AMERICAN ACADEMY OF OTOLARYNGOLOGY–HEAD AND NECK SURGERY FOUNDATION

Original Research—General Otolaryngology

Otolaryngology–Head and Neck Surgery
2024, Vol. 170(6) 1519–1526

DOI: 10.1002/ohn.759
http://otojournal.org
WILEY

Performance and Consistency of ChatGPT-4 Versus Otolaryngologists: A Clinical Case Series

Jérôme R. Lechien, MD, PhD, MS[1,2,3,4], Mattheuw R. Naunheim, MD, MBA[1,5], Antonino Maniaci, MD, PhD[1,6], Thomas Radulesco, MD, PhD, MS[1,7], Alberto M. Saibene, MD, MA[1,8], Carlos M. Chiesa-Estomba, MD, PhD, MS[1,9]*, and Luigi A. Vaira, MD[1,10,11]*

Abstract

Objective. To study the performance of Chatbot Generative Pretrained Transformer-4 (ChatGPT-4) in the management of cases in otolaryngology–head and neck surgery.

Study Design. Prospective case series.

Setting. Multicenter University Hospitals.

Methods. History, clinical, physical, and additional examinations of adult outpatients consulting in otolaryngology departments of CHU Saint-Pierre and Dour Medical Center were presented to ChatGPT-4, which was interrogated for differential diagnoses, management, and treatment(s). According to specialty, the ChatGPT-4 responses were assessed by 2 distinct, blinded board-certified otolaryngologists with the Artificial Intelligence Performance Instrument.

Results. One hundred cases were presented to ChatGPT-4. ChaGPT-4 indicated a mean of 3.34 (95% confidence interval [CI]: 3.09, 3.59) additional examinations per patient versus 2.10 (95% CI: 1.76, 2.34; $P = .001$) for the practitioners. There was strong consistency ($k > 0.600$) between otolaryngologists and ChatGPT-4 for the indication of upper aerodigestive tract endoscopy, positron emission tomography and computed tomography, audiometry, tympanometry, and psychophysical evaluations. Primary diagnosis was correctly performed by ChatGPT-4 in 38% to 86% of cases depending on subspecialty. Additional examinations indicated by ChatGPT-4 were pertinent and necessary in 8% to 31% of cases, while the treatment regimen was pertinent in 12% to 44% of cases. The performance of ChatGPT-4 was not influenced by the human-reported level of difficulty of clinical cases.

Conclusion. ChatGPT-4 may be a promising adjunctive tool in otolaryngology, providing extensive documentation about additional examinations, primary and differential diagnoses, and treatments. The ChatGPT-4 is more effective in providing a primary diagnosis, and less effective in the selection of additional examinations and treatments.

Keywords

artificial intelligence, ChatGPT-4, head neck surgery, otolaryngology, performance

Received November 2, 2023; accepted March 17, 2024.

[1]Research Committee of Young Otolaryngologists of the International Federation of Otorhinolaryngological Societies (IFOS), Paris, France
[2]Division of Laryngology and Broncho-Esophagology, Department of Otolaryngology–Head Neck Surgery, EpiCURA Hospital, UMONS Research Institute for Health Sciences and Technology, University of Mons (UMons), Mons, Belgium
[3]Department of Otorhinolaryngology and Head and Neck Surgery, Foch Hospital, Phonetics and Phonology Laboratory (UMR 7018 CNRS, Université Sorbonne Nouvelle/Paris 3), Paris Saclay University, Paris, France
[4]Department of Otorhinolaryngology and Head and Neck Surgery, CHU Saint-Pierre, Brussels, Belgium
[5]Department of Otolaryngology, Massachusetts Eye and Ear, Harvard Medical School, Boston, Massachusetts, USA
[6]Department of medicine and surgery, Faculty of Medicine and Surgery, University of Enna "Kore", Enna, Italy
[7]ENT-HNS Department, APHM, CNRS, IUSTI, La Conception University Hospital, Aix Marseille Univ, Marseille, France
[8]Otolaryngology Unit, Department of Health Sciences, ASST Santi Paolo E Carlo, Università Degli Studi Di Milano, Milan, Italy
[9]Department of Otorhinolaryngology–Head and Neck Surgery, Hospital Universitario Donostia, San Sebastian, Spain
[10]Maxillofacial Surgery Operative Unit, Department of Medicine, Surgery and Pharmacy, University of Sassari, Sassari, Italy
[11]Department of Biomedical Sciences, PhD School of Biomedical Sciences, University of Sassari, Sassari, Italy
*These authors contributed equally to this article and may be joined as co-senior authors.

Corresponding Author:
Jérôme R. Lechien, MD, PhD, MS, Division of Laryngology and Broncho-Esophagology, Department of Otolaryngology–Head Neck Surgery, EpiCURA Hospital, UMONS Research Institute for Health Sciences and Technology, University of Mons (UMons), Mons, Belgium.
Email: Jerome.Lechien@umons.ac.be

The development of artificial intelligence (AI)-powered language models, such as Chatbot Generative Pretrained Transformer (ChatGPT), is emerging in medicine. ChatGPT has been used to respond to both simple and complicated questions related to medical examinations, and clinical vignettes, and may improve scientific reports through spelling correction or referencing.[1-3] The accessibility and popularity of ChatGPT may encourage patients to use them for education prior to a medical consultation,[4] and young practitioners may consider ChatGPT-4 as an adjunctive clinical tool for improving their practice. To date, only few small cohort studies investigated the consistency and performance of ChatGPT in the management of common otolaryngological conditions.[1,2,4]

In this study, we investigated the performance of ChatGPT-4 in the diagnostic and therapeutic management of true clinical cases and, consequently, its consistency in providing medical information.

Methods

Patients and Setting

Adult patients consulting in the Departments of Otolaryngology–Head Neck Surgery of Saint-Pierre University Hospital (Brussels, Belgium) and Dour Medical Center (Dour, Belgium) were consecutively recruited from July 2, 2023 to October 12, 2023. This study adhered to the Strengthening the Reporting of Observational Studies in Epidemiology guidelines for observational studies to ensure transparency and replicability of our findings.[5] All participants prospectively enrolled were aged 18 years or older and provided written informed consent. The exclusion criteria included incomplete data records, such as ongoing disease diagnosis exploration or without a clear diagnosis, patients who declined to participate or withdrew consent. To minimize selection bias, we used consecutive sampling, where every presenting patient meeting the inclusion criteria was invited to participate until the predetermined sample size was reached.

The following data were collected: demographics, clinical history, complaints, comorbidities, medication, additional examinations, and treatments. All patients underwent a complete ear, nose, and throat examination, including otoscopy, nasolaryngoscopy, oral examination, and neck palpation.

Data of complete medical records were anonymized and entered in the ChatGPT-4 (open AI) interface, which is accessible through the API (https://chat.openai.com). On October 15, 2023, ChatGPT-4 was systematically interrogated for differential diagnoses, additional examinations, and potential treatment(s) with the following questions: *What are your primary and differential diagnoses*?; *What are your additional examinations to find the diagnosis*?; and *What are your treatment(s) for the primary diagnosis*? The responses of ChatGPT-4 were collected in a database by an assistant physician. For each subspecialty of cases (laryngology-head and neck, rhinology, or otology), the ChatGPT-4 responses were judged by 2 board-certified otolaryngologists chosen based on their experience and for having completed a fellowship in that specific specialty. Judges used all possible national and international consensus statements or guidelines[6-15] to establish what is the appropriate management, which will be used to rate the ChatGPT-4 responses. The study was approved by the institutional review board of CHU Saint-Pierre (n°BE0762023230708).

ChatGPT-4 Performance

The performance of ChatGPT-4 in the management of otolaryngological cases was evaluated with the Artificial Intelligence Performance Instrument (AIPI), which is a validated and reliable instrument used to assess the performance and consistency of artificial intelligence chatbots.[2] AIPI was developed by the AI Study Group of the Young-Otolaryngologists of the International Federation of Otorhinolaryngological Societies, which includes board-certified otolaryngologists. AIPI provides a comprehensive approach to otolaryngological cases and has high intra- and interrater reliabilities.[2] AIPI includes 9 items that assess the ability of chatbot to consider medical and surgical history; symptoms; physical examination; diagnosis; additional examinations; management plan; and treatments in the overall management of real clinical cases (**Figure 1**). AIPI is subdivided into the 4 following sub-scores associating common items: patient feature score (/6), diagnosis score (/7), additional examination score (/5), and treatment score (/3). The final AIPI score ranges from 0 (inadequate management) to 20 (excellent management).

Level of Complexity of the Case

Each clinical case included 4 basic elements: the medical history; the clinical examination; the technical diagnostic findings (from additional examinations); and the treatment,[4,16] all of which involve the psychosocial context of the patient. The findings of these features lead to a variation in the degree of complexity, as the complexity may increase when the case contains distracting information. Many scoring systems were developed to rate the degree of complexity of real clinical cases. In this study, the raters used a modified version of the General Items off the Amsterdam Clinical Challenge Scale (ACCS) test[4,16] to rate the complexity of clinical cases submitted to ChatGPT-4. The baseline version of ACCS included 6 generic items: previous history/actual context; problem presented; communication (patient complaints and responses); physical examination (typical vs atypical signs); patient management (adequate vs complicated management); and prevention. Items were rated on a 5-point Likert scale, ranging from 1 (easy) to 5 (difficult). The extremes of each item were defined in general terms. For example, the item "problem presented" is scored 1 when the problem is straightforward, not likely

Outcomes of Artificial Intelligence Performance Instrument (AIPI)	Practitioner evaluation			Item score	Subscores
1. Consideration of medical and surgical history in the AI management:	Fully (2)	Partly (1)	Not (0)	…../2	Patient
2. Consideration of symptoms of patients in the AI management	Fully (2)	Partly (1)	Not (0)	…../2	feature score
3. Consideration of physical findings reported by practitioner(s)	Fully (2)	Partly (1)	Not (0)	…../2	…...../6
4. The differential diagnoses provided by AI are:	Complete and plausible (3) Incomplete but plausible (2) Incomplete and not plausible for one or several (1) Absent (0)			…../3	
5. The primary diagnosis of AI was:	Correct (3) Plausible (2) Not plausible (1) Absent (0)			…../3	Diagnosis score
6. The management plan of AI included potential physical/additional examinations for determining the diagnosis		Yes (1)	No (0)	…../1	…...../7
7. The additional examinations proposed by AI are/include	All pertinent and necessary examinations (3) All pertinent but partialy necessary examinations (2) An association of pertinent, necessary, and inadequate examinations (1) An association of inadequate examinations (0)			…../3	Additional Examination Score
8. AI identified the most relevant additional examination to perform first	Yes (1) No, AI provided a list without stratification (0)			…../1	…...../5
9. The treatments proposed by AI are/include	All pertinent and necessary therapeutic findings (3) All pertinent but incomplete therapeutic findings (2) An association of pertinent, necessary, and inadequate therapeutic findings (1) No adequate therapeutic approach (0)			…../3	Treatment score …...../3
				Total AIPI	…...../20

Figure 1. Artificial Intelligence Performance Instrument (AIPI). AIPI is a valid and reliable score assessing the performance of artificial intelligence (AI)-powered language models. AIPI score ranges from 0 (inadequate management) to 20 (adequate management).

to be serious and of a limited nature, and 5 when it is vague, difficult to define, probably serious, or complex.[4] For example, patients with atypical disease presentation or poor therapeutic response to an evidence-based treatment may be assessed as 5/5 in examination and management. A clinical case has a score ranging from 6 to 30.[4] As proposed in a recent study, the scores ranging from 6 to 14, 15 to 23, and 24 to 30 may be considered as easy, moderate, and difficult, respectively.[4] The clinical case complexity was evaluated by the practitioner who conducted the consultation with ACCS at the end of each patient consultation (Supplemental Appendix S1, available online).

Statistical Analyses

Statistical analyses were performed using the Statistical Package for the Social Sciences for Windows (SPSS version 22.0; IBM Corp). Additional examinations indicated by otolaryngologists and ChatGPT-4 were coded with a predefined number in a matrix, which facilitated the evaluation of consistency between the physician's findings versus ChatGPT-4 (Cohen's κ analysis). The total number of additional examinations indicated by ChatGPT-4 and otolaryngologists were compared with the Mann-Whitney U test. The interrater reliability was evaluated for the AIPI score of both experts in each subspecialty.

The consistency was considered as low, moderate, and strong for $\kappa < 0.40$, 0.40 to 0.60, and $\kappa > 0.60$, respectively. The association between the difficulty level of cases and the AIPI score of judges was evaluated with the Spearman coefficient. A level of significance of $P < .05$ was used.

Results

The data of 100 patients with otolaryngological conditions were presented to ChatGPT-4 for clinical management (**Figure 2**). The primary diagnoses are reported in **Table 1**. The mean age of patients was 50.5 ± 15.7 years. The mean level of clinical case complexity for these cases, as measured by the ACCS, was 15.4 ± 5.7. Considering subspecialties, ACCS mean scores for laryngological, head and neck, rhinological, and otological cases were 18.7 ± 6.5, 15.2 ± 4.2, 13.5 ± 5.3, and 14.1 ± 6.0, respectively. The mean AIPI score was 12.8 ± 3.9. Considering subspecialities, AIPI scores were 13.7 ± 3.5, 13.3 ± 2.8, 10.9 ± 4.2, and 15.1 ± 4.4 for laryngological, head and neck, rhinological, and otological cases, respectively. There was no significant association between the level of complexity of clinical case (ACCS) and the performance of ChatGPT-4 (AIPI). The interrater reliability values between experts were 0.883, 0.838, and 0.901 for laryngology-head and neck surgery, rhinology, and otology, respectively.

The primary diagnosis provided by ChatGPT-4 was correct in 62% of cases. ChatGPT-4 performance in correctly diagnosing the case significantly varied between subspecialty cases according to the AIPI score (**Table 2**). Head and neck cases were correctly diagnosed most frequently (86%), and

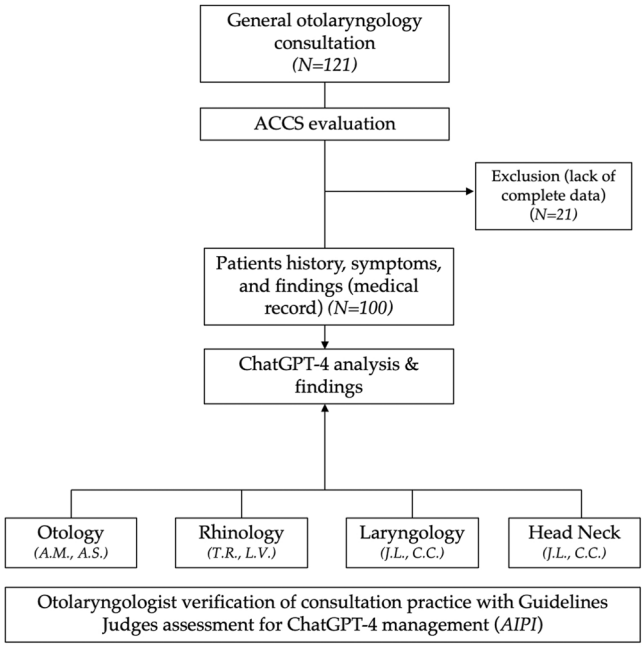

Figure 2. Chart flow. ACCS, Amsterdam Clinical Challenge Scale; AIPI, Artificial Intelligence Performance Instrument; ChatGPT, Chatbot Generative Pretrained Transformer.

laryngology was least frequently correct (38%). However, there was no statistical difference between subspecialties in terms of correctness of the diagnosis.

The recommended diagnostic workup by ChatGPT was rated as both pertinent and necessary in 18% of cases. ChatGPT-4 recommended more diagnostic examinations than clinicians, with a mean of 3.3 ± 1.3 additional examinations per patient versus 2.1 ± 1.3 ($P = .001$) for the practitioners. The consistency analysis reported strong consistencies ($\kappa > 0.600$) between otolaryngologists and ChatGPT-4 for the indication of upper aerodigestive tract endoscopy, positron emission tomography and computed tomography, audiometry, tympanometry, and psychophysical evaluations (**Table 3**); it demonstrated lower consistency for rhinomanometry, blood testing, voice quality assessment, and neck tomodensitometry.

The treatment recommended by ChatGPT was rated as pertinent and necessary in 25% of cases, ranging from 12% to 44% based on subspecialty. These differences were statistically significant with head and neck and otology cases having more frequently pertinent and necessary treatment recommendations. An example of ChatGPT-4 inputs/outputs is available in Supplemental Appendix S2, available online. The most common failures of ChatGPT-4 are reported in Supplemental Appendix S3, available online, and included propositions of H2 blockers or increased proton-pump inhibitors in case of alkaline laryngopharyngeal reflux (N = 6); proposition of radiation or medical treatment in postradiation cancer requiring salvage surgery (N = 2), or medical treatments for benign lesions of the vocal folds requiring phonosurgery (N = 2).

Table 1. Primary Diagnoses of Patients

Primary diagnosis	N (%)
Upper aerodigestive tract cancer	28
Laryngopharyngeal reflux disease	9
Postviral olfactory dysfunction	6
Eustachian tube dysfunction	3
Rhinopharyngitis	3
Allergic rhinitis	3
Acute maxillary rhinosinusitis	3
Chronic otitis media	2
Nasal septal deviation	2
Odontogenic maxillary rhinosinusitis	2
Chronic rhinosinusitis with nasal polyps	2
Chronic rhinosinusitis without nasal polyps	2
Unilateral or bilateral vocal cord paralysis	2
Vocal fold nodules	2
Benign paroxysmal vertigo	1
Presbycusis	1
Ear cerumen wax	1
Eagle syndrome	1
Patulous Eustachian tube	1
External ear duct stenosis	1
Obstructive sleep apnea syndrome	1
Posttraumatic anosmia	1
Tornwald cyst	1
Sphenoid mucocele	1
Sinus osteoma	1
Vasomotor rhinitis	1
Medicamentosa rhinitis	1
Empty nose syndrome	1
Reinke edema	1
Vocal fold hemorrhage	1
Syphilitic tonsillitis	1
Laryngeal hypersensitivity	1
Infectious laryngitis	1
Vocal fold polyp	1
Superior laryngeal nerve paralysis	1
Glottic insufficiency	1
Medicamentosa laryngitis	1
Laryngocele	1
Iatrogenic dysphagia (postneurosurgical spine surgery)	1
Vocal fold scar	1
Psychogenic aphonia	1
Esophageal scleroderma	1
Recurrent tonsil abscess	1
Salivary lithiasis	1
Parotid lymphoepithelial cyst	1

Discussion

The primary findings of the present study support that ChatGPT-4 performance is more effective for proposing plausible primary and differential diagnoses rather than for the indication of additional examinations or

Table 2. Performance of ChatGPT-4

AIPI outcomes	Otology N = 11	Rhinology N = 26	Laryngology N = 34	Head and Neck N = 29	Total	*P* value
Primary diagnosis [N (%)]						
Correct	7 (67)	17 (65)	13 (38)	25 (86)	62 (62)	NS
Plausible	2 (18)	6 (23)	6 (18)	4 (14)	18 (18)	
Not plausible	2 (18)	3 (12)	15 (44)	0 (0)	20 (20)	
Absent	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	
Relevant additional examination						
Pertinent and necessary	3 (27)	2 (8)	4 (12)	9 (31)	18 (18)	.001
Pertinent and not all necessary	2 (18)	7 (27)	4 (12)	14 (48)	27 (27)	
Pertinent, necessary, and inadequate	6 (55)	15 (57)	24 (70)	6 (21)	51 (51)	
Only inadequate examinations	0 (0)	2 (8)	2 (6)	0 (0)	4 (4)	
Treatment regimen						
Pertinent and necessary	4 (36)	4 (15)	4 (12)	13 (44)	25 (25)	.007
Pertinent and incomplete	3 (27)	13 (50)	7 (21)	8 (28)	31 (31)	
Association of pertinent/necessary and inadequate	4 (36)	7 (27)	23 (67)	8 (28)	42 (42)	
No adequate strategy	0 (0)	2 (8)	0 (0)	0 (0)	2 (2)	

The AIPI was scored by 2 board-certified otolaryngologists for each specialty (agreement). The comparison of AIPI findings between subspecialities was carried out by Kruskal-Wallis test.

Abbreviations: AIPI, Artificial Intelligence Performance Instrument; ChatGPT, Chatbot Generative Pretrained Transformer; NS, nonsignificant.

Table 3. Additional Examination Consistency

Additional examinations	OTO (N)	ChatGPT-4 (N)	κ
Sinus tomodensitometry	24	35	0.502
PET-CT	22	15	0.770
Biopsy	20	30	0.579
Neck tomodensitometry	19	42	0.267
Neck magnetic resonance	17	28	0.521
Voice quality assessment	16	2	0.194
Blood biology	10	21	0.216
Rhinomanometry	9	2	0.342
Psychophysical olfactory testing	9	9	0.634
Audiometry	9	6	0.784
Upper aerodigestive tract endoscopy (GA)	9	8	0.807
Tympanometry	8	4	0.648

The consistency analysis was carried out on the 12 most prescribed additional examinations by practitioners.

Abbreviations: ChatGPT-4, Chatbot Generative Pretrained Transformer-4; GA, general anesthesia; OTO, otolaryngologists; PET-CT, positron emission tomography and computed tomography.

treatments. ChatGPT-4 provides a correct or plausible diagnosis for 80% of cases tested but infrequently recommends pertinent and necessary diagnosis (18%) and treatment (25%). Thus, ChatGPT-4 works as a virtual encyclopedia, proposing lists of additional examinations without selecting the most adequate tests for the clinical situation. In practice, most patients are not subjected to a single large battery of tests at one time but underwent the fewest possible additional examination to be cost-effective. This observation corroborates those found in small cohorts of theoretical clinical vignettes or true clinical cases.[1,2,4] In a preliminary study of 15 clinical vignettes, Vaira et al observed that ChatGPT-4 responses were entirely or nearly entirely correct in 87.2% of cases. At the same time, ChatGPT-4 provided a fully or nearly fully correct diagnosis in 81.7% of cases.[1] For treatment, the proposed therapeutic procedure was judged to be complete in 56.7% of cases, which was substantially higher compared to our data. Ayoub et al[17] reported that ChatGPT indicated correct medical advice in 68.2% of clinical scenarios.[4]

Though the performance of ChatGPT-4 appeared to vary between subspecialty cases in our study, the level of difficulty of the clinical case did not influence the performance of the chatbot, which corroborated recent findings.[18] Indeed, Dallari et al reported in a cross-sectional study that the difficulty of clinical vignettes did not influence the performance of ChatGPT in both consistency and correctness scores.[18] The similar performance of ChatGPT for managing easy versus difficult clinical cases corroborates the findings of a preliminary study,[4] in which the performance of ChatGPT did not vary between mild, moderate, and difficult clinical cases at the ACCS.[4] This finding was, however, not observed in a recent neurological study.[19] Galetta and Meltzer investigated the accuracy of ChatGPT-4 in the management of neurologic case vignettes. In a series of 29 clinical vignettes, ChatGPT-4 identified the good diagnosis in less than 50% but GPT-4 was more accurate with localization and diagnosis of easier versus harder cases.[19]

The knowledge about the performance of ChatGPT in the management of real clinical cases is important for

practitioners and patients. Indeed, medical students, residents, or other young practitioners may consider using ChatGPT-4 in the management of complicated cases. Our findings indicate that large language models may pose a substantial risk if routinely applied to guide workup and treatment without further scrutiny. The availability of ChatGPT-3.5 for potential use by patients prior to the consultation is another important issue. Indeed, some patients using ChatGPT may question practitioners' judgment and propositions, which may theoretically impair trust in the doctor-patient relationship. This point may apply more to real clinical cases (patient symptoms and findings) than to theoretical information since ChatGPT-4 appears to be performant for patient education questions related to general otolaryngology.[4]

While ChatGPT has demonstrated promising results in assisting with the diagnostic and therapeutic framing of head and neck pathologies,[1] it is crucial to integrate psychosocial complexity into its use. Given the complex interplay between psychological and social dynamics in head and neck conditions, the impact on both diagnosis and treatment is profound. Consequently, although ChatGPT is adept at synthesizing medical data and literature to offer valuable insights, health care professionals are responsible for integrating these AI-generated insights with a comprehensive and nuanced understanding of each patient's distinctive psychosocial landscape. This integration is crucial, as it ensures that the diagnostic and therapeutic approaches are not only informed by clinical evidence but are also finely attuned to the individualized psychosocial factors that can influence patient outcomes. By doing so, clinicians may deliver more personalized, empathetic, and effective care that addresses the multifaceted needs of patients with otolaryngological conditions.

The main strengths of the present study are the originality and consideration of real clinical cases, which makes this study the largest case series evaluating the consistency and performance of ChatGPT-4 in otolaryngology. Compared to theoretical clinical vignettes, the real clinical cases reflect the reality of practitioners who may encounter communication or other clinical management difficulties. The use of a validated AI instrument and the evaluation by 2 blinded judges are additional strengths.

The lack of information about the hyperparameters of ChatGPT-4 is a limitation because most large language models are nondeterministic and, as demonstrated in our study, their outputs may vary with each run, which may be curtailed by fine-tuning specific hyperparameters. Specifically, hyperparameter tuning may influence the ChatGPT-4 findings through preventing overfitting and improving the improvement of speed of response, performance, or balancing resources. The hyperparameters are still unknown. The test-retest reliability performed herein between regenerated ChatGPT-4 answers addressed these findings. Still, the lack of knowledge about the hyperparameters of ChatGPT-4 may limit the understanding of the system's responses and proposition.

The existence of some interjudge reliability discordances in the ChatGPT-4 performance assessment is a limitation but it is related to the lack of consensus in the management of some prevalent diseases found in this study, such as laryngopharyngeal reflux. For example, some judges estimated adequate the ChatGPT-4 management with pH study, gastrointestinal endoscopy, and proton-pump inhibitor prescription, while others recommended hypopharyngeal-esophageal multichannel intraluminal impedance-pH monitoring or nasopharyngeal pH monitoring in case of Eustachian/nasal involvement, and alginate therapy to treat nonacid reflux. In this study, we observed that ChatGPT poorly performed in the indication of some unusual additional examinations, including hypopharyngeal-esophageal multichannel intraluminal impedance-pH testing or psychophysical olfactory testing. Moreover, the ChatGPT performance was evaluated through practitioner inputs. The performance could be lower in case of unclear or incomplete information provided by non-health care professionals.

Another limitation of this study is that the complexity assessment of cases was conducted by a single individual, which may introduce a degree of subjectivity, limiting the generalizability of this part of the analysis. Furthermore, although the AIPI provides a valuable framework for evaluating ChatGPT-4's performance, this tool may not fully encapsulate the nuances of AI's role in clinical decision-making. The limitations of AIPI, particularly in reflecting the complex dynamics of real-world clinical scenarios, should be carefully considered in interpreting our findings. Another limitation is the short-term variability in ChatGPT-4's responses. This challenge underscores the fluctuating nature of AI models and their outputs, which could affect the consistency and reliability of AI-assisted clinical decisions. Finally, a limitation of our study is the potential lack of generalizability of our findings to broader patient populations, given the specific composition of our sample. Our sample included a disproportionately high number of cases involving upper aerodigestive tract cancers (28%) and did not encompass a range of cases that might be more commonly encountered in various clinical practice environments.

Conclusion

ChatGPT-4 may be a promising adjunctive tool in otolaryngology, providing information about recommended workups, primary and differential diagnoses, and treatments for clinical cases. ChatGPT-4 is more consistent in diagnosis than in the selection of the most appropriate additional examinations and treatments.

Author Contributions

Jérôme R. Lechien, design, acquisition of data, data analysis and interpretation, drafting, final approval, and accountability for the work, final approval of the version to be published, agreement to be accountable for all aspects of the work in

ensuring that questions related to the accuracy or integrity of any part of the work are appropriately investigated and resolved; **Mattheuw R. Naunheim**, final approval, and accountability for the work, final approval of the version to be published, agreement to be accountable for all aspects of the work in ensuring that questions related to the accuracy or integrity of any part of the work are appropriately investigated and resolved; **Antonino Maniaci**, design, acquisition of data, data analysis and interpretation, drafting, final approval, and accountability for the work, final approval of the version to be published, agreement to be accountable for all aspects of the work in ensuring that questions related to the accuracy or integrity of any part of the work are appropriately investigated and resolved; **Thomas Radulesco**, design, acquisition of data, data analysis and interpretation, drafting, final approval, and accountability for the work, final approval of the version to be published, agreement to be accountable for all aspects of the work in ensuring that questions related to the accuracy or integrity of any part of the work are appropriately investigated and resolved; **Alberto M. Saibene**, design, acquisition of data, data analysis and interpretation, drafting, final approval, and accountability for the work, final approval of the version to be published, agreement to be accountable for all aspects of the work in ensuring that questions related to the accuracy or integrity of any part of the work are appropriately investigated and resolved; **Carlos M. Chiesa-Estomba**, statistics, final approval, and accountability for the work, final approval of the version to be published, agreement to be accountable for all aspects of the work in ensuring that questions related to the accuracy or integrity of any part of the work are appropriately investigated and resolved; **Luigi A. Vaira**, design, acquisition of data, data analysis and interpretation, drafting, final approval, and accountability for the work, final approval of the version to be published, agreement to be accountable for all aspects of the work in ensuring that questions related to the accuracy or integrity of any part of the work are appropriately investigated and resolved.

Disclosures

Competing interests: The authors have no conflict of interest.

Funding source: Nothing to disclose.

Supplemental Material

Additional supporting information is available in the online version of the article.

References

1. Vaira LA, Lechien JR, Abbate V, et al. Accuracy of ChatGPT-generated information on head and neck and oromaxillofacial surgery: a multicenter collaborative analysis. *Otolaryngol Head Neck Surg*. 2023. doi:10.1002/ohn.489
2. Lechien JR, Maniaci A, Gengler I, Hans S, Chiesa-Estomba CM, Vaira LA. Validity and reliability of an instrument evaluating the performance of intelligent chatbot: the Artificial Intelligence Performance Instrument (AIPI). *Eur Arch Otrhinolaryngol*. 2023;281:2063-2079. doi:10.1007/s00405-023-08219-y
3. 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. doi:10.1002/ohn.526
4. Lechien JR, Georgescu BM, Hans S, Chiesa-Estomba CM. ChatGPT performance in laryngology and head and neck surgery: a clinical case-series. *Eur Arch Otrhinolaryngol*. 2023;281:319-333. doi:10.1007/s00405-023-08282-5
5. von Elm E, Altman DG, Egger M, Pocock SJ, Gøtzsche PC, Vandenbroucke JP. The Strengthening the Reporting of Observational Studies in Epidemiology (STROBE) statement: guidelines for reporting observational studies. *J Clin Epidemiol*. 2008;61(4):344-349.
6. Stachler RJ, Francis DO, Schwartz SR, et al. Clinical practice guideline: hoarseness (dysphonia) (update). *Otolaryngol Head Neck Surg*. 2018;158(1_suppl):S1-S42. doi:10.1177/0194599817751030
7. Hellings PW, Fokkens WJ, Orlandi R, et al. The EUFOREA pocket guide for chronic rhinosinusitis. *Rhinology*. 2023;61(1):85-89. doi:10.4193/Rhin22.344
8. Bhattacharyya N, Gubbels SP, Schwartz SR, et al. Clinical practice guideline: benign paroxysmal positional vertigo (update). *Otolaryngol Head Neck Surg*. 2017;156(3_suppl):S1-S47. doi:10.1177/0194599816689667
9. Lechien JR, Vaezi MF, Chan WW, et al. The Dubai definition and diagnostic criteria of laryngopharyngeal reflux: the IFOS consensus. *Laryngoscope*. 2023;134:1614-1624.
10. Hellings PW, Fokkens WJ, Bachert C, et al. Positioning the principles of precision medicine in care pathways for allergic rhinitis and chronic rhinosinusitis—a EUFOREA-ARIA-EPOS-AIRWAYS ICP statement. *Allergy*. 2017;72(9):1297-1305. doi:10.1111/all.13162
11. Chandrasekhar SS, Tsai Do BS, Schwartz SR, et al. Clinical practice guideline: sudden hearing loss (update). *Otolaryngol Head Neck Surg*. 2019;161(1_suppl):S1-S45. doi:10.1177/0194599819859885
12. Tunkel DE, Anne S, Payne SC, et al. Clinical practice guideline: nosebleed (epistaxis). *Otolaryngol Head Neck Surg*. 2020;162(1_suppl):S1-S38. doi:10.1177/0194599819890327
13. Pynnonen MA, Gillespie MB, Roman B, et al. Clinical practice guideline: evaluation of the neck mass in adults. *Otolaryngol Head Neck Surg*. 2017;157(2_suppl):S1-S30. doi:10.1177/0194599817722550
14. Seidman MD, Gurgel RK, Lin SY, et al. Clinical practice guideline: allergic rhinitis. *Otolaryngol Head Neck Surg*. 2015;152(1 suppl):S1-S43. doi:10.1177/0194599814561600
15. Rosenfeld RM, Schwartz SR, Cannon CR, et al. Clinical practice guideline: acute otitis externa. *Otolaryngol Head Neck Surg*. 2014;150(1 suppl):S1-S24. doi:10.1177/0194599813517083
16. Gercama AJ, de Haan H, van der Vleuten V. Reliability of the Amsterdam Clinical Challenge Scale (ACCS): a new instrument to assess the level of difficulty of patient cases in medical education. *Med Educ*. 2000;34(7):519-524.
17. 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*. 2023. doi:10.1002/ohn.465

18. Dallari V, Sacchetto A, Saetti R, Calabrese L, Vittadello F, Gazzini L. Is artificial intelligence ready to replace specialist doctors entirely? ENT specialists vs ChatGPT: 1-0, ball at the center. *Eur Arch Otrhinolaryngol*. 2023;281:995-1023. doi:10.1007/s00405-023-08321-1
19. Galetta K, Meltzer E. Does GPT-4 have neurophobia? Localization and diagnostic accuracy of an artificial intelligence-powered chatbot in clinical vignettes. *J Neurol Sci*. 2023;453:120804. doi:10.1016/j.jns.2023.120804