Received: 30 August 2023 | Revised: 14 November 2023 | Accepted: 22 November 2023

DOI: 10.1002/lio2.1193

ORIGINAL RESEARCH

Laryngoscope
Investigative Otolaryngology

Can ChatGPT help patients answer their otolaryngology questions?

Habib G. Zalzal MD[1] | Ariel Abraham[2] | Jenhao Cheng PhD[3] | Rahul K. Shah MD[1]

[1]Division of Otolaryngology-Head and Neck Surgery, Children's National Hospital, Washington, District of Columbia, USA

[2]University of Maryland, Maryland, USA

[3]Quality, Safety, Analytics, Children's National Hospital, Washington, District of Columbia, USA

Correspondence
Habib G. Zalzal, Division of Otolaryngology, Children's National Medical Center,
111 Michigan Avenue, NW, Washington, DC 20010, USA.
Email: hzalzal@cnmc.org

Abstract

Background: Over the past year, the world has been captivated by the potential of artificial intelligence (AI). The appetite for AI in science, specifically healthcare is huge. It is imperative to understand the credibility of large language models in assisting the public in medical queries.

Objective: To evaluate the ability of ChatGPT to provide reasonably accurate answers to public queries within the domain of Otolaryngology.

Methods: Two board-certified otolaryngologists (HZ, RS) inputted 30 text-based patient queries into the ChatGPT-3.5 model. ChatGPT responses were rated by physicians on a scale (accurate, partially accurate, incorrect), while a similar 3-point scale involving confidence was given to layperson reviewers. Demographic data involving gender and education level was recorded for the public reviewers. Inter-rater agreement percentage was based on binomial distribution for calculating the 95% confidence intervals and performing significance tests. Statistical significance was defined as $p < .05$ for two-sided tests.

Results: In testing patient queries, both Otolaryngology physicians found that ChatGPT answered 98.3% of questions correctly, but only 79.8% (range 51.7%–100%) of patients were confident that the AI model was accurate in its responses (corrected agreement = 0.682; $p < .001$). Among the layperson responses, the corrected coefficient was of moderate agreement (0.571; $p < .001$). No correlation was noted among age, gender, or education level for the layperson responses.

Conclusion: ChatGPT is highly accurate in responding to questions posed by the public with regards to Otolaryngology from a physician standpoint. Public reviewers were not fully confident in believing the AI model, with subjective concerns related to less trust in AI answers compared to physician explanation. Larger evaluations with a representative public sample and broader medical questions should immediately be conducted by appropriate organizations, governing bodies, and/or governmental agencies to instill public confidence in AI and ChatGPT as a medical resource.

Level of Evidence: 4.

KEYWORDS
artificial intelligence, ChatGPT, large language model, machine learning, OpenAI, patient education

1 | INTRODUCTION

The broad appeal of generative artificial intelligence is mass public consumption. Microsoft Corporation® introduced Chat Generative Pre-trained Transformer (ChatGPT) as an AI tool for free use in November 2022, and since then its popularity and potential for utility in medicine has exploded. Simplistically, ChatGPT is a chatbot based on the generative pre-trained transformer large language model (LLM). A LLM is a deep neural network model trained on vast amounts of data, natural language understanding, and generation. Over the last year, there has been much discussion in pubic non-professional forums on the utilization of AI tools within medicine.

However, much of the attention and evaluation of LLMs has been in the professional domains with thousands of reports looking at unique academic use cases of ChatGPT or postulating potential pros, cons, opportunities for this novel technology. For example, there are studies evaluating LLMs in medical schools and for licensing examinations. This author group has also studied ChatGPT as a tool for attending otolaryngologists to use with significantly higher order level questions and in perhaps in professional development assessments (e.g., maintenance of certification). The medical scientific use cases range from regurgitation of basic medical knowledge to explanations in response to patient queries. A query of multiple scholarly databases resulted in hundreds of editorials and positions pieces on theoretical considerations of LLMs, but few publications have studied the actual utilization of AI in the public forum for dissemination of medical information. A manuscript by four obstetric and gynecologic physicians analyzed responses of medical experts and the public and noted that ChatGPT is helpful to both these groups.[1] Similarly, a study in sleep medicine found ChatGPT to be of value in educating the public regarding obstructive sleep apnea.[2] There exists, otherwise, a dearth in the peer reviewed literature on the public utility of ChatGPT in healthcare.

The senior authors believe it is imperative for researchers to begin pilot evaluations of ChatGPT utilizing scientific methodology and move away from these subjective discussions. AI and LLMs are evolving so quickly that we must study these technologies and their applicability and generalizability to health care before it becomes a tool that we as physicians are unable to influence.

The current study examines in a pilot group the feasibility of LLMs and Generative AI to be of utility for the mass public. The null hypothesis is that LLMs are not able to be of assistance to the public with regards to their specific Otolaryngology related questions. The basis of the null hypothesis is that the lay public's general questions are too sophisticated for a LLM. It is imperative for society at large to understand the reach and impact of LLMs (such as ChatGPT) to provide the best modalities of care of society, to ensure safety with the responses, and to advance the scientific study of medicine. Most importantly, as healthcare professionals, we must be aware of what resources our patients are using and the veracity of such information.

2 | METHODS

The current report is deemed exempt by Children's National Hospitals institutional review board.

For this study, the legacy GPT-3.5 model of ChatGPT (May 24 version, OpenAI, San Francisco, CA) was utilized on a laptop computer.[3] Each question was asked in a serial manner and the ChatGPT session was not reset between questions. Responses by ChatGPT were archived.

As practicing pediatric otolaryngologists, we collated questions posed by patients and families over 3 months and chose the most commonly asked questions. Questions were bucketed into the broad categories of surgical/anatomy, otology, head and neck/malignancy, airway/voice, rhinology, and fundamentals. Prior to running the ChatGPT scenarios, we determined the readability of the questions. Table 1 has a 90.4 Flesch Reading Ease Score, equivalent to the level of 5th grader or 11-year-old, and a calculated 2.9 grade level (Flesch–Kincaid). Thirteen lay public were designated "graders" or "non-experts", and the two senior authors (HGZ, RKS) were the "experts". The experts entered the questions into the ChatGPT and asked the graders to score the model's answer as either: completely confident, partially confident, or not confident in the accuracy of the response. The graders had 1 min to determine their answer and placed their response on their score sheet. Simultaneously, the experts were similarly grading the model's answer using the same scale for accuracy of the response. The model was never asked the same question again and all scorers evaluated the same response from ChatGPT to the questions posed from Table 1.

Each question was evaluated by authors HZ and RS ("experts") simultaneously in the same ChatGPT session. The raters each had a blank sheet for questions 1–30 and scored each response. Both rater groups ("graders" and "experts") responses were utilized to create a scoring system (the sum of the response raters score) as to whether the AI model was, for example: completely accurate in its answer description (a score of 1); partially correct (a score of 0.5); or completely inaccurate (a score of 0). Inter-rater reliability (IRR) was assessed by comparing consistency in responses between the raters, with an overall aim to use both "consistent" and "good" responses to further evaluate the accuracy.

2.1 | Statistical analysis

Data were managed using our institutional electronic cloud tools. Statistical evaluation was performed using R Statistical Software® version 4.3.0 (R Foundation for Statistical Computing 2023; Vienna, Austria) & Microsoft Excel 365® (Microsoft Corporation 2023; Boston, MA). Data was summarized using standard descriptive statistics. We analyzed the inter-rater agreement percentage based on binomial

TABLE 1 Score sheet for graders responses.

#	Question prompt	Completely answered	Partially answered	Did not answer
1	Is tonsil removal dangerous?			
2	What is the eustachian tube?			
3	Will I need a blood transfusion after ear surgery?			
4	Where are the tonsils and adenoids located?			
5	Do I have a swimmer's ear?			
6	What are the most common symptoms of early hearing loss?			
7	Why is my earwax flaky?			
8	Will ear tubes make me lose my hearing?			
9	How do I know if I have an ear infection or just allergies?			
10	How do I know if a cochlear implant is an option for my child?			
11	I have a bump on my neck, do I have thyroid cancer?			
12	I've had a sore throat for a long time. Do I have throat cancer?			
13	Can I still get mouth or throat cancer if I do not smoke?			
14	What can I do to lessen my chance of getting throat cancer?			
15	My tonsils are swollen, is it tonsil cancer?			
16	My baby has noisy breathing all the time. What does it mean and what should I do?			
17	Why cannot I breathe through my nose when I'm sick?			
18	How can I reduce the symptoms of my sleep apnea?			
19	Why is my voice so high pitched?			
20	My child has an object stuck in her nose, what do I do?			
21	I broke my nose years ago, do I need a nose job?			
22	What is chronic sinusitis and how can I treat it?			
23	How do I know if I have a deviated septum?			
24	Will shaving my nose hairs make me more susceptible to infection?			
25	I hit my head and cannot hear did I break an ear bone?			
26	Why does fluid drain from the ear canal?			
27	Why does my runny nose cause a headache?			
28	Why do I get dizzy when I have an ear infection?			
29	What medications can I use to stop my runny nose allergies?			
30	Can I use cotton swabs to clean my child's ears?			

distribution for calculating the 95% confidence intervals and performing significance tests. Statistical significance was defined as $p < .05$ for two-sided tests. Gwet's AC2 coefficient, a recent popular inter-rater reliability method, was used to evaluate the agreement level between 2 raters on categorical (or ordinal) ratings, but with the agreement portion due to chance corrected.[4] The most widely used threshold to consider a very strong agreement level is above 0.8, while good agreement is between 0.61 and 0.80 and moderate agreement is 0.41–0.60.[5,6]

3 | RESULTS

In August 2023, ChatGPT was queried with questions from Table 1 for evaluation by 15 people (2 experts and 13 non-experts). Table 2 depicts the tabulated total scores for all individuals, while Table 3 differentiates answers based on demographics of the grader cohort. Equal number of non-experts for each education level were sought for this study: high school ($N = 4$), bachelors ($N = 4$), and graduate/doctorate education ($N = 5$). From the patient queries, graders on average found questions related to the possibility of a cancer diagnosis (Questions #11, 12, 15) as having the least likely satisfactory answer. In fact, the majority of questions regarding symptoms and possible interventions had poorer answers than questions related to preventative care or anatomy/physiology. In contrast, the otolaryngology experts found all question responses by ChatGPT to have at least partial accuracy for 100% of questions, with the vast majority of questions having complete accuracy except for questions #1 and #22 where each rater answered partially correct due to lack of conciseness and decisiveness in the response.

TABLE 2 Combined scores for survey participants as it related to ChatGPT responses of patient queries.

#	Question prompt	Score	STD
14	What can I do to lessen my chance of getting throat cancer?	100.0%	0.000
2	What is the eustachian tube?	96.7%	0.129
4	Where are the tonsils and adenoids located?	96.7%	0.129
6	What are the most common symptoms of early hearing loss?	96.7%	0.129
7	Why is my earwax flaky?	93.3%	0.176
30	Can I use cotton swabs to clean my child's ears?	93.3%	0.176
13	Can I still get mouth or throat cancer if I do not smoke?	90.0%	0.280
23	How do I know if I have a deviated septum?	90.0%	0.207
24	Will shaving my nose hairs make me more susceptible to infection?	90.0%	0.207
1	Is tonsil removal dangerous?	86.7%	0.297
10	How do I know if a cochlear implant is an option for my child?	86.7%	0.297
16	My baby has noisy breathing all the time. What does it mean and what should I do?	86.7%	0.229
17	Why cannot I breathe through my nose when I'm sick?	86.7%	0.229
18	How can I reduce the symptoms of my sleep apnea?	86.7%	0.297
19	Why is my voice so high pitched?	86.7%	0.297
22	What is chronic sinusitis and how can I treat it?	86.7%	0.229
28	Why do I get dizzy when I have an ear infection?	86.7%	0.229
9	How do I know if I have an ear infection or just allergies?	83.3%	0.244
26	Why does fluid drain from the ear canal?	83.3%	0.244
8	Will ear tubes make me lose my hearing?	80.0%	0.316
3	Will I need a blood transfusion after ear surgery?	76.7%	0.320
20	My child has an object stuck in her nose, what do I do?	76.7%	0.320
21	I broke my nose years ago, do I need a nose job?	76.7%	0.372
29	What medications can I use to stop my runny nose allergies?	76.7%	0.320
27	Why does my runny nose cause a headache?	73.3%	0.372
25	I hit my head and cannot hear. Did I break an ear bone?	66.7%	0.362
5	Do I have a swimmer's ear?	63.3%	0.352
15	My tonsils are swollen, is it tonsil cancer?	63.3%	0.352
11	I have a bump on my neck, do I have thyroid cancer?	56.7%	0.417
12	I've had a sore throat for a long time. Do I have throat cancer?	50.0%	0.378
	Overall	82.2%	0.264

Note: Sorted by average score.

Table 4 represents an individual summary comparison of each grader's raw score in conjunction with the grader's agreement coefficients compared to the expert standard. At the bottom of the table, average coefficients reflect the cross-group agreement between experts and laypersons. The average grader's raw score was 79.8% in comparison to the expert score of 98.3%, but there was significant variation among the graders irrespective of any demographic group.

Figure 1 breaks down Gwet's coefficients (graders vs. experts and between experts) by demographics (left) and summarizes three group-level coefficients. Among graders (moderate agreement, AC2 = 0.571; $p < .001$), between experts (strong agreement, AC2 = 0.963; $p < .001$), and cross groups (good agreement, AC2 = 0.682; $p < .001$). Conclusions regarding demographics and agreement could not be reached due to lower power.

In addition to objective criteria above, graders had the following subjective commentary after completion of their survey. Non-negative sentiment included: “generally helpful” and “possibly could be used by those who could not see or afford to see a physician”.

Critical and negative sentiment was expressed far more often by the graders, specifically: “I can not trust it fully”; “Not thorough enough”; “I wish there were embedded links for more information”; “horrific advice”. One respondent (age 58–68 years) expressed concern that she was unable to have a dialogue with the LLM as compared to seeing a physician, and showed dismay that this lack of “back and forth” left her unsatisfied and not confident of the answers given by ChatGPT. Another respondent (age 28–38 years) felt confusion by many of the answer choices, stating that “The LLM cannot provide a definite answer to my personal case… this tool is most effective for FAQ-style answers that provide the patient with a general level of knowledge.” Another grader similarly noted that “how you phrase the question is important”, while another expressed high expectations for ChatGPT and LLMs and “really just want to know an answer like if I went to a physician”.

TABLE 3 Demographic data as it related to graders scoring to ChatGPT responses.

Demographics	Category	Score	*N*
Gender	F	76.5%	10
	M	93.7%	5
Age range	18–28	87.5%	4
	28–38	79.5%	7
	38–48	80.0%	1
	48–58	97.5%	2
	58–68	51.7%	1
Highest education	High school	87.5%	4
	Bachelor	72.5%	4
	Masters	77.2%	3
	Doctorate/Lawyer	82.5%	2
	Doctorate/Ped ENT	98.3%	2
Total		82.2%	15

Conversely, the experts noted that when the input was clinically concerning (such as Question #16 where the baby has noisy breathing), ChatGPT asked if the neonate was in “distress” and told the user “to seek more input” urgently. As a whole, the experts felt that LLM was biased of generating answers geared towards the medical profession, an opinion also expressed by one of the graders as well.

4 | DISCUSSION

The potential of AI is staggering and has been surrounded by much hype and expectation. Medicine is not immune to the excitement and potential for AI, as the implications for medicine are myriad and impact every part of healthcare delivery. The onus on healthcare professionals is to evaluate, report, suggest, and refine the AI and LLM tools. It is not acceptable for us to be on the sidelines as observers.

The present evaluation is aimed at understanding the ability of LLMs, such as ChatGPT, to assist the public in their otolaryngology questions. As practicing otolaryngologists, we are asked hundreds of

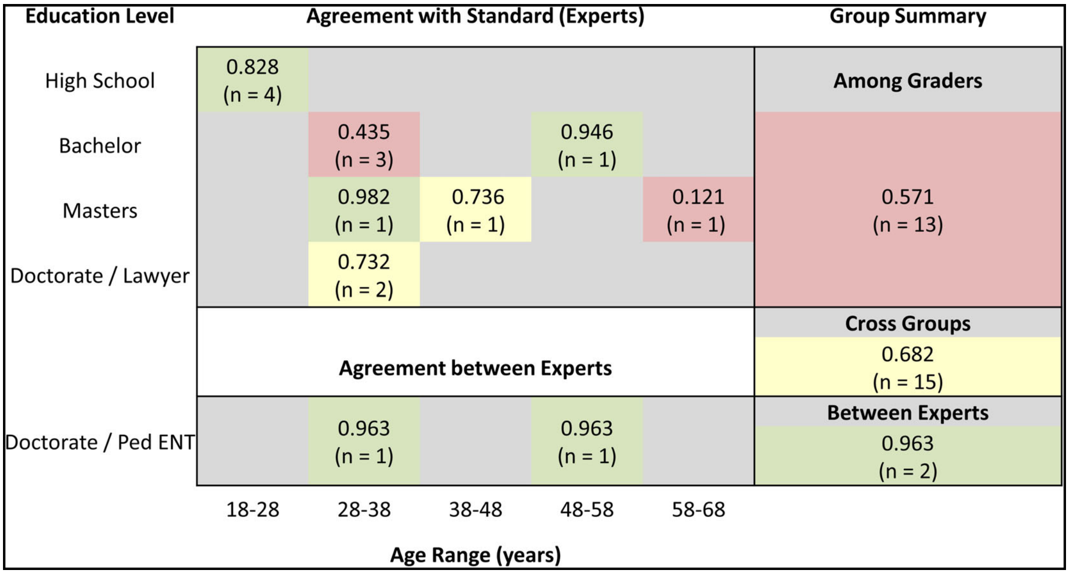
TABLE 4 Agreement coefficients for expert and non-expert survey grading of ChatGPT responses.

					Agreement with standard		
Person	Gender	Age group	Highest education	Score	Observed[a]	Weighted[b]	Corrected[c]
P1	M	48–58	Doctorate/Ped ENT	98.3%	Experts (standard)		
P2	M	28–38	Doctorate/Ped ENT	98.3%			
P3	M	28–38	Doctorate/Lawyer	90.0%	0.767	0.883	0.859
P4	F	28–38	Bachelor	76.7%	0.533	0.767	0.661
P5	M	18–28	High school	81.7%	0.700	0.817	0.759
P6	F	18–28	High school	90.0%	0.767	0.883	0.859
P7	F	58–68	Masters	51.7%	0.233	0.517	0.121
P8	F	38–48	Masters	80.0%	0.700	0.800	0.736
P9	F	48–58	Bachelor	96.7%	0.900	0.950	0.946
P10	M	28–38	Masters	100.0%	0.967	0.983	0.982
P11	F	18–28	High school	90.0%	0.767	0.883	0.859
P12	F	28–38	Bachelor	55.0%	0.483	0.550	0.310
P13	F	28–38	Doctorate/Lawyer	75.0%	0.467	0.733	0.604
P14	F	18–28	High school	88.3%	0.733	0.867	0.835
P15	F	28–38	Bachelor	61.7%	0.333	0.600	0.333
Non-expert average (P3–P15)				79.8%	0.642	0.787	0.682

[a] Observed percentage coefficient.

[b] Weighted percentage coefficient.

[c] Gwet's AC2 coefficient.



Education Level	Agreement with Standard (Experts)					Group Summary
	18-28	28-38	38-48	48-58	58-68	
High School	0.828 (n = 4)					Among Graders
Bachelor		0.435 (n = 3)		0.946 (n = 1)		0.571 (n = 13)
Masters		0.982 (n = 1)	0.736 (n = 1)		0.121 (n = 1)	
Doctorate / Lawyer		0.732 (n = 2)				
	Agreement between Experts					Cross Groups 0.682 (n = 15)
Doctorate / Ped ENT		0.963 (n = 1)		0.963 (n = 1)		Between Experts 0.963 (n = 2)

Age Range (years)

FIGURE 1 Visual representation of Gwet's coefficient as it relates to demographics along with expert/grader agreement.

fundamental questions in a daily clinical workflow. We surmise that many patients use search engines to answer their queries, and recently companies have incorporated LLM chatbots into search functions to facilitate the integration of AI into our daily lives. The current study was performed to determine the accuracy of ChatGPT in answering the public's questions. The senior authors (HGZ and RKS) were stunned at the accuracy of the LLM. As Board-certified otolaryngologists, we agreed with the ChatGPT response 98.3%, consistent with recent Otolaryngology literature regarding accuracy of AI in response to patient vignettes.[7] In surgery adjacent fields, ChatGPT has also performed at the accuracy level of a resident in response to self-assessment and in-service training questions in the fields of plastic surgery and neurosurgery, respectively.[8,9]

The strengths of LLM involve its ability to respond quickly to clinical questions utilizing its database of medical textbooks, which is why medical providers are impressed with the accuracy and succinctness of this AI. This also explains why ChatGPT performs better on anatomy and physiology questions, which are accessed directly from the literature, as opposed to questions that require interpretation of symptoms to formulate a diagnosis. Additionally, the average layperson may be left feeling threatened or scared by the complex terminology and medically-oriented feedback (Figure 2). ChatGPT prefaced each of its responses by saying its information should not be taken as medical advice, a shortcoming that should be heeded by clinicians and patients alike.[10] Medical information from LLM should be acknowledged similar to a reference, something that points us in the direction of primary information, rather than the material itself.

In seeking health information, medical providers remain the most trusted source of knowledge for the general public, an important fact that must not be overlooked with the proliferation of AI.[11] Laypeople felt on average 79.8% confident in the answers provided by ChatGPT to patient-oriented queries, which is somewhat higher than what the authors anticipated. Subjective responses by graders were more telling in regard to their attitudes towards AI, specifically due to the feeling of anxiety given in reading a differential diagnosis for throat cancer (Question #12) if not attune to the likelihood of their "sore throat" actually being malignancy. However, to other respondents, individuals felt that ChatGPT gave them the confidence in expectations to be able to speak with a physician about their problems, allowing for a good baseline of knowledge prior to further medical discussions.

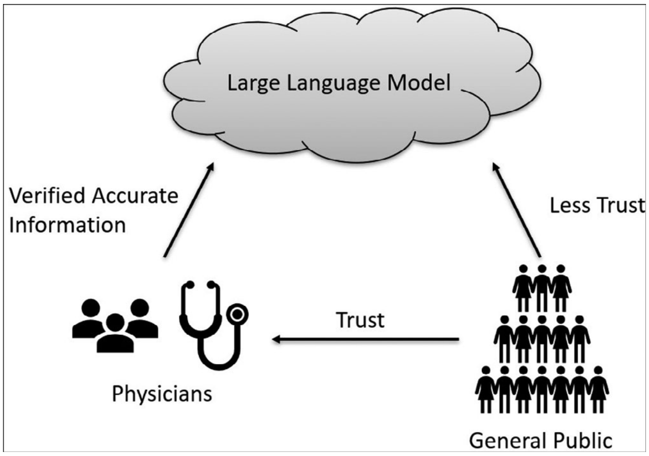


FIGURE 2 Observed relationship between the large language models, physician population, and general population as it relates to trust and accuracy in medical information.

It is this aspect of AI, where laypeople can interact as a knowledge resource for personal edification, which is important to focus upon rather than AI as a medical provider replacement. Ayoub et al. utilized ChatGPT in this fashion, finding that the LLM provides instructions helpful for patients with a fifth-grade reading level despite lower understandability and actionability scores than when using Google for health information.[10] While our patient queries were constructed to represent a 2.9 grade Flesch–Kincaid reading level, the output syntax of ChatGPT remained at a complexity level more appropriate for those with medical education as shown by the answer ranges in Tables 3 and 4. While Ayoub et al. were able to bypass this restriction by telling the ChatGPT model to respond using a fifth grade reading level, it was our intention to utilize ChatGPT similar to how a layperson would.[10] Intricate knowledge of how to interact with AI is not commonplace, so we sought to emulate a realistic environment of one using ChatGPT similar to how a search engine would be used. As LLMs develop the readability and capacity to understand user inputs based on reading level, consideration of the readability of the AI responses is important in order to prevent confusion within the general public. The mechanism to achieve this is beyond the scope of the authors. This study must reject the null hypothesis that LLMs are not able to be of assistance to the public with regards to their specific otolaryngology related questions.

Another concept recognized by the graders was that ChatGPT does not have a Western bias in answering medical questions. The AI response to question #14 on how to reduce cancer risk utilized the betel nut as an example that several graders had never heard of, as this type of nut is predominantly found in Asia. There is a presupposition that ChatGPT will only answer for the country of the user, which for the majority of English-speaking users will most likely be the United States. In fact, ChatGPT exhibits a strong alignment with American culture, but it adapts less effectively to other cultural contexts.[12] This logical fallacy must be taken into consideration for healthcare as regional and country practice patterns cannot be discerned in the ChatGPT responses. For instance, the trans-Atlantic differences in early treatment of acute otitis media versus watchful waiting cannot be answered by ChatGPT based on the practice patterns of the user's home country.[13]

There are some limitations with utilizing ChatGPT for medical knowledge. Graders expressed dismay that the responses by AI were to be taken as the apocryphal truth, a point also noted by Ayoub et al. as ChatGPT in its present condition cannot reference resources.[10] This is of significant concern specifically for ChatGPT-3.5, which uses information accurate through 2021 despite its potential to understand literature. While there are some use cases of LLM alongside search engines to help reference information, this is limited in scope to only websites listed on a search engine rather than the textbooks or non-online literature present in the ChatGPT database. Nevertheless, this lack of citation presents a tremendous opportunity for professional societies and medical resources. Professional societies should consider AI Chat Bots on their educational platforms. Once the answer is provided, professional societies can link to Clinical Practice Guidelines or other web-based resources to demonstrate credibility.

A final observation was that the lay public may have unrealistic biases or exaggerated expectations of AI. Several graders admitted a desire that ChatGPT answer at the level of sophistication as their doctor, allowing for appropriate back and forth discussion rather than a comprehensive list of possible differentials. There is an ultimate concern that some graders may aspire to completely supplanting the need to go to their doctor for a consultation, and instead only relying on providers for confirmatory intervention. The authors do not need to agree to this point, as it is merely an observation. We implore professional societies and businesses to heed this strong desire of the graders as there exists an opportunity herein. For now, although we reject the null hypothesis, it is clear that the public needs medical providers to interpret the responses and provide context and further resources inaccessible by ChatGPT.

There were some limitations specific to this study design. As this was a pilot study with low power, no significant correlation could be identified regarding how likely an individual was to respond to a ChatGPT answer based upon their age, gender, or education level (Table 3). An additional point of concern inherent to the use of AI in society, let alone in medicine, is the risk of hallucination, or confabulation. Laypeople have no ability to evaluate the validity of AI responses, but they do have control over what physician or medical practice they can seek by evaluating credentials before an appointment so one can reasonably trust their answers and diagnosis. LLMs in a medical context are potentially dangerous for this reason—the patient has no ability to discern if the LLM is hallucinating and will likely inherently trust its answer.

5 | CONCLUSION

The authors believe that ChatGPT has potential to be a useful medical information tool for society. Physicians found the information presented by the AI model to be on the whole medically accurate and comprehensive (98.3%), but laypersons did not sense that they could trust the results confidently (79.8%) compared to if the information came from their own personal physician. Organized medicine needs to move rapidly to continue to evaluate LLMs and AI, ensure the safety of these platforms for our patients, and help utilize this amazing technology to drive improvements in quality of care and outcomes.

AUTHORS CONTRIBUTIONS

All authors contributed to the study conception and design. Material preparation, data collection and analysis were performed by Habib G. Zalzal, Ariel Abraham, and Rahul K. Shah. Statistical analysis was provided by Jenhao Cheng. The first draft and subsequent revisions of the manuscript were written by Habib G. Zalzal, Ariel Abraham, and Rahul K. Shah. All authors commented on previous versions of the manuscript. All authors read and approved the final manuscript.

CONFLICT OF INTEREST STATEMENT

The authors have no financial or proprietary interests in any material discussed in this article.

DATA AVAILABILITY STATEMENT

On request.

ORCID

Habib G. Zalzal https://orcid.org/0000-0002-0777-977X

REFERENCES

1. Grunebaum A, Chervenak J, Pollet SL, Katz A, Chervenak FA. The exciting potential for ChatGPT in obstetrics and gynecology. *Am J Obstet Gynecol*. 2023;228(6):696-705.
2. 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;19(12):1989-1995.
3. OpenAI. ChatGPT: OpenAI. 2023 updated May 24, 2023. Available from: https://chat.openai.com/chat
4. Gwet KL. *Handbook of Inter-Rater Reliability: the Definitive Guide to Measuring the Extent of Agreement among Raters*. 4th ed. Advanced Analytics, LLC; 2014.
5. Landis JRKG. The measurement of observer agreement for categorical data. *Biometrics*. 1977;33(1):159-174.
6. Altman DG. In: Lindsey CCJZJ, ed. *Practical Statistics for Medical Research (Chapman & Hall/CRC Texts in Statistical Science)*. 1st ed. Chapman and Hall; 1991.
7. Qu R, Qureshi U, Petersen G, Lee S. Diagnostic and management applications of ChatGPT in structured otolaryngology clinical scenarios. *OTO Open*. 2023;7(3):e67.
8. Humar P, Asaad M, Bengur FB, Nguyen V. ChatGPT is equivalent to first year plastic surgery residents: evaluation of ChatGPT on the plastic surgery in-service exam. *Aesthet Surg J*. 2023;43:NP1085-NP1089.
9. Hopkins BS, Nguyen VN, Dallas J, et al. ChatGPT versus the neurosurgical written boards: a comparative analysis of artificial intelligence/machine learning performance on neurosurgical board-style questions. *J Neurosurg*. 2023;139:904-911.
10. 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(6):556-558.
11. Harris E. Public had most trust in advice from physicians, nurses during pandemic. *JAMA*. 2023;329(13):1053.
12. Cao Y, Zhou L, Lee S, Cabello L, Chen M, Hershcovich D. Assessing cross-cultural alignment between chatgpt and human societies: an empirical study. *arXiv Preprint* arXiv:230317466. 2023.
13. Ovnat Tamir S, Shemesh S, Oron Y, Marom T. Acute otitis media guidelines in selected developed and developing countries: uniformity and diversity. *Arch Dis Child*. 2017;102(5):450-457.

How to cite this article: Zalzal HG, Abraham A, Cheng J, Shah RK. Can ChatGPT help patients answer their otolaryngology questions? *Laryngoscope Investigative Otolaryngology*. 2024;9(1):e1193. doi:10.1002/lio2.1193