol*. 2023;30(8):5284-5285.
27. Seth I, Cox A, Xie Y, et al. Evaluating chatbot efficacy for answering frequently asked questions in plastic surgery: a ChatGPT case study focused on breast augmentation. *Aesthet Surg J*. 2023;43(10):1126-1135.
28. Shah NH, Entwistle D, Pfeffer MA. Creation and adoption of large language models in medicine. *JAMA*. 2023;330(9):866-869.
29. Statcounter. Search Engine Market Share Worldwide—March 2023. Published March 2023. Accessed April 28, 2023. https://gs.statcounter.com/search-engine-market-share
30. Thirunavukarasu AJ, Ting DSJ, Elangovan K, et al. Large language models in medicine. *Nat Med*. 2023;29:1930-1940.
31. Weinstein SL. The burden of musculoskeletal conditions. *J Bone Joint Surg Am*. 2016;98(16):1331.
32. World Health Organization. WHO/Europe explores collaborations to improve quality of health information online. Updated June 16, 2023. Accessed August 9, 2023. https://www.who.int/europe/news/item/16-06-2023-who-europe-explores-collaborations-to-improve-quality-of-health-information-online#:~:text=WHO%2FEurope%20explores%20collaborations%20to%20improve%20quality%20of%20health%20information%20online,-16%20June%202023&text=The%20WHO%20Office%20on%20Quality,States'%20efforts%20in%20this%20area