© Modeling and Artificial Intelligence in Ophthalmology 2024; 6:1-11
Original article
https://doi.org/10.35119/maio.v6i1.137

Artificial intelligence in practice: measuring its medical accuracy in oculoplastics consultations

Adam J. Neuhouser[1], Alisha Kamboj[1], Ali Mokhtarzadeh[1], Andrew R. Harrison[1,2]

[1]Department of Ophthalmology and Visual Neurosciences, University of Minnesota, Minneapolis, MN, USA; [2]Department of Otolaryngology and Head and Neck Surgery, University of Minnesota, Minneapolis, MN, USA

Abstract

Purpose: The aim of this study was to investigate the medical accuracy of responses produced by Chat Generative Pretrained Transformer 4 (Chat GPT-4) and DALLE-2 in relation to common questions encountered during oculoplastic consultations.
Methods: The 5 most frequently discussed oculoplastic procedures on social media were selected for evaluation using Chat GPT-4 and DALLE-2. Questions were formulated from common patient concerns and inputted into Chat GPT-4, and responses were assessed on a 3-point scale. For procedure imagery, descriptions were submitted to DALLE-2, and the resulted images were graded for anatomical and surgical accuracy. Grading was completed by 5 oculoplastic surgeons through a 110-question survey.
Results: Overall, 87.3% of Chat GPT-4's responses achieved a score of 2 or 3 points, denoting a good to high level of accuracy. Across all procedures, questions about pain, bruising, procedure risk, and adverse events garnered high scores. Conversely, responses regarding specific case scenarios, procedure longevity, and procedure definitions were less accurate. Images produced by DALLE-2-were notably subpar, often failing to accurately depict surgical outcomes and realistic details.
Conclusions: Chat GPT-4 demonstrated a creditable level of accuracy in addressing common oculoplastic procedure concerns. However, its limitations in handling case-based scenarios suggests that it is best suited as a supplementary source of information rather than a primary diagnostic or consultative tool. The current state

Correspondence: Adam Neuhouser, MD, Department of Ophthalmology and Visual Neurosciences, University of Minnesota, Phillips Wangensteen Building, 9th Floor, 516 Delaware Street SE, Minneapolis, MN 55455, USA
E-mail: neuho005@umn.edu

of medical imagery generated by means of artificial intelligence lacks anatomical accuracy. Significant technological advancements are necessary before such imagery can complement oculoplastic consultations effectively.

Keywords: artificial intelligence, Chat GPT, DALLE, oculoplastics, patient information

1. Introduction

In recent years, the demand for oculoplastic services and related educational resources has significantly increased across the United States.[1] In accordance with this trend, the American Society of Ophthalmic Plastic and Reconstructive Surgery has produced patient education brochures on many oculoplastic conditions and pathologies to inform patients of their diagnoses and treatment options. However, studies demonstrate that these resources and other online oculoplastic educational materials may be written at inappropriate reading levels and demonstrate low accountability.[2,3]

Prospective oculoplastics patients often resort to social media platforms to review and discuss various cosmetic and reconstructive surgeries.[4,5] In today's digitally connected world, social media has emerged as a primary source of information for individuals seeking health advice, second opinions, and information about physicians. However, the prevalence of misinformation and unreliable sources on social media platforms may misguide patients, potentially leading them away from making scientifically and medically informed decisions.[5]

The Chat Generative Pretrained Transformer 4 (Chat GPT-4) is the latest artificial intelligence (AI) chatbot developed by OpenAI, an AI research laboratory. This large language model leverages deep-learning neural network software to generate contextually relevant responses to user prompts, basing its responses on a vast majority of digitally accessible, text-based information up to its last training date, September 2021. Chat GPT-4's ability to generate detailed responses to complex inquiries has propelled it to become the fastest-growing consumer application in history.[6] Given its increasing popularity, it is anticipated that patients will likely resort to Chat GPT-4 for medical advice.

Prospective patients also often query social media and other websites to view before-and-after procedure photographs, which are cited as one of the most influential factors in deciding whether to undergo a procedure.[7] Some patients may soon turn to generative AI models like DALLE-2—a text-to-image AI system developed by OpenAI—to create novel before-and-after photographs.

To the authors' knowledge, there has not yet been an investigation into the medical accuracy of responses generated by Chat GPT-4 and DALLE-2 with regards to oculoplastics concerns. In this study, we present an analysis of Chat GPT-4-gen-

erated responses regarding the 5 most commonly discussed oculoplastics procedures and evaluate the accuracy of AI-generated pre- and post-procedure photographs, aiming to assess the medical accuracy of OpenAI technology.

2. Methods

Based on a cross-sectional study by Schmuter *et al.*,[3] we selected the top 5 most frequently discussed oculoplastic procedures on social media to represent the topics most likely to be queried on Chat GPT-4. These procedures included facial filler, botulinum toxin injection, lower blepharoplasty, upper blepharoplasty, and ptosis repair. For each procedure, 9 categories of prompts were formulated based on common patient questions or concerns during clinic encounters. Four prompts corresponded to preoperative assessment questions (*e.g.*, related to procedure definition, risks, cost, and condition etiology), 4 related to postoperative outcome inquiries (*e.g.*, related to procedure pain, bruising, scarring, and longevity), and one was a case-based question. All prompts were written from the patient's perspective.

All questions were posed to Chat GPT-4 in a new session to mitigate any potential learning or contextualization from previous queries. This approach was used to simulate a first-time interaction for each question. To account for variability and randomness in responses, each question was inputted into Chat GPT-4 twice. Each output was recorded for analysis. Responses were evaluated based on a 3-point grading scale:

- 3 points for detailed and highly accurate answers that covered all aspects of the question.
- 2 points for answers that were mostly accurate but may have minor omissions.
- 1 point for those that provided some accurate information but missed several key points.
- 0 points for answers that were largely inaccurate or failed to address important aspects of the question.

To evaluate the accuracy of AI-generated procedure imagery, a short, nonspecific, text description of an image was written for each procedure. The descriptions varied between before-and-after photos and postoperative recovery images. The prompts were submitted to DALLE-2 in independent sessions, and it produced 4 images corresponding to the given description. The 4 images were submitted for analysis and graded on a 3-point scale:

- 3 points for accurate, clear, and anatomically realistic representations of surgical outcomes.
- 2 points for adequate depictions with satisfactory anatomical realism.
- 1 point for limited or vague portrayals with minimal anatomical accuracy.
- 0 points for images that inadequately represented the surgical outcome.

A 110-question, web-based, point-based survey was created using Google Forms (Alphabet Inc., Mountain View, CA, USA) and distributed by e-mail to 5 oculoplastic surgeons certified by the American Society of Ophthalmic Plastic and Reconstructive Surgery (ASOPRS). These surgeons independently assessed the suitability of all 90 Chat GPT-4 responses and 20 DALLE-2 images, producing a total of 550 evaluations. The survey data was collected in an anonymous fashion. The statistics were performed using DATAtab: Online Statistics Calculator (datatab.net, Graz, Austria).[8] When applicable, the alpha level was selected to be 0.05.

We calculated the overarching percentages of scores attributed to all questions and the scoring distribution by procedure. A detailed breakdown by prompt category for each procedure was evaluated to highlight specific areas of strength or concern in Chat GPT-4 responses. Inter-rater reliability among the oculoplastic surgeons was assessed using Kendall's coefficient of concordance, a measure well-suited for ordinal data such as our 3-point grading scale. To determine if there were consistent differences between 2 sets of responses, the Wilcoxon test was used to analyze the variability of scores from duplicate Chat GPT-4 responses. The Jaccard similarity coefficient, calculated using an online algorithm from Tilores (tilores.io, Berlin, Germany), quantified the text similarity of the duplicated responses from Chat GPT-4.[9] Finally, the Friedman test compared the scores across images generated by DALLE-2 to determine whether there were notable differences in ratings for images associated with the same prompt. This study design did not require ethics review by an institutional review board.

3. Results

The full list of question and image prompts, and associated responses, can be accessed in the Appendix. A summary of average scores for Chat GPT-4 and DALLE-2 responses to the prompts utilized in this investigation is provided in Table 1.

Of the 450 evaluations of Chat GPT-4 responses, 172 evaluations (38.2%) received a score of 3 points, while 221 evaluations (49.1%) received a score of 2 points, 52 evaluations (11.6%) received a score of 1 point, and 5 evaluations (1.1%) received a score of 0 points.

Overall, Chat GPT-4 scores were comparable for all procedures, ranging between 2.13 and 2.38 points. The highest average score pertained to the answer given regarding upper blepharoplasty pain and bruising (2.7 points), while the lowest score was associated with a facial filler case-based question (0.9 points). Chat GPT-4 scored the highest on prompts addressing "procedure pain" (2.52 points), "risks/adverse events" (2.4 -points), and "procedure bruising" (2.46 points). In contrast, prompts related to "case scenarios" (1.80 points), "procedure longevity" (2.02 points), and "procedure definition" (2.12 points) received the lowest average scores (Table 1).

Table 1. Chat GPT-4 prompt category and procedure average scores (0–3 points)

Procedure	Facial filler	Botulinum toxin injection	Lower blephar-oplasty	Upper blephar-oplasty	Ptosis r epair	**Category average score**
Prompt category						
Procedure definition	2.3	1.9	2.2	2.2	2.0	2.12
Risks/adverse events	2.6	2.2	2.5	2.6	2.4	2.46
Etiology of condition	2.4	2.1	2.2	2.6	2.4	2.34
Cost of procedure	2.4	2.3	2.1	1.9	2.4	2.22
Procedure pain	2.6	2.4	2.5	2.7	2.4	2.52
Procedure bruising	2.4	2.4	2.5	2.7	2.3	2.46
Procedure scarring	2.2	2.5	2.4	2.5	1.7	2.26
Procedure longevity	2.4	1.7	2.3	1.8	1.9	2.02
Case scenario	0.9	1.8	2.2	2.4	1.7	1.80
Procedure average score	2.24	2.14	2.32	2.38	2.13	

The same responses were evaluated by multiple reviewers. Inter-rater reliability analysis revealed a Kendall's coefficient of concordance of 0.57, indicating moderate agreement among the reviewers.

A Wilcoxon test determined the consistency of Chat GPT-4's responses. There was no significant difference between the scores for the first and second set of responses for the same question ($W = 1149$, $p = 0.77$), highlighting the stability of Chat GPT-4's responses. The Jaccard similarity coefficient further emphasized this consistency ($J = 0.84$), indicating that 84% of the observed attributes overlapped between the 2 response sets.

For image-generator prompts, 4 unique images were generated from each submission. As an example, DALLE-2's image-generated interpretation of the prompt, "Full face before-and-after photos of a 50-year-old person who underwent tear trough filler" is viewable in Figure 1. Of the 100 total evaluations, the overall

Fig. 1. DALLE-2's image-generated interpretation of the prompt "Full face before-and-after photos of a 50-year-old person who underwent tear trough filler."

Table 2. DALLE-2 image set average scores (0–3 points)

Procedure	Facial filler	Botulinum toxin injection	Lower blephar-oplasty	Upper blephar-oplasty	Ptosis repair	Total average score
Image Set Average Score	0.50	0.45	0.70	0.45	0.25	0.47

average score was 0.47 points, with individual image sets ranging from 0.25 to 0.70 points (Table 2).

A Friedman test assessed score differences across the 4 images within each prompt. It revealed no significant score variations among images from the same prompt ($p > 0.05$ for all), indicating consistent reviewer perceptions of the images' realism.

5. Discussion

As interest in AI integration within the broader healthcare system increases, understanding its capabilities and limitations is vital to prevent potential misinformation leading to confusion or harmful outcomes. The efficiency and convenience of Chat GPT-4 has contributed to its rapid adoption, but its trustworthiness to provide accurate medical information remains under scrutiny. This survey aimed to evaluate the accuracy and reliability of OpenAI technology in answering common oculoplastic procedure questions and expand upon its potential application in the clinical setting. It is also the first cross-sectional study assessing the accuracy of pre- and post-procedure, AI-generated images.

Overall, the vast majority (87.3%) of Chat GPT-4's responses achieved a score of 2 or 3 points, denoting a good to high level of accuracy, with minor omissions of essential details. Comparable scores were also found for the overall average response scores of each procedure. This suggests that Chat GPT-4 has the potential to provide fairly accurate general medical information. It may serve as a valuable resource for information in regions lacking immediate medical expertise.

Across the literature, several surveys have assessed the accuracy of Chat GPT's responses to medical questions, often employing a point-system or Likert-scale for evaluation. These studies commonly state that the AI-generated responses are generally of acceptable quality. Reported rates of Chat GPT-4 answers scoring "appropriate", "very good," or "good" in other studies ranged from 73% to 96%.[10-16] The results of this survey align with these rates, which highlight Chat GPT-4's potential as a supplementary resource for healthcare professionals and as an educational aid for patients. However, these studies, including our own, have identified instances of misinformation, suggesting a need for vigilant verification of AI-generated content. The consensus among researchers is that AI-generated information requires additional validation studies on accuracy and patient safety before its integration into clinical practice can proceed.

The highest scored responses pertained to anticipated pain and bruising for upper blepharoplasty (2.7 points). Across all procedures, prompts related to pain and bruising scored high (2.52 and 2.46 points, respectively), as did prompts regarding procedure risks and potential adverse events (2.46 points). The higher scores in these areas may result from a higher quantity and standardization of

associated data pertaining to these topics in the medical literature. As pain, bruising, and risks of a procedure are common concerns for patients, Chat GPT-4's accuracy in these areas is reassuring; this platform may be able to effectively provide basic information in these areas.

On the other hand, responses regarding specific case scenarios (average score of 1.80 points), procedure longevity (2.0 points), and procedure definition (2.12 points) received the lowest scores. The poor performance with specific case scenarios highlights Chat GPT-4's limitations and the ongoing necessity for human medical judgment and regulation. In the majority of cases, Chat GPT-4 avoided directly answering the patients' question, instead offering general information and suggesting that the patient pursue a professional consultation. Recommendations for patient-specific scenarios require nuanced insights, which at this moment appear to be a limitation in Chat GPT-4's skill set.

Chat GPT-4 is designed to introduce variability in responses to identical inputs. This is due to the inherent randomness introduced during the sampling process when the model produces outputs. Both the Wilcoxon test and the Jaccard similarity coefficient offered a multidimensional perspective on the consistency and similarity of Chat GPT-4's outputs. The Wilcoxon test found no significant difference in scores given by reviewers between the first and second response to a prompt ($W = 1149$, $p = 0.77$), and a Jaccard similarity coefficient of 0.84 implied that 84% of the textual content of the paired responses overlapped. This consistency enhances the reliability of Chat GPT-4 and assures patients and medical professionals that repeated inquiries would yield similar quality responses.

Overall, DALLE-2's AI-generated images scored low; the average score among all images was 0.47 points. The average score for the 5 image sets ranged from 0.25 points to 0.70 points. These images failed to depict accurate surgical outcomes and lacked realistic detail. Given the importance of before-and-after images in setting patient expectations, this finding highlights a gap in DALLE-2's present capabilities. Common shortcomings of these photos included failure to show all relevant aspects of the face, failure to maintain ipsilaterality for linked images, and reversal of the order of before-and-after photos. Patients relying on such images to make an informed decision may be misled by portrayed surgical outcomes or postoperative states.

In assessing the perceived realism of AI-generated photos across the 4 images per prompt, the Friedman test revealed no significant differences in scores among the images from the same prompt ($p > 0.05$ for all). This suggests, the images were perceived as similarly poor, and the reviewers did not consistently rank one image as more realistic than the others. This also implies a degree of consistency among the reviewers in their assessments of the photos.

Training data for AI largely influences the quality and characteristics of the generated images. The limited availability of real pre- and postoperative images due to protections regarding patient health information might be contributing to the shortfall in pre- and post-procedure image realism.[17] It is also possible that the complexities of postoperative healing are too intricate for DALLE-2's current rendition.

The primary limitation of this study was the limited number of reviewers. While all 5 reviewers were board-certified ASOPRS surgeons, a small number of reviewers raises concerns about potential biases. Reviewers were not blinded to the source of the generated answers and photos. This knowledge was inevitable given the nature of the study, and preconceived notions about OpenAI technology may have influenced their responses. Additionally, the study's narrow focus on oculoplastic procedures limits the results' generalizability to other fields of medicine. Another limitation was the use of a single IP address for all question sessions. While new user sessions were initiated for each query to simulate first-time interactions, the consistent IP address may have influenced the AI model's learning and response generation.

While Chat GPT-4 demonstrated acceptable responses for several generic prompts, it fell short in offering personalized advice. Meanwhile, DALLE-2's poor ability to generate medical imagery underscores the challenges with creating accurate and realistic medical images. These findings highlight the current capabilities of OpenAI technology, which support its use as a basic educational tool for helping patients learn about oculoplastic procedures. The model's performance on case-based scenarios revealed a key opportunity for improvement. Both Chat GPT-4 and DALLE-2 should be viewed as supplementary tools rather than primary diagnostic or consultative platforms. In their current form, these AI tools cannot replace evaluation and guidance by a physician.

Future collaborations with medical institutions and feedback from medical professionals may enhance AI's performance and image generation capabilities. With additional research in machine learning and the use of larger datasets, these tools may eventually be able to provide second opinions for oculoplastic patients, especially those living in remote locations. The goal is for AI to empower healthcare providers and guide patients with reliable and up-to-date information to make evidenced-based decisions. This study offers a preliminary evaluation of the accuracy and consistency of information generated by Chat GPT-4 and DALLE-2 in the context of common oculoplastic concerns. Further research with a more expansive dataset is essential to develop a definitive understanding of their capabilities and limitations as medical tools.

Declarations

Ethics approval and consent to participate
Not required.

Competing interests
Adam J. Neuhouser, Alisha Kamboj, and Ali Mokhtarzadeh have no competing interests to declare. Andrew R. Harrison is speaker and consultant for Horizon Pharmaceuticals and RVL Pharmaceuticals.

Funding
None to declare.

Acknowledgements
None to declare.

References

1. Akosman S, Qi L, Pakhchanian H, Foos W, Maliakkal J, Raiker R, Belyea DA, Geist C. Using infodemiology metrics to assess patient demand for oculoplastic surgeons in the United States: insights from Google Search Trends. Orbit. 2022 Nov 12;1-7. https://doi.org/10.1080/01676830.2022.2142945
2. Cohen SA, Tijerina JD, Kossler A. The Readability and Accountability of Online Patient Education Materials Related to Common Oculoplastics Diagnoses and Treatments. Semin Ophthalmol. 2023;38(4):387-393. https://doi.org/10.1080/08820538.2022.2158039
3. Chen J, Wang Y. Social Media Use for Health Purposes: Systematic Review. Journal of Medical Internet Research. 2021;23(5):e17917. https://doi.org/10.2196/17917
4. Arab K, Barasain O, Altaweel A, et al. Influence of Social Media on the Decision to Undergo a Cosmetic Procedure. Plastic and Reconstructive Surgery Global Open. 2019;7(8):e2333. https://doi.org/10.1097/GOX.0000000000002333
5. Schmuter G, North VS, Kazim M, Tran AQ. Medical Accuracy of Patient Discussions in Oculoplastic Surgery on Social Media. Ophthalmic Plastic and Reconstructive Surgery. 2023;39(2):132-135. https://doi.org/10.1097/IOP.0000000000002257
6. Bartz D, Bartz D. As ChatGPT's popularity explodes, U.S. lawmakers take an interest. Reuters. 2023 Feb 13.
7. Nayak LM, Linkov G. Social Media Marketing in Facial Plastic Surgery: What Has Worked? Facial Plastic Surgery Clinics of North America. 2019;27(3):373-377. https://doi.org/10.1016/j.fsc.2019.04.002
8. DATAtab Team. Cite DATAtab: DATAtab: Online Statistics Calculator. DATAtab e.U. Graz, Austria; 2023.
9. Tilores.io. Jaccard Similarity Coefficient Algorithm Online Tool. [Accessed September 5, 2023]. Available from: https://tilores.io/jaccard-similarity-coefficient-algorithm-online-tool
10. Mago J, Sharma M. The Potential Usefulness of ChatGPT in Oral and Maxillofacial Radiology. Cureus. 2023;15(7):e42133. https://doi.org/10.7759/cureus.42133

11. Hu X, Ran AR, Nguyen TX, et al. What can GPT-4 do for Diagnosing Rare Eye Diseases? A Pilot Study. Ophthalmology Therapy. 2023. https://doi.org/10.1007/s40123-023-00789-8
12. Lahat A, Shachar E, Avidan B, et al. Evaluating the use of large language model in identifying top research questions in gastroenterology. Sci Rep. 2023;13:4164. https://doi.org/10.1038/s41598-023-31412-2
13. Biswas S, Logan NS, Davies LN, Sheppard AL, Wolffsohn JS. Assessing the utility of ChatGPT as an artificial intelligence-based large language model for information to answer questions on myopia. Ophthalmic and Physiological Optics. 2023;43(6):1562-1570. https://doi.org/10.1111/opo.13207
14. Samaan JS, Yeo YH, Rajeev N, Hawley L, Abel S, Ng WH, Samakar K. Assessing the Accuracy of Responses by the Language Model ChatGPT to Questions Regarding Bariatric Surgery. Obesity Surgery. 2023;33(6):1790-1796. https://doi.org/10.1007/s11695-023-06603-5
15. Johnson D, Goodman R, Patrinely J, et al. Assessing the Accuracy and Reliability of AI-Generated Medical Responses: An Evaluation of the Chat-GPT Model. Preprint. Research Square. 2023 Feb 28. https://doi.org/10.21203/rs.3.rs-2566942/v1
16. Sarraju A, Bruemmer D, Van Iterson E, Cho L, Rodriguez F, Laffin L. Appropriateness of Cardiovascular Disease Prevention Recommendations Obtained From a Popular Online Chat-Based Artificial Intelligence Model. JAMA. 2023;329(10):842-844. https://doi.org/10.1001/jama.2023.1044
17. Karako K, Song P, Chen Y, Tang W. New Possibilities for Medical Support Systems Utilizing Artificial Intelligence (AI) and Data Platforms. Bioscience Trends. 2023;17(3):186-189. https://doi.org/10.5582/bst.2023.01138