MARMARA
MEDICAL JOURNAL

Original Article

https://dergipark.org.tr/tr/pub/marumj

ChatGPT versus strabismus specialist on common questions about strabismus management: a comparative analysis of appropriateness and readability

Didem DIZDAR YIGIT[1], Mehmet Orkun SEVIK[1], Aslan AYKUT[1], Eren CERMAN[2]

[1] Department of Ophthalmology, School of Medicine, Marmara University, Istanbul, Turkey
[2] Department of Ophthalmology, Donaustadt Hospital, Vienna, Austria
Corresponding Author: Didem DIZDAR YIGIT
E-mail: drdidemdizdar@gmail.com

Submitted: 20.04.2024 **Accepted:** 22.05.2024

ABSTRACT

Objective: Patients widely use artificial intelligence-based chatbots, and this study aims to determine their utility and limitations on questions about strabismus. The answers to the common questions about the management of strabismus provided by Chat Generative Pre-trained Transformer (ChatGPT)-3.5, an artificial intelligence-powered chatbot, were compared to answers from a strabismus specialist (The Specialist) in terms of appropriateness and readability.

Patients and Methods: In this descriptive, cross-sectional study, a list of questions from strabismus patients or caregivers in outpatient clinics about treatment, prognosis, postoperative care, and complications were subjected to ChatGPT and The Specialist. The answers of ChatGPT were classified as appropriate or not, considering the answers of The Specialist as the reference. The readability of all the answers was assessed according to the parameters of the Readable online toolkit.

Results: All answers provided by ChatGPT were classified as appropriate. The mean Flesch Kincaid Grade Levels of the respective answers given by ChatGPT and The Specialist were 13.75±1.55 and 10.17±2.17 ($p<0.001$), higher levels indicating complexity; and the mean Flesch Reading Ease Scores of which higher scores indicated ease, were 23.86±9.38 and 44.54±14.66 ($p=0.002$). The mean reading times were 15.6±2.85 and 10.17±2.17 seconds for ChatGPT and The Specialist, respectively ($p=0.003$). The overall reach of the answers by ChatGPT and The Specialist was 56.87±11.67 and 81.67±12.80 ($p<0.001$).

Conclusion: Although, ChatGPT provided appropriate answers to all compiled strabismus questions, those were complex or very difficult to read for an average person. The readability scores indicated a college graduation degree would be required to understand the answers provided by ChatGPT. However, The Specialist gave similar information in a more readable form. Therefore, physicians and patients should consider the limitations of such similar platforms for ocular health-related questions.

Keywords: Artificial intelligence, ChatGPT, Readability, Strabismus, Strabismus surgery

1. INTRODUCTION

As the field of artificial intelligence (AI) advances, several AI-powered search platforms have been developed to serve in various sectors, including healthcare. Those platforms are proposed to potentially assist patients and their caregivers with medical conditions and treatment options [1-3]. One of the AI-powered search platforms, Chat Generative Pre-trained Transformer (ChatGPT), a language model-based bot developed and released in November 2022 by OpenAI, has already surpassed 100 million users by January 2023 [3,4].

ChatGPT-4 is the latest version of the bot with improved performance, released in March 2023 [4,5]. However, due to the subscription-based paid nature of the latest version, most people still prefer using the open-access ChatGPT-3.5, which is still comparably reliable [5]. Nevertheless, despite its impressive capabilities, there are concerns about the reliability, readability, and comprehensiveness of the specific responses of the chatbot to common questions asked by patients as well as caregivers in several health-related conditions [1-3].

Various treatment options are available to address strabismus management in children, including non-surgical approaches like patching and refractive error correction, as well as surgical options. It is shown that parental stress is an issue that should be considered when treating children with ophthalmological disorders. To lower stress and achieve an optimal treatment environment for the patient, patients/caregivers should be well-informed about their conditions. Nowadays, it is common

How to cite this article: Yigit Dizdar D, Sevik O M, Aykut A, Cerman E. ChatGPT versus strabismus specialist on common questions about strabismus management: a comparative analysis of appropriateness and readability. Marmara Med J 2024;37(3): doi: 10.5472/marumj.1571218

http://doi.org/10.5472/marumj.1571218
Marmara Med J 2024;37(3): 323-326

ISSN: 1309-9469

knowledge that people seek detailed answers to their questions online, even after doctor visits. Therefore, health workers should be aware of the advantages/risks of online data [2].

Several studies have already investigated the accuracy and readability of health-related online data freely or by AI-based chatbots [1-3, 6-8].

However, to our knowledge, no previous studies are available regarding the appropriateness and readability of content provided by online AI-based systems on common questions asked by patients or caregivers on managing strabismus. Therefore, this study aims to evaluate the answers to the common questions about strabismus management provided by ChatGPT-3.5 and to compare them to the responses of a strabismus specialist in terms of appropriateness and readability.

2. MATERIALS and METHODS

This cross-sectional study was exempted from ethics committee approval by the Marmara University Institutional Review Board since it did not involve human subjects. This study was conducted using the open-access language model ChatGPT-3.5 [4] and the open-access online readability tool Readable [9] in August 2023.

The authors compiled a list of the 15 most common questions asked by the patients or their caregivers at Marmara University Pendik Training and Research Hospital, Istanbul, Turkey, outpatient Pediatric Ophthalmology and Strabismus Clinic about the surgical and non-surgical treatment options for strabismus, the prognosis of strabismus, postoperative care of the patients, and strabismus surgery-related complications. The compiled questions are listed in Table I.

Table I. *The compiled common questions asked by the patients or their caregivers in our outpatient pediatric ophthalmology and strabismus clinic.*

1.	Is there a possibility to correct strabismus with glasses or medication?
2.	What is the rate of success in strabismus surgery?
3.	Will the visual acuity improve with the strabismus surgery?
4.	Is there a risk of strabismus recurrence?
5.	Is there a risk of blindness after the strabismus surgery?
6.	Is there a risk of strabismus worsening after the surgery?
7.	Are we going to continue wearing glasses after the surgery?
8.	Will the lazy eye improve after the strabismus surgery?
9.	If recurrence occurs after the initial surgery, is there a chance for another surgery?
10.	Is it better to have the strabismus surgery after the age of eighteen?
11.	Is there a risk of double vision after the strabismus surgery?
12.	Is the strabismus surgery performed with lasers?
13.	How long will the patient stay in the hospital after the strabismus surgery?
14.	Will medication be required after the surgery?
15.	How frequently will follow-up appointments be scheduled after the surgery?

Appropriateness

One of the strabismus specialists (The Specialist; D.D.Y.) answered all compiled questions to provide accurate, evidence-based, and comprehensible responses intended to be understandable by the public audience. Those answers were set as the reference for appropriateness evaluation. The questions were asked two times on the ChatGPT-3.5 platform. For the second time, ChatGPT was asked to summarize the first answers into a shorter and more readable version, i.e., the final version. The final version of the answers to each question provided by ChatGPT were classified as "appropriate" or "inappropriate" according to the reference answers provided by The Specialist.

Readability

The readability of all answers ChatGPT and The Specialist provided was assessed using seven indices from the Readable online toolkit [9]. Those indices included the Flesch Reading Ease Score (FRES), Flesch-Kincaid Reading Grade Level (FKRGL) Scores, Gunning Fog Index, Coleman-Liau Index, Simple Measure of Gobbledygook (SMOG) Index, Reading Time (in seconds), and Overall Reach.

Flesch readability tests (FRES and FKRGL) use mathematical formulas based on the sentence length, word count, and number of syllables for each word [10,11]. The FRES ranges between 0 (unreadable) and 100 (very easy to read), with higher numbers indicating easier readability. The corresponding score ranges between very difficult (i.e., scientific papers), difficult (i.e., academic papers), fairly difficult, standard (i.e., easily understandable by 13 – to 15-year-olds), fairly easy, easy, and very easy (i.e., comics) to read texts are 0-30, 30-50, 50-60, 60-70, 70-80, 80-90, and 90-100 points, respectively [10,12]. The scores obtained from FKGRL approximately correspond to the United States grade levels (i.e., a text with a FKGRL score of 8.2 can be interpreted as understandable by an average person who graduated from 8th Grade) [11,12].

The Gunning Fog Index assesses the average length of sentences in combination with the rate of polysyllabic words. The index score ranges from 0 to 20 and measures clarity and simplicity. The Coleman-Liau index is usually used in addition to other indices and is especially useful in medical documents. It is based on sentence length and an average number of letters per 100 words. The SMOG index uses the frequency of polysyllabic words in a sample of sentences 13]. It is instrumental in health care and measures comprehensiveness. The results from the latter three indices correspond to the school grade level of a person to understand a piece of text, such as FKGRL. The lower the Gunning Fog index, Coleman-Liau index, and SMOG index scores are, the easier the text to read and comprehend.

Overall reach measures the proportion of the target audience that can read given content easily. It is currently calibrated against the literate general public, so a reach of 100% means your content is readable by about 85% of the public (the literate percentage).

Statistical Analysis

Descriptive statistics were given as mean, standard deviation, median, minimum, maximum, frequency, and ratio values where relevant. The distribution of data was analyzed with the Kolmogorov-Smirnov test. Independent sample t-test and Mann-Whitney U test were used to analyze quantitative independent data. Wilcoxon test was used in the analysis of dependent quantitative data. The chi-square test was used to analyze qualitative independent data, and the Fischer test was used when the chi-square test conditions were not met. The analysis was made using SPSS software for IOS, version 28.0 (SPSS Inc, Chicago, IL). The significance level was set as $p<0.05$.

3. RESULTS

The appropriateness evaluation showed 100% agreement between The ChatGPT and The Specialist.

The evaluated readability indices of The ChatGPT and The Specialists are given in Table II.

Table II. *The comparison of the answers provided by The Specialist and ChatGPT in terms of readability.*

Readability Indices	The Specialist Mean ± SD Median (Min-Max)	ChatGPT Mean ± SD Median (Min-Max)	P-value†
FRES	44.54 ± 14.66 43.96 (14-67)	23.86 ± 9.38 20 (12-39)	**0.002**
FKRGL	10.17 ± 2.17 10.36 (6-13)	13.75 ± 1.55 13.93 (11-15)	**<0.001**
Gunning Fog Index	13.32 ± 3.0 13.44 (8-17)	18.34 ± 2.74 18.21 (13-23)	**<0.001**
Coleman-Liau Index	12.19 ± 1.87 11.93 (8-15)	16.41 ± 1.64 16.57 (12-18)	**<0.001**
SMOG Index	12.70 ± 1.83 13.02 (8-15)	15.90 ± 1.74 15.65 (13-19)	**<0.001**
Reading Time‡‡	10.17 ± 2.17 9 (4-23)	15.6 ± 2.85 16 (10-20)	**0.003**
Overall Reach	81.67 ± 12.80 82 (59-100)	56.87 ± 11.67 56 (40-77)	**<0.001**

FKRGL: Flesch-Kincaid Reading Grade Level, FRES: Flesch Reading Ease Score, SD: Standard Deviation, SMOG: Simple Measure of Gobbledygook. † Wilcoxon Signed-Rank Test, ‡ Evaluated in seconds, Bold values indicate statistical significance.

The mean FKGL of the answers given by ChatGPT and The Specialist were for university sophomore and high school sophomore grade level audiences, respectively. The mean FRES results indicated reading the answers of The Specialist was nearly twice as easy as ChatGPT's. It is noted that the responses of ChatGPT were very difficult; meanwhile, the responses of The Specialist were merely difficult to read. All three Gunning Fog, Coleman-Liau, and SMOG indices scores indicated clearer and easier-to-understand responses by The Specialist, with significantly less reading time.

4. DISCUSSION

Recently, there has been an increasing interest in using AI-based systems. However, only a small portion of opportunities in clinical practices were discovered. There are a lot of advantages and pitfalls waiting to be evaluated while adapting our understanding of healthcare to the future [1,14].

In a study evaluating the performance of ChatGPT on various ophthalmology examinations, it was stated that the accuracy of the AI-based model was 59.4% and 49.2% [5]. They also noted that this outcome was noteworthy and promising in ophthalmology; as the questions get more specific on a subject, the accuracy falls [5,15]. In our study, we asked commonly asked questions by patients and caregivers and tried to have simple, not-too-detailed, and understandable responses. We observed that ChatGPT-3.5 was able to provide appropriate answers to those of a strabismus specialist regarding strabismus management.

We used five indices, as proposed by Momenaei et al., to evaluate the readability of every response given by ChatGPT-3.5 and the strabismus specialist [2]. The mean FRES for The Specialist was between 30 and 50, indicating an appropriate level for merely 33% of the population. However, the score of ChatGPT was even lower, between 0 and 30, meaning it was a very hard level of reading, which was readable for only 4.5% of the US population [2,16]. The FKRGL scores ranged from 0 to 30 for both ChatGPT and The Specialist, indicating that the responses were very hard to understand and that college graduation would be required to understand all the answers given.

The mean Gunning Fog Index of The Specialist was 13.32, meaning an average first-year college student could understand the answers. However, the mean Gunning Fog index of ChatGPT was 18.34, implying that the answers were readable for post-graduates [16,17]. Meanwhile, the mean Coleman Linau Index scores, showing the grade level in the US school system required to understand a text, were between 10-13 for The Specialist (12.19) and higher than 13 for ChatGPT (16.14). Those results indicated a reading ability necessitating a high school and college education for the answers of The Specialist and ChatGPT, respectively. Also, the mean SMOG Index results, showing year estimates of schooling needed to comprehend a text, were higher for the answers given by ChatGPT (12.70) than The Specialist (15.90). This index was stated as a measure of the readability of consumer-focused healthcare materials [6].

The Overall Reach score of ChatGPT was also lower than that of The Specialist, indicating that answers generated by ChatGPT-3.5 seem too complicated for the general population.

This study has some limitations, like the variability of the responses of AI-based systems as the training data and the version of the system change. The current database is based on internet data until September 2021, which limits its ability to give more recent data. The results found in this study also will not apply to the newer ChatGPT versions. There are other popular language models, and they could have been included in the comparison with ChatGPT. Also, the accuracy of the responses were not scored according to a scale, but only categorized as

http://doi.org/10.5472/marumj.1571218
Marmara Med J 2024;37(3): 323-326

appropriate or inappropriate. On the other hand, the strengths of this study are the comparison of the publicly available version of the ChatGPT and the evaluation of readability by multiple indices.

In conclusion, this study showed that ChatGPT-3.5 might provide highly accurate answers to common questions about surgical and non-surgical treatment options for strabismus, prognosis, postoperative care, and surgery-related complications. However, clinicians should be aware of that the responses given by AI-based models to medical inquiries are still more challenging to read and understand than the responses of the medical specialists by the general audience.

Compliance with Ethical Standards

Ethical approval: This cross-sectional study was exempted from ethics committee approval by the Marmara University Institutional Review Board since it did not involve human subjects.

Financial support: This study received no specific grant from any funding agency in the public, commercial or not-for-profit sectors.

Conflict of interest: The authors declare that they have no potential conflict of interest regarding the investigation, authorship, and/or publication of this article.

Author contributions: DDT: Study planning, writing, editing, AA and EC: Study planning, editing, MOS: Writing, Editing. All authors read and approved the final version of the artical.

REFERENCES

[1] Korngiebel DM, Mooney SD. Considering the possibilities and pitfalls of Generative Pre-trained Transformer 3 (GPT-3) in healthcare delivery. NPJ Digit Med 2021; 4: 93. doi:10.1038/s41746.021.00464-x.

[2] Momenaei B, Wakabayashi T, Shahlaee A, et al. Appropriateness and readability of ChatGPT-4-generated responses for surgical treatment of retinal diseases. Ophthalmol Retina 2023; 7: 862-8. doi:10.1016/j.oret.2023.05.022.

[3] Sarraju A, Bruemmer D, Van Iterson E, et al. Appropriateness of cardiovascular disease prevention recommendations obtained from a popular online chat-based artificial intelligence model. JAMA 2023; 329: 842-4. doi:10.1001/jama.2023.1044.

[4] OpenAI. ChatGPT. Computer software. 2022. https://openai.com/blog/ChatGPT. Accessed on 03 December, 2023.

[5] Teebagy S, Colwell L, Wood E, et al. Improved performance of ChatGPT-4 on the OKAP examination: A comparative study with ChatGPT-3.5. J Acad Ophthalmol (2017) 2023; 15: e184-e187. doi:10.1055/s-0043.177.4399.

[6] Fitzsimmons PR, Michael BD, Hulley JL, et al. A readability assessment of online Parkinson's disease information. J R Coll Physicians Edinb 2010; 40: 292-6. doi:10.4997/JRCPE.2010.401

[7] Kloosterboer A, Yannuzzi NA, Patel NA, et al. Assessment of the quality, content, and readability of freely available online information for patients regarding diabetic retinopathy. JAMA Ophthalmol 2019; 137: 1240-5. doi:10.1001/jamaophthalmol.2019.3116.

[8] Patel AJ, Kloosterboer A, Yannuzzi NA, et al. Evaluation of the content, quality, and readability of patient accessible online resources regarding cataracts. Semin Ophthalmol 2021; 36: 384-91. doi:10.1080/08820.538.2021.1893758.

[9] AddedBytes. Readable. In, 2011-2023.

[10] Flesch R. A new readability yardstick. J Appl Psychol 1948; 32: 221-33. doi:10.1037/h0057532.

[11] Kincaid P, Fishburne RP, Rogers RL, Chissom BS. Derivation of New Readability Formulas (Automated Readability Index, Fog Count and Flesch Reading Ease Formula) for Navy Enlisted Personnel. 1975. Institute for Simulation and Training. 56. https://stars.library.ucf.edu/istlibrary/56 Accessed on 10 January, 2024

[12] Jindal P, MacDermid JC. Assessing reading levels of health information: uses and limitations of flesch formula. Educ Health (Abingdon) 2017; 30: 84-8. doi:10.4103/1357-6283.210517.

[13] McLaughlin GH. SMOG grading: A new readability formula. J Read 1969; 12: 639-46.

[14] Nath S, Marie A, Ellershaw S, et al. New meaning for NLP: the trials and tribulations of natural language processing with GPT-3 in ophthalmology. Br J Ophthalmol 2022; 106: 889-92. doi:10.1136/bjophthalmol-2022-321141.

[15] Kung TH, Cheatham M, Medenilla A, et al. Performance of ChatGPT on USMLE: Potential for AI-assisted medical education using large language models. PLOS Digit Health 2023; 2: e0000198. doi:10.1371/journal.pdig.0000198.

[16] Flesch RF. Art of readable writing. Pennsylvania: The Haddon Craftsmen, 1949.

[17] Hamat A, Jaludin A, Mohd-Dom TN et al. Diabetes in the news: readability analysis of malaysian diabetes corpus. Int J Environ Res Public Health 2022; 19:6802. doi:10.3390/ijerph19116802