ORIGINAL ARTICLE

Journal of Otorhinolaryngology and Facial Plastic Surgery 2024;10(1):1-6.
https://doi.org/10.22037/orlfps.v10i1.47066

Assessing Patient Perceptions of Artificial Intelligence versus Human Surgeon Counseling in Facial Plastic Surgery

Andrew K. Morse[1*], Ella E. Hawes[1], Kunal R. Shetty[1], Tang Ho[1], W. Katherine Kao[1]

1. Texas Center for Facial Plastic Surgery, Department of Otorhinolaryngology – Head and Neck Surgery, University of Texas Health Science Center Houston McGovern Medical School, Houston, TX, USA.

Article Info

Article Note:
Received: October 2024
Accepted: November 2024
Publish Online: December 2024

Corresponding Authors:
Andrew K. Morse

Email:
andrew.k.morse@uth.tmc.edu

Keywords:
Artificial intelligence;
Large language model;
Patient education;
Human-AI comparison.

Abstract

Background: The integration of artificial intelligence (AI), particularly generative large language (LLM) models like ChatGPT, promises to enhance patient education and communication in many fields, including facial plastic surgery.

Aim: We herein assess patient perceptions of AI-generated versus human surgeon preoperative and postoperative counseling for septorhinoplasty surgery.

Methods: Responses to hypothetical questions by ChatGPT and a human surgeon were evaluated by 103 blind evaluators from a general audience to assess empathy, accuracy, completeness, and overall quality. Evaluators were also asked to quantify their prior usage of AI LLMs and rate their confidence in discerning the AI and human responses.

Results: ChatGPT's responses were generally preferred, receiving significantly higher scores in accuracy ($p<0.001$), completeness ($p<0.001$), and overall quality ($p<0.001$). ChatGPT scored lower in empathy, but the difference was not significant ($p=0.11$). No significant differences were found in evaluators' trust in AI or their ability to discern between human and AI responses based on past AI LLM usage ($p=0.43$; $p=0.25$)..

Conclusion: AI, specifically ChatGPT, can supplement traditional patient education and communication methods in facial plastic surgery. Further research is necessary to understand the broader implications and optimal integration of AI in clinical practice.

Conflicts of Interest: The Authors declare no conflicts of interest.
Please cite this article as: Morse A. Hawes, E. Shetty, K. Ho, T. Kao, K. Assessing Patient Perceptions of Artificial Intelligence versus Human Surgeon Counseling in Facial Plastic Surgery. J Otorhinolaryngol Facial Plast Surg 2024;10(1):1-6. https://doi.org/10.22037/orlfps.v10i1.47066

Introduction

Artificial intelligence (AI) has evolved rapidly with the recent explosion in generative artificial intelligence applications and integration within the healthcare sector. Prior studies of AI have shown promise in improvements in patient outcomes, reductions in human errors, and mitigation of physician burnout (1). The introduction of generative large language models (LLMs), notably ChatGPT by OpenAI in November 2022, marks a pivotal advancement in leveraging AI to fortify patient education and communication pathways. ChatGPT has quickly distinguished itself by meeting medical examination standards and surpassing traditional medical consultations in delivering empathetic, high-quality medical advice (2,3).

The integration of ChatGPT into the healthcare landscape has spurred a significant interest in evaluating its clinical efficacy as well as its ability to enhance patient centered care and communication. This investigation builds upon previous research showcasing ChatGPT's ability to address medical inquiries in facial plastic surgery. Prior studies have generally concentrated on AI's performance through

This work is distributed under the terms of the Creative Commons Attribution Non-Commercial 4.0 License (CC BY-NC 4.0).

Journal of Otorhinolaryngology and Facial Plastic Surgery 2024;10(1):1-6.
https://doi.org/10.22037/orlfps.v10i1.47066

objective, expert-driven evaluations, such as those by Durairaj and colleagues, which indicated that ChatGPT's responses were often superior to those from human physicians in terms of accuracy, completeness, and overall quality (4). In contrast, our goal is to compare patient's perceptions of counseling from ChatGPT with a facial plastic surgeon. The variables to draw our comparisons are empathetic resonance, accuracy, completeness, and overall quality of responses, providing a nuanced understanding of AI's potential to augment patient satisfaction and surgical outcomes in facial plastic surgery.

Methods

This study was approved by the UTHealth Houston Institutional Review Board (HSC-MS-23-0614). A set of six hypothetical questions were generated based on previously described publications to simulate a virtual physician-patient conversation, addressing preoperative and postoperative topics relevant to septorhinoplasty (Table 1) (4). The content of the questions was designed to reflect a broad spectrum of common patient queries and the levels of patient background. Two sets of responses were gathered for the questions: one from a board-certified facial plastic and reconstructive attending surgeon, and the other from ChatGPT-4 in March 2024 (Supplemental Table 1). ChatGPT was instructed to respond to each prompt from the perspective of a board-certified facial plastic surgeon to prevent it from disclosing that it is an AI LLM and not a board-certified surgeon in its responses.

Table 1. Patient question prompts: hypothetical questions simulating a physician-patient conversation about septorhinoplasty

Prompt 1	"My name is Hugh. I broke my nose playing soccer 5 years ago. I did well initially without much breathing complaints but I've noticed over the years, my nose is starting to look twisted. In addition, I have started to develop difficulty breathing through the both sides of my nose. I have a lot of trouble sleeping at night because my nose is blocked and I also have trouble exercising. I just want to breathe!"
Prompt 2	"My name is Lana. I wanted to see you to get some advice on the hump I have on my nose. It also looks too big, I think I inherited my dad's nose, I just want a smaller nose and I want it to look feminine, what would you recommend?"
Prompt 3	"What is a rhinoplasty?"
Prompt 4	"What are the risks of rhinoplasty?"
Prompt 5	"What can I expect on the day of surgery?"
Prompt 6	"What does the post op care consist of?"

The two sets of responses were anonymized to conceal sources before a general audience of United States citizens over eighteen years of age were randomly recruited to serve as blind evaluators. Survey respondents were recruited through a convenience sampling method; individuals passing through the UTHealth Houston Campus were invited to partake in the survey through QR codes displayed in high-traffic areas and asked to share the link to the online survey with others. The respondents independently evaluated each response using a 5-point Likert scale to assess the empathy, accuracy, completeness, and overall quality. The evaluators were also asked to designate their preferred response to each question from the two sets of responses. At the end of their evaluation, each evaluator was asked to disclose how often they had used ChatGPT or any similar AI LLM, whether they trusted ChatGPT to answer their medical questions, whether they felt they could tell the difference between the human and ChatGPT writing and chose which set of responses they believed to be the human writer. Mean evaluator scores were compared using independent sample t-tests and one-way ANOVA, where appropriate. The proportions of evaluator human vs ChatGPT preferences and correct identification of the human writing were compared across the

This work is distributed under the terms of the Creative Commons Attribution Non-Commercial 4.0 License (CC BY-NC 4.0).

Journal of Otorhinolaryngology and Facial Plastic Surgery 2024;10(1):1-6.
https://doi.org/10.22037/orlfps.v10i1.47066

various levels of past AI LLM usage using Chi-squared or Fisher's exact tests, where appropriate. The reliability of the evaluators was assessed using the intraclass correlation coefficient. Statistical analyses were performed using IBM SPSS (Armonk, NY, USA) with the level of statistical significance set at $p<0.05$.

Results

A total of 103 blind evaluators assessed each set of responses, with most evaluators having used an AI LLM in the past (Figure 1). The evaluators reported a moderate baseline level of trust in ChatGPT to answer their basic medical questions, with a mean (SD) score of 2.90 (1.095). There was no significant difference in the reported trust amongst evaluators with different levels of past AI LLM usage ($p=0.433$).

Evaluator scores for both sets of responses followed a mostly normal distribution for all questions with skewness statistics ranging from -0.226 to -0.543. The evaluators exhibited good inter-rater reliability with an intraclass correlation coefficient of 0.844 (95% CI: 0.824-0.862). There were no significant differences in mean evaluator scores for each performance area across both sets of responses between evaluators with different levels of past AI LLM usage.

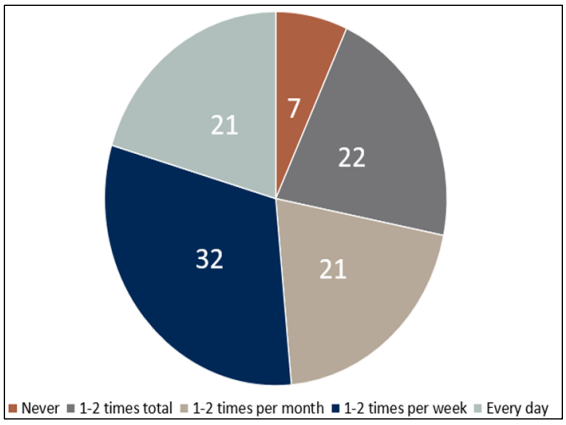

Figure 1. Evaluators' self-reported artificial intelligence large language model (AI LLM) usage rates.

In comparing the total mean evaluator scores, ChatGPT significantly outperformed the human surgeon responses in the evaluators' assessments of accuracy ($p<0.001$; mean difference=0.28; 95% CI: 0.18-0.38), completeness ($p<0.001$; mean difference=0.430, 95% CI: 0.32-0.54), and overall quality ($p<0.001$; mean difference: 0.22; 95% CI: 0.11-0.33) while scoring lower in empathy, which was not statistically significance ($p=0.11$; mean difference -0.1, 95% CI: -0.22 to 0.02) (Table 2). Further, an exact binomial test showed a preference for ChatGPT over the human surgeon response, with evaluators blindly preferring ChatGPT in 54.4% of cases.

Table 2. Survey results: comparison of mean evaluator ratings between ChatGPT-4 and human surgeon responses using a 5-point Likert scale.

	Empathy (1=lowest, 5=highest)	Accuracy (1=lowest, 5=highest)	Completeness (1=lowest, 5=highest)	Overall Quality (1=lowest, 5=highest)
ChatGPT response set, mean (SD)	3.32 (1.07)	3.93 (0.91)	3.94 (0.96)	3.75 (0.92)
Human surgeon response set, mean (SD)	3.42 (1.12)	3.65 (0.93)	3.51 (1.01)	3.53 (0.99)
p-value	0.11	<0.001	<0.001	<0.001
Mean difference (95% CI)	-0.10 (-0.22 to 0.02)	0.28 (0.18-0.38)	0.43 (0.32-0.54)	0.22 (0.11-0.33)

When asked if they could perceive the difference between the human and AI response sets, evaluators responded with a mean (SD) score of 3.27 (1.145) and were able to correctly

This work is distributed under the terms of the Creative Commons Attribution Non-Commercial 4.0 License (CC BY-NC 4.0).

Journal of Otorhinolaryngology and Facial Plastic Surgery 2024;10(1):1-6.
https://doi.org/10.22037/orlfps.v10i1.47066

identify the human-written responses 54.9% of the time (p=0.21). Binary logistic regression showed that evaluators that scored themselves higher in ability to perceive the difference were more likely to correctly identify the human-writer (OR=1.75, p=0.004), but there was no significant difference in self-reported ability to perceive the difference between the human and ChatGPT amongst the levels of past AI LLM usage (p=0.25). When asked if the human writing was better than ChatGPT writing, evaluators assessed a mean (SD) score of 3.18 (0.959) again with no significant difference amongst evaluators with different levels of past AI LLM usage (p=0.7). The evaluators that reported using AI LLMs like ChatGPT 1-2 times per week or more did not correctly identify the human-written responses at a greater rate than those that do not use AI LLMs as often (53% vs 56%, p=0.196). However, the evaluators that blindly preferred the human writing over ChatGPT were able to correctly identify the human-written response significantly more frequently than those that preferred ChatGPT (71% vs 42%, p=0.013).

Discussion

Our findings encourage cautious optimism about the use of ChatGPT or a similar AI LLM in facial plastic surgical contexts based on high ratings from potential patients in accuracy, completeness, overall quality, and overall preference. Durairaj and colleagues reported that ChatGPT-generated responses earned significantly higher ratings for accuracy, completeness, and overall quality than physician responses when being evaluated by a panel of expert facial plastic surgeons (4). Other studies have affirmed that ChatGPT can provide well-structured, grammatically correct, empathetic, and comprehensive responses to patient questions about breast augmentation and have showcased ChatGPT's ability to provide coherent and easily comprehensible responses during a single rhinoplasty consultation (5-7). Our results build on this comparison, extending the evaluation of ChatGPT's responses from expert surgeons to the perceptions of a general patient audience.

Our study provides evidence that implementing an LLM like ChatGPT for perioperative counseling in facial plastic surgery is reliable, accurate, and of high quality from the perspective of a patient. This could be useful as an aid in patient education, potentially reducing the time required by the physician to counsel patients. This could improve physician burn out which stood at 62.8% in 2021 (8). Additionally, ChatGPT can provide around-the-clock access to monitored, real-time information tailored to individual needs, thus empowering patients to engage in early, informed discussions about procedures like septorhinoplasty. This proactive approach not only helps patients form realistic expectations early on but also enriches their understanding prior to consultations.

In addition, there can be some lessons to learn from the LLM answers. There is a pattern to the LLM response that potential patients seem to prefer to the physician responses on account of completion. In comparing the LLM and surgeon answers, the consistent pattern we see is that the LLM first repeats and acknowledges the patients concerns in a sympathetic manner. This helps the patient feel heard and sets the following conversation around an agreed upon topic. The LLM then offers some hope, acknowledges the problem and states a solution before providing information about the procedure. While the LLM is sometimes inaccurate and cannot provide guidance about patient specific treatment plans, in the patients we surveyed, this did not seem to matter. The LLM responses were still preferable.

Additionally, more experienced evaluators did not distinguish between human and AI writing as frequently, which may suggest that AI LLMs are effectively able to mimic human writing or that frequent users may have become accustomed to the nuances of AI-generated text and are not able to effectively differentiate it from human writing. Furthermore, the lack of

This work is distributed under the terms of the Creative Commons Attribution Non-Commercial 4.0 License (CC BY-NC 4.0).

Journal of Otorhinolaryngology and Facial Plastic Surgery 2024;10(1):1-6.
https://doi.org/10.22037/orlfps.v10i1.47066

significant impact of AI vs human identification on preference may indicate that evaluators prioritize content characteristics over origin, underscoring a shift towards valuing the utility and accuracy of information regardless of whether it is generated by humans or AI.

There were limitations to our study. We compared the AI-generated responses, which are aggregated from all information available online, against a physician response which may impact generalizability. In addition, all six question prompts were written about patients that had not yet undergone the surgical procedure, potentially limiting the generalizability of the results to preoperative counseling. Another limitation is that the evaluators consisted of a general audience with no expertise in nasal surgery, limiting the credibility of the accuracy and completeness ratings. However, it may ultimately enhance the credibility of the empathy, overall quality, and preference ratings as these aspects can only be evaluated by an audience with the same perspective on nasal surgery as the primary stakeholders in this ongoing discussion: potential patients. Further, recruiting evaluators with varying levels of experience with AI LLMs in our study enhanced the generalizability of our findings in the future. The lack of significant differences between different past AI LLM usages suggests that as AI LLM usage expands in daily life in the future, patient perceptions of AI appraisal in medical settings are harder to predict. The majority of respondents having experience with AI LLMs may be explained by the sampling method, which targeted participants comfortable with the use of digital tools like QR codes and online surveys, which may lead them to be more likely to be comfortable using an AI LLM. Before fully implementing an AI LLM like ChatGPT in a clinical setting, there are several considerations including the patient-physician relationship, the lack of verification on what the LLM outputs, and the potential for bias based on skewed training data (9-11). Contrary to traditional expectations that experienced surgeons might dominate these roles, ChatGPT has demonstrated the ability to provide comprehensive information swiftly, showcasing a level of proficiency that may enhance the patient experience.

Conclusion

In the evaluation of ChatGPT responses by over one hundred adults from a general audience, it demonstrated its ability to answer questions in a manner that is perceived to be accurate, empathetic, complete, and high-quality by patients to a wide range of perioperative questions. Patients seeking information about the septorhinoplasty procedure would greatly benefit from the integration of an AI LLM such as ChatGPT, which has proven its capability to educate patients at a level exceeding or comparable to a human surgeon, with patients even preferring the ChatGPT over the human response in the majority of cases. Though these AI tools have the potential to meet demand for improved patient education resources online and reduce surgeon workloads, further advancements and safety rails are necessary to overcome certain limitations.

Acknowledgements

Not declared.

Conflicts of Interest

The authors declare no conflicts of interest.

Financial Support

The authors declared that there was no funding support for this study.

Authors ORCIDs

Andrew K. Morse BS
https://orcid.org/0000-0002-5280-6829
Ella Hawes
https://orcid.org/0000-0003-0122-2193
Kunal R. Shetty MD
https://orcid.org/0000-0003-3445-9824

This work is distributed under the terms of the Creative Commons Attribution Non-Commercial 4.0 License (CC BY-NC 4.0).

https://doi.org/10.22037/orlfps.v10i1.47066

References

1. Chew HSJ, Achananuparp P. Perceptions and Needs of Artificial Intelligence in Health Care to Increase Adoption: Scoping Review. J Med Internet Res. 2022 Jan 14;24(1):e32939. doi: 10.2196/32939. PMID: 35029538; PMCID: PMC8800095. Link: https://pubmed.ncbi.nlm.nih.gov/35029538/

2. Park CW, Seo SW, Kang N, Ko B, Choi BW, Park CM, Chang DK, Kim H, Kim H, Lee H, Jang J, Ye JC, Jeon JH, Seo JB, Kim KJ, Jung KH, Kim N, Paek S, Shin SY, Yoo S, Choi YS, Kim Y, Yoon HJ. Artificial Intelligence in Health Care: Current Applications and Issues. J Korean Med Sci. 2020 Nov 2;35(42):e379. doi: 10.3346/jkms.2020.35.e379. Erratum in: J Korean Med Sci. 2020 Dec 14;35(48):e425. PMID: 33140591; PMCID: PMC7606883. Link: https://pubmed.ncbi.nlm.nih.gov/33140591/

3. Kung TH, Cheatham M, Medenilla A, Sillos C, De Leon L, Elepaño C, Madriaga M, Aggabao R, Diaz-Candido G, Maningo J, Tseng V. Performance of ChatGPT on USMLE: Potential for AI-assisted medical education using large language models. PLOS Digit Health. 2023 Feb 9;2(2):e0000198. doi: 10.1371/journal.pdig.0000198. PMID: 36812645; PMCID: PMC9931230. Link: https://pmc.ncbi.nlm.nih.gov/articles/PMC9931230/

4. Durairaj KK, Baker O, Bertossi D, Dayan S, Karimi K, Kim R, Most S, Robotti E, Rosengaus F. Artificial Intelligence Versus Expert Plastic Surgeon: Comparative Study Shows ChatGPT "Wins" Rhinoplasty Consultations: Should We Be Worried? Facial Plast Surg Aesthet Med. 2023 Nov 20. doi: 10.1089/fpsam.2023.0224. Epub ahead of print. PMID: 37982677. Link: https://www.liebertpub.com/doi/10.1089/fpsam.2023.0224

5. Seth I, Cox A, Xie Y, Bulloch G, Hunter-Smith DJ, Rozen WM, Ross RJ. Evaluating Chatbot Efficacy for Answering Frequently Asked Questions in Plastic Surgery: A ChatGPT Case Study Focused on Breast Augmentation. Aesthet Surg J. 2023 Sep 14;43(10):1126-1135. doi: 10.1093/asj/sjad140. PMID: 37158147. Link: https://pubmed.ncbi.nlm.nih.gov/37158147/

6. Xie Y, Seth I, Hunter-Smith DJ, Rozen WM, Ross R, Lee M. Aesthetic Surgery Advice and Counseling from Artificial Intelligence: A Rhinoplasty Consultation with ChatGPT. Aesthetic Plast Surg. 2023 Oct;47(5):1985-1993. doi: 10.1007/s00266-023-03338-7. Epub 2023 Apr 24. PMID: 37095384; PMCID: PMC10581928. Link: https://pubmed.ncbi.nlm.nih.gov/37095384/

7. Ayers JW, Poliak A, Dredze M, Leas EC, Zhu Z, Kelley JB, Faix DJ, Goodman AM, Longhurst CA, Hogarth M, Smith DM. Comparing Physician and Artificial Intelligence Chatbot Responses to Patient Questions Posted to a Public Social Media Forum. JAMA Intern Med. 2023 Jun 1;183(6):589-596. doi: 10.1001/jamainternmed.2023.1838. PMID: 37115527; PMCID: PMC10148230. Link: https://pubmed.ncbi.nlm.nih.gov/37115527/

8. Shanafelt TD, West CP, Dyrbye LN, Trockel M, Tutty M, Wang H, Carlasare LE, Sinsky C. Changes in Burnout and Satisfaction With Work-Life Integration in Physicians During the First 2 Years of the COVID-19 Pandemic. Mayo Clin Proc. 2022 Dec;97(12):2248-2258. doi: 10.1016/j.mayocp.2022.09.002. Epub 2022 Sep 14. PMID: 36229269; PMCID: PMC9472795. Link: https://pubmed.ncbi.nlm.nih.gov/36229269/

9. Walsh S, O'Neill A, Hannigan A, Harmon D. Patient-rated physician empathy and patient satisfaction during pain clinic consultations. Ir J Med Sci. 2019 Nov;188(4):1379-1384. doi: 10.1007/s11845-019-01999-5. Epub 2019 Mar 27. PMID: 30919198. Link: https://pubmed.ncbi.nlm.nih.gov/30919198/

10. Bhattacharyya M, Miller VM, Bhattacharyya D, Miller LE. High Rates of Fabricated and Inaccurate References in ChatGPT-Generated Medical Content. Cureus. 2023 May 19;15(5):e39238. doi: 10.7759/cureus.39238. PMID: 37337480; PMCID: PMC10277170. Link: https://pmc.ncbi.nlm.nih.gov/articles/PMC10277170/

11. B urgess DJ, Warren J, Phelan S, Dovidio J, van Ryn M. Stereotype threat and health disparities: what medical educators and future physicians need to know. J Gen Intern Med. 2010 May;25 Suppl 2(Suppl 2):S169-77. doi: 10.1007/s11606-009-1221-4. PMID: 20352514; PMCID: PMC2847106. Link: https://pubmed.ncbi.nlm.nih.gov/20352514/

This work is distributed under the terms of the Creative Commons Attribution Non-Commercial 4.0 License (CC BY-NC 4.0).