exposure to keratitis and obstructive sleep apnea.[5] It has been reported that 98% of cases are due to mutations in FGFR2, which result in either amino acid substitutions Ser252Trp or Pro253 Arg.[8]

In the case presented, characteristic features of Apert syndrome are evident in the early clinical and radiologic images (Figs. 1 and 2). These images provide a glimpse into the advent of 3-dimensional computed tomography reconstructions for craniofacial surgical planning. This technique had only been used experimentally in the 1980's, and now forms an indispensable role in the management of craniofacial conditions.[9] Subsequent images provide insight into the postoperative progression of the skeletal structures and overlying soft tissues (Figs. 1 and 2). Postoperative radiologic images show progressive resorption of the frontal bone, which unsurprisingly occurred most significantly in the first few years following fronto-orbital advancement as an adult. There was little observable change to the frontal bone over the subsequent 3 decades.

Similarly, there appears to be evidence of progressive resorption of the maxilla. It is worth noting that access to dental and orthodontic interventions for Apert syndrome would have been limited in the 1950s, and likely different from what is available today. Full-mouth dental extractions were likely a historical means of managing challenging dentition. Although management has continued to improve, achieving stable long-term orthodontic outcomes remains a challenge. While the observable resorption of the maxilla may be partly due to the patient being edentulous, the underlying genetic defect and surrounding soft tissue envelope are also likely to play a role.[10] Nonetheless, clinical photographs demonstrate a reasonably stable mid-face advancement, with little appreciable change from the initial postoperative period to now.

The management of Apert syndrome requires a multidisciplinary protocol-based approach, which continues to evolve. Tessier advanced the surgical correction of facial deformity in Apert syndrome, with the goal of addressing obstructive sleep apnea, deepening the orbits to reduce ophthalmic complications, and improving the facial appearance.[1,6] Various surgical modifications have since been described, such as monobloc advancement, Le Fort III advancement, distraction osteogenesis, and staged multilevel mid-face osteotomies.[1,5] Given the complexity of Apert syndrome and the various surgical approaches described, significant controversy regarding best management persists. Despite these efforts, rates of relapse tend to be high, and this has been hypothesized to be secondary to the underlying genetic defect and soft tissue envelope.[10]

The case presented is unusual in that the patient was already 39 years of age at presentation. At the time of her birth in the 1950s, craniofacial surgery was not yet available. The decision to operate in adulthood was driven by the patient's desire to reduce the stigma encountered secondary to Apert syndrome. When the patient presented during the 1980s, the procedures were staged according to the available knowledge and unit protocols at the time. A 2-stage approach was utilized to reduce the risk of major complications, including infection, extradural collection, cerebrospinal fluid leak, and frontal bone flap loss. However, fronto-orbital advancements performed in adulthood are associated with increased bony resorption, as demonstrated in this case. Conversely, patients who undergo fronto-orbital advancement in infancy tend to have very high reoperation rates.[5] This is consistent with a previously published review of 94 cases of Apert syndrome at our institution, which determined that results from fronto-orbital advancements are more likely to be stable if performed after 18 months of age.[1] Similar results were found when mid-face advancement was delayed until adolescence.

Serial imaging over more than 3 decades demonstrates that a degree of relapse is to be expected even if surgical intervention is delayed into adulthood. The patient's mid-face remains small and abnormal in shape, with razor-sharp infraorbital rims and a vertically short maxilla. This has likely been exacerbated by the removal of all teeth as a teenager, although this may have been the simplest option at the time, given the complexity of malocclusion in Apert syndrome. Perhaps the patient's likelihood of sustaining a facial bone fracture was also amplified by the age-related decrease in bone density combined with the underlying genetic defect and craniofacial abnormalities described.

CONCLUSION

The management of Apert syndrome remains challenging. We present the case of a patient with Apert syndrome who underwent surgical management in adulthood, with more than 3 decades of subsequent follow-up. Clinical and radiologic images demonstrate a degree of resorption and relapse that is likely unavoidable. This provides insight into the long-term skeletal stability and esthetic outcomes for patients with Apert syndrome.

REFERENCES

1. Breik O, Mahindu A, Moore MH, et al. Apert syndrome: surgical outcomes and perspectives. *J Craniomaxillofac Surg* 2016;44:1238–1245
2. Wilkie AO, Slaney SF, Oldridge M, et al. Apert syndrome results from localized mutations of FGFR2 and is allelic with Crouzon syndrome. *Nat Genet* 1995;9:165–172
3. Conrady CD, Patel BC, Sharma S. *Apert syndrome*. StatPearls, Treasure Island (FL). 2023
4. Warren SM, Shetye PR, Obaid SI, et al. Long-term evaluation of midface position after Le Fort III advancement: a 20-plus-year follow-up. *Plast Reconstr Surg* 2012;129:234–242
5. Allam KA, Wan DC, Khwanngern K, et al. Treatment of apert syndrome: a long-term follow-up study. *Plast Reconstr Surg* 2011;127:1601–1611
6. Tessier P. The definitive plastic surgical treatment of the severe facial deformities of craniofacial dysostosis. Crouzon's and Apert's diseases. *Plast Reconstr Surg* 1971;48:419–442
7. Tovetjarn R, Tarnow P, Maltese G, et al. Children with Apert syndrome as adults: a follow-up study of 28 Scandinavian patients. *Plast Reconstr Surg* 2012;130:572e–576ee
8. Das S, Munshi A. Research advances in Apert syndrome. *J Oral Biol Craniofac Res* 2018;8:194–199
9. Vannier M, Warfield S, Yoshida H Abdominal imaging. Computational and clinical applications: 5th International Workshop, Held in Conjunction with MICCAI 2013, Nagoya, Japan, September 22, 2013, Proceedings. 2013: 1 online resource (XIV, 300 pages 147 illustrations)
10. Wong GB, Kakulis EG, Mulliken JB. Analysis of fronto-orbital advancement for Apert, Crouzon, Pfeiffer, and Saethre-Chotzen syndromes. *Plast Reconstr Surg* 2000;105:2314–2323

Evaluating the Success of ChatGPT in Addressing Patient Questions Concerning Thyroid Surgery

Şamil Şahin, MD, Mustafa Said Tekin, MD,† Yesim Esen Yigit, MD,‡ Burak Erkmen, MD,* Yasar Kemal Duymaz, MD,‡ and İlhan Bahşi, MD, PhD§*

Objective: This study aimed to evaluate the utility and efficacy of ChatGPT in addressing questions related to thyroid surgery, taking into account accuracy, readability, and relevance.

Methods: A simulated physician-patient consultation on thyroidectomy surgery was conducted by posing 21 hypothetical questions to ChatGPT. Responses were evaluated using the DISCERN score by 3 independent ear, nose and throat specialists. Readability measures including Flesch Reading Ease), Flesch-Kincaid Grade Level, Gunning Fog Index, Simple Measure of Gobbledygook, Coleman-Liau Index, and Automated Readability Index were also applied.

Results: The majority of ChatGPT responses were rated fair or above using the DISCERN system, with an average score of 45.44 ± 11.24. However, the readability scores were consistently higher than the recommended grade 6 level, indicating the information may not be easily comprehensible to the general public.

Conclusion: While ChatGPT exhibits potential in answering patient queries related to thyroid surgery, its current formulation is not yet optimally tailored for patient comprehension. Further refinements are necessary for its efficient application in the medical domain.

Key Words: ChatGPT, DISCERN score, patient education, readability measures, thyroid surgery

In the past few years, there has been a significant advancement in the field of health care with regard to artificial intelligence (AI).[1] This progress has resulted in the emergence of many applications, ranging from diagnostic tools to virtual assistants that offer health-related information. Artificial intelligence–driven chatbots refer to software systems that possess the capability to engage in conversations with individuals using natural language and offer a range of services, typically through textual means.[2,3] It is also known that the use of various AI tools in many stages such as diagnosis, diagnosis, follow-up, and even academic writing in the field of medicine has recently increased rapidly.[4,5] The ChatGPT model, which is based on OpenAI's GPT-4, serves as a notable illustration of this phenomenon. These robots have the potential to be utilized in several domains, encompassing the dissemination of specialized knowledge, addressing inquiries, delivering customer assistance, and establishing simulated conversational spaces for therapeutic purposes.[6,7]

Thyroid surgery is a common surgical procedure utilized to treat a variety of thyroid disorders, including goiter, hyperthyroidism, and thyroid cancer.[8] The provision of accurate and timely information by health care personnel is crucial in addressing the numerous questions and concerns that patients commonly have before undergoing medical treatments. This practice holds significant importance as it contributes to enhancing patient outcomes.[9] Nevertheless, it may not always be feasible for patients to initiate communication with the surgeon or supporting staff to address their inquiries. Currently, chatbots that utilize AI possess a significant capability to provide patients with the required information.

In this research, we will assess the efficacy of ChatGPT in providing responses to a diverse array of questions regarding thyroid surgery. We will evaluate the precision, readability, and relevance of the responses. Furthermore, an examination of the outcomes will be conducted to analyze the potential consequences for both individuals seeking medical treatment and professionals working in the health care industry.

METHODS

On March 5, 2023, we asked ChatGPT 21 different hypothetical questions simulating physician-patient consultation about thyroidectomy surgery. The answers received were evaluated by 3 independent ear, nose and throat specialists with the DISCERN score. The final DISCERN score was determined by taking the arithmetic average of the scores given by 3 physicians. The readability of the answers was calculated using Flesch Reading Ease (FRE), Flesch-Kincaid Grade Level (FKGL), Gunning Fog Index (Gunning FOG), Simple Measure of Gobbledygook (SMOG), Coleman-Liau Index (CLI) and Automated Readability Index (ARI). As the study included publicly available data, there was no need for an ethics committee.

Mean, SD, median, minimum, maximum value frequency, and percentage were used for descriptive statistics. The distribution of variables was checked with the Kolmogorov-Simirnov test. Spearman Rank Correlation Coefficient was used in the correlation analysis. Intraclass correlation was used in the reliability analysis. SPSS 28.0 was used for statistical analyses.

RESULTS

A total of 21 hypothetical questions were posed, replicating a simulated interaction between a physician and a patient on the topic of thyroidectomy surgery. The questions are displayed in Supplemental Table 1 (Supplemental Digital Content, Table 1, http://links.lww.com/SCS/G399). The responses provided by ChatGpt to the questions are documented in Supplemental Material 1 (Supplemental Digital Content 2, http://links.lww.com/SCS/G400).

According to the DISCERN Scoring System, there was no response at the very poor level. Six were poor, 6 were fair, 8 were good, and 1 was excellent. The average DISCERN Score was 45.44 ± 11.24 (Supplemental Digital Content, Table 2, http://links.lww.com/SCS/G401). Supplemental Table 3 (Supplemental Digital Content, Table 3, http://links.lww.com/SCS/

From the *Ear Nose and Throat Specialist, Private Practice; †Department of Otolaryngology, Medipol Mega Hospital, Medipol University; ‡Department of Otolaryngology, Umraniye Training and Research Hospital, University of Health Sciences, Istanbul; and §Department of Anatomy, Faculty of Medicine, Gaziantep University, Gaziantep, Turkey.

Received May 10, 2024.

Accepted for publication May 15, 2024.

Address correspondence and reprint requests to İlhan Bahşi, MD, PhD, Department of Anatomy, Faculty of Medicine, Gaziantep University, Gaziantep TR-27310, Turkey; E-mail: dr.ilhanbahsi@gmail.com

During the preparation of this work, the authors used ChatGPT to translate parts of the manuscript. After using this tool/service, the authors reviewed and edited the content as needed and take full responsibility for the content of the publication.

The authors report no conflicts of interest.

Supplemental Digital Content is available for this article. Direct URL citations are provided in the HTML and PDF versions of this article on the journal's website, www.jcraniofacialsurgery.com.

ISSN: 1049-2275

DOI: 10.1097/SCS.0000000000010395

G402) displays a significant [$r = 0.966$ (0.991–0.998)] strong ($P = 0.05$) correlation between the DISCERN scores of 3 experts.

The mean FRE score for the responses to 21 questions was 45.74 ± 9.66. The mean FKGL score was 11.03 ± 2.33, the mean Gunning FOG score was 14.36 ± 2.41, the mean SMOG score was 10.46 ± 1.66, the mean CLI score was 14.28 ± 1.10, and the mean ARI score was 10.98 ± 2.66 (Supplemental Digital Content, Table 4, http://links.lww.com/SCS/G403).

A statistically significant correlation ($P < 0.05$, $r = 0.991$) was found between the DISCERN score and the FRE score. A statistically significant negative correlation ($P < 0.05$, $r = -0.918$) was found between the DISCERN score and FKGL score. A statistically significant negative correlation ($P < 0.05$, $r = -0.891$) was found between the DISCERN score and the Gunning FOG score. A statistically significant negative correlation ($P < 0.05$, $r = -0.912$) was found between the DISCERN score and the SMOG score. A statistically significant negative correlation ($P < 0.05$, $r = -0.711$) was found between the DISCERN score and the CLI score. A statistically significant negative correlation ($P < 0.05$, $r = -0.847$) was found between the DISCERN score and the ARI score (Supplemental Digital Content, Table 5, http://links.lww.com/SCS/G404).

DISCUSSION

The prevalence of chatbots utilizing AI has grown in tandem with advancements in the area of AI. Artificial intelligence–driven chatbots have gained significant traction in several industries such as banking, customer service, and education, due to their ability to enhance service efficiency through user interactions. In recent times, there has been a noticeable rise in its utilization within the health care industry.[10]

Regarding patient health, the accuracy of the information provided is unquestionably crucial. To educate and inform patients, it is essential that the information be accurate and easily understood by the patients. This study aimed to assess the responses provided by ChatGPT in a simulated scenario of 21 patient-physician consultations concerning common questions regarding thyroidectomy. In addition, the intelligibility of these responses was analyzed using FKGL, FRE, Gunning FOG, SMOG, CLI, and ARI.

Scoring with DISCERN is a method for assessing the quality of health-related information and resources. This scoring system is predominantly utilized in the presentation of treatment or health information and provides users with a sense of the information's reliability and usability. It contains a total of 16 questions organized into three sections. In the first section, 8 questions evaluate the reliability of the source of the information, in the second section, 7 questions examine the efficacy of the offered treatment methods and whether they are evidence-based, and in the final section, 1 question evaluates the information's usefulness and utility. Each question is assigned a point value between 0 and 5, and the cumulative score is determined. Scores are typically divided into 5 categories based on their total score range[11]: "excellent, good, fair, poor, and very poor." According to the findings of our study, the DISCERN scoring of ChatGPT responses to the questions yielded no very poor responses and only one excellent response. There were 6 responses at the poor and fair levels, respectively. There were 8 responses rated as good. Therefore, the majority of responses were at or above the fair level.

The quality of Internet resources is important for patient education, and the readability of the text provided is a major factor in determining this quality. The American Medical Association and the National Institute of Health recommend that educational materials be written at a sixth-grade level.[12,13] There are 6 popular tools for assessing readability in the field of otorhinolaryngology. The FRE was created in 1948 and is widely used to evaluate medical materials written for adults. The FKGL is used to evaluate academic materials. The Gunning Fog Index (GFI) and SMOG use multisyllabic word count and average sentence length to determine readability level, but SMOG evaluates the entire text. The CLI determines the score level of a document based on sentence length and number of characters. Automatic Readability Index is calculated using the number of characters, words, and sentences.[14–16] The 6 most popular tools were used in our study. The responses generated by ChatGPT were well above the desired readability level in all of these tools. Information with high DISCERN scores was more readable than those with low DISCERN scores. But these were far from the recommended level of grade 6.

The investigations conducted on this topic and published in the academic literature have yielded varying outcomes. Bellinger and colleagues compared the patient education materials regarding benign paroxysmal positional vertigo obtained through Google Search to those transmitted through ChatGPT with the DISCERN score. They concluded that the quality of responses generated by ChatGPT was inadequate. They also concluded that the content produced by ChatGPT had higher readability scores and presented a greater level of complexity, rendering it more challenging to comprehend.[17] Cocci et al[18] examined the use of ChatGPT in the context of providing information to patients in the field of urology. The researchers reached the conclusion that the DISCERN scores pertaining to the material generated by ChatGPT were indicative of low quality. In addition, the readability levels were found to be comparable to those expected of college graduates, hence posing significant challenges for patients in terms of comprehension. Whiles and colleagues conducted a research study wherein a series of 13 inquiries pertaining to the field of urology were posed to ChatGPT, then assessing the responses using the Discern score. Consequently, the researchers reached the determination that ChatGPT now lacks the capability to adequately provide information to patients, exhibits misinterpretations of recommendations, and disregards significant contextual information.[19] Mishra et al[20] inquired about ChatGPT for treatment-related information pertaining to a variety of neurosurgical diseases. The replies were assessed using the DISCERN scoring method. The researchers noted that a portion of the responses exhibited inadequacy however their degrees of readability were notably higher. Consequently, patients have reached the conclusion that ChatGPT is not currently suitable for safe utilization within the patient information domain. Golan et al[21] conducted an assessment of the online content related to shock wave treatment in erectile dysfunction provided by ChatGPT. The researchers determined that the quality of this content, as measured by DISCERN scoring, did not meet adequate standards. The researchers reached the conclusion that ChapGPT has limitations in its ability to effectively distinguish data sources of lower quality. In a further investigation, the accuracy of ChatGpt's responses pertaining to bladder cancer, prostate cancer, kidney cancer, benign prostatic hypertrophy, and urinary stones was assessed using the DISCERN score. The findings revealed that the majority of the answers supplied by ChatGpt surpassed the average level of accuracy. Despite this, they argued that one should be careful about the information provided by ChatGPT.[22] Hurley and colleagues conducted an evaluation of the data generated by ChatGPT in relation to the issue of "shoulder stabilization surgery," utilizing the DISCERN tool. The researchers discovered that the majority of the outcomes were satisfactory or

exceeded expectations. However, it was also discovered that the readability levels of this material were significantly higher. One potential drawback associated with the information obtained through ChatGPT is the lack of mentioned sources for the information provided.[23]

After examining the published studies, it becomes evident that there is a lack of consensus about the material generated by ChatGPT. Certain studies exhibit higher DISCERN scores, whereas others have lower scores. In the current study, a more moderate outcome was seen. The predominant focus of the replies was on the moderate level, even with a somewhat higher number of responses indicating a good level. Nevertheless, an agreement exists throughout the scholarly literature about readability levels. The resulting responses exhibit a far higher level of readability than what is often advised.

Before and after surgery, it is reasonable for patients to have questions about their surgical procedure. Patients will not be able to promptly ask the physician questions that they missed asking during the interview or that occur to them later. However, ChatGPT is a resource that patients can access at any time. Even though the majority of answers provided by ChatGPT in our study were fair and above in DISCERN scores, the high readability scores prevent patients from using ChatGPT as a source of information. With the enhancements that will be made to ChatGPT, particularly in this regard, we believe the responses will be of higher quality and more legible. This may require collaboration between software developers and health care professionals. Having a tool that patients can use at any moment to ask inquiries about their surgery increases patients' adherence to treatment and simplifies physicians' work. But ChatGPT does not seem to be able to do this in its current form.

Limitations

First, the study used hypothetical questions, but real-world patients' questions may vary in complexity and specificity. Second, the evaluation is based on the DISCERN score, which is subject to comments. Third, readability measures may not capture the full context and relevant scope of medical information. Finally, the study was conducted in English and does not contain provisions regarding answers given in other languages.

CONCLUSION

Although ChatGPT has the potential as a tool to answer patient questions about thyroid surgery, it is not capable of doing this properly in its current form. More research and development efforts are needed to improve the performance of AI chatbots such as ChatGPT in the health care industry.

REFERENCES

1. Habal MB. Brave new surgical innovations: the impact of bioprinting, machine learning, and artificial intelligence in craniofacial surgery. *J Craniofac Surg* 2020;31:889–890
2. Aggarwal A, Tam CC, Wu D, et al. Artificial intelligence-based chatbots for promoting health behavioral changes: systematic review. *J Med Internet Res* 2023;25:e40789
3. Gaikwad T. Artificial intelligence-based chatbot. *Int J Res Appl Sci Eng Technol* 2018;6:2305–2306
4. Bahsi I, Balat A. The role of AI in writing an article and whether it can be a co-author: what if it gets support from 2 different AIs like ChatGPT and Google Bard for the same theme? *J Craniofac Surg* 2024;35:274–275
5. Uranbey Ö, Ayrancı F, Erdem BE. ChatGPT guided diagnosis of ameloblastic fibro-odontoma: a case report with eventful healing. *Eur J Ther* 2024;30:240–247
6. Verma S, Sharma R, Deb S, et al. Artificial intelligence in marketing: Systematic review and future research direction. *Int J Inform Manag Data Insights* 2021;1:100002.
7. Vlačić B, Corbo L, Costa e Silva S, et al. The evolving role of artificial intelligence in marketing: a review and research agenda. *J Bus Res* 2021;128:187–203
8. Carsuzaa F, Delbreil A, Chabrillac E. Thyroid surgery, complications and professional liability. *Gland Surg* 2023;12:1025–1027
9. Leonard P. Exploring ways to manage healthcare professional-patient communication issues. *Support Care Cancer* 2017;25:7–9
10. Naqvi WM, Sundus H, Mishra G, et al. AI in medical education curriculum: the future of healthcare learning. *Eur J Ther* 2024;30:e23–e25
11. Charnock D, Shepperd S, Needham G, et al. DISCERN: an instrument for judging the quality of written consumer health information on treatment choices. *J Epidemiol Community Health* 1999;53:105–111
12. Ad Hoc Committee on Health Literacy for the Council on Scientific Affairs AMA. Health Literacy: Report of the Council on Scientific Affairs. *JAMA* 1999;281:552–557
13. The National Library of Medicine (MedlinePlus). *How to Write Easy-to-Read Health Materials*. 2022. Accessed September 5, 2024. https://medlineplus.gov/pdf/health-education-materials-assessment-tool.pdf
14. Wong K, Levi JR. Readability trends of online information by the American Academy of Otolaryngology-Head and Neck Surgery Foundation. *Otolaryngol Head Neck Surg* 2017;156:96–102
15. Svider PF, Agarwal N, Choudhry OJ, et al. Readability assessment of online patient education materials from academic otolaryngology-head and neck surgery departments. *Am J Otolaryngol* 2013;34:31–35
16. Kim JH, Grose E, Philteos J, et al. Readability of the American, Canadian, and British otolaryngology-head and neck surgery societies' patient materials. *Otolaryngol Head Neck Surg* 2022;166:862–868
17. Bellinger JR, De La Chapa JS, Kwak MW, et al. BPPV information on Google versus AI (ChatGPT). *Otolaryngol Head Neck Surg* 2023;170:1504–1511.
18. Cocci A, Pezzoli M, Lo Re M, et al. Quality of information and appropriateness of ChatGPT outputs for urology patients. *Prostate Cancer Prostatic Dis* 2024;27:103–108
19. Whiles BB, Bird VG, Canales BK, et al. Caution! AI bot has entered the patient chat: ChatGPT has limitations in providing accurate urologic healthcare advice. *Urology* 2023;180:278–284
20. Mishra A, Begley SL, Chen A, et al. Exploring the intersection of artificial intelligence and neurosurgery: let us be cautious with ChatGPT. *Neurosurgery* 2023;93:1366–1373
21. Golan R, Ripps SJ, Reddy R, et al. ChatGPT's ability to assess quality and readability of online medical information: evidence from a cross-sectional study. *Cureus* 2023;15:e42214
22. Szczesniewski JJ, Tellez Fouz C, Ramos Alba A, et al. ChatGPT and most frequent urological diseases: analysing the quality of information and potential risks for patients. *World J Urol* 2023;41:3149–3153
23. 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. e726

Streamlining Iliac Bone Harvest for Maxillary Reconstruction With Novel Use of Arthrex Osteochondral Autograft Transfer System