NORTHWESTERN MEDICAL JOURNAL

Northwestern Med J. 2026;6(1):41-51
DOI: 10.54307/NWMJ.2026.182

RESEARCH ARTICLES

Evaluation of the accuracy of ChatGPT-generated information in the field of general audiology

Gökçe Saygı Uysal[1], Aykut Özdoğan[1], Zülküf Küçüktağ[1], Esma Altan[1]

[1]Department of Otorhinolaryngology-Head and Neck Surgery, Ministry of Health Etlik City Hospital, Ankara, Türkiye

Cite as: Saygı Uysal G, Özdoğan A, Küçüktağ Z, Altan E. Evaluation of the accuracy of ChatGPT-generated information in the field of general audiology. Northwestern Med J. 2026;6(1):41-51.

ABSTRACT

Aim: This study evaluates the accuracy and reliability of ChatGPT's responses to open-ended questions in otology and audiology, focusing on its potential use in training ear, nose, and throat (ENT) professionals. As artificial intelligence (AI) applications like ChatGPT become more accessible to healthcare professionals and the public, ensuring that the information provided is reliable, accurate, and reproducible is crucial, especially in the medical field.

Materials and Methods: In March 2024, 60 audiology-related questions, categorized as 'general audiology,' 'hearing,' and 'balance,' were posed twice using ChatGPT (version 4) on the same computer to assess reproducibility. The responses were recorded as the '1st' and '2nd' answers. Three ENT specialists independently evaluated the answers to ensure accuracy, with a third reviewer specializing in audiology assessing the agreement between the responses. Answers were categorized as 1 (completely correct), 2 (partially correct), 3 (mixed accuracy), or 4 (incorrect). Analyses were conducted separately for each subgroup.

Results: Statistically significant difference was found between the two responses in general audiology questions ($p = 0.008$) and across all responses collectively ($p = 0.002$), while no significant difference was observed in hearing and balance questions ($p > 0.05$). The second responses had higher accuracy rates, with 65%, 80%, and 70% accuracy for general audiology, hearing, and balance areas, respectively.

Conclusion: ChatGPT's second responses were more accurate and reliable, making it a valuable resource for clinicians despite occasional misleading answers. With continued advancements, AI is expected to become a more reliable tool in audiology.

Keywords: answer, artificial intelligence, audiology, Chat-GPT, head and neck surgery

Corresponding author: Gökçe Saygı Uysal **E-mail:** gcsaygi@gmail.com
Received: 04.03.2025 **Accepted:** 27.06.2025 **Published:** 31.01.2026

INTRODUCTION

Artificial intelligence applications have recently developed rapidly and have become available in the field of health and medical sciences as well as in many other fields. Since these applications can be accessed by health professionals and all segments of the society, it is thought that they may indirectly affect human health. Therefore, it is very important that the information provided by such applications is reliable, accurate and reproducible for its applicability in medical fields.

One of the most recent artificial intelligence applications is the Chat Generative Pre-trained Transformer (ChatGPT) program.

ChatGPT is a new artificial intelligence model designed to generate human-like conversational dialog and can generate answers to textual inputs/questions from a large database of information (websites, books and recent articles) (1-3). ChatGPT, which has a large language model trained by Open AI (Artificial Intelligence) on internet-sourced data and a very large information source, has a highly developed ability to answer questions and translate between languages, as well as the ability to create conversations on various topics. Thanks to this feature, it can be used in both medical and non-medical fields (1-4).

Since applications in the medical field and the information sources accessed require high responsibility and reliability, it is of great importance to develop an artificial intelligence model with accurate and reliable medical knowledge (5). There are studies evaluating the ChatGPT application in terms of medical exams, clinical evaluation and diagnosis, and article writing (6-9).

While some researchers consider the medical information provided by ChatGPT to be valuable, others have distanced themselves from this issue due to misuse during medical writing, security, plagiarism, inability to ensure the accuracy of the information and legal problems that may arise (10-12).

When the studies in the field of Ear, Nose and Throat Diseases (ENT) are reviewed; artificial intelligence studies have been reported in clinical grading systems, evaluation of cochlear implant function, clinical management of parathyroid gland diseases, prediction of clinical prognosis of some diseases, and determination of accuracy and reliability of information about head and neck cancers (4,6,13). In the field of audiology, there are a limited number of studies on ChatGPT (14,15).

Our aim in this study is to determine the accuracy and reliability of ChatGPT's answers to open-ended questions posed in the field of otology and audiology and subdivided into subgroups within the field and to determine whether this artificial intelligence application can be used in the training of ENT professionals.

MATERIAL AND METHOD

In March 2024, 60 questions related to the field of audiology were asked to ChatGPT (Chat GPT version 4) in 3 sub-headings.

The questions to be used in the study were categorized in 3 basic areas: 'general audiology', 'hearing' and 'balance'. The questions were adapted from general ENT and audiology reference books and questions used in board examinations. A total of 60 questions were prepared, 20 open-ended questions in English in each field.

An example question for each subheading area is provided below:

1. What is the concept of audiometric zero? Please explain.
2. Please explain the clinical signs and symptoms of vestibular neuritis.
3. What are the current treatment modalities for sudden hearing loss?

The questions were asked twice, one after the other, from the same computer, each time with a 'new chat' function to assess reproducibility. The answers were recorded as '1st and 2nd answer' (shown in the Supplementary File).

All answers from the ChatGPT were evaluated by 3 different, active, ENT specialists (reviewers) who were not in contact with each other during the evaluation phase. All questions were checked simultaneously by 3 different reviewers to exclude individual factors, to minimize the margin of error of the examiners and to ensure that the answers were evaluated accurately. The agreement between the 1st and 2nd answers was assessed by the 3rd reviewer (an expert with specific studies in audiology).

In order to ensure standardization in the evaluation of the answers, the answer categorization determined in a previous study by Kuşcu et al. was used (4). According to this scheme, answers were categorized as 1 (Completely correct), 2 (Partially correct), 3 (A mix of accurate and inaccurate/misleading), 4 (Completely incorrect/ irrelevant). All analyses were performed separately for all questions according to the subheadings in each of the three groups.

Statistical analysis

The analysis of the data included in the study was performed with SPSS (Statistical Program in Social Sciences) 27 program. Descriptive statistics were calculated as number, percentage, mean, standard deviation, median and min-max.

Inter-measurement consistencies and analyses were determined by intraclass correlation coefficient (ICC, intraclass correlation coefficient: 0.80-1.0 very high correlation) and Kruskall Wallis H Test and Wilcoxon Sign Test (p; statistical significance, $p<0.05$; there is a statistically significant difference between groups).

RESULTS

Harmonization between reviewers

When the responses received from ChatGPT were analyzed by three reviewers, the intra-group ICC was used to examine the consistency of the reviewers' decisions. For answer 1, the ICC is 0.994, while for answer 2, the ICC is 0.991, indicating that the raters are highly consistent in their decisions for both answers. (Table 1)

Table 1 presents the classification of responses to questions by three independent reviewers,. Both the first and second responses were evaluated separately, and inter-reviewer agreement was analyzed using p-values and intraclass correlation coefficients (ICC).

For the first responses, the proportion of responses deemed *completely correct* ranged from 65.0% to 66.7% across reviewers. *Partially correct* responses accounted for 26.7% to 28.3%, while *mixed* responses were minimal (1.7%), and *completely incorrect* responses were only 5.0%. The p-value was calculated as 0.918, indicating no statistically significant differences among reviewers' classifications. Furthermore, the ICC value of 0.994 demonstrates an exceptionally high level of inter-rater reliability for the first responses.

Table 1. Harmonization between reviewers

Questions			Completely correct, n%	Partially correct, n%	Mix of accurate - inaccurate/ misleading, n%	Completely incorrect/ irrelevant, n%	p	ICC
Total	1. Answer	1	39(65.0)	17(28.3)	1(%1.7)	3(5.0)	.918	.994
		2	40(66.7)	16(26.7)	1(%1.7)	3(5.0)		
		3	40(66.7)	16(26.7)	1(%1.7)	3(5.0)		
	2. Answer	1	43(71.7)	13(21.7)	4(6.7)		.980	.991
		2	43(71.7)	12(20.0)	5(8.3)			
		3	43(71.7)	12(20.0)	5(8.3)			

ICC: intraclass correlation coefficients.

Similarly, for the second responses, *completely correct* answers were consistently rated at 71.7% by all reviewers. *Partially correct* classifications ranged between 20.0% and 21.7%, while *mixed* responses varied slightly from 6.7% to 8.3%. Notably, *completely incorrect or irrelevant* responses were absent in this round. The p-value of 0.980 again indicates no statistically significant differences between reviewers, and the ICC value of 0.991 reflects a strong level of agreement.

Repeatability

The questions were asked twice to the Chat GPT and it was examined whether the 1st and 2nd answers were compatible with each other, in other words, the repeatability of the application. This evaluation was done separately for each question category. The Wilcoxon sign test was used to compare whether there was a statistically significant difference between the 1st and 2nd answers in all areas. The rates of agreement in the answers to the questions asked in each of the three fields were 65% (13 compatible, 7 incompatible answers), 90% (18 compatible, 2 incompatible) and 80% (16 compatible, 4 incompatible answers) for general audiology, hearing and balance fields, respectively.

When evaluated in terms of all categorized answers, a statistically significant difference was found between the 1st and 2nd answers (p=0.002). When analyzed according to the sub-headings, a statistically significant difference was found between both answers, especially in general audiology questions (p=0.008). No significant difference was found in the other two sub-headings (p>0.05) details are shown in table 2.

Table 2 illustrates the concordance between ChatGPT's first and second answers (harmony) and the accuracy rates of those answers according to expert evaluations (controller compliance). For each subheading (General Audiology, Hearing, and Balance), it reports both the number and percentage of concordant answers, the accuracy of each answer, and the p values for their comparison.

Concordance Rates (1st vs. 2nd answers):

- General Audiology: 65% concordance (13/20), p = 0.008 → Significant difference
- Hearing: 90% concordance (18/20), p = 0.157 → Not significant
- Balance: 80% concordance (16/20), p = 0.083 → Not significant
- Total: 78% concordance (47/60), p = 0.002 → Overall significant difference

Accuracy According to Expert Evaluations (1st & 2nd answers):

- The 1st answers were found to be 95–100% accurate across all areas.
- The 2nd answers showed similarly high accuracy rates (97–100%).

Table 2. Concordance analysis of 1st and 2nd answers from ChatGPT

Subject	Harmony	Controller Compliance	
	n(%)	1. Answer, n(%)	2. Answer, n(%)
General Audiology	13(%65)	20(%100)	19(%95)
P	.008*	1.00	.992
Hearing	18(%90)	20(%100)	20(%100)
P	.157	1.00	1.00
Balance	16(%80)	19(%95)	20(%100)
P	.083	.950	1.00
Total	47(%78)	59(%98)	58(%97)
P	.002*	.979	.998

- No significant differences were detected between these accuracy rates ($p > 0.9$).

Overall, ChatGPT's first and second answers exhibit a high level of concordance (78%). However, in the General Audiology domain there is a statistically significant difference between the two rounds of answers ($p = 0.008$), suggesting slightly lower repeatability in this area. Expert evaluations confirm that both sets of answers are largely accurate, underscoring the high quality of content. In conclusion, ChatGPT's responses are generally consistent and reliable, though some subdomains—especially General Audiology—may require closer attention to repeatability.

Evaluation of answers according to categorization - accuracy rates

All three reviewers evaluated 2 answers each as completely wrong (4) for the 1st answers in the general audiology domain, no completely wrong (4) evaluation was made for the 2nd answers. In the field of hearing, no completely wrong (4) assessment was made for answers 1 and 2. In the field of balance, 1 answer was evaluated as completely wrong (4) by all 3 controllers for the 1st answers (5%). In the 2nd answers, no completely wrong (4) assessment was made.

It was observed that the accuracy rates of the 2nd answers were higher for all subheadings. Based on the 2nd answers, the accuracy rates were 65%, 80% and 70% for general audiology, hearing and balance areas, respectively.

Details are shown in Table 3.

Commentary on Table 3:

- **General Audiology:**
  - In the 1st answers, 65% were completely correct (category 1) and 10% were completely incorrect (category 4).
  - In the 2nd answers, the completely correct rate remained at 65%, while completely incorrect responses dropped to 0%.
  - This suggests that critical errors present in the first round were corrected in the second.
  - ICC = 0.985 and p = 0.944 confirm high inter rater consistency and no significant change in category distribution.
- **Hearing:**
  - In the 1st answers, 75% were completely correct and none were completely incorrect.
  - In the 2nd answers, the completely correct rate rose to 80%, with 0% completely incorrect.
  - ICC = 1.00 and p = 1.00 indicate perfect agreement and no distributional differences between rounds.
- **Balance:**
  - In the 1st answers, 55% were completely correct and 5% completely incorrect.
  - In the 2nd answers, the completely correct rate increased to 70%, and completely incorrect responses were eliminated.
  - ICC = 0.992 and p = 0.950 again demonstrate very high agreement and stable category distributions.

The elimination of completely incorrect responses and the stable or improved rates of completely correct answers in the second round indicate that ChatGPT's likelihood of critical errors decreases on repeat questioning. The consistently high ICC values across all fields further underscore strong inter rater reliability. In sum, Table 3 shows that the second answers are at least as accurate—and often more accurate—than the first ones, with a significantly reduced error rate.

DISCUSSION

Artificial intelligence applications have started to be used in many fields such as technology, industry, software and health. The use of artificial intelligence applications in the field of health constitutes an area where not only health professionals who are trained and serve in this field, but also individuals from all segments of society can easily access information because it is easily accessible and can be concluded quickly. Fast

Table 3. Evaluation of ChatGPT's responses by category in each field

Questions			1 n(%)	2 n(%)	3 n(%)	4 n(%)	p	ICC
General Audiology	1. Answer	1	13(65.0)	5(25.0)		2(10.0)	1.00	1.00
		2	13(40.0)	5(26.7)		2(13.3)		
		3	13(40.0)	5(26.7)		2(13.3)		
	2. Answer	1	13(65.0)	4(20.0)	3(15.0)		.944	.985
		2	13(65.0)	3(15.0)	4(20.0)			
		3	13(65.0)	3(15.0)	4(20.0)			
Hearing	1. Answer	1	15(75.0)	5(25.0)			1.00	1.00
		2	15(75.0)	5(25.0)				
		3	15(75.0)	5(25.0)				
	2. Answer	1	16(80.0)	3(15.0)	1(5.0)		1.00	1.00
		2	16(80.0)	3(15.0)	1(5.0)			
		3	16(80.0)	3(15.0)	1(5.0)			
Balance	1. Answer	1	11(55.0)	7(35.0)	1(5.0)	1(5.0)	.950	.992
		2	12(60.0)	6(30.0)	1(5.0)	1(5.0)		
		3	12(60.0)	6(30.0)	1(5.0)	1(5.0)		
	2. Answer	1	14(70.0)	6(30.0)			1.00	1.00
		2	14(70.0)	6(30.0)				
		3	14(70.0)	6(30.0)				

ICC: Intraclass Correlation Coefficient, p; Kruskall Wallis H Test p value.

and easy access to this summary information may also bring along misguidance of patients / patient relatives and possible ethical problems.

In the literature, there are studies on whether ChatGPT application can be a source of information for clinicians in different medical fields, health professionals in the learning process and patients / patient relatives, and in the study of Ayoub et al. It was noted that ChatGPT may contradict basic knowledge when giving medical advice, and this may create problems in terms of patient safety (16).

In the field of ENT, there are studies on the applicability of Chat GPT in different areas such as evaluation of clinical prognosis, staging of the disease, and diagnosis (1-6,13,17 -20).

It has the potential to provide rapid access to topics related to their fields for professionals receiving ENT specialty training and to be a resource for exams. For this reason, the study by Park et al. demonstrated that the interest in ChatGPT has increased in ENT education as in other medical fields (21). In this study, the effectiveness of ChatGPT in supporting clinical decisions, patient education, assisting in research and literature review was investigated and it was concluded that ChatGPT provides important information in clinical applications, but it has disadvantages such as data reliability, inability to perform physical examination and inaccurate information (21).

In the study of Qu et al. (6), it was reported that ChatGPT was inadequate as a diagnostic tool, in the study of Brennan et al. (22) it was reported to be a good supplementary source in ENT education, but in the study of Hoch et al. (8) it was reported to have

low accuracy rates in multiple-choice questions in ENT board exams. As a result, it has been reported in various studies that Chat GPT as a diagnostic tool shows a lower accuracy rate in diagnosis and triage compared to clinicians (6,8,23,24).

In the literature, there are studies evaluating the accuracy of answers by posing questions to artificial intelligence, as well as studies comparing different search engines (24,25).

In a study on balance, benign paroxysmal positional vertigo (BPPV) patient education materials obtained from traditional search engines (Google) and ChatGPT were compared and as a result it was found that the information in ChatGPT was harder to read, of lower quality, and more difficult to understand compared to the information in Google searches (25).

In another study comparing chatbots, Bellinger et al. (26) asked ChatGPT, ChatGPT Plus, Google Bard and Microsoft Bing questions about six rhinology topics such as epistaxis, chronic sinusitis, sinus infections, allergic rhinitis, allergies and nasal polyps and evaluated the answers with criteria such as readability, quality, understandability and applicability. As a result of this study, Bard and Bing provided higher readability, while ChatGPT Plus stood out in terms of quality and accuracy. Both chatbots showed advantages in providing understandable and actionable information for patients (25,26).

There are few studies on hearing and more have evaluated whether the Chat GPT can be a suitable source of patient information in this area (14,15). For example, one study reported that it could be used as an aid to medical documentation in cases of eustachian tube dysfunction (15).

In an article discussing the future applications of chatbots in hearing health, the possible use of these tools by patients, clinicians and researchers was discussed, and from the perspective of patients, Swanepoel et al. stated that chatbots can be used for initial screening, making recommendations for interventions, patient education and support, but the accuracy of the information should be ensured (27).

Patel et al. evaluated the performance of ChatGPT 3.5 and 4.0 on ENT (Rhinology) Standardized Board Examination questions in comparison with residents. They stated that ChatGPT4, which is a higher version, performed much better and that artificial intelligence applications may be useful in ENT education (28).

Chiesa-Estomba et al. stated that ChatGPT can guide the clinician in decision-making processes (planning of cialoendoscopy) (7).

In their study, Sireci et al. stated that Chat-GPT can guide the choice of optimal treatment (29).

Ayoub et al. compared whether ChatGPT and Google Search engine can be a resource for the rules and instructions that patients and their relatives should follow after pediatric ENT surgeries. It was emphasized that ChatGPT was inferior to Google Search engine, but it could be advantageous for both patients and clinicians when alternative sources of information are limited (16).

In their study, Riestra-Ayora et al. (GPT-3.5) employed a question-based evaluation approach to assess the accuracy and reliability of responses generated by ChatGPT on rhinologic pathologies. Their findings suggest that ChatGPT can serve as a valuable information source for health professionals (30). Kuşçu et al. examined the accuracy and reproducibility of ChatGPT's answers to the questions asked in the field of head and neck cancers and stated that the answers were largely accurate and reproducible (4).

Workman et al. stated that, in general, the ChatGPT answered more than 80% of the questions correctly and could be a resource (31).

In our study, ChatGPT was asked detailed, open-ended, interpretative and specific knowledge-based questions in the field of audiology to investigate whether this application can help in the education process or exam preparations in the field of ENT and audiology.

In the evaluation of the responses, the response categorization determined in the study by Kuşcu et al. (4).

Another scale that can be used in such studies is the Likert scale (32). However, we preferred to use the categorization of Kusçu et al.

In the Likert scale, which is another classification similar to this classification, the responses are categorized as (1 = extremely unsatisfactory, 2 = unsatisfactory, 3 = neutral, 4 = satisfactory and 5 = extremely satisfactory) (32).

As in our study, it is thought that a new ChatGPT evaluation scale can be created by basing future studies on this or similar classifications. For example, (33-35) evaluated the accuracy and reproducibility of ChatGPT's responses to a series of questions about tinnitus. Three open-ended questions (divided into two groups, comprising basic and more comprehensive information) were posed again at three and six months. It was found that the majority of responses received at three months were of a higher quality, and no change was observed in the responses at six months compared to those at three months.

In our study, the evaluation of the accuracy and reproducibility of ChatGPT's answers to the questions asked in the field of audiology was investigated by asking the questions 2 times in a row. Similar to the study of Jedrzejczak WW et al. it was observed that the accuracy rates of the answers to the questions asked for the second time were higher. This result is consistent with the result we found in our study. The advantage of our study is that we have more questions (35).

Since the accuracy rate was higher in the second answers to the questions asked twice to the ChatGPT, it can be concluded that the questions should be asked repeatedly to the ChatGPT to reach the correct answer. It was observed that the correct answers were higher in hearing and balance than in general audiology. This may be due to data overload or it may be interpreted as inadequacy in terms of repeatability and reliability.

Chat-GPT repeatability for hearing and balance questions was found to be high (90% and 80%, respectively). In the study conducted by Kuşcu et al. (4), the repeatability was 94.1%, which is comparable to the results observed in the hearing and balance domain in our study. The reproducibility in the general audiology domain was found to be 65% in our study, which was lower than that reported by Kuşcu et al (4).

This result may support that more objective data can be obtained by grouping the questions even within the field and this can be considered as another positive aspect of our study.

Another strength of our study is that there is no previous study specific to the field of audiology and the number of questions is the highest in this field. However, the preference for open-ended questions instead of multiple-choice questions may be a disadvantage since it requires interpretation and knowledge. This may negatively affect the accuracy rates of the answers to the questions.

For example, in the general audiology category, the 13th question;

"What do you think about a patient who underwent surgery due to middle ear pathology in the right ear, showing lateralization of Weber test to the left and bilateral positive Rinne tests on postoperative examination?" was evaluated as "Completely incorrect/irrelevant (4)" for the 1st answer and "Mix of accurate - inaccurate/ misleading (3)" for the 2nd answer.

Open ended questions, which do not specify a single correct answer and require interpretation and creativity, expand the model's probabilistic generation space. This increases the likelihood that, rather than reproducing high probability patterns from its training data, the model will deviate from context and introduce incomplete, misleading, or fabricated details ("hallucinations"). Moreover, because open ended prompts can admit many valid approaches, the model's outputs are harder to verify rigorously and lack clear source grounding, further raising error risk. Finally, higher sampling temperature and similar hyperparameter settings amplify response diversity at the expense of consistency, thereby increasing the chance of incorrect or incoherent answers.

In the literature, studies with open-ended and case discussion questions were found. This makes the

results of our study supportable with the literatüre (31,36).

Riestra-Ayora et al. (30) used 65 questions (rhinology questions), Habib G. Zalzal et al. (37) used 30 questions (pediatric ENT questions to inform patients and their relatives) and Ziya Karimov et al. (36) used 25 questions (case presentation). In our study, a total of 60 open-ended questions, 20 questions in each of the 3 areas, were used.

In the study of Zalzal et al. (37), it was observed that while the rate of complete accuracy was 56.7% in the first answers to open-ended questions, it increased to 96.7% in the answers received for the second time, and at the same time, it was stated that asking the questions twice enabled reaching the correct answer, and this result is similar to our study.

Hoch et al. reported that 57% of 2576 questions (479 multiple-choice and 2097 single-choice) were answered correctly and that single-choice questions were associated with a significantly higher rate of correct answers than multiple-choice questions. It was also emphasized that accuracy rates were lower for questions in audiology (71% incorrect answers) compared to other fields (8).

In our study, we found that the accuracy rate for questions in the field of general audiology was lower than in other fields.

CONCLUSION

While providing access to basic information in a specialized field such as audiology, caution may be required as accurate information may be confused with errors that non-expert users may find difficult to recognize. This information can be used for educational purposes, under the control of experts in the field.

It is indisputable that artificial intelligence will become an increasingly reliable source with the developments in technology and the studies to be carried out especially in the field of audiology. In addition to being a potential source of information for the time being, further studies are needed for it to become a useful tool in the decision-making process of clinicians in diagnosis and treatment.

Limitations

This study focuses on a specific version of ChatGPT, called GPT-4, and the responses it provided during a specific period. It is important to note that the results may vary with model updates. Comparative studies with different AI models may also be conducted in the future.

The responses were categorized based on accuracy, but the study did not analyze critical factors such as incorrect diagnosis and treatment due to the potential clinical consequences of inaccurate information.

The questions posed were open-ended, resulting in lengthy answers, and the study had a limited number of questions for ease of evaluation. Future studies with more diverse questions could be beneficial.

It is also important to highlight that this study specifically provides guidance for healthcare professionals, and further research with guiding questions for patients and their relatives would contribute to the literature.

Ethical approval

In this study, since no patient or experimental animal data were used, ethical committee approval was not required.

Author contribution

Concept: GSU; Design: GSU, AÖ; Data Collection or Processing: GSU, AÖ, ZK, EA; Analysis or Interpretation: GSU, ZK; EA; Literature Search: GSU; Writing: GSU, AÖ. All authors reviewed the results and approved the final version of the article.

Source of funding

The authors declare the study received no funding.

Conflict of interest

The authors declare that there is no conflict of interest.

REFERENCES

1. Munoz-Zuluaga C, Zhao Z, Wang F, Greenblatt MB, Yang HS. Assessing the accuracy and clinical utility of ChatGPT in laboratory medicine. Clin Chem. 2023; 69(8): 939-40. [Crossref]
2. The Lancet Digital Health. ChatGPT: friend or foe? Lancet Digit Health. 2023; 5(3): e102. [Crossref]
3. Van Dis EAM, Bollen J, Zuidema W, van Rooij R, Bockting CL. ChatGPT: Five priorities for research. Nature. 2023; 614(7947): 224-6. [Crossref]
4. Kuşcu O, Pamuk AE, Sütay Süslü N, Hosal S. Is ChatGPT accurate and reliable in answering questions regarding head and neck cancer? Front Oncol. 2023; 13: 1256459. [Crossref]
5. Masters K. Artificial intelligence in medical education. Med Teach. 2019; 41(9): 976-80. [Crossref]
6. 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. [Crossref]
7. Chiesa-Estomba CM, Lechien JR, Vaira LA, et al. Exploring the potential of Chat-GPT as a supportive tool for sialendoscopy clinical decision making and patient information support. Eur Arch Otorhinolaryngol. 2024; 281(4): 2081-6. [Crossref]
8. Hoch CC, Wollenberg B, Lüers JC, 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; 280(9): 4271-8. [Crossref]
9. D'Amico RS, White TG, Shah HA, Langer DJ. I asked a ChatGPT to write an editorial about how we can incorporate chatbots into neurosurgical research and patient care.... Neurosurgery. 2023; 92(4): 663-4. [Crossref]
10. Amann J, Vayena E, Ormond KE, Frey D, Madai VI, Blasimme A. Expectations and attitudes towards medical artificial intelligence: A qualitative study in the field of stroke. PLoS One. 2023; 18(1): e0279088. [Crossref]
11. Else H. Abstracts written by ChatGPT fool scientists. Nature. 2023; 613(7944): 423. [Crossref]
12. Haupt CE, Marks M. AI-generated medical advice-GPT and beyond. JAMA. 2023; 329(16): 1349-50. [Crossref]
13. Knoedler L, Baecher H, Kauke-Navarro M, et al. Towards a reliable and rapid automated grading system in facial palsy patients: Facial palsy surgery meets computer science. J Clin Med. 2022; 11(17): 4998. [Crossref]
14. Nielsen JPS, von Buchwald C, Grønhøj C. Validity of the large language model ChatGPT (GPT4) as a patient information source in otolaryngology by a variety of doctors in a tertiary otorhinolaryngology department. Acta Otolaryngol. 2023; 143(9): 779-82. [Crossref]
15. Kim HY. A case report on ground-level alternobaric vertigo due to eustachian tube dysfunction with the assistance of conversational generative pre-trained transformer (ChatGPT). Cureus. 2023; 15(3): e36830. [Crossref]
16. Ayoub NF, Lee YJ, Grimm D, Divi V. Head-to-head comparison of ChatGPT versus Google search for medical knowledge acquisition. Otolaryngol Head Neck Surg. 2024; 170(6): 1484-91. [Crossref]
17. Crowson MG, Dixon P, Mahmood R, et al. Predicting postoperative cochlear implant performance using supervised machine learning. Otol Neurotol. 2020; 41(8): e1013-23. [Crossref]
18. Wang B, Zheng J, Yu JF, et al. Development of artificial intelligence for parathyroid recognition during endoscopic thyroid surgery. Laryngoscope. 2022; 132(12): 2516-23. [Crossref]
19. Lim SJ, Jeon ET, Baek N, et al. Prediction of hearing prognosis after intact canal wall mastoidectomy with tympanoplasty using artificial intelligence. Otolaryngol Head Neck Surg. 2023; 169(6): 1597-605. [Crossref]
20. Arambula AM, Bur AM. Ethical considerations in the advent of artificial intelligence in otolaryngology. Otolaryngol Head Neck Surg. 2020; 162(1): 38-9. [Crossref]
21. Park I, Joshi AS, Javan R. Potential role of ChatGPT in clinical otolaryngology explained by ChatGPT. Am J Otolaryngol. 2023; 44(4): 103873. [Crossref]
22. Brennan L, Balakumar R, Bennett W. The role of ChatGPT in enhancing ENT surgical training - a trainees' perspective. J Laryngol Otol. 2024; 138(5): 480-6. [Crossref]
23. Levine DM, Tuwani R, Kompa B, et al. The diagnostic and triage accuracy of the GPT-3 artificial intelligence model: An observational study. Lancet Digit Health. 2024; 6(8): e555-61. [Crossref]
24. Rao A, Kim J, Kamineni M, Pang M, Lie W, Succi MD. Evaluating ChatGPT as an adjunct for radiologic decision-making. medRxiv. 2023; 02.02.23285399. [Crossref]
25. Bellinger JR, De La Chapa JS, Kwak MW, Ramos GA, Morrison D, Kesser BW. BPPV information on Google versus AI (ChatGPT). Otolaryngol Head Neck Surg. 2024; 170(6): 1504-11. [Crossref]
26. Bellinger JR, Kwak MW, Ramos GA, Mella JS, Mattos JL. Quantitative comparison of chatbots on common rhinology pathologies. Laryngoscope. 2024; 134(10): 4225-31. [Crossref]
27. Swanepoel D, Manchaiah V, Wasmann JW. The rise of AI chatbots in hearing health care. The Hearing Journal. 2023; 76(04): 26,30,32. [Crossref]
28. Patel EA, Fleischer L, Filip P, et al. Comparative performance of ChatGPT 3.5 and GPT4 on rhinology standardized board examination questions. OTO Open. 2024; 8(2): e164. [Crossref]

29. Sireci F, Lorusso F, Immordino A, et al. ChatGPT as a new tool to select a biological for chronic rhino sinusitis with polyps, "Caution Advised" or "Distant Reality"? J Pers Med. 2024; 14(6): 563. **[Crossref]**

30. Riestra-Ayora J, Vaduva C, Esteban-Sánchez J, et al. ChatGPT as an information tool in rhinology. Can we trust each other today? Eur Arch Otorhinolaryngol. 2024; 281(6): 3253-9. **[Crossref]**

31. Workman AD, Rathi VK, Lerner DK, Palmer JN, Adappa ND, Cohen NA. Utility of a LangChain and OpenAI GPT-powered chatbot based on the international consensus statement on allergy and rhinology: Rhinosinusitis. Int Forum Allergy Rhinol. 2024; 14(6): 1101-9. **[Crossref]**

32. Goodman RS, Patrinely JR, Stone CA, et al. Accuracy and reliability of chatbot responses to physician questions. JAMA Netw Open. 2023; 6(10): e2336483. **[Crossref]**

33. Deiana G, Dettori M, Arghittu A, Azara A, Gabutti G, Castiglia P. Artificial intelligence and public health: Evaluating ChatGPT responses to vaccination myths and misconceptions. Vaccines (Basel). 2023; 11(7): 1217. **[Crossref]**

34. Patil NS, Huang RS, van der Pol CB, Larocque N. Comparative performance of ChatGPT and bard in a text-based radiology knowledge assessment. Can Assoc Radiol J. 2024; 75(2): 344-50. **[Crossref]**

35. Jedrzejczak WW, Skarzynski PH, Raj-Koziak D, Sanfins MD, Hatzopoulos S, Kochanek K. ChatGPT for tinnitus information and support: Response accuracy and retest after three and six months. Brain Sci. 2024; 14(5): 465. **[Crossref]**

36. Karimov Z, Allahverdiyev I, Agayarov OY, Demir D, Almuradova E. ChatGPT vs UpToDate: Comparative study of usefulness and reliability of Chatbot in common clinical presentations of otorhinolaryngology-head and neck surgery. Eur Arch Otorhinolaryngol. 2024; 281(4): 2145-51. **[Crossref]**

37. Zalzal HG, Cheng J, Shah RK. Evaluating the current ability of ChatGPT to assist in professional otolaryngology rducation. OTO Open. 2023; 7(4): e94. **[Crossref]**