The Journal of Laryngology & Otology

cambridge.org/jlo

The utility of Chat Generative Pre-trained Transformer as a patient resource in paediatric otolaryngology

Walter M Jongbloed[1] and Nancy Grover[2]

[1]Department of Otolaryngology – Head & Neck Surgery, UConn Health, Farmington, CT, USA and [2]Connecticut Children's Department of Otolaryngology, Hartford, CT 06106, USA

Main Article

Nancy Grover takes responsibility for the integrity of the content of the paper

Presented at the 2023 Society of Ear, Nose and Throat Advancement in Children Annual Meeting, 1 December 2023, Charleston, SC, USA.

Cite this article: Jongbloed WM, Grover N. The utility of Chat Generative Pre-trained Transformer as a patient resource in paediatric otolaryngology. *J Laryngol Otol* 2024;**138**: 1115–1118. https://doi.org/10.1017/S0022215124001154

Received: 18 December 2023
Revised: 25 March 2024
Accepted: 9 April 2024
First published online: 4 December 2024

Keywords:
artificial intelligence;
patient education as topic;
obstructive sleep apnea

Corresponding author:
Walter M. Jongbloed;
Email:jongbloed@uchc.edu

Abstract

Objectives. This study aimed to determine the feasibility of using Chat Generative Pre-trained Transformer (ChatGPT) (https://chatgpt.com) as a patient resource for paediatric otolaryngology conditions and assess the quality of responses generated by ChatGPT when compared with information available on the internet.

Method. ChatGPT responses to common paediatric otolaryngology conditions were compared with top internet pages for readability (Flesch Reading Ease score, word count), expediency (time taken to generate response), validity (comparison of recommendations to the American Academy of Otolaryngology Head and Neck Surgery guidelines) and consistency (changes in recommendations based on question variation).

Results. ChatGPT was more expeditious in generating responses with fewer words, albeit with higher reading scores. When compared with accredited guidelines, there was no difference in validity between these sources (internet sources and ChatGPT). Consistent responses were obtained with question variation.

Conclusion. ChatGPT may be a valuable source for patients and families in providing valid information comparable to internet materials.

Introduction

Large language models, or chatbots, such as Chat Generative Pretrained Transformer (ChatGPT) use language processing to generate conversational responses to written inputs.[1] ChatGPT, a free online tool trained on millions of pages of data from across the internet with data current to September 2021, has made substantial inroads into the field of medicine, even proving its ability to pass the US Medical Licensing Examination.[2] It is the fastest growing consumer application to date, having reached over 100 million users by January 2023.[3]

Approximately 80 per cent of internet users search online for health information.[4] Although ChatGPT seems to have the potential to upend medical care, providing patients with more data than a simple online search in a language that a layman can understand, it has limitations: it can give users a different answer depending on input phrasing, it may write plausible-sounding but incorrect or nonsensical answers and it has been noted to perpetuate disparities and biases in race, sex and culture.[5–7] As medical providers, we should be up to date with the online tools available to our patients and be able to provide our opinion on the information available, and we anticipate patients may be seeking information on ChatGPT.

Responses generated by ChatGPT in response to clinical vignettes have also been compared with those of physicians in terms of diagnostic accuracy and treatment plans, specifically within the otolaryngology field. Physicians tended to highly agree with the differential diagnoses and treatment plans generated by ChatGPT.[8]

With the further integration of artificial intelligence (AI) into our society, especially via large language models such as ChatGPT, this study intended to determine its utility as a patient resource. Patient information and resources available to families can be overwhelming and daunting. A language model designed to respond in a conversational tone to any question promises great potential in simplifying the patient experience. Confirming that the recommendations and advice align with recommendations from accredited sources is necessary before endorsing this resource to patients and families.

Materials and methods

This study aimed to investigate the utility of ChatGPT as a patient resource. Four common paediatric otolaryngology conditions were studied: snoring, sleep apnoea, treatment of sleep apnoea and ear wax (cerumen) impaction. Two questions for each condition were entered into ChatGPT version 3.5, with the questions for each condition varied slightly to test for consistency. The responses generated by ChatGPT were then compared with the

top internet page recommended by an online search engine in the following domains: readability (Flesch Reading Ease score and word count), expediency (time taken to generate response), validity (comparison of recommendations to American Academy of Otolaryngology Head and Neck Surgery (AAO-HNS) guidelines) and consistency (changes in recommendations based on alterations in the question).

On 19 May 2023, inputs in the form of questions were entered into ChatGPT and an online search engine (Google) as shown in Table 1. Two independent otolaryngologists tested the validity of the responses against the AAO-HNS recommendations and inter-relater reliability was assessed using Cohen's kappa test. Descriptive statistics were summarised with means and standard deviations for continuous variables. The Mann–Whitney U test was used to assess differences between the responses generated by ChatGPT and the top webpage recommended by the search engine. The Mann–Whitney U test was designed to test whether there were significant differences in the distribution of a continuous variable (Flesch Reading Ease score, word count, time taken (seconds) to generate response and validity score of 0–3 based on comparison with recommendations from the AAO-HNS) between generated responses from ChatGPT and the top web pages recommended by the search engine. A p value of less than 0.05 was considered statistically significant.

This study was exempt from review by the Connecticut Children's Medical Center Institutional Review Board because it does not constitute human subject research.

Results and analysis

Outputs from ChatGPT and the top web page recommended by the search engine were obtained on 19 May 2023.

Readability was characterised by two measures, the Flesch Reading Ease score and word count. The mean Flesch Reading Ease score for ChatGPT was 44.9 (college level),

Table 1. Validity and consistency of responses

Questions	Top internet webpage for each topic	ChatGPT validity score user 1	ChatGPT validity score user 2	Internet validity score user 1	Internet validity score user 2
Snoring in children	https://www.luriechildrens.org/en/blog/snoring-in-children-toddlers-when-to-worry/				
– Is snoring in children worrisome?		3	2	3	3
– Should I be worried if my child snores?		2	2	3	3
Sleep apnoea in children	https://www.mayoclinic.org/diseases-conditions/pediatric-sleep-apnea/symptoms-causes/syc-20376196#:~:text=Pediatric%20obstructive%20sleep%20apnea%20is,or%20is%20blocked%20during%20sleep				
– What is sleep apnoea in children?		3	3	3	3
– Does my child have sleep apnoea?		3	3	3	3
Treatment of sleep apnoea	https://www.mayoclinic.org/diseases-conditions/pediatric-sleepapnea/diagnosis-treatment/drc-20376199				
– What is the treatment of sleep apnoea in children?		3	3	3	3
– How is sleep apnoea in children treated?		3	3	3	3
Ear wax in children	https://www.choa.org/parent-resources/caring-for-yourkid-at-home/ear-cleaning-in-kids#:~:text=Steam%20from%20the%20shower%20or,to%20remove%20any%20excess%20earwax				
– How should I clean my child's ear wax?		3	3	3	3
– What should I do if my child has ear wax?		3	3	3	3
Validity scores					
– Validity		$p = 0.23$			
– Consistency		$p = 0.43$			
– Inter-relater reliability (Cohen's kappa)		Moderate agreement (95.83%) (Cohen's kappa = 0.48)			

ChatGPT = Chat Generative Pre-trained Transformer

https://doi.org/10.1017/S0022215124001154 Published online by Cambridge University Press

Table 2. Readability of responses

	Mean score	Standard deviation	Grade level	Significance
Flesch Reading Ease score				
– ChatGPT	44.90	8.05	College	$p = 0.032$
– Internet source	57.55	10.46	10th- to 12th-grade or high school	
Word count				
– ChatGPT	344	37.70		$p = 0.00194$
– Internet source	728.75	235.4		

ChatGPT = Chat Generative Pre-trained Transformer

with a standard deviation of 8.05. The mean Flesch Reading Ease score for the internet-generated sources was 57.55 (10th- to 12th-grade or high school) with a standard deviation of 10.46. ChatGPT had a significantly more difficult Flesch Reading Ease score than the internet sources (Table 2). ChatGPT also generated significantly fewer words (Table 2).

Expediency was measured by time taken to generate a response for ChatGPT and the time taken to reach top internet search engine recommended webpage. ChatGPT was more expeditious in generating a response (Table 3).

Validity was measured by comparison of responses to guidelines from the AAO-HNS (Table 1). The guidelines were analysed for key components. Three key components were determined for each condition, with one point assigned for each component, such that a score of 3 suggested complete validity. Two independent otolaryngologists generated responses from ChatGPT for each condition and assigned validity scores for both the ChatGPT responses and the search engine recommended web pages.

For the topics 'snoring in children' and 'sleep apnoea in children', the following components were considered necessary for full validity: (1) an accurate definition of obstructive sleep apnoea, (2) an accurate list of symptoms and causes for concern, and (3) validated treatments and a recommendation to see a provider. For 'treatment of sleep apnoea', the following three components were deemed necessary for full validity: (1) an accurate explanation of surgical treatments, (2) an accurate explanation of medical treatments, and (3) a recommendation to see a provider. For 'ear wax in children', the following three components were considered necessary for full validity: (1) an accurate definition of ear wax and/or cerumen, (2) a recommendation to see a provider, and (3) a warning against home remedies.

The mean validity score for ChatGPT was 2.75, with a standard deviation of 0.45, and the mean validity score for the internet sources was 3, with a standard deviation of 0. There was no statistically significant difference between the validity of the responses ($p = 0.234$), meaning both sources provided valid responses. Inter-relater reliability was measured using Cohen's kappa test and moderate agreement (95.83 per cent) was found between the two resources (Cohen's kappa = 0.48), meaning that there was general agreement between users on the validity of the ChatGPT responses and the web pages.

To assess the consistency between the responses, the input into ChatGPT was varied slightly for each topic, as shown in Table 1. The validity scores were then compared between the initial question and the varied question. The mean validity score for the initial question was 2.875, with a standard deviation of 0.354, and the mean validity score for the varied question was 2.625, with a standard deviation of 0.518. There was no significant difference in the validity of responses generated for the slightly varied questions, indicating consistency in responses ($p = 0.430$).

Table 3. Time taken to generate response

Resource	Mean time (s)	Standard deviation (s)	Significance
ChatGPT	1.49	0.35	$p = 0.0009$
Internet source	6.13	0.94	

ChatGPT = Chat Generative Pre-trained Transformer

Discussion

This study examined the utility of ChatGPT as a resource for patients and their families. In comparison with recommendations from the AAO-HNS, ChatGPT responses demonstrated validity on a par with the top recommended webpages on the internet. The integration of large language models, such as ChatGPT, has elevated the role of AI in disseminating healthcare information. These findings support the notion that ChatGPT can serve as a reliable patient resource.

Not only did the ChatGPT responses compare favourably to internet material, but they also consistently aligned with accredited recommendations from the AAO-HNS. The use of ChatGPT as a patient resource is substantiated by existing literature, but it is not without limitations. While ChatGPT's post-operative instructions for specific procedures have been found to be equivalent to institutional recommendations, they were found to be less understandable and actionable.[9] Hence, it is crucial to emphasise that ChatGPT should not be used as a replacement for a physician's guidance.

ChatGPT holds promise as a source of information for patients, provided it is used judiciously. It has demonstrated its ability to exercise clinical judgment and offer medical diagnoses and treatment plans when presented with clinical vignettes incorporating medical jargon, relevant history, physical examinations and diagnostic findings.[8] These capabilities have been observed to yield highly accurate differential diagnoses and reasonable treatment plans.[8]

In terms of accessibility and timeliness, this study affirms that ChatGPT is an accessible and user-friendly platform. Its ability to generate concise yet valid responses is advantageous for patients and families. However, it is worth noting that the readability level of ChatGPT is significantly higher than that of the top recommended internet materials, potentially limiting its accessibility for individuals who have not pursued higher education. This presents a notable limitation.

https://doi.org/10.1017/S0022215124001154 Published online by Cambridge University Press

Although large language models such as ChatGPT have shown promise as a patient resource, there are limitations. It generates responses based on patterns learned from extensive datasets, which are only up to date as of September 2021 at this time. Consequently, there is a risk of ChatGPT providing outdated information or not reflecting the latest recommendations. It is important to emphasise that during this study, every response from ChatGPT recommended consulting a healthcare professional. Similar practices have been observed in other healthcare studies involving ChatGPT.[10]

It is evident from current literature that ChatGPT serves as a valuable tool in medicine but that it should not replace the expertise and clinical judgment of medical professionals. For example, the quality of responses from ChatGPT was inferior to that of a second-year resident in terms of both accuracy and completeness when responding to clinical questions and scenarios in the subspecialty of head and neck surgery.[10]

Ethical considerations also come into play. ChatGPT has demonstrated bias in previous studies, potentially perpetuating stereotypes and misinformation, and should be used with caution.[5–7] User privacy is also a concern, especially as the model incorporates prior questions into future responses and can process sensitive healthcare information, including personal details and medical records if entered by patients into the chatbox.[11]

As large language models and AI continue to evolve, particularly in the field of medicine, it becomes imperative to establish guidelines and quality control measures for AI-driven healthcare. One of the medicolegal implications that requires attention is accountability in cases where incorrect information or recommendations lead to patient harm.[12]

This study affirms that ChatGPT is a valid resource for patients, demonstrating comparability with the top internet-recommended sources and AAO-HNS guidelines in the ENT areas of snoring in children, sleep apnoea in children, treatment of sleep apnoea and earwax impaction. However, there are several limitations of this study. The investigation focused on only four highly specific topics within the field of paediatric otolaryngology, limiting the generalisability even within the field of ENT. Moreover, the questions posed to ChatGPT were straightforward, mirroring the types of questions patients and families are likely to ask.

- Chat Generative Pre-trained Transformer (ChatGPT) has been shown to be an effective patient resource
- ChatGPT delivers concise, quick and valid responses to commonly asked patient questions in paediatric ENT, but responses are generated at a higher reading level than that found in online resources
- ChatGPT is an accessible and user-friendly platform that can provide tailored responses to simple questions posed by patients and families

Future research should explore more complex, high-level inquiries to better assess validity, but for the purpose of this study, basic questions were chosen to test the utility of ChatGPT as a patient resource. The number of outputs which were analysed totalled 16, considerably limiting the power of this study. Further studies that analyse a larger collection of responses are needed to validate this resource.

Conclusion

This study represents one of the first efforts to assess the validity of ChatGPT as a resource for patient information in otolaryngology. It highlights the potential of AI integration in healthcare to streamline information delivery and provide tailored, prompt responses to patients and families. While AI, such as ChatGPT, has yet to fully replicate the clinical expertise, judgment and skill of trained physicians, it is making significant strides in the field of medicine. This progress invites critical examination of ethical, medicolegal and scientific aspects of this resource.

Competing interests. None declared

References

1 Open AI. ChatGPT. In: https://chat.openai.com [17 October 2023]
2 Gilson A, Safranek CW, Huang T, Socrates V, Chi L, Taylor RA *et al.* How does ChatGPT perform on the United States Medical Licensing Examination? The implications of large language models for medical education and knowledge assessment. *JMIR Med Educ* 2023;**9**:e45312
3 Milmo, D. ChatGPT reaches 100 million users two months after launch. *The Guardian*, 2 February 2013. In: https://www.theguardian.com/technology/2023/feb/02/chatgpt-100-million-users-open-ai-fastest-growing-app#:~:text=ChatGPT%2C%20the%20popular%20artificial%20intelligence,analysis%20by%20data%20firm%20Similarweb
4 Yee D, Modiri O, Shi VY, Hsiao JL. Readability, quality, and timeliness of online health resources for rosacea. *Int J Dermatol* 2021;**60**(3):e90–2
5 Li H, Moon JT, Purkayastha S, Celi LA, Trivedi H, Gichoya JW. Ethics of large language models in medicine and medical research. *Lancet Digital Health* 2023;**5**(6):e333–5
6 Lucy L, Bamman D. Gender and representation bias in GPT-3 generated stories. In: *Proceedings of the Third Workshop on Narrative Understanding.* : Association for Computational Linguistics, 2021;48–55
7 Abid A, Farooqi M, Zou J. Large language models associate Muslims with violence. *Nat Mach Intell* 2021;**3**:461–3
8 Qu RW, Qureshi U, Petersen G, Lee SC. Diagnostic and management applications of ChatGPT in structured otolaryngology clinical scenarios. *OTO Open* 2023;**7**(3):e67
9 Ayoub NF, Lee YJ, Grimm D, Balakrishnan K. Comparison between ChatGPT and Google search as sources of postoperative patient instructions. *JAMA Otolaryngol Head Neck Surg* 2023;**149**:556–8
10 Vaira LA, Lechien JR, Abbate V, Allevi F, Audino G, Beltramini GA *et al.* Accuracy of ChatGPT-generated information on head and neck and oromaxillofacial surgery: a multicenter collaborative analysis. *Otolaryngol Head Neck Surg* 2024;**170**(6):1492–503
11 Zhang J, Zhang Z. Ethics and governance of trustworthy medical artificial intelligence. *BMC Med Inform Decis Mak* 2023;**23**:7
12 Dave T, Athaluri SA, Singh S. ChatGPT in medicine: an overview of its applications, advantages, limitations, future prospects, and ethical considerations. *Front Artif Intell* 2023;**6**:1169595

https://doi.org/10.1017/S0022215124001154 Published online by Cambridge University Press