ChatGPT: is it able to guide parents and caregivers on paediatric tonsillectomy-related questions?

Original Article

Authors

Maria José Lucas dos Santos
Unidade Local de Saúde Santa Maria, Lisboa, Portugal

Tomás Carvalho
Unidade Local de Saúde Santa Maria, Lisboa, Portugal

Tiago Fuzeta Eça
Unidade Local de Saúde Santa Maria, Lisboa, Portugal

Leonel Luís
Unidade Local de Saúde Santa Maria, Lisboa, Portugal

Correspondence:
Maria José Lucas dos Santos
mjlucasdossantos@gmail.com

Article received on April 25, 2024.
Accepted for publication on June 1, 2024.

Abstract

Aim: To assess the accuracy of ChatGPT in answering parents and caregivers' questions about pediatric tonsillectomy, comparing its versions 3.5 and 4.0.
Study Design: Instrument validation study of ChatGPT in answering questions about pediatric tonsillectomy.
Material & Methods: We prompted ChatGPT versions 3.5 and 4.0 with 21 questions. The answers were assessed in accordance with the latest American Academy of Otolaryngology's guideline on Tonsillectomy in Children. The assessment of the responses generated by the two versions were compared using McNemar's test (exact version).
Results: Of the 21 ChatGPT-generated answers, 13 (61.9%) were deemed accurate using version 3.5, and 19 (90.5%) using version 4.0 (p=0.031). The inter-rater agreement was very good - Cohen's Kappa=0.97 (v3.5) and 0.83 (v4.0).
Conclusion: ChatGPT version 3.5 should not be regarded as a sufficiently accurate tool for guiding caregivers on pediatric tonsillectomy-related questions. Version 4.0 seems to be a significantly more reliable tool.

Keywords: ChatGPT; OpenAI; Artificial Intelligence; Pediatric tonsillectomy; Parents and caregivers' questions

Introduction

Chat Generative Pre-Trained Transformer (ChatGPT), developed by OpenAI, is a language model based on Artificial Intelligence (AI). This tool can generate coherent text from prompts inserted by users. It is a chatbot optimized for dialogue and mimics human conversation.[1,2] Its latest versions of 3.5 (v3.5) and 4.0 (v4.0) were released in November 2022 and March 2023, respectively. So far, v4.0 requires a monthly subscription, while v3.5 is free.

This tool has become increasingly popular since its launch, and its potential applications

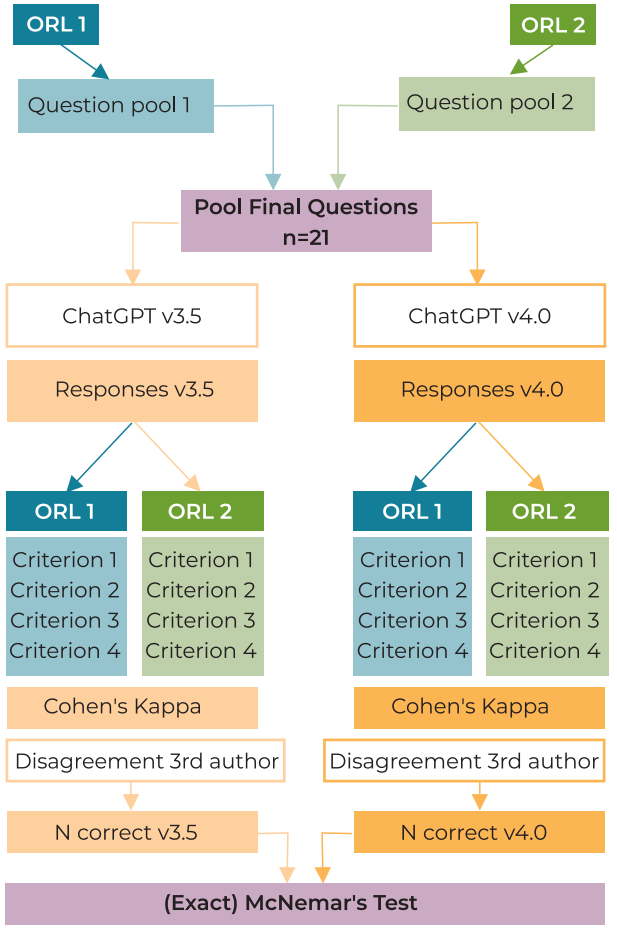
in numerous fields, including health, have been widely discussed. Its easy access allows the general population to search for information on medical topics, ask health-related questions, and obtain concise and understandable responses. However, it is crucial to determine if the responses generated by ChatGPT are correct and valid.[3-5] The growing interest in understanding its implications in medicine has led to a substantial number of scientific publications on ChatGPT. Several papers aiming to assess the software's performance in answering questions and clinical cases[6-9] related to the otorhinolaryngology (ORL) field, or its contribution to clinical research have been published.[10-13] Other authors have explored the role of this tool in informing patients about ORL pathologies or surgeries, with promising results.[14-19] Tonsillectomy is one of the most frequently performed surgeries worldwide, with the majority being performed in pediatric patients.[20] Therefore, several parents and caregivers are expected to use information available online, including tools such as ChatGPT, to answer their questions regarding the procedure, indications, and post-operative care. The aim of this study was to assess the accuracy and validity of ChatGPT responses to questions from parents and caregivers about pediatric tonsillectomy, according to the best scientific evidence available. Versions 3.5 and 4.0 of the software were also compared.

Materials and Methods

Two otorhinolaryngologists independently developed questions from the perspective of parents and caregivers, based on the latest version of the pediatric tonsillectomy guidelines published by the American Academy of Otolaryngology (AAO).[20] After discussion, a final set of 21 questions was selected. Then, they were entered into ChatGPT v3.5 and v4.0. The questions were formulated and entered into the software in English. No prompts were incorporated to restrict the references used by the tool or to direct the response according to the user's characteristics.

We created two new ChatGPT accounts, one for each version assessed. Two authors independently evaluated the two sets of responses, according to the AAO guidelines. Each question was assessed according to four criteria: conformity to the guidelines; citation or reference to the guidelines; indication to discuss the content of the response with the attending physician, and clarity of the response. Disagreements between the evaluators were resolved by a third author. The degree of agreement between the evaluators was analyzed using Cohen's Kappa test. Responses from v3.5 and v4.0 were assessed and compared using the McNemar test (exact version). Statistical analysis was performed using IBM SPSS Statistics v.29 software. The results were considered statistically significant for p-values ≤ 0.05. Figure 1 shows the flowchart of the study.

Figure 1
Study flowchart



Results

From the final set of 21 questions chosen by the authors to be entered into the software, the first group (questions 1 to 7) addressed topics related to surgical indications, and the second group (questions 8 to 21) focused on aspects related to the procedure. The second group was further subdivided into questions related to the operative results (question 8), risks of the procedure (questions 15 to 18), and post-operative care (questions 9 to 14, 20, and 21). Among the 21 responses obtained from ChatGPT, 13 (61.9%) were considered correct in v3.5 and 19 (90.5%) in v4.0. The difference was considered statistically significant ($p = 0.031$) using the McNemar test. The remaining answers were deemed incorrect or incomplete. (Table 1). The full list of responses generated by ChatGPT in shown in the Supplementary Table. ChatGPT v4.0 generated two (9.5%) responses that were considered incomplete or incorrect by the authors, both referring to aspects related to the surgical procedure: question 16 regarding the risk of post-tonsillectomy hemorrhage (PAH); and question 20 on the need for dietary restriction in the post-operative period. Conversely, eight (38.1%) of the answers given by v3.5 were The answers to these questions were evasive and not very specific, holding the doctor responsible for the decision and failing to refer users to the content of the AAO guidelines.

None of the responses from v3.5 mentioned or referenced the AAO guidelines and only two from v4.0 directly mentioned them, which were the responses to questions 1 and 2, both on the indication for tonsillectomy in children with frequent tonsillitis. Most responses mentioned, when relevant, the importance of discussing the information presented with the attending physician (88.2% in v3.5; 100% in v4.0). All responses generated by both ChatGPT versions were considered clear and presented in a language that was appropriate for parents and caregivers (100% for both v3.5 and v4.0). The degree of agreement between the evaluators was very good (Cohen's Kappa = 0.97 and 0.83 for v3.5 and v4.0, respectively).

Discussion

AI is developing rapidly, with new technologies and tools emerging daily to facilitate and expedite various everyday processes across numerous fields. Since its launch in November 2022, ChatGPT has become increasingly popular, reaching more than 180 million users worldwide. It can be valuable in the field of medicine, as it is an accessible, fast, and potentially free tool for users to ask questions about their health. ChatGPT is therefore expected to change the medical profession and doctor-patient relationship. Consequently, it is crucial to assess the quality of the information it provides, ensuring that it is reliable and aligns with the latest evidence.

The present study aimed to determine whether ChatGPT versions 3.5 and 4.0 could correctly inform parents and caregivers who asked questions about pediatric tonsillectomy. Accordingly, the responses generated by the tool were compared with the latest AAO guidelines, one of the most solid and consensual documents on this topic in the world. After evaluation by two otorhinolaryngologists, only 61.9% of the responses generated by ChatGPT v3.5 were considered accurate, according to the content of the AAO guidelines. This percentage increased to 90.5% with v4.0. Thus, the results seem to suggest that the paid version of the software (v4.0) generates significantly more reliable information than its free and widely-used version (v3.5). These findings agree with those of other studies previously published in the fields of ORL and Ophthalmology.[12, 13, 21]

When comparing the performance of the two versions of ChatGPT, v3.5 responded incompletely or incorrectly to five of the seven questions on tonsillectomy indications. This is because the responses of this version of the tool were clearly more evasive, did not commit to specific guidelines, and referred the answer to the physician. Conversely, v4.0 correctly responded to all questions on the surgical indications, and even referred to the text of the AAO guidelines when responding to Questions 1 and 2. The performance of both the versions

Table 1
Questions asked to v3.5 and v4.0 of ChatGPT and assessment of answers

Question entered into ChatGPT	Guideline	Version	Assessment of the response generated by ChatGPT				
			#1 Conformity with the guidelines	#2 Citation or reference to the guidelines	#3 Recommends clarifying with the attending physician	#4 Clear response	Agreement between the evaluators
#1 My child has a history of 3 throat infections in each of the past 2 years. Should an otolaryngologist refer him for tonsillectomy?	KAS 1	3.5			✓	✓	4/4
		4.0	✓	✓	✓	✓	4/4
#2 My child has a history of 6 documented tonsillitis in each of the past 2 years. Should an otolaryngologist refer him for tonsillectomy?	KAS 2	3.5			✓	✓	4/4
		4.0	✓	✓	✓	✓	4/4
#3 What are the criteria for defining a documented tonsillitis?	KAS 2	3.5	✓			✓	3/4
		4.0	✓		✓	✓	3/4
#4 Should physicians assess a child with recurrent throat infections for modifying factors that would favor tonsillectomy?	KAS 3	3.5			N/a	✓	4/4
		4.0	✓		✓	✓	4/4
#5 Should physicians assess children with obstructive sleep-disordered breathing for comorbid conditions that may improve with tonsillectomy?	KAS 4	3.5			N/a	✓	4/4
		4.0	✓		✓	✓	4/4
#6 When should a clinician refer a child for polysomnography, before performing tonsillectomy?	KAS 5/6	3.5				✓	4/4
		4.0	✓		✓	✓	3/4
#7 My child has obstructive sleep apnea documented by polysomnography. Should an otolaryngologist refer him for tonsillectomy?	KAS 7	3.5	✓		✓	✓	4/4
		4.0	✓		✓	✓	4/4
#8 Will children with obstructive sleep-disordered breathing be cured after tonsillectomy?	KAS 8	3.5	✓		✓	✓	4/4
		4.0	✓		✓	✓	4/4
#9 Will my child be needing pain medication after tonsillectomy?	KAS 9	3.5	✓		✓	✓	4/4
		4.0	✓		✓	✓	3/4
#10 What should I do if I can not control my child's pain after tonsillectomy?	KAS 9	3.5	✓		✓	✓	4/4
		4.0	✓		✓	✓	4/4
#11 Will my child be needing antibiotics after tonsillectomy?	KAS 10	3.5			✓	✓	4/4
		4.0	✓		✓	✓	4/4
#12 What are the indications for inpatient monitoring children after tonsillectomy?	KAS 12	3.5	✓		✓	✓	4/4
		4.0	✓		✓	✓	4/4
#13 Which pain medications can be used to control children pain after tonsillectomy?	KAS 13	3.5	✓		✓	✓	4/4
		4.0	✓		✓	✓	3/4
#14 Is codeine an option for pain control after tonsillectomy in children?	KAS 14	3.5	✓		✓	✓	4/4
		4.0	✓		✓	✓	4/4
#15 What are the possible complications of tonsillectomy?	Introdução	3.5	✓		✓	✓	4/4
		4.0	✓		✓	✓	4/4
#16 What are the rates of post-tonsillectomy bleeding?	KAS 15	3.5			N/a	✓	4/4
		4.0			✓	✓	4/4
#17 What should I do if my child has a post-tonsillectomy bleeding?	KAS 15	3.5	✓		✓	✓	4/4
		4.0	✓		✓	✓	4/4
#18 What is obstructive sleep-disordered breathing?	Introdução	3.5	✓		N/a	✓	4/4
		4.0	✓		N/a	✓	4/4
#19 Does tonsillectomy negatively affect the immune function?	Introdução	3.5	✓		✓	✓	4/4
		4.0	✓		✓	✓	4/4
#20 Does my child need to restrict his diet after tonsillectomy?	Table 8	3.5			✓	✓	4/4
		4.0			✓	✓	4/4
#21 How long is the recovery after tonsillectomy?	Table 8	3.5	✓		✓	✓	4/4
			✓		✓	✓	4/4

Legend: (✓): criterion met; N/a: not applicable; KAS: Key action statement.

of the tool was similar when responding to questions about the surgical procedure (results, risks and recommendations): Three responses from v3.5 (11, 16, and 20) and two from v4.0 (16 and 20) were incorrect. This seemed to occur because the second set of questions assesses knowledge objectively explained in various sources, whereas questions in the first set require integration and interpretation of clinical pictures with the information in the literature before formulating a response.[17]

This study revealed that most responses given by both versions of ChatGPT recommend discussing the information provided with the attending physician/otorhinolaryngologist (88.2% in v3.5; 100% in v4.0). This makes it safer for parents and caregivers to use the tool and encourages them to refer to professionals who can integrate the information presented with the actual clinical condition. This trend follows the results of previous studies.[15,19]

The present study differs from previously published studies in the field of ORL because it evaluated both versions of ChatGPT as instruments for informing the population by comparing the information generated with the latest AAO guidelines on tonsillectomy. In fact, the responses were validated based on an objective, solid, and unique instrument, which brings together the best and latest evidence available on the subject. The aim was to minimize the possible biases introduced by other assessment methodologies based on expert opinions or unstructured bibliographic research. The results of this study should be interpreted carefully as it was performed with a limited sample of 21 questions about a single ORL surgical procedure. Furthermore, there was a lack of assessment of the clarity of ChatGPT responses by laypeople, as the responses were only assessed by physicians. ChatGPT also has limitations inherent to its operation which influenced the study design. Of particular note is the importance of the quality of wording of the questions to ensure more adequate responses. This may compromise the practical applicability of the results as parents and caregivers may be less detailed and specific in their questions, potentially leading to less useful or less clear responses. However, ChatGPT does not spontaneously provide bibliographic references that can be consulted to verify the origin of the information. Since the tool accesses a huge amount of scientific information with different degrees of robustness (textbooks, guidelines, and scientific articles vs. web pages without peer review), users must be able to consult the sources of information. It is also fundamental to program the software to prioritize the most reliable information. Therefore, further studies are necessary to formally validate ChatGPT as a patient education tool in ORL.

Conclusion

This study, although based on a limited sample, suggests that the free and widely disseminated version of ChatGPT (v3.5) cannot be considered a reliable source of medical information on tonsillectomy in children. Conversely, the paid version (v4.0) appears to be a significantly more reliable tool for informing parents and caregivers, with most responses adhering to the content of the AAO guidelines. Given the widespread use of ChatGPT in everyday life, further studies should be conducted to assess and validate this and other AI tools in medicine.

Conflict of Interests

The authors declare that they have no conflict of interest regarding this article.

Data Confidentiality

The authors declare that they followed the protocols of their work in publishing patient data.

Human and animal protection

The authors declare that the procedures followed are in accordance with the regulations established by the directors of the Commission for Clinical Research and Ethics and in accordance with the Declaration of Helsinki of the World Medical Association.

Privacy policy, informed consent and Ethics committee authorization

All the processed data were based in published reports that fulfilled privacy policy and ethical considerations.

Financial support

This work did not receive any grant contribution, funding or scholarship.

Scientific data availability

There are no publicly available datasets related to this work.

References

1. De Angelis L, Baglivo F, Arzilli G, Privitera GP, Ferragina P, Tozzi AE, et al. ChatGPT and the rise of large language models: the new AI-driven infodemic threat in public health. Front Public Health. 2023 Apr 25:11:1166120. doi: 10.3389/fpubh.2023.1166120.
2. Roumeliotis KI, Tselikas ND. ChatGPT and open-AI models: a preliminary review. Future Internet. 2023; 15(6):192. https://doi.org/10.3390/fi15060192
3. Carlbring P, Hadjistavropoulos , Kleiboer A, Andersson G. A new era in Internet interventions: the advent of Chat-GPT and AI-assisted therapist guidance. Internet Interv. 2023 Apr 11:32:100621. doi: 10.1016/j.invent.2023.100621.
4. Vaishya R, Misra A, Vaish A. ChatGPT: Is this version good for healthcare and research? Diabetes Metab Syndr. 2023 Apr;17(4):102744. doi: 10.1016/j.dsx.2023.102744.
5. Liévin V, Hother CE, Motzfeldt AG, Winther O.Can large language models reason about medical questions? Patterns (N Y). 2024 Mar 1;5(3):100943. doi: 10.1016/j.patter.2024.100943
6. Qu RW, Qureshi U, Petersen G, Lee SC. Diagnostic and management applications of ChatGPT in structured otolaryngology clinical scenarios. OTO Open. 2023 Aug 22;7(3):e67. doi: 10.1002/oto2.67
7. Hoch CC, Wollenberg B, Lüers JC, Knoedler S, Knoedler L, Frank K. et al. ChatGPT's quiz skills in different otolaryngology subspecialties: an analysis of 2576 single-choice and multiple-choice board certification preparation questions. Eur Arch Otorhinolaryngol. 2023 Sep;280(9):4271-4278. doi: 10.1007/s00405-023-08051-4.
8. Long C, Lowe K, Zhang J, Santos AD, Alanazi A, O'Brien D. et al. A novel evaluation model for assessing ChatGPT on otolaryngology-head and neck surgery certification examinations: performance study. JMIR Med Educ. 2024 Jan 16:10:e49970. doi: 10.2196/49970.
9. Radulesco T, Saibene AM, Michel J, Vaira LA, Lechien JR. ChatGPT-4 performance in rhinology: A clinical case series. Int Forum Allergy Rhinol. 2024 Jan 24. doi: 10.1002/alr.23323.
10. Lechien JR, Gorton A, Robertson J, Vaira LA. Is ChatGPT-4 accurate in proofread a manuscript in otolaryngology-head and neck surgery? Otolaryngol Head Neck Surg. 2023 Sep 17. doi: 10.1002/ohn.526.
11. Nachalon Y, Broer M, Nativ-Zeltzer N. Using ChatGPT to generate research ideas in dysphagia: a pilot study. Dysphagia. 2024 Jun;39(3):407-411. doi: 10.1007/s00455-023-10623-9.
12. Frosolini A, Franz L, Benedetti S, Vaira LA, de Filippis C, Gennaro P. et al. Assessing the accuracy of ChatGPT references in head and neck and ENT disciplines. Eur Arch Otorhinolaryngol. 2023 Nov;280(11):5129-5133. doi: 10.1007/s00405-023-08205-4.
13. Lechien JR, Briganti G, Vaira LA. Accuracy of ChatGPT-3.5 and -4 in providing scientific references in otolaryngology–head and neck surgery. Eur Arch Otorhinolaryngol. 2024 Apr;281(4):2159-2165. doi: 10.1007/s00405-023-08441-8.
14. 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 Jun 1;149(6):556-558. doi: 10.1001/jamaoto.2023.0704.
15. Moise A, Centomo-Bozzo A, Orishchak O, Alnoury MK, Daniel SJ. Can ChatGPT guide parents on tympanostomy tube insertion? Children (Basel). 2023 Sep 30;10(10):1634. doi: 10.3390/children10101634.
16. Campbell DJ, Estephan LE, Mastrolonardo EV, Amin DR, Huntley CT, Boon MS. Evaluating ChatGPT responses on obstructive sleep apnea for patient education. J Clin Sleep Med. 2023 Dec 1;19(12):1989-1995. doi: 10.5664/jcsm.10728.
17. Zalzal HG, Abraham A, Cheng J, Shah RK. Can ChatGPT help patients answer their otolaryngology questions? Laryngoscope Investig Otolaryngol. 2023 Dec 9;9(1):e1193. doi: 10.1002/lio2.1193.
18. Soto-Galindo GA, Capelleras M, Cruellas M, Apaydin F. Effectiveness of ChatGPT in identifying and accurately guiding patients in rhinoplasty complications. Facial Plast Surg. 2023 Dec 27. doi: 10.1055/a-2218-6984.
19. Langlie J, Kamrava B, Pasick LJ, Mei C, Hoffer ME. Artificial intelligence and ChatGPT: an otolaryngology patient's ally or foe? Am J Otolaryngol. 2024 May-Jun;45(3):104220. doi: 10.1016/j.amjoto.2024.104220.
20. Mitchell RB, Archer SM, Ishman SL, Rosenfeld RM, Coles S, Finestone SA. et al. Clinical practice guideline: tonsillectomy in children (update)- executive summary. Otolaryngol Head Neck Surg. 2019 Feb;160(2):187-205. doi: 10.1177/0194599818807917.
21. Taloni A, Borselli M, Scarsi V, Rossi C, Coco G, Scorcia V. et al. Comparative performance of humans versus GPT-4.0 and GPT-3.5 in the self-assessment program of American Academy of Ophthalmology. Sci Rep. 2023 Oct 29;13(1):18562. doi: 10.1038/s41598-023-45837-2.