Original Article

Microsoft Copilot Provides More Accurate and Reliable Information About Anterior Cruciate Ligament Injury and Repair Than ChatGPT and Google Gemini; However, No Resource Was Overall the Best

Suhasini Gupta, B.S., Rae Tarapore, M.D., Brett Haislup, M.D., and Allison Fillar, M.D.

Purpose: To analyze and compare the quality, accuracy, and readability of information regarding anterior cruciate ligament (ACL) injury and reconstruction provided by various artificial intelligence AI interfaces (Google Gemini, Microsoft Copilot, and OpenAI ChatGPT). **Methods:** Twenty questions regarding ACL reconstruction were inputted into ChatGPT 3.5, Gemini, and the more precise subinterface within Copilot and were categorized on the basis of the Rothwell criteria into Fact, Policy, and Value. The answers generated were analyzed using the DISCERN scale, JAMA benchmark criteria, and Flesch-Kincaid Reading Ease Score and Grade Level. The citations provided by Gemini and Copilot were further categorized by source of citation. **Results:** All 3 AI interfaces generated DISCERN scores ($\geq$50) demonstrating "good" quality of information except for Policy and Value by Copilot which were scored as "excellent" ($\geq$70). The information provided by Copilot demonstrated greater reliability, with a JAMA benchmark criterion of 3 (of 4) as compared with Gemini (1) and ChatGPT (0). In terms of readability, the Flesch-Kincaid Reading Ease Score scores of all 3 sources were $<$30, apart from Fact by Copilot (31.9) demonstrating very complex answers. Similarly, all Flesch-Kincaid Grade Level scores were $>$13, indicating a minimum readability level of college level or college graduate. Finally, both Copilot and Gemini had a majority of references provided by journals (65.6% by Gemini and 75.4% by Copilot), followed by academic sources, whereas Copilot provided a greater number of overall citations (163) as compared with Gemini (64). **Conclusions:** Microsoft Copilot was a better resource for patients to learn about ACL injuries and reconstruction compared with Google Gemini or OpenAI ChatGPT in terms of quality of information, reliability, and readability. The answers provided by LLMs are highly complex and no resource was overall the best. **Clinical Relevance:** As artificial intelligence models continually evolve and demonstrate increased potential for answering complex surgical questions, it is important to investigate the quality and usefulness of the responses for patients. Although these resources may be helpful, they should not be used as a substitute for any discussions with health care providers.

Anterior cruciate ligament (ACL) rupture is one of the most common sports-related injuries, with a substantial global disease burden and the cause of an increasing rate of ACL reconstruction and repair.[1-4] There is an increasing rate of these injuries in the young patient population, with an incidence of ACL tears at 68.6 per 100,000 people, peaking between 19 and 25 years old in male patients and between 14 and 18 years old in female patients.[5] This patient population especially has been shown to have increased access to and interest in obtaining health-information from online information sources outside of their surgeon, such as different social-networking and media sites such as YouTube.[6,7] Despite this plethora of available information, previous studies have demonstrated a variable quality and readability of online orthopaedics information available to patients.[8,9]

Given that the availability of different language-learning models (LLMs), medical research has been expanding to assess the possibility of artificial intelligence (AI) platforms providing personalized health

From UMass Chan Medical School, Worcester, Massachusetts, U.S.A. (S.G.); and Department of Orthopaedic Surgery, MedStar Union Memorial Hospital, Baltimore, Maryland, U.S.A. (R.T., B.H., A.F.), U.S.A.

Received July 6, 2024; accepted October 30, 2024.

Address correspondence to Rae Tarapore, M.D., 777 S. Eden St. Apt 726, Baltimore, MD 21202, U.S.A. E-mail: raetarapore@gmail.com

2666-061X/241062

https://doi.org/10.1016/j.asmr.2024.101043

education to patients regarding their injury, procedure, and rehabilitation.[10,11] The most popularly used and most widely studied LLM has been ChatGPT, which already had a 100 million user-base in January 2023, after its launch 2 months before in November 2022.[10,12] As such, some studies have assessed the information provided by ChatGPT in regard to ACL injury and reconstruction and have demonstrated its ability to provide accurate information with noted limitations in terms of reliability and readability.[11,13]

There is an increased availability of various other AI platforms with different features, including citations, ability to switch between different versions of the software to obtain different kinds of information, and hyperlinks to view images and videos related to the topic being discussed. There has been an increased interest in analyzing and comparing the differences between these different platforms in terms of health literacy and patient education in nonorthopaedics medical research. Three studies from the past year compared ChatGPT with Microsoft and Google AI products in various medical fields. The results from these studies were mixed, with no singular AI tool consistently displaying superiority.[14-16] Notably, all studies cited concerns regarding readability and reliability in regards to all LLM models when used in a medical setting.

The purpose of this study is to analyze and compare the quality, accuracy, and readability of information regarding ACL injury and reconstruction provided by various AI interfaces (Google Gemini, Microsoft Copilot, and OpenAI ChatGPT). We hypothesized that although all 3 AI platforms would be able to supplement perioperative information regarding ACL injury and reconstruction, they would be unable to be the sole source of adequate information for patient education.

Methods

Twenty questions regarding ACL injury and reconstruction (Table 1) were obtained from a previous study that compiled the questions through literature search, consensus statements in the field, and commonly asked patient questions.[11] These were inputted into ChatGPT 3.5, Gemini, and the more precise subinterface within Copilot. The more precise subinterface in Copilot was selected, as it is best for "fact-finding" as per the platform's instructions.

The questions were categorized on the basis of the Rothwell criteria into Fact, Policy, and Value. Under this criterion, a "Fact" question is one that asks whether something is true, and to what extent.[17,18] A "Policy" question asks whether a certain course of action should be taken to solve a problem. A "Value" question asks for evaluation of an idea, object, or event. The answers generated by the platforms were then analyzed using the DISCERN scale to assess the quality of information, Journal of the American Medical Association (JAMA) benchmark criteria to assess the accuracy, and Flesch-Kincaid Reading Ease Score (FRES) and grade Level (FKGL) to assess the readability.

The DISCERN score scale is used by similar orthopaedics studies to access the quality of the source of information and comprises of 16 questions that judge the reliability, quality of information provided on treatment choices and overall rating of the publication.[11,17-19] Each question is scored from 1 to 5, with 1 corresponding to the poorest quality or "no" and 5 corresponding with the highest quality or "yes." The overall summation of the 16 questions is 80, with a score >50 indicating a "good" source of information and >70 indicating an "excellent" source of information. In this study, there were 2 assessors (S.G. and R.T.) who independently scored each LLM. The discrepancies between the 2 scorers were minimal; any discrepancies were later discussed amongst the two scorers until a consensus was reached.

The JAMA benchmark criteria assess the reliability of a publication providing health information on the basis of 4 sections, with an overall maximum score of 4.[17,20] The first section, "authorship," identifies the person or group that produces the information, "attribution" assesses if citations and references are provided, "disclosures" assesses if information regarding who "owns" the information is available, and "currency" assesses whether dates for when the information was last reviewed or posted are provided.[20,21]

The FRES and FKG were used to assess readability, as have been done in similar studies, as well as by the American education system.[18] The FRES score is of 100, with a lower score corresponding to a more complex test, and a FKGL score assesses the minimum grade reading level needed to understand a text, with a lower score indicating a lower minimum grade reading level. The formula to calculate FKGL is: 206.835 − 1.015 * (total words/total sentences) − 84.6 * (total syllables/total words), and the formula to calculate FRES is 0.39 * (total words/total sentences) + 11.8 * (total syllables/total words) − 15.59.

Finally, the citations provided by both Google Gemini and Microsoft Copilot were categorized by source of citation, including nonmedical media site, commercial, single surgeon practice, academic, medical information site, journal, and government. These categories have been used by similar orthopaedics studies to assess the type of websites patients use to seek answers regarding questions related to shoulder arthroplasty.[22] Citations could not be assessed for ChatGPT, as they were not provided. Statistical analysis was performed using an unpaired T-test to calculate differences in mean DISCERN, FRES, and FKGL

Table 1. Commonly Asked Questions Anterior Cruciate Ligament Injury and Repair, Categorized By Rothwell Criteria

Rothwell Criteria	Anterior Cruciate Ligament Injury and Repair
Fact	1. What movement patterns and what muscle groups should be trained to avoid anterior cruciate ligament injury and/or anterior cruciate ligament graft rupture?
	2. What strategies should be used to counteract kinesophobia?
	3. What are the most important risk factors for postoperative knee stiffness following anterior cruciate ligament reconstruction?
	4. What are the risks and benefits of stump preservation in anterior cruciate ligament reconstruction?
	5. What are the indications for anterior cruciate ligament injuries in children and when should they be treated?
	6. Where should the femoral tunnel be placed in the setting of anterior cruciate ligament reconstruction?
	7. What are the indications for anterior cruciate ligament repair and when should it be performed and for which patient category?
	8. What are the indications for anterior cruciate ligament repair and when should it be performed and for which patient category?
	9. What does accelerated rehabilitation mean?
Policy	10. What are the indications for use of different anterior cruciate ligament graft types?
	11. When should a lateral augmentation procedure be added to an anterior cruciate ligament reconstruction?
	12. When can anterior cruciate ligament injury be treated nonoperatively?
	13. What is the most ideal timing for anterior cruciate ligament reconstruction in the setting of concomitant meniscus bucket handle tear and locked knee?
	14. How can patellofemoral problems following anterior cruciate ligament reconstruction be avoided?
	15. When should a medial collateral ligament reconstruction be performed in the setting of anterior cruciate ligament injury?
	16. When should osteotomy be performed in the setting of revision anterior cruciate ligament reconstruction?
Value	17. On the basis of the current knowledge is it worthwhile to biologically enhance anterior cruciate ligament surgery (using platelet-rich plasma, stem cells, microfragmented adipose tissue, plasma clot matrix, fibrin clot, hyaluronic acid)?
	18. What is the most cost effective treatment for a patient with anterior cruciate ligament tear? If surgery is chosen is there a difference in cost between treatments and what is it?
	19. Should the same surgical technique and graft type be used in revision anterior cruciate ligament reconstruction as in primary anterior cruciate ligament reconstruction?
	20. From a risk factor standpoint, what would be the worst outcome case scenario leading to anterior cruciate ligament injury?

scores between LLMs, and a P value $< .05$ was considered statistically significant.

Results

For the ChatGPT answers, the DISCERN score for Fact was 58, Policy was 59, and Value was 59. For the Microsoft Copilot answers the DISCERN score for Fact was 67, Policy was 70, and Value was 70. For the Google Gemini answers the DISCERN score for Fact was 62, Policy was 65, and Value was 65 (Table 2). T-test analysis yielded statistically significant differences between the LLMs when comparing the mean DISCERN scores across Fact, Policy, and Value questions (Appendix Table 1, available at www.arthroscopyjournal.org).

The JAMA benchmark criteria for Fact, Policy, and Value questions were 0 of 4 for ChatGPT (as no information was provided regarding Authorship, Disclosures, Currency or Attribution), 1 of 4 for Google Gemini (as Authorship information was provided), and 3 of 4 for Microsoft Copilot (as Authorship, Currency and Attribution information was provided).

For the answers provided by ChatGPT (Table 3), the FRES score for Fact was 11.78, Policy was 6.67, and Value was 2.65. The FKGL score for Fact was 17.17, Policy was 17.82, and Value was 19.32.

For the answers provided by Microsoft Copilot, the FRES score for Fact was 31.9, Policy was 21.5, and Value was 28.5. The FKGL score for Fact was 13.1, Policy was 15.6, and Value was 14.3.

For the answers provided by Google Gemini, the FRES score for Fact was 21.6, Policy was 13.3, and Value was 18.9. The FKGL score for Fact was 15.5, Policy was 17.1, and Value was 16.1.

t-test analysis demonstrated statistically significant differences between the LLMs when comparing the mean FRES scores across Fact, Policy, and Value questions, for Copilot versus ChatGPT and ChatGPT versus Gemini, and for mean FKGL between Copilot versus ChatGPT (Appendix Table 1).

There were a total of 64 citations provided by Google Gemini and a total of 163 citations provided by Copilot (Table 4). Both Copilot and Gemini had a majority of references provided by Journals (65.6% by Gemini and 75.4% by Copilot), followed by Academic sources (20.31% by Gemini and 8.58% by Copilot). The lowest number of citations were from Government and Non-Medical Mediate site sources (0% for Google Gemini)

Table 2. DISCERN Scores for Questions, Categorized by Rothwell Criteria

	Google Gemini			Microsoft Copilot			ChatGPT		
DISCERN Scale	Fact	Policy	Value	Fact	Policy	Value	Fact	Policy	Value
Section 1: Is this publication reliable									
Are the aims clear?	3	3	3	3	3	3	3	3	3
Does it achieve its aims?	4	4	4	4	4	4	3	3	3
Is it relevant?	5	5	5	5	5	5	5	5	5
Is it clear what sources of information were used to compile the publication (other than the author or producer)	3	3	3	5	5	5	1	1	1
Is it clear when the information used or reported in the publication was produced?	2	2	2	3	3	3	1	1	1
Is it balanced and unbiased?	3	3	3	4	4	5	3	3	3
Does it provide details of additional sources of support and information?	3	3	3	3	3	3	1	1	1
Does it refer to areas of uncertainty?	5	5	5	5	5	5	5	5	5
Section 2: How good is the quality of information on treatment choices?									
Does it describe how each treatment works?	5	5	5	5	5	5	5	5	5
Does it describe the benefits of each treatment?	5	5	5	5	5	5	5	5	5
Does it describe the risks of each treatment?	5	5	5	5	5	5	5	5	5
Does it describe what would happen if no treatment is used?	3	5	3	3	5	3	3	5	3
Does it describe how the treatment choices affect overall quality of life?	4	4	5	4	4	4	5	4	5
Is it clear that there may be more than one possible treatment choice?	4	4	5	4	4	5	4	4	5
Does it provide support for shared decision- making?	5	5	5	5	5	5	5	5	5
Section 3: Overall rating of the publication									
On the basis of the answers to all the aforementioned questions, rate the overall quality of the publication as a source of information about treatment choices	3	4	4	4	5	5	4	4	4
Total	62	65	65	67	70	70	58	59	59

and from Single Surgeon Practice, Medical Practice and Government sources (0% for Microsoft Copilot).

Discussion

Information provided by Copilot demonstrated better quality as compared with Gemini and ChatGPT. All 3 AI interfaces generated "good" DISCERN scores (scores >50), demonstrating good quality of information provided, except for the Policy and Value questions answered by Copilot which were scored to be "excellent" quality of information (scores >70). The information provided by Copilot demonstrated higher reliability, with a JAMA benchmark criterion of 3 (of 4) as compared with ChatGPT (score of 0) and Gemini (score of 1). The readability of the answered generated by all 3 platforms was considered complex (FRES scores <30), except for Fact questions by Copilot. Finally, it is important to note that all 3 LLMs provided answers that encouraged the reader to reach out to their surgeon to obtain more accurate and personalized information.

Quality Assessment

The DISCERN analysis for ChatGPT, Copilot, and Gemini answers indicated that the AI interfaces were "good" sources of information (scores >50), except for the Policy and Value questions answered by Copilot, which were scored as "excellent" quality of information (scores >70). These findings contribute to varying data depicted by other studies in nonorthopaedics focused literature that have compared different LLMs and have shown greater DISCERN scores for Copilot in comparison with ChatGPT.[15,23,24] In these studies, Google BARD (now Gemini) has also demonstrated greater DISCERN scores than ChatGPT and Bing AI (also produced from Microsoft similar to Copilot).[15,23,24]

In addition, in this study, similar DISCERN scores categorized by Rothwell questions were noted for all 3 AI interfaces, with Fact questions having the lowest score (58 for ChatGPT, 67 for Copilot, and 62 for Gemini) and Value and Policy questions having the same score (59 for ChatGPT, 70 for Copilot, 65 for Gemini). Although several studies both in orthopaedics and nonorthopaedics literature have used the DISCERN score and Rothwell classification system, only Warren et al.[19] have carried out an analysis of both and demonstrated (in regard to rotator cuff repair surgery) similarly lowest DISCERN scores for Fact questions (51 by ChatGPT), and greatest DISCERN scores for Value questions (55 by ChatGPT). This might indicate a pattern of LLMs being able to answer Value questions with increasing accuracy as compared with Fact questions, which is especially interesting to consider, given the vast availability of fact-based information these AI platforms have.

Table 3. FRES and FKGL Scores, Categorized by Rothwell Criteria

	OpenAI ChatGPT			Google Gemini			Microsoft Copilot		
Scores	Fact	Policy	Value	Fact	Policy	Value	Fact	Policy	Value
FRES	11.78	6.67	2.65	21.6	13.3	18.9	31.9	21.5	28.5
FKGL	17.17	17.82	19.32	15.5	17.1	16.1	13.1	15.6	14.3

FKGL, Flesch-Kincaid Grade Level; FRES, Flesch-Kincaid Reading Ease Score.

Accuracy Assessment

As per the results of this study, the JAMA benchmark score for Copilot was 3 of 4, indicating greatest reliability, followed by Gemini (1 of 4), and then ChatGPT (0 of 4). The authors of other orthopaedics literature have found similarly low JAMA (score = 0) for ChatGPT output when obtaining information regarding shoulder stabilization and rotator cuff repairs.[18,19,25]

In regard to ACL injury and reconstructions, other studies have analyzed the accuracy of non-AI websites and sources. These studies have demonstrated that although the average JAMA score was relatively high for Academic websites (3.16), the average combined score for medical and governmental websites was lower ($P < .001$) than nonmedical websites.[8] In addition, a Health On the Net Code of Conduct (HONcode) certification obtained by certain websites to increase transparency also contributed to a greater JAMA score (3.5 ± 0.7) compared with websites that did not have the same (2.2 ± 1.2).[26] This would indicate that Copilot's reliability is similar to HONcode certified websites specific to provide ACL reconstruction information, but the reliability of Gemini and ChatGPT (3.5) are much lower.

Readability Assessment

The readability of the answers generated by all 3 platforms was considered too complex (FRES scores <30, except for Fact questions by Copilot). This finding was corroborated by a minimum reading level of an "academic text" for the answers generated (FKGL<20). These findings align with other studies in literature that have found similarly low readability of information regarding ACL injury and treatment from ChatGPT 3.5 with a mean reading grade of 18.08, and from ChatGPT 4 with a mean reading grade of 17.9, both far exceeding the average eighth-grade reading level of American patients.[13] Mu et al.[16] carried out a similar analysis between different LLMs in regard to information specific to melanoma and also found no significant difference between Google Gemini, Bing AI, similar to this study.

Citation Analysis

Of the 163 citations provided by Microsoft Copilot, most citations were from Journal (75.46%), Academic (8.58%), and Medical Information sites (7.36%), as compared with Nonmedical media (5.52%) and Commercial websites (3.06%). Although most citations provided by Google Gemini were from Journals (65.62%) and Academic website (20.31%), the remainder of the citations were more variably distributed amongst Commercial, Single Surgeon Practice, Medical Practice, and Medical Information websites.

Notably, in considering reliability, although Gemini's answers included hyperlinks of websites with more information and images as references, an instruction to provide the references in a list format had to be inputted, unlike Copilot, which provided the same by default. This along, with the more variable distribution of citations and overall lower number of citations, indicates that Copilot likely demonstrates a greater quality of references. In comparison with ChatGPT, nonorthopaedics literature (focused on urological cancer) has found similarly greater DISCERN values for Copilot and BARD (now Gemini), in comparison with ChatGPT because of the availability of citations provided by the former two.[23,27]

Table 4. Categorization of Citation Sources provided by Microsoft Copilot and Google Gemini

Types of Citation	Google Gemini	Microsoft Copilot
Nonmedical media site	0 (0%)	9 (5.52%)
Commercial	2 (3.12%)	5 (3.06%)
Single-surgeon practice	3 (4.68%)	0 (0%)
Medical practice	1 (1.56%)	0 (0%)
Academic	13 (20.31%)	14 (8.58%)
Medical information site	3 (4.68%)	12 (7.36%)
Journal	42 (65.62%)	123 (75.46%)
Government	0 (0%)	0 (0%)
Total number of citations	64	163

Potential for LLM Uses in Clinical Orthopaedic Surgery

Although the primary objective of this study was to compare varying AI models using ACL injury and reconstruction as clinical context, this study also allows us to assess the limits of LLM use in clinical practice. Patients understandably have questions in the preoperative and postoperative window that arise while they are not in the direct care of a physician. Therefore, patient phone calls often create an additional burden to the health care system (for attending physicians, physician assistants, nurse practitioners, resident physicians, and office staff). Being able to use AI platforms as an alternative or augment to call centers for patient questions would greatly offload a time burden for health care workers. Given the results of our study, current AI models are not at the point at which this would be an acceptable option.

This is for a variety of factors; first and foremost, patient variability always dictates the attention and care of a licensed practitioner for many health-related questions. Although gross generalizations and guidelines exist throughout medicine, every patient and ACL rupture presents originality that would be difficult for AI to take into account. Furthermore, our study demonstrated that all AI models use information with necessity for greater levels of education to comprehend, far beyond the average reading level in this country.[21] An understated role of health care workers is to synthesize complex information and present it in a manner that is digestible and appropriate for each specific patient. This is an aspect that AI platforms would need to improve upon greatly before they can be recommended by physicians as an alternative to calling the provider's office for answers to questions.

At present, a strategy that many surgeons use to provide patients with information is to create an information packet with most commonly asked questions and the surgeon's preferred response. Although this strategy has notable limitations (for example, many patients do not like sifting through large packets of information for answers), we would still endorse this option over recommending patients to use AI interfaces for answers to their questions regarding ACL injury and reconstruction.

Limitations

This study is not without limitations. LLMs can alter their output on the basis of the specifications given as input. As such, one of the major limitations of this study is that a target audience, reading level, or level of complexity of information was not specified in inputting the questions into the AI platforms, to best mimic how a patient would use one of these interfaces. Therefore, it is possible that when entered from the patient's perspective, these questions would be altered in verbiage, which subsequently might alter the answers.

Conclusions

Microsoft Copilot was a better resource for patients to learn about ACL injuries and reconstruction compared with Google Gemini or OpenAI ChatGPT in terms of quality of information, reliability, and readability. The answers provided by LLMs are highly complex and no resource was overall the best.

Disclosures

All authors (S.G., R.T., B.H., A.F.) declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

1. Swenson DM, Collins CL, Best TM, Flanigan DC, Fields SK, Comstock RD. Epidemiology of knee injuries among U.S. high school athletes, 2005/2006-2010/2011. *Med Sci Sports Exerc* 2013;45:462-469.
2. Mall NA, Chalmers PN, Moric M, et al. Incidence and trends of anterior cruciate ligament reconstruction in the United States. *Am J Sports Med* 2014;42:2363-2370.
3. Herzog MM, Marshall SW, Lund JL, Pate V, Mack CD, Spang JT. Trends in incidence of ACL reconstruction and concomitant procedures among commercially insured individuals in the United States, 2002-2014. *Sports Health* 2018;10:523-531.
4. Garrett WE, Swiontkowski MF, Weinstein JN, et al. American Board of Orthopaedic Surgery Practice of the Orthopaedic Surgeon: Part-II, certification examination case mix. *J Bone Joint Surg Am* 2006;88:660-667.
5. Sanders TL, Maradit Kremers H, Bryan AJ, et al. Incidence of anterior cruciate ligament tears and reconstruction: A 21-year population-based study. *Am J Sports Med* 2016;44:1502-1507.
6. Perrin A. Social Media Usage: 2005-2015. Pew Research Centre. https://www.pewresearch.org/internet/2015/10/08/social-networking-usage-2005-2015/. Accessed September 30, 2024.
7. Cassidy JT, Fitzgerald E, Cassidy ES, et al. YouTube provides poor information regarding anterior cruciate ligament injury and reconstruction. *Knee Surg Sports Traumatol Arthrosc* 2018;26:840-845.
8. Castle JP, Khalil LS, Tramer JS, et al. Indications for surgery, activities after surgery, and pain are the most commonly asked questions in anterior cruciate ligament injury and reconstruction. *Arthrosc Sports Med Rehabil* 2023;5:100805.
9. Cassidy JT, Baker JF. Orthopaedic patient information on the world wide web: An essential review. *J Bone Joint Surg Am* 2016;98:325-338.
10. Rengers TA, Thiels CA, Salehinejad H. Academic surgery in the era of large language models: A review. *JAMA Surg* 2024;159:445-450.
11. Kaarre J, Feldt R, Keeling LE, et al. Exploring the potential of ChatGPT as a supplementary tool for providing

orthopaedic information. *Knee Surg Sports Traumatol Arthrosc* 2023;31:5190-5198.
12. Eysenbach G. The role of ChatGPT, generative language models, and artificial intelligence in medical education: A conversation with ChatGPT and a call for papers. *JMIR Med Educ* 2023;9:e46885.
13. Fahy S, Oehme S, Milinkovic D, Jung T, Bartek B. Assessment of quality and readability of information provided by ChatGPT in relation to anterior cruciate ligament injury. *J Pers Med* 2024;14(1):104.
14. Tepe M, Emekli E. Assessing the responses of large language models (ChatGPT-4, Gemini, and Microsoft Copilot) to frequently asked questions in breast imaging: A study on readability and accuracy. *Cureus* 2024;16:e59960.
15. Seth I, Lim B, Xie Y, et al. Comparing the efficacy of large language models ChatGPT, BARD, and Bing AI in providing information on rhinoplasty: An observational study. *Aesthetic Surg J Open Forum* 2023;5:ojad084.
16. Mu X, Lim B, Seth I, et al. Comparison of large language models in management advice for melanoma: Google's AI BARD, BingAI and ChatGPT. *Skin Health Dis* 2024;4:e313.
17. Hurley ET, Crook BS, Lorentz SG, et al. Evaluation high-quality of information from ChatGPT (artificial intelligence-large language model) artificial intelligence on shoulder stabilization surgery. *Arthroscopy* 2024;40:726-731.e6.
18. Warren E Jr, Hurley ET, Park CN, et al. Evaluation of information from artificial intelligence on rotator cuff repair surgery. *JSES Int* 2023;8:53-57.
19. Charnock D, Shepperd S, Needham G, Gann R. DISCERN: An instrument for judging the quality of written consumer health information on treatment choices. *J Epidemiol Community Health* 1999;53:105-111.
20. Silberg WM, Lundberg GD, Musacchio RA. Assessing, controlling, and assuring the quality of medical information on the Internet: *Caveant lector et viewor*—Let the reader and viewer beware. *JAMA* 1997;277:1244-1245.
21. Zhang S, Fukunaga T, Oka S, et al. Concerns of quality, utility, and reliability of laparoscopic gastrectomy for gastric cancer in public video sharing platform. *Ann Transl Med* 2020;8:196.
22. McCormick JR, Kruchten MC, Mehta N, et al. Internet search analytics for shoulder arthroplasty: What questions are patients asking? *Clin Shoulder Elb* 2023;26:55-63.
23. Szczesniewski JJ, Ramos Alba A, Rodríguez Castro PM, Lorenzo Gómez MF, Sainz González J, Llanes González L. Quality of information about urologic pathology in English and Spanish from ChatGPT, BARD, and Copilot. *Actas Urol Esp (Engl Ed)* 2024;48:398-403 [in English, Spanish].
24. Şahin MF, Ateş H, Keleş A, et al. Responses of five different artificial intelligence chatbots to the top searched queries about erectile dysfunction: A comparative analysis. *J Med Syst* 2024;48:38.
25. Eng E, Mowers C, Sachdev D, et al. Despite an advanced readability for the general population, ChatGPT-3.5 can effectively supplement patient-related information provided by the treating surgeon regarding common questions about rotator cuff repair. *Arthroscopy* 2025;41:42-52.
26. Guzman AJ, Dela Rueda T, Williams N, et al. Online patient education resources for anterior cruciate ligament reconstruction: An assessment of the accuracy and reliability of information on the internet over the past decade. *Cureus* 2023;15:e46599.
27. Musheyev D, Pan A, Loeb S, Kabarriti AE. How well do artificial intelligence chatbots respond to the top search queries about urological malignancies? *Eur Urol* 2024;85:13-16.